Kvantitatív szövegelemzés és szövegbányászat a politikatudományban

Kvantitatív szövegelemzés és szövegbányászat a politikatudományban

Szerkesztette
Sebők Miklós

L'Harmattan

2016

A kiadvány a Nemzeti Kulturális Alap (témaszám: 3437/02240) és a Magyar Tudományos Akadémia Társadalomtudományi Kutatóközpontja támogatásával készült.

Szerkesztette: Sebők Miklós
A szerkesztő munkatársai: Balázs Ágnes, Zágoni Bella
A tárgymutatót készítette: Balázs Ágnes
Szerzők: Balázs Ágnes, Kubik Bálint György, Molnár Csaba, Sebők Miklós, Szabó Gabriella, Vancsó Anna, Zágoni Bella, Zorigt Burtejin
Szakmai lektor: Kovács Borbála
Lektor: Kovács László

A borítón látható szófelhőt a II.1. fejezetben tárgyalt szózsák modellt alkalmazó www.tagxedo.com weboldal segítségével készítettük a Comparative Agendas Hungary projekt (Boda – Sebők, 2015) keretében elkészült, a cap.tk.mta.hu oldalon található interpelláció adatbázis alapján, az Országgyűlésben 1990 és 2014 között elhangzott interpellációk címeiből.

ISBN 978-963-414-229-4

A kiadásért felel Gyenes Ádám, a L'Harmattan Kiadó igazgatója.

A kiadó kötetei megrendelhetők, illetve kedvezménnyel megvásárolhatók:
L'Harmattan Könyvesbolt
Tel.: +36-1-267-5979
1053 Budapest, Kossuth L. u. 14–16.
harmattan@harmattan.hu
www.harmattan.hu

Párbeszéd Könyvesbolt
Tel.: +36-1-445-2775
1085 Budapest, Horánszky utca 20.
parbeszedkonyvesbolt@gmail.com
www.konyveslap.hu

TARTALOM

Sebők Miklós

BEVEZETÉS

A fejezet bemutatja a kötet tárgyát, szerkezetét, valamint a kvantitatív szövegelemzés és szövegbányászat pozícióját a társadalomtudományon belül. A kötet nagyobb blokkjai lefedik a szövegek kvalitatív és kvantitatív társadalomtudományi elemzésének módszertani alapjait, a szövegbányászat legfontosabb feladatait, illetve néhány gyakori eljárását. A kötet két fő célcsoportjaként a társadalomtudományi kutatói és felsőoktatási közösséget határozzuk meg, valamint rögzítjük, hogy a kvantitatív szövegelemzés területén belül elsődlegesen a dokumentum- és tartalomelemzési módszertanhoz kapcsolódunk.

A jelen kötet legfontosabb célja az, hogy bevezetést nyújtson a nemzetközi politikatudomány egy kurrens irányzatába, *a szövegek kvantitatív elemzésébe (quantitative text analysis – QTA)*. A szövegek és más minőségi adatok (filmek, képek) elemzése annyiban különbözik a mennyiségi (kvantitatív) adatokétól, hogy nyers formájukban még nem alkalmasak arra, hogy statisztikai, illetve ökonometriai elemzés alá vessük őket, s így további módszertani problémákat vetnek fel, melyek speciális tárgyalása szükséges.

A kötet kiindulópontját a *politikai szövegek társadalomtudományi elemzésének* problémája adja. E feladatot – eltérő jellegű kutatási eredményekkel, de – kvalitatív és kvantitatív módszertani eszközökkel egyaránt meg lehet oldani, így a könyv négy nagyobb blokkja közül az első e módszerek relatív hasznosságát és fogalmi alapjait tárgyalja. Mivel a QTA magyarországi politikatudományi felhasználása e sorok írásakor még messze nem tekinthető általánosnak, az első fejezetben érintjük ennek olyan kutatástervezési problémáit is, mint a felfedezés logikája, illetve a kutatás folyamata (I.1.). Ennek során külön kitérünk a számítógépes támogatással, illetve gépi elemzéssel elvégezhető munkafolyamatokra, mint a 21. századi társadalomtudomány gyorsan fejlődő és lassan megkerülhetetlenné váló területére (I.2.).

A kötet következő két blokkja (II–III.) már a szűken vett kvantitatív szövegelemzés és a vele nagy átfedést mutató szövegbányászat feladataiba nyújt betekintést. Ezek közül a szöveg először tisztázza az adatok visszakeresésének

és kinyerésének alapjait (II.1.), majd bevezetést nyújt a talán legelemibb szövegbányászati feladatba, a névelem-felismerésbe (II.2.). Ezt követően olyan haladóbb feladatokkal ismerkedünk meg, mint a deduktív logikát követő osztályozás (III.1.), az ennek egy speciális alkalmazásának tekinthető véleményelemzés (III.2.), illetve az induktív elven működő csoportosítás (III.3.). A kötet negyedik blokkja a leggyakrabban használt szövegbányászati megoldások gyakorlati alkalmazását tárgyalja. Ennek kapcsán foglalkozunk a félig automatizált felügyelt tanulás osztályozási megközelítésével (IV.1.), míg a csoportosítási feladat klasszikus megoldásaként pedig a felügyelet nélküli tanulási algoritmusokat mutatjuk be (IV.2). A könyvet záró V. fejezet tágabb kitekintést nyújt a kvantitatív szövegelemzés tudományterületére, illetve határvidékeire, segítve a téma iránt érdeklődő olvasót a további tájékozódásban.

***

E rövid szerkezeti összefoglalóból részben már következik, hogy milyen új ismeretanyagokra tud támaszkodni az olvasó, miután letette a kötetet. Egyfelől egy általános áttekintést kap a társadalomtudományi szövegelemzés teljes területéről, s ezen belül is a dokumentum- és tartalomorientált megközelítésekről. Képes lehet eldönteni, hogy vizsgálati tárgyához, kérdéséhez és a rendelkezésre álló adatforrásokhoz milyen kutatási logikát érdemes alkalmaznia, s mindebben hogyan tud támaszkodni a számítógépes adatelemzés nyújtotta előnyökre.

Az ilyen általános szövegelemzési ismeretek mellett az olvasó egy bevezető szinten képes lesz alkalmazni a kvantitatív szövegelemzés és szövegbányászat legalapvetőbb eljárásait saját kutatására. Deduktív vagy induktív felfedező logikája fényében dönthet az adatelemzés módjáról, és a felkínált menüből kiválaszthatja a kutatási tervéhez legjobban illeszkedő megoldásokat. A kötetet végigkísérő konkrét példák segítségével pedig akár reprodukálhatja is ezen eljárásokat saját kutatásában.

Ezen ismeretanyag szinte automatikusan meghatározza azt a két célcsoportot, melyre figyelemmel a szerkezeti vázlatot tartalommal töltöttük fel: a kutatási szférát és a felsőoktatást. Ennek eredménye a reményeink szerint a magyar politikatudományi kutatási gyakorlatot segítő tárgyalásmód, illetve a szöveg oktatási kellékekkel való felvértezése.

Kezdve az elsővel, a kötetben a hangsúly kifejezetten a QTA gyakorlati alkalmazásán, s nem az elvont elméleti ismereteken van. Ez utóbbiak egyfelől könnyebben hozzáférhetőek a magyar olvasó számára is, másrészt pedig csak általános útmutatást adnak a gyakran egészen prózai (pl. kódolási) problémákkal szembesülő kutatónak. Ebből is következik, hogy könyvünk egyik legfontosabb célcsoportját a magyarországi társadalomtudományos kutatóközösség, s közöttük is a pozitivista, illetve (az I. blokk esetében) a posztpozitivista és

megalapozott elméleti paradigmában dolgozó kollégák jelentik. Nem volt célunk így számos fontos, de a politikatudományban korlátozottan használt és/vagy számítógépes támogatást kevésbé használó kvalitatív adatelemzési módszertan bemutatása (pl. részt vevő megfigyelés, dekonstrukció stb.).

Az aktív kutatók mellett könyvünk hasonlóan fontos célcsoportjának tartjuk a *felsőoktatási hallgatókat és oktatókat*. Ennek megfelelően a kötet szerkezetében keverednek a kézikönyv és a tankönyv elemei. A politikatudományi kutatóképzés tartalmának megújítása nélkül nehezen elképzelhető a legfejlettebb kvalitatív és kvantitatív kutatási módszerek hazai elterjedése. A fejezetek e cél érdekében számos illusztrációt hoznak a QTA-technikák gyakorlati alkalmazásai közül, ami reményeink szerint még a legszkeptikusabb olvasót is meggyőzik majd e megoldások hasznosságáról.

A tárgyalás áttekinthetőségét szolgálja az egységes fejezetszerkezet, a legfontosabb fogalmak magyar és angol nyelvű szószedete, valamint a további olvasásra ajánlott szakirodalom szerepeltetése. Az oktatásban való közvetlen alkalmazást segíthetik továbbá a fejezetek végén megadott vizsgakérdések, illetve a kötet honlapján (qta.tk.mta.hu) szereplő további információk: gyakorlófeladatok (megoldásokkal), az egyes feladatokra alkalmazható scriptek és kereskedelmi programok bemutatása, a témával kapcsolatos prezentációk és további ajánlott irodalmak.

***

Miközben reményeink szerint a jelen kötet önmagában is megfelelő bevezetést nyújt a kvantitatív szövegelemzés alapjaiba, fontos hangsúlyozni, hogy a magunk számára kitűzött feladatot sok szempontból korlátoztuk. Nem célunk így az, hogy a szöveg nemzetközi és magyar kiadók által kiadott általános vagy kvalitatív kutatás-módszertani művekkel (ld. pl. Creswell, 2013; Esterberg, 2002; Berg – Lune, 2014; illetve Héra – Ligeti, 2014; Babbie, 2003) vagy a speciálisan a kvalitatív adatok (s benne a szövegek) elemzésére fókuszáló kézikönyvekkel (Grbich, 2013; Miles et al., 2013) versenyezzen. Másfelől gyűjtőköre általánosabb, mint a kifejezetten egy-egy programmal elvégezhető kvalitatív adatelemzési feladatokat bemutató könyveknek (mint amilyen Friese, 2014; illetve Bazeley és Jackson, 2013 műve). Ugyanígy elhatárolható a kötet tárgya a „statisztikai adat- és szövegelemzés", illetve a számítógépes nyelvészet tárgykörétől (ld. a 2000-es évek óta MSZNYK néven megrendezésre kerülő konferenciákat). Művünk célja ugyanis határozottan alkalmazott jellegű, fókusza a más tudományterületeken már kidolgozott eljárások és technikai megoldások „átültetése" politikatudományi problémákra.

A könyv fejezeteinek megírását vezérlő szempont mindezek fényében a *dokumentum-* és *tartalomközpontúság volt*. Ez egyfelől lefedi az ún. beavatko-

zásmentes vizsgálatok (Babbie, 2003: 351) körét. A dokumentumok (szövegek, multimédiás tartalmak stb.) tartalomelemzésének tipikus feladatai jelentették kiindulópontunkat, melytől különböző irányokba terjeszkedtünk tovább. E szerkesztői döntésből fakadóan így például nem foglalkozunk a fókuszcsoportok mint fontos kvalitatív adatelemzési módszertan lebonyolításának kérdéseivel – számos fejezet tartalmaz ugyanakkor hasznos tudnivalókat egy már átírt fókuszcsoport-beszélgetés elemzéséhez. Ennyiben pedig gyűjtőkörünk már kiterjedt a beavatkozásmentes vizsgálatokon túli adatfelvételi módszereket alkalmazó kutatásokra, legalábbis amennyiben ezek kvantitatív szövegelemzésre alkalmas alapanyagot hoznak létre (maga a kvalitatív adatgyűjtési módszer ugyanakkor ebben az esetben sem tárgya könyvünknek).

Utolsó megjegyzésünk a könyv tárgyával kapcsolatban a terminológiára vonatkozik. Amiben a kötet a tematikus korlátozások mellett is hozzáadott értéket adhat a már elérhető irodalomhoz képest, az a speciálisan a magyarországi kontextusra, s ezen belül is a politikatudományi alkalmazásra irányuló szöveg. Mivel a magyar politikatudományban (s talán általában a társadalomtudományokban) a kvantitatív szövegelemzés és szövegbányászat még gyerekcipőben jár, ezért reflektálnunk kellett arra a tényre, hogy a magyar nyelvű terminológia még sok esetben meglehetősségen képlékeny (ld. pl. a *sentiment analysis*szel kapcsolatos megjegyzéseinket a III.2. fejezetben). Itt követtük a hazai, kialakulóban lévő – de a társadalomtudományi alkalmazásoknál mindenképpen előrébb járó – számítógépes nyelvészeti irodalmat (ld. pl. Tikk, 2007 szövegbányászati bevezetését). Mivel ezen irodalom bizonyos alapfogalmakat konzekvensen magyarított (pl. az osztályozás mint a *classification* és a csoportosítás mint a *clustering* magyar megfelelője), így a mindenképpen hasznos közös terminológia kialakítása érdekében követtük megoldásaikat. A magyar nyelvű QTA-szaknyelv megerősítése érdekében a szövegben – ahol van elterjedt fordítás – a szakkifejezéseket magyarul, kurziválva jelöltük, míg angol megfelelőjüket a fejezetvégi szószedetekben közöltük.

***

A kötet közvetlen előzménye az MTA Társadalomtudományi Kutatóközpontjában (TK) az OTKA támogatásával elindított Comparative Agendas Project (CAP) kutatás. E projekt keretében Boda Zsolt projektvezető és Sebők Miklós kutatásvezető irányításával 2013-tól kezdődően folyt a magyar politikával kapcsolatos kvalitatív források vizsgálata kézi kódolás segítségével. Az újságcímlapok, törvények, parlamenti beszédek és költségvetési sorok többéves közpolitikai célú kódolása során érett meg bennünk az a felismerés, hogy a nemzetközi CAP projektben elvárt hosszú idősorok és a kutatási kérdésekhez kötött speciális feladatok megoldása érdekében szükség lesz a QTA-technikák alkalmazásá-

ra. Ebben megerősítettek a projekthez és más kutatóhálózatokhoz kapcsolódó nemzetközi konferenciákon szerzett tapasztalatok, melyek egyértelművé tették e módszerek gyakorlati hasznát a nagy méretű kutatási projektben.

A kötet gondolata végül a szerkesztő által a Bibó István Szakkollégiumban tartott bevezető jellegű kvalitatív adatelemző kurzuson született meg. A hallgatók visszajelzései alapján is hasznosnak tűnt, ha nemcsak előadásjegyzetek, de egy kerek, s egyben felhasználóbarátabb kézirat is segíti a kvalitatív adatelemzési technikák elsajátítását. A szerzőgárda jelentős részben szintén a CAP projektből, az említett kurzus résztvevőiből, illetve az MTA TK PTI munkatársaiból verbuválódott. Valamennyi szerző tudott saját kutatói tapasztalataira támaszkodni fejezetének megírásakor, ami reményeink szerint hozzájárulhat a felsőoktatási oktatók mellett a társadalomtudományi projektekben részt vevő kutatók munkájához is.

A kötet szerzői köszönettel tartoznak az MTA TK PTI vezetésének, Körösényi Andrásnak és Boda Zsoltnak a projekthez nyújtott erkölcsi és anyagi támogatásért, illetve a Nemzeti Kulturális Alapnak a könyv kiadásához nyújtott támogatásáért. Szintén köszönettel tartozunk Papp Zsófiának, valamint az MTA TK Politikatudományi Intézetében lefolytatott vita résztvevőinek, akik hasznos tanácsokkal segítették a kötet elkészültét. Külön köszönet illeti Kovács Borbálát, a kötet szakmai lektorát, aki alapvető fontosságú megjegyzésekkel, valamint számtalan kisebb pontosítással egyaránt hozzájárult a kézirat minőségének javításához. Szintén köszönjük Kovács László lektori munkáját. A fennmaradó hibákért természetesen csak a szerkesztő felelős.

I. A SZÖVEGEK TÁRSADALOMTUDOMÁNYI ELEMZÉSE

Sebők Miklós – Zorigt Burtejin – Zágoni Bella

I.1. A SZÖVEGEK TÁRSADALOMTUDOMÁNYI ELEMZÉSE: EGY ELMÉLETI ÁTTEKINTÉS

A fejezet bevezetést nyújt a szövegek társadalomtudományi elemzésének fogalmi alapjaiba. Ennek keretében először elhatároljuk a kvalitatív és kvantitatív adatforrásokat, illetve módszertant. Ezt követően vizsgálatunkat a szövegek társadalomtudományi elemzésére szűkítjük, és meghatározzuk e kutatási terület fő paradigmáit. A fejezet zárásaként – egy újabb tematikus szűkítés nyomán – meghatározzuk a kötet tárgyát: a szövegek tartalmának kvantitatív elemzését és a szövegbányászatot.

A modern társadalomtudományok egyik legfontosabb kutatási területe az *adatok elemzése*. Miközben az adatelemzésről létezik egy általános, sztereotipikus kép (bonyolult műveletek végrehajtása numerikus adatokat tartalmazó táblázatok vagy grafikonok segítségével), a valóságban e leegyszerűsítésnél sokkal színesebb a társadalomtudományok által felhasznált adatforrások sora és módszertani eszköztára. Az adatforrások esetében ugyanis nemcsak kvantitatív-mennyiségi, de kvalitatív-minőségi adatokat is vizsgálhatunk: ilyen minőségi adatforrás a szöveg vagy a kép. Módszertani értelemben pedig adatelemzésnek tekinthetőek bizonyos kvalitatív eljárások is a szokványos kvantitatív-statisztikai eszköztár mellett.

E fejezetben a *kvalitatív adatok,* és ezen belül is elsősorban a *szövegek* társadalomtudományi elemzésének elméleti-módszertani alapjait tárgyaljuk. Gondolatmenetünk dióhéjban a következő. A szövegek és más kvalitatív adatforrások (videók, hanganyagok stb.) ugyanúgy vizsgálhatóak kvalitatív, mint kvantitatív társadalomtudományi módszertannal, illetve ezek valamilyen keverékével. Írásunk tárgyát elsődlegesen a *szövegek* képezik, így a következő lépésben azt kell meghatároznunk, hogy hogyan közelít az ilyen dokumentumokhoz a kvalitatív és a kvantitatív módszertan.

A kvalitatív módszertant követő kutatók *egy része* a szövegre mint kvalitatív adatra tekint (más részük a szöveget interpretálni vagy „olvasni" akarja – erre a kettőségre visszatérünk). A kvalitatív adatként felfogott szöveg esetében nem törekszünk az adatforrások számszerű formátumra alakítására: fő tevékenységünk a szöveg *kódolása,* azaz elemeinek elkülönítése és csoportosítása. A kvan-

titatív érdeklődésű kutató ezzel szemben a szöveget statisztikai elemzésre alkalmazott formára hozva *keres vissza* vagy *nyer ki* információkat a szövegből.

Az alábbiakban először a társadalomtudományban használt adatforrásokkal, illetve módszertani megközelítésekkel foglalkozunk. Ezt követően vizsgálatunkat a *szövegek elemzési* módjaira szűkítjük, majd egy újabb tematikus korlátozás után meghatározzuk a könyv tárgyát: a *szövegek tartalmának* kvantitatív társadalomtudományi elemzési módszereit, a kvantitatív szövegelemzést és a szövegbányászatot.

A társadalomtudományok adatforrásai

A társadalomtudományokban *adat* alatt információk együttesét értjük (Schreiber, 2008: 185). Az adatforrások jellegük szerint lehetnek kvantitatívak-mennyiségiek, illetve kvalitatívak-minőségiek. A mennyiségi és minőségi adattípusok elválasztása ugyanakkor korántsem magától értetődő. Jól jellemzi e problémát a következő példa, mely egy labdarúgó-mérkőzés beszámolóját dolgozza fel (Winter, 1991).

I.1.1. táblázat – Példa a mennyiségi és minőségi adattípusok elválasztására

Wimbledon 0 – Liverpool 0	Izgalmasabb dolgok történtek a parkolóban, mint a pályán.

A táblázat bal oldali oszlopában láthatjuk a mérkőzés eredményét, ami egy mennyiségi információt nyújt számunkra, míg jobb oldali oszlopában egy minőségi jelentéssel bíró véleményt olvashatunk. Az, hogy a két információ közül melyik adat bír számunkra fontosabb információtartalommal, nem máson, mint érdeklődésünkön múlik. Ha például a klub edzői karának vagyunk a tagja, akkor valószínűleg jobban fog érdekelni minket a mérkőzés eredménye. Laikusként vagy kívülállóként viszont érdekesebb információ lehet számunkra, hogy milyen volt a mérkőzést körülvevő hangulat (Dey, 1993). Tehát a kvantitatív, illetve kvalitatív adatforrásoknak nincs „önértéke": relatív hasznuk leginkább az érdeklődésünktől („kutatói felfogásunktól") és konkrét kutatási kérdésünktől függ.

Ahogy fentebb meghatároztuk, az adat kifejezés általánosságban a mennyiségi vagy minőségi változók felvett értékeinek összességére utal. A kvantitatív adatok mennyiségekre vonatkoznak, vagyis numerikus („számszerű") információt hordoznak. Hány törvény születik egy parlamenti ciklus során? Mekkora a női képviselők parlamenti aránya? Az ilyen és ehhez hasonló kérdések megválaszolása az adatok minősége (így a törvények tartalma; a nők politikában be-

töltött tényleges súlya) kapcsán nem szolgál többletinformációval. Ez utóbbiak vizsgálatához át kell fogalmazni kutatási kérdéseinket is: „milyen a törvények közpolitikai tartalma?"; vagy „a női képviselők eltérő ügyeket képviselnek-e, mint a férfi képviselők?". Ezek már sajátosan a kvalitatív adatok elemzéséhez kapcsolódó kérdések, melyeket nem tudunk, vagy – módszertani okok miatt – nem tartunk szerencsésnek mennyiségi mutatókkal megválaszolni.

Kutatásunk tervezése során fontos azt is mérlegelni, hogy a kvalitatív adatok „beszerzése" rendszerint jobban terhelt módszertani problémákkal, mint a kvantitatív adatoké, még akkor is, ha ugyanolyan forrásból (pl. egy kérdőívből) szerzik be őket. Ennek egyik eleme, hogy míg a kvantitatív adatok jól lehatárolható típusokba sorolhatóak a mérési skála alapján (mint pl. nominális vagy ordinális), addig a minőségi adatok esetében egy hasonlóan elemi kategorizálási szinten (pl. szóbeli-verbális/nem verbális) még a konkrét kutatástól távoli absztrakciókkal (és általában egy sokkal szélesebb adattípus-palettával) dolgozunk.

A kvalitatív adatok elemzésének további módszertani problémáját az adatok strukturálatlansága adja, ugyanis a kvalitatív (minőségi) adatok – szemben a kvantitatív (mennyiségi) adatokkal – nyers formájukban még nem alkalmasak arra, hogy statisztikai elemzéseket végezzünk velük. A kvalitatív adatelemzés ráadásul elsősorban az adatforrás explicit (avagy „manifeszt") tartalmát vizsgálja, magyarán a szöveget szó szerint értelmezi. A módszer nehezebben kezeli a rejtett jelentéseket vagy az átvitt értelmek értelmezését. Szintén nem képes még teljesen kezelni az iróniát vagy a szarkazmust, gyakori élőnyelvi előfordulásuk ellenére. Gondoljunk csak az interpellációkban, közösségi oldalakon írt bejegyzésekben gyakran megjelenő gúnyos hangnemre, amelyeknek a szöveg értelmezése és magyarázása szempontjából jelentősége lehet. Harmadszor a kvalitatív kutatások gyakori kritikája, hogy átláthatatlanok és szubjektívek lehetnek. Ezért törekedni kell arra, hogy a kvalitatív adatelemzés átlátható, nyomon követhető és ellenőrizhető módon történjen. Így az ilyen típusú adatok elemzése a kutatók részéről nagyfokú figyelmet és érzékenységet kíván az adatok iránt.

A szövegek társadalomtudományi elemzése: kutatási stratégiák és módszertan

Az *adat* fogalmának definiálása után a következő feladat az *adatelemzés* tevékenységének leírása, mely egy többlépéses folyamat. Amint megtörtént az adatgyűjtés, a kutató adatait a kézi vagy gépi elemzésre alkalmas formára hozza (Schreiber, 2008: 185). Ez a forma lehet egy táblázat vagy jegyzetlapok gyűjteménye, a lényeg, hogy egyszerűsítsék a kutató munkafolyamatát. Nyers

adathalmazát a kvalitatív módszertant követő kutató *memókban,* avagy kutatási naplóban írja le, illetve kódolhatja egyes elemeit. Ezt követően empirikus anyagát „párbeszédbe hozza" a szakirodalommal és a kutatás elméleti fogalomkészletével (Van den Hoonard – Van den Hoonard, 2008: 186).

A konkrét elemzési eljárások részben attól is függenek, hogy hogyan tekint a kutató a vizsgált kvalitatív adatokra, illetve szövegre. Többféle kategorizálási rendszer létezik ennek kapcsán, melyek közül itt röviden kettőt ismertetünk. A szakirodalomban rendre használják a *dokumentum* fogalmát (Prior, 2008: 230). Ennek főbb típusai a következők: a leírt szövegek, a vizuális-, illetve hanganyagok, térképek, filmek vagy fotók, valamint mindezek interneten megfigyelhető „folyamai". A dokumentumoktól eltérő tulajdonsággal rendelkezik a *beszélgetés*, melynek kontextusfüggő és dinamikus (időben kibomló) elemzésére külön részterület jött létre.

A megfigyelés és tapasztalás útján „felvehető" minőségi adatokat másfelől gyakran az adatgyűjtés módszertana szerint csoportosítják (kérdőívek, fókuszcsoportok, interjúk, összegyűjtött webes tartalmak). Részben ezen adatforrásokhoz is kapcsolódik, hogy milyen kutatási stratégiát illetve „szövegfelfogást" választanak a kutatók vizsgálatukhoz.

Általánosságban négy nagy ilyen társadalomtudományi szövegértelmezést különíthetünk el eltérő ismeretelméleti feltevéseik és felhasznált módszertanuk alapján: a szó-tér modelleket, valamint a narratív, retorikai és diszkurzív elemzéseket (Bauer et al., 2014: 25). E kutatási stratégiák részletes elemzésére nincs hely, így elsődlegesen a számunkra legfontosabb szó-tér modellek és a másik három megközelítés eltéréseivel foglalkozunk. A szövegek tartalmát a benne szereplő szavak mennyiségével és pozíciójával vizsgáló szó-tér modellek a szövegeket elsősorban (kutatási céljaik érdekében) *használják,* nem pedig *olvassák* (i. m.: 8–9). Megközelítésük strukturális (a szöveg szerkezetére irányul) és nem interpretatív jellegű. Az ilyen klasszikus *tartalomelemzés* a szöveg elemeit kategóriákba sorolja (azaz „lekódolja"), melyek már megszámolható értékekkel rendelkező, s ezért kvantitatív módszerekkel elemezhető változókat adnak.

Továbbhaladva kutatási témánk pontosítása felé, a tartalomelemzés általános kategóriáján belül még igen eltérő tudományfilozófiai és kutatásmetodológiai hátterű megközelítéseket határolhatunk el. Végezhetjük elemzésünket *induktív* módon, mely során fogalmi kategóriákat azonosítunk közvetlenül az empirikus anyagból kiindulva vagy *deduktívan*, amely során a nyers adatokat előre meghatározott kategóriákba soroljuk, amelyeket a szakirodalomból, ennek elméleti kereteiből merítünk. A harmadik lehetőség a fent említett módszerek együttes alkalmazása, mely esetben az egyes felfedezési módszereknek az elemzés különböző szakaszaiban lesz szerepe. E hibrid megoldás valószínűleg azért is kezd egyre népszerűbbé válni, mert egyszerre veszi igénybe a kutató elméleti érzékenységét és a jelentéstartalom önálló meghatározását, valamint

az összehasonlíthatóságot más kutatásokkal a sztenderdizált kódkönyveknek megfelelően.

Rátérve már az egyes analitikus stratégiák tartalmára a – jellemzően (poszt) pozitivista ihletésű, *deduktív* megközelítés esetében az egyik lehetőség egy előre meghatározott szótár (azaz a szövegben előforduló kulcsszavak és az elméleti kategóriák közötti átváltási kulcs) alkalmazása. A magyar politikatudományban egy példa erre a nemzetközi sztenderdek szerinti magyar adatbázisokat fejlesztő Comparative Agendas Project (CAP – cap.tk.mta.hu), mely előzetesen meghatározott közpolitikai témakategóriákba sorolja be többek között a törvényeket, interpellációkat, illetve médiamegjelenéseket.

Ezzel a stratégiával szemben az *induktív* megközelítés, ezen belül is az ún. *megalapozott elméleti* megközelítés állhat. Erre a *szózsák* (ld. II.1. fejezet) vagy a klaszteralapú kategorizálás hozható fel példának (III.3. fejezet). A szózsák módszerével az egyes szavak, kifejezések gyakoriságát vizsgálhatjuk a korpuszban, melyből további következtetéseket vonhatunk le. Példa lehet erre egy olyan kutatás, mely azt vizsgálja, milyen *témákra*, ügyekre helyezi a hangsúlyt beszédeiben két versengő miniszterelnök-jelölt. A beszédeket e módszerrel vizsgálva látni fogjuk, hogy melyek azok a szavak, kifejezések, amelyeket leggyakrabban használnak, melyből következtetéseket vonhatunk le például politikai stratégiájukra nézve.

A klaszteralapú induktív kategorizálás egy másik logikát követ. Tegyük fel, hogy van egy adatbázisunk az összes magyarországi törvényjavaslattal 1990 és 2014 között, mely tartalmazza a benyújtók bizottsági tagságát és a javaslat közpolitikai témáját. A klaszterek statisztikai vizsgálatával egyértelművé válhat, hogy bizonyos bizottságok és bizonyos témák együtt mozognak, s ezáltal meghatározhatóak a több dimenzióban is megjelenő közpolitikai alrendszerek (pl. az oktatási alrendszer), anélkül, hogy előzetes feltevésekkel rendelkeztünk volna ezek listájáról vagy jellemzőiről. Ugyanezen kérdés vizsgálható ugyanakkor a deduktív felfedezési elv mentén is, mely esetben a két megközelítés eredményének összehasonlítása tovább gazdagíthatja elemzésünket.

Az adatforrások, kutatási stratégiák és metodológiák áttekintése után végezetül különbséget kell tennünk a módszertan, illetve az alkalmazott kutatási technológia között. Az adatelemzés történhet kvalitatív vagy kvantitatív módszertan segítségével (ez az adatforrások kvalitatív vagy kvantitatív jellegétől független!). Másfelől pedig technológiai értelemben történhet kézzel, gépi támogatással vagy automatizáltan. E két dimenzióban meghatározható kutatási eljárásokat ad példát az I.1.2. táblázat.

I.1.2. táblázat – Az adatelemzés módszertanának és technológiájának függetlensége

Kézi elemzés		Az adatelemzés technológiája		
		Gépi támogatású elemzés	Gépi elemzés	
Az adatelemzés módszertana	Kvantitatív	-	Felügyelt gépi tanulás	Felügyelet nélküli gépi tanulás
	Kvalitatív	Pl. Interjúk kézi diszkurzív elemzése (kódolással vagy anélkül)	Pl. Interjúk diszkurzív elemzése CAQDAS segítséggel	„Automatikus kódolás" CAQDAS szoftverben

A szövegek társadalomtudományi elemzési módjainak feltérképezése során az egyik legfontosabb lépés a kvalitatív és kvantitatív módszertan megkülönböztetése. A kvalitatív szövegelemzéssel nagy átfedést mutat a *kvalitatív tartalomelemzés,* illetve a *számítógéppel támogatott szövegelemzés* (Popping, 2000) területe (így ezeket a kötetben egy kutatási területként fogjuk fel). A *kvantitatív szövegelemzést* (Mehl, 2006) más néven szokták *szövegbányászatnak,* szöveganalitikának (ill. számítógépes analitikának, vö. üzleti analitika) a „szöveg mint adatnak" (Grimmer – Stewart, 2013; Slapin – Proksch, 2008: 709), vagy „szavak mint adatnak" (Slapin – Proksch, 2014; Laver et al., 2003) is nevezni.

A két terület elválasztásának kulcsa a szöveg mint adat jellegének eltérő felfogása. Az egyik esetben a szöveget jegyzetekkel és kódokkal ugyan ellátjuk, de magán a szövegen nem végzünk beavatkozást. A másik esetben ugyanakkor az eredendően kvalitatív információforrást kvantitatív módon leképezzük (pl. egy szövegben a szavak gyakoriságát numerikusan megmutató táblázatban), azaz kvantifikáljuk.

Mindezek alapján nézzük a táblázat módszertan-technológia kombinációit! A kvantitatív módszertan alapján történő tartalomelemzéshez felhasználhatunk a kutató által „felügyelt" gépi segítséget, melynek során a számítógépen futtatott algoritmusok közötti választás folyamatos segítséget igényel (ld. IV.1. fejezet). Ennek alternatívája a teljesen automatizált (felügyelet nélküli) kódolás (IV.2. fejezet). Áttérve a kvalitatív módszertanra, egy szöveg tartalmának elemzését elvégezhetjük kézzel, a folyamatot gyorsító és áttekinthetőbbé tevő gépi támogatással vagy bizonyos szabályok előzetes megadása mellett automatikus (de még nem a szöveg kvantifikálásával, numerikus formára alakításával elvégzett) kódolással.

Az így meghatározott egyes adatelemzési módszertanok és technológiák megfelelő együttesének kiválasztása elsődlegesen a kutatási kérdéstől és a ren-

delkezésre álló (vagy választott) adatforrásoktól függ. A kötet első részében (azaz a jelen és a következő fejezetben) általában a szövegek társadalomtudományi elemzésével foglalkozunk. A kötet további négy részében ugyanakkor fókuszunkat már a kvantitatív szövegelemzésre és szövegbányászatra szűkítjük. E választást indokolja, hogy a kvantitatív szövegelemzésben rejlő potenciál még messze nincs kihasználva a társadalomtudományokban. E potenciál egyik legfontosabb forrását a „Big Data" jelenségéhez kapcsolódó gépi adatelemzési vagy adatfeldolgozási technikák és programok jelentik. Az olyan sztenderdizált QTA-feladatok, mint a *szózsák,* a *névelem-felismerés,* a *klasszifikáció* vagy a *klaszterezés* mind új dimenziót nyernek a gépi támogatással mind a feldolgozható adatok mennyisége, mind az eljárás *megbízhatósága* és *belső/külső érvényessége* kapcsán.

Ellenőrző kérdések

- Ha arra lenne kíváncsi, hogy a nemzetiségi háttérrel rendelkező vagy a fogyatékossággal élő országgyűlési képviselők inkább képviselik-e a nemzetiségek, illetve a fogyatékossággal élők érdekeit, mint a magukat a két csoporthoz nem soroló választókét, kvantitatív vagy kvalitatív módszerrel tenné ezt? Miért?
- Ha azt szeretné megvizsgálni, hogy az országgyűlésben lévő női képviselők száma befolyásolja-e a nőkkel kapcsolatos ügyek parlamenti napirenden való megjelenésének gyakoriságát, kvantitatív vagy kvalitatív módszerrel tenné ezt? Miért?
- Mondjon példát kvantitatív adatok kvantitatív elemzésére!
- Mondjon példát kvalitatív adatok kvantitatív elemzésére!
- Milyen induktív megközelítést alkalmazó módszerek határozhatóak meg a kvantitatív szövegelemzés terén?
- Milyen deduktív megközelítést alkalmazó módszerek határozhatóak meg a kvantítatív szövegelemzés terén?
- Milyen módszertani előnyei, illetve hátrányai vannak a kvantitatív, illetve a kvalitatív szövegelemzésnek?

Szószedet

Magyar	Angol
Adatelemzés	Data analysis
Adatfelvétel	Data collection
Adatfelvételi módszertan	Data collection method
Belső érvényesség	Internal validity
Beszélgetéselemzés, társalgáselemzés	Conversation analysis (CA)
Deduktív megközelítés	Deductive approach
Dokumentum	Document
Elemzési módszertan	Method of analysis
Induktív megközelítés	Inductive approach
Kategorizálási feladatok	Categorizing tasks
Klaszterezés	Clustering
Külső érvényesség	External validity
Kvalitatív kutatásmódszertan	Qualitative research method
Kvalitatív tartalomelemzés	Qualitative content analysis
Kvantitatív szövegelemzés	Quantitative text analysis
Megalapozott elmélet	Grounded theory
Megbízhatóság	Reliability
Névelem-felismerés	Named entity recognition (NER)
Statisztikai gépi tanulás	Statistical machine learning
Számítógéppel támogatott szövegelemzés	Computer-assisted text analysis
Szózsák	Bag of words
Szöveg mint adat	Text as data
Szövegbányászat	Text mining

Magyar	Angol
Témakör, szövegtartomány	Domain
Ügy	Issue

Ajánlott irodalom

A társadalomtudományi kutatás módszertana kapcsán a legáltalánosabban használt szöveg Babbie (2008) klasszikusnak számító műve. Ha a számítógéppel támogatott szövegelemzésben szeretnénk elmélyedni, Popping (2000) kiváló kiindulási pontot jelent. A kvantitatív szövegelemzésben Mehl (2006), továbbá Grimmer és Stewart (2013) segítenek eligazodni. Ha a szövegbányászatot szeretnénk részletesebben megismerni a Tikk (2007) által szerkesztett magyar nyelvű mű jelenthet igazodási pontot.

Molnár Csaba

I.2. SZÖVEGKÓDOLÁS A GYAKORLATBAN: KÉZI, GÉPPEL TÁMOGATOTT ÉS GÉPI MEGOLDÁSOK

A fejezetben a kézi, a géppel támogatott, valamint a gépi kódolási módszertan közötti választáshoz adunk szempontokat. A kézi kódolás legfőbb előnye az érvényességében, a szöveg lehető legmélyebb értelmezésében rejlik, ugyanakkor a módszer lassú, potenciálisan inkonzisztens, és komolyan képzett kutatókat igényel. A géppel támogatott módszerek a kézi kódolás inkonzisztenciájára nyújtanak részleges gyógyírt, azonban a különböző programok kialakítása más és más kutatások elvégzésére teszi alkalmassá azokat. A gépi kódolás a leggyorsabb eljárásmód, emellett a legkövetkezetesebb is. Hátránya ugyanakkor, hogy alkalmazása speciális informatikai szakértelmet igényel, a szöveg gépi értelmezése leegyszerűsítő jellegű lehet, illetve a használt algoritmus eredményekre gyakorolt befolyása miatt validálási problémák is felmerülhetnek.

Az előző fejezetben megismerkedtünk a kvalitatív adatok elemzésének főbb típusaival és módszertani alapelveivel. Nem tértünk ki ugyanakkor azokra a döntési elvekre, melyekkel kiválaszthatjuk a kutatásunknak leginkább megfelelő módszertant. E fejezet amellett, hogy bemutatja a kézi, a géppel támogatott és a gépi kódolás legfontosabb előnyeit és hátrányait, egy konkrét kutatás példáján keresztül feltárja e fontos kutatásszervezési döntés lépéseit, valamint a mérlegelendő szempontok körét is. A számítógépes szoftverek fejlődése ellenére a kvalitatív kutatási feladatok (így pl. a tartalomelemzés) elvégzése továbbra is feltételezi a kutatók folyamatos módszertani döntéseit. Sőt az élő nyelv gépi nyelvre történő leképezése egyes speciális feladatok kapcsán még kifejezetten gyerekcipőben jár (ld. pl. az irónia érzékelését egy fórumbejegyzésben – vö. III.2. fejezet). Ennek fényében hasznos ismernünk az egyes kódolási eljárásokat, illetve ezek kutatás-módszertani előnyeit és hátrányait. Az alábbiakban a gyakorlati kutatásokat segítendő ezeket egy nemzetközi és egy magyar példa segítségével tárgyaljuk.

Mikor kódoljunk kézzel, és mikor használjunk gépi eljárásokat?

A politikatudományi kutatások sok esetben nagy *szövegkorpuszokkal* dolgoznak, melyek tartalma még az anyagot behatóan ismerő kutató számára is nehezen áttekinthető. A probléma egyik megoldása az, ha a szöveg különböző szakaszait címkékkel látjuk el, azaz lekódoljuk. Fontos leszögezni, hogy ezzel önmagában nem alakul át a szerzett ismeretanyag kvantitatív módon elemezhetővé: a kvalitatív politikatudományi kutatások egyik meghatározó módszertana a kódolási eljárás. A kódok megalkotásának alapvetően két irányát különböztethetjük meg. Az *induktív kódolás* során kutatásunk közben alakítjuk ki kódjainkat, az előzetes kódkönyvet folyamatosan alakítjuk. A *deduktív kódolás* eleve adott, korábban elkészített, a kutatás közben nem változtatható kódkönyv alapján történik (Zhang – Wildemuth, 2009: 310).

Ezen definíciós bevezetés után módszertani áttekintésünket kezdjük a három adatfelvételi technika meghatározásával. A *kézi kódolás* során a kutató a korpusz minden egyes szövegét elolvassa, és azokhoz saját értelmezése szerint (a nyomtatott lap szélén vagy egy táblázatban) kódokat rendel. A *géppel támogatott kódolás*ok során egy speciális program felhasználói felületén belül látja el a kódoló megfelelő címkékkel a szöveg általa meghatározott hosszúságú szakaszait. Itt tehát elképzelhető, hogy egyetlen szó kap egy kódot, de az is, hogy akár több oldallal történik ugyanez (például egy politikai beszédben előfordulhat, hogy egy közpolitikai témakört a szónok csupán néhány szóban érint, míg más ügyekről hosszasan értekezik).

A *gépi kódolás* a harmadik vizsgált módszer, melynek két fő típusa van: az *osztályozás (klasszifikáció)* és a *csoportosítás (klaszterezés)*. Az előbbi csoportba sorolhatóak a szótáralapú és a felügyelt tanulási eljárások, melyek alkalmazása során a kutatónak ki kell dolgoznia egy háttéranyagot (ami lehet egy fontos szavakat magába foglaló szószedet vagy egy lekódolt, a korpuszból vett minta is), mely alapján a program csoportokba rendezi a korpusz teljes állományát. A csoportosítás módszereinek alkalmazása esetén a kutató feladata a megfelelő csoportosító mechanizmus kiválasztása, mely alapján egy program elvégzi a szövegek különböző kategóriákba sorolását. A hasonló szövegeket tömörítő csoportok elnevezésének lépése csak ezután következik. Mindezek fényében az I.2.1. táblázat ismerteti a három módszercsoport előnyeit és hátrányait (ezeket a példák kapcsán részletesebben is demonstráljuk).

I.2.1. táblázat – A három módszer előnyei és hátrányai

	Előnyök	Hátrányok
Kézi kódolás	A gépinél nagyobb eséllyel ismeri fel a költői képeket és alakzatokat, értelmezi helyesen a szövegkörnyezetet és a célzásokat Kutatási kérdéstől függően mélyebb elemzésre képes	A megbízható eredményekhez egynél több kódoló munkavégzésére lehet szükség Jelentős előerő bevonását igényli, relatíve lassú Inkonzisztenciaprobléma (kódolások egyezése) Tapasztalt, kiképzett kutatók szükségesek hozzá
Géppel támogatott kódolás	A gépinél nagyobb eséllyel ismeri fel a költői képeket és alakzatokat, értelmezi helyesen a szövegkörnyezetet és a célzásokat Kutatási kérdéstől függően mélyebb elemzésre képes, áttekinthetőbb Alkalmasabb nagyobb korpuszok feldolgozására Javíthatja a konzisztenciát Hasznos kiegészítő funkciók Kvantitatív elemzésekkel könnyebben egybekapcsolható Csoportmunka szervezésére alkalmasabb	A megbízható eredményekhez egynél több kódoló munkavégzésére lehet szükség Jelentős előerő bevonását igényli, relatíve lassú Túlzott sztenderdizáláshoz vezethet A programok használata betanulási idővel jár A különböző programok eltérő kutatásokban használhatóak inkább
Gépi kódolás	Jól alkalmazható homogén szövegek elemzésére (osztályozás) Heterogén, ismeretlen korpuszok feldolgozására is alkalmas (csoportosítás) Konzisztens Gyors Kevés vagy viszonylag kevés humán kapacitást igényel	A nyelv leegyszerűsítő értelmezése Kevésbé alkalmas értelmezési feladatok elvégzésére Az eredmény nem független a használt algoritmustól Szükség lehet szótár és/vagy tanítóhalmaz elkészítésére A programok használata betanulási idővel jár

Ezen áttekintés után nézzük részletesebben is az egyes eljárásokat!

Kézi kódolás

Noha a nyelvről való gondolkodás, az implicit nyelvfilozófia már a szókratikus korban megjelent (gondoljunk csak Platón Kratüloszára vagy Arisztotelész Herméneutikájára), önálló filozófiai ággá csupán a 18. században vált. Az igazi fordulatot a kutatási terület történelmében az analitikus filozófia megjelenése hozta, mely a nyelv szerepét felértékelte a világ megismerésének folyamatában (Ambrus, 2007: 1065). A paradigma alapító atyja, Gottlob Frege, a kezdetektől állította, hogy a megértés alapja a korábbi felfogással szemben nem a szó, hanem a mondat kell legyen, hiszen előbbi izoláltan értelmezhetetlen (i. m.: 1067).

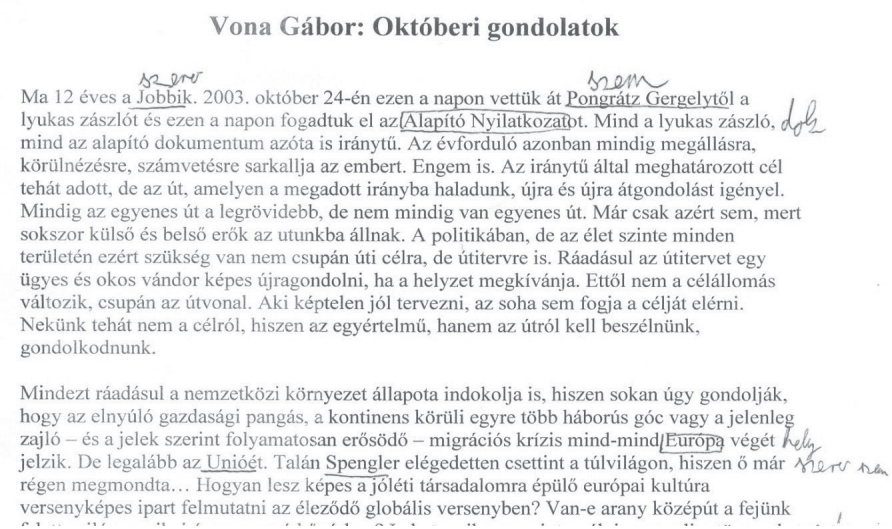
Az irányzat első generációjának meggyőződése szerint a hétköznapi nyelv nem igazán érthető meg annak – felfogásuk szerint – zavaros volta miatt, olyannyira, hogy emiatt szükségesnek tartották egy ideális nyelv megalkotását.

Noha a később felbukkanó hétköznapinyelv-filozófia vitatta ezeket a sarkos kijelentéseket (i. m.: 1127), kétségtelen, hogy nem csupán a beszélt, hanem az írott nyelv jelentése is számtalan nehézséget rejt magában. A számítógépes szövegelemzés jelenlegi ismereteink szerint óhatatlanul leegyszerűsítéssel, modellezéssel jár együtt, ami az analitikus nyelvfilozófia felfogása szerint az eleve ingoványos talajon álló szövegértelmezés során könnyen juthat pontatlan következtetésekre. Mindezek következtében egészen addig, amíg nem sikerül kifejleszteni olyan eljárásokat, melyek az emberi elme módjára tudnak szövegeket értelmezni, a gépi megértés, azaz a szerző és az olvasó közötti azonosulás (Farkas – Kelemen, 2007: 1279), lehetőségei korlátozottnak tekinthetőek.

Erre a filozófiai kitérőre azért volt szükségünk, hogy bemutassuk, miért is foglalhatja el a kézi kódolás eljárása az *„aranystandard"* szerepkörét a szövegkódolási módszerek között (ld. Albaugh et al., 2014). Tévedés lenne persze azt hinnünk, hogy a leegyszerűsítő gépi kódolás elvetésével máris garantált számunkra a hibátlan munkavégzés lehetősége: a géppel támogatott és a gépi kódolás nem a trehány kutatók módszere, ahogy a kézi kódoláshoz sem csupán az igényes, helyes eredményre jutó vizsgálatok folyamodnak. Az élő személyek által végzett szövegelemzés sem vezet minden esetben eredményre: noha a számítógépeknél nagyobb valószínűséggel értik meg a szövegkörnyezetet, a költői képeket és alakzatokat, a célzásokat, könnyebben észrevehetik az eltérő szövegkörnyezetek és szövegösszefüggések különbségeit, ők sem dolgoznak hibátlanul.

A kutatói elfogultságok és tévedések lehetősége fennáll: egy közpolitikai kódkönyv alapján dolgozó kódolóval könnyen megeshet, hogy félreértelmez egy felszólalást, esetleg hiányos ismeretei miatt rosszul sorol be egy tételt, netán erőltetetten nem a helyes csoportba helyez el egy szöveget. Egy kézi kódolással dolgozó kutató a gépeknél sokkal precízebb, mélyebb elemzésre képes, ugyanakkor lehet sokkal inkonzisztensebb is. Ahogy kutatásában halad előre, megváltoznak előfeltevései, egy idő után a hasonló szövegeket elkezdi tendenciaszerűen kódolni, ami magában rejti a tévedés növekvő kockázatát. A lassú és nehézkesen haladó kézi kódolás közben változhatnak a kódolási szokások, és igen komoly munkafegyelem, illetve szigorúan lefektetett kutatási elvek kellenek, hogy ez ne vezessen az adatminőség romlásához. Az I.2.1. ábra egy politikusi írás kézi kódolásának eredményét mutatja (a kódolás célja a tulajdonnevek azonosítása az írás témájának meghatározása során).

I.2.1. ábra – Példa a kézi kódolás alkalmazására



Vona Gábor: Októberi gondolatok

Ma 12 éves a Jobbik. 2003. október 24-én ezen a napon vettük át Pongrátz Gergelytől a lyukas zászlót és ezen a napon fogadtuk el az Alapító Nyilatkozatot. Mind a lyukas zászló, mind az alapító dokumentum azóta is iránytű. Az évforduló azonban mindig megállásra, körülnézésre, számvetésre sarkallja az embert. Engem is. Az iránytű által meghatározott cél tehát adott, de az út, amelyen a megadott irányba haladunk, újra és újra átgondolást igényel. Mindig az egyenes út a legrövidebb, de nem mindig van egyenes út. Már csak azért sem, mert sokszor külső és belső erők az utunkba állnak. A politikában, de az élet szinte minden területén ezért szükség van nem csupán úti célra, de útitervre is. Ráadásul az útitervet egy ügyes és okos vándor képes újragondolni, ha a helyzet megkívánja. Ettől nem a célállomás változik, csupán az útvonal. Aki képtelen jól tervezni, az soha sem fogja a célját elérni. Nekünk tehát nem a célról, hiszen az egyértelmű, hanem az útról kell beszélnünk, gondolkodnunk.

Mindezt ráadásul a nemzetközi környezet állapota indokolja is, hiszen sokan úgy gondolják, hogy az elnyúló gazdasági pangás, a kontinens körüli egyre több háborús góc vagy a jelenleg zajló – és a jelek szerint folyamatosan erősödő – migrációs krízis mind-mind Európa végét jelzik. De legalább az Unióét. Talán Spengler elégedetten csettint a túlvilágon, hiszen ő már régen megmondta… Hogyan lesz képes a jóléti társadalomra épülő európai kultúra versenyképes ipart felmutatni az éleződő globális versenyben? Van-e arany középút a fejünk

A kézi kódolás további nehézségét jelenti, hogy gyakran menet közben ismerjük fel a kódkönyv hibáit, bővítjük azt, így ezek visszamenőleges kijavítása a már lekódolt anyagban mindenképpen nehézkes. E nehézségek leküzdésében segítséget nyújt, a kódolás megbízhatóságát növeli, ha több kutató párhuzamosan kódolja ugyanazon szövegeket (illetve ezt egy kutató is elvégezheti, bár sokkal időigényesebben, ld. Seale, 1999: 467–469). Ennek híján nehezebben tekinthető megbízhatónak a kézi kódolás eredménye. Jól látható tehát, hogy az „aranystandard" előállítása semmivel sem egyszerűbb az arany kibányászásánál: rengeteg erőforrást igényel megvalósítása.

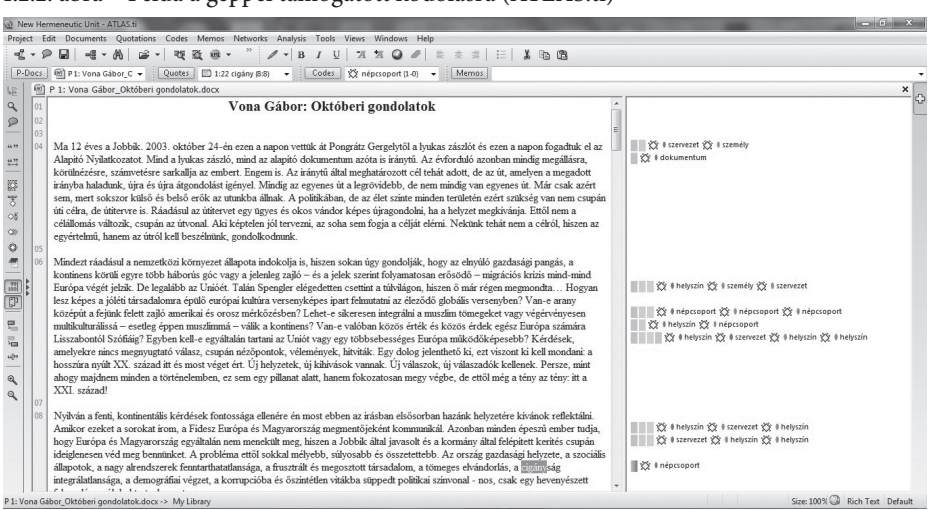
A kézi kódolás utolsó nehézsége a kódolók személyében rejlik. Az elfogultságok és tévedések valószínűségének csökkentésére, a megbízható munkavégzésük lehetővé tételére csupán nagyon mélyreható és alapos kiképzésük után nyílhat lehetőség, illetve a kutatóknak folyamatosan reflexíven kell viszonyulniuk kutatásuk tárgyához, ellenőrizniük, s ha szükséges, felül kell vizsgálniuk magukat. Ennek híján egyéni meglátásaik kikerülhetetlenül tévutakra (vagy legalábbis egymástól eltérő utakra) vezetik őket, ami kutatásuk eredményeinek értékét jelentősen csökkentheti. Persze a megfelelő kiképzés sem nyújt feltétlenül optimális eredményt: senkitől sem várható el, hogy Németh László-i minőségben értelmezze például egy irodalmi korpusz szövegeit, ugyanakkor bizonyos hermeneutikai alapismeretek feltétlenül szükségesek.

Géppel támogatott kódolás

A géppel támogatott kódolás képes a kézi kódolás egyes előnyeinek megőrzése mellett kiaknázni a gépi módszerek konzisztenciájában és gyorsaságában rejlő lehetőségeket. Ehhez valamilyen kvalitatív adatelemzésre használható számítógépes programot érdemes kiválasztani (pl. ATLAS.ti, MAXqda, NVivo – Maietta, 2008: 107). Mivel a jelen kötet a továbbiakban elsősorban a gépi kódolási technikákra összpontosít, röviden érdemes áttekintenünk az ilyen programok működésének alapelveit (az egyes termékek konkrét használatának ismertetésére ugyanakkor nem vállalkozhatunk).

Minden idesorolható szoftver rendelkezik három alapvető „rendszer" feldolgozására alkalmas elemmel. Ez a három rendszer az eredeti szövegeket megőrző *dokumentumrendszer,* a hozzá fűzött megjegyzéseket, az esetleges átalakításokat tartalmazó *emlékeztetőrendszer,* valamint az adatokat két szinten (a szövegek attribútumai és a szövegrészletek kódjai szintjén) rendszerező *kategóriarendszer* (Maietta, 2008: 103). Míg a dokumentumrendszer tartalmán a program segítségével nem lehet változtatni (így a szöveg eredeti integritása is fennmarad), az emlékeztetőrendszer és a kategóriarendszer könnyen, gyorsan kezelhető és módosítható. Az I.2.2. ábra példájának középső paneljén látható az eredeti, elemzésre váró politikai szöveg, míg jobb oldalán az egyes szövegrészletekhez társított kódok.

I.2.2. ábra – Példa a géppel támogatott kódolásra (ATLAS.ti)



A számítógéppel támogatott kódolás megőrzi a kézi kódolás legfőbb előnyét, a mélyebb összefüggések feltárásának lehetőségét (Bhowmick, 2006: 4). A módszer fő előnyei a nagyobb áttekinthetőségből fakadnak. A kódok könnyen összegyűjthetőek, rendszerezhetőek, a lekódolt szövegrészek kódonként csoportosíthatóak, így sokkal egyszerűbb és pontosabb lehet a kódonkénti elemzésünk. Emiatt jobban sztenderdizálható és konzisztensebb a munkafolyamat, a kódok könnyebben módosíthatóak, illetve áttekinthetőségükből fakadóan csoportmunka szervezésére is sokkal alkalmasabbak. A programok számtalan hasznos kiegészítő funkcióval rendelkeznek, melyek miatt sokkal könnyebb nagyobb korpuszt feldolgozni, mint kézi kódolás esetén. A munkavégzés így gyorsabb, valamint kvantitatív elemzésekkel is könnyen egybekapcsolható. A kódolást elősegítő eszközökön túl egyéb funkciók is a rendelkezésünkre állnak, melyek a kódolás áttekinthetőségét segítik elő. Ilyen a szövegrészletekhez rendelve, valamint külön is vizsgálható emlékeztető, a kód, a szűrő, valamint a kódok metszeteit feltáró együttes előfordulás eszköze. Az eredmények felhasználására szolgálnak a különböző ábrázoló eszközök (térképek, modellek, folyamatábrák, hálózatok), a kvalitatív és a kvantitatív kutatások összekapcsolását lehetővé tevő integráló eszközök, valamint a csoportos kódolást elősegítő eszközök (például megjegyzések – ld. Maietta, 2008: 106–107). Ilyen a kézi kódolás szövegkiemelésére emlékeztető *vizuális segítség,* az azonosan lekódolt szövegrészleteket összegyűjtő *gyűjtőeszköz,* a megjegyzéseket kódkönyvszerűen összeszedő *szakaszcímkék,* illetve a megjelölt részletek ábrákba rendezését lehetővé tevő *adatprofil-gyűjtemény* is (i. m.: 104–105).

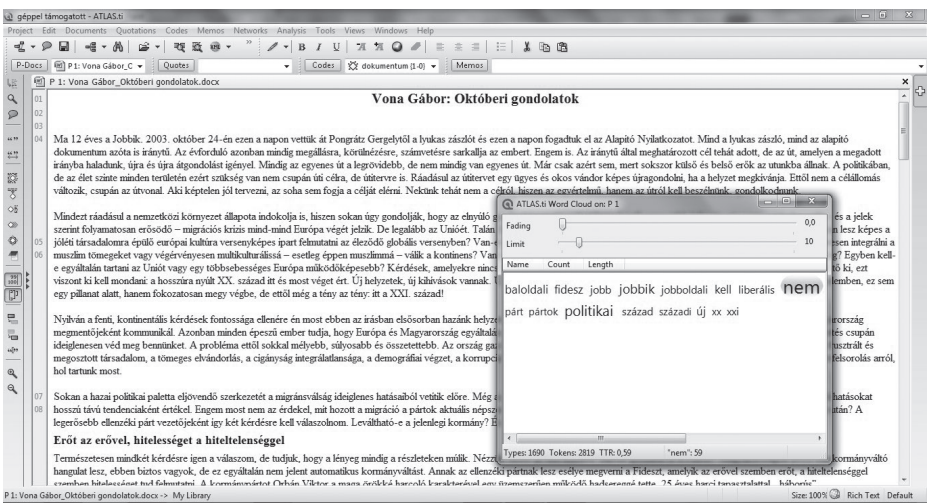
Másfelől a módszer hátrányai is részben ezen átláthatóságban gyökereznek. A kódok áttekinthetősége túlzott sztenderdizálásra, valamint a kvantitativitásba való átcsúszásra csábíthat, fokozottabb mértékben, mint a kézi kódolás során kialakuló reflexszerű besorolás. Szükséges elsajátítani a programok működését használatukhoz, valamint mivel minden programot konkrét eljárásmódra, elméletre vagy problématípusra írtak, ezért különböző kutatásokban eltérő mértékben használhatóak (ld. Bhowmick, 2006: 6 táblázatát). Megmarad továbbá a gépi támogatású kódoláson belül is jórészt a kézi kódolás lassúsága, valamint jelentős élőerőigénye.

Gépi kódolás

A leggyakoribb gépi kódolási megoldások az osztályozás és a csoportosítás feladatainak logikáját követik (ld. III.). Mindkettő közös tulajdonsága, hogy a kézi munkavégzésnél összehasonlíthatatlanul gyorsabban dolgoznak fel hatalmas korpuszokat is, illetve sokkal kisebb élőerő bevonását teszik szükségessé. Ráadásul a felügyelt tanulási eljárásokat leszámítva egyetlen kutató is

megbízható eredményekre juthat. Az automatizált megoldások miatt megszűnik a kézi kódolás esetlegessége, a kutatói elfogultságok és tévedések következtében előforduló hibás kódolás. A felügyelt tanulási megoldás (lásd IV.1. fejezet) a vizsgált szövegekből kiindulva egy viszonylag homogén korpusz megfelelő elemzésére képes, nagy valószínűséggel sikerül felismernie azokat a mintákat, melyek alapján a kódoló is az adott kódot írta (volna) be. A nem túl bő szókincsű korpuszokat emellett szótáralapú megoldásokkal is jól elemezhetjük, mert ezekben kicsi a valószínűsége, hogy körülírják a szótárunkban meghatározott kifejezések valamelyikét (Schwartz – Ungar, 2015: 80). A csoportosítás kiegészíti ezeket a megoldásokat: segít eligazodni azokban a korpuszokban is, melyek tartalmáról olyan keveset tudunk, hogy érdemi vizsgálatuk megkezdéséhez máskülönben hosszas tanulmányozásukra lenne szükségünk. A gépi kódolás az ilyen korpuszok elemeinek csoportosítását is lehetővé teszi. A I.2.3. ábra a szavak gyakoriságára épülő szövegjellemzés gyakorlati eredményére nyújt egy példát: a szógyakoriság alapján meghatározható, hogy a szöveg a politikai pártokkal és ideológiákkal foglalkozik.

I.2.3. ábra – Egy gépi kódolási megoldás (ATLAS.ti)



Ugyanakkor a kézi kódolás „aranystandardjával" szemben a gépi kódolásnak vannak hátrányai is. Eredményeinek validálása nem magától értetődő, amire megoldást jelenthet az úgynevezett *csoportos gépi kódolás,* azaz egyszerre több, akár tucatnyi algoritmus lefuttatása, melyek kölcsönösen validálják egymást. A munkavégzésen belül az elemzés folyamata rendkívüli mértékben felgyorsul, ugyanakkor az osztályozás esetében megjelennek új, előkészítő jellegű feladatok (pl. a *tanítóhalmaz* és a *szótár* elkészítése), melyek továbbra is jelentős erőfeszí-

tést igényelnek (bár lehet, hogy elmarad a másik két megoldásnál tapasztalt mértéknél). A módszer hátránya, hogy az alkalmazott kategorizálási algoritmus befolyásolhatja a vizsgálat eredményét, míg e módszer esetében már fokozottan érvényesül a nyelv gépi megértésének korábban részletezett problémája. Összegzésképpen annyit mondhatunk, hogy a számítógép nem képes a nyelvet emberi minőségben értelmezni, csupán annak leegyszerűsítő, egy sajátosan gépi értelmezését képes biztosítani (pl. a szavak, kifejezések gyakoriságának mérésével). Másfelől pedig hangsúlyozandó, hogy gépi kódolás alkalmazásának betanulási költsége, ideje jelentősen meghaladja a „normál", kézi kódoláshoz szükséges mennyiséget.

Ezen áttekintés után két példa segítségével illusztráljuk a módszertani eljárások alkalmazását (ld. 1. és 2. keret). Első példánk egy, a nemzetközi politikatudományban bevett kutatási irányhoz, a választási ígéretek teljesülésének vizsgálatához kapcsolódik (kapcsolódó hazai publikációkért ld. Soós – Körösényi, 2013, illetve Soós, 2015). Amennyiben valaki kísérletet tesz választási programok elemzésére, óhatatlanul szembesül a kódolás problematikájával. Hogyan dolgozzuk fel a meglévő szövegeket? Mi alapján azonosítsuk az egyedi ígéreteket? Hogyan csoportosítsuk őket közpolitikai területük szerint? Joaquín Artés (2011: 154) tanulmánya, mely a két legnagyobb spanyolországi párt választási ígéreteinek 1989 és 2004 közötti megvalósulását vizsgálja, jó példát nyújt az ilyen és ehhez hasonló módszertani dilemmákra. A szerző – hasonlóan a fent említett magyar kutatókhoz – a kézi kódolás eszközét választotta munkájához.

Másik példánk egy magyar kutatás, melyben a Sebők Miklós – Berki Tamás szerzőpáros a rendszerváltás utáni magyar költségvetéseket vizsgálta a gépi kódolás eszközeit alkalmazva. Ők többek között olyan kérdésekre keresték a választ, hogy hogyan lehet százezres nagyságrendű szövegsort érvényes módon lekódolni? Mivel az egyes módszerek gyakorlati használatát a IV. fejezet részletesen taglalja, itt a kutatói döntés hátterét helyezzük a középpontba: arra keressük a választ, hogy milyen okokból döntöttek a kutatók egy adott kódolási eljárás mellett.

I.2.1. példa

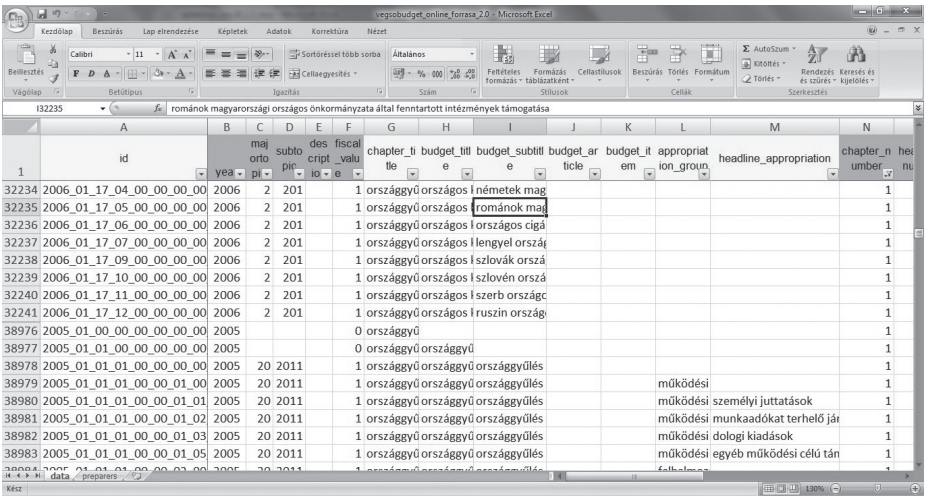
A spanyol Joaquín Artés (2011: 154) az 1989–2004 közötti spanyol politika két meghatározó pártja, a Néppárt és a Spanyol Szocialista Munkáspárt nyolc választási programját vizsgálta meg abból a szempontból, hogy valóra váltották-e a bennük megfogalmazott ígéreteket. Kutatócsoportjával ehhez a kézi kódolás megoldását választotta. Miért dönthetett így? A pártprogramok vizsgálatán belül a szerző kiemelt és összegyűjtött tizenkét gazdaságpolitikával foglalkozó ígérettípust. A kutatás, mivel viszonylag kis korpusszal foglalkozott, nem

igényelte a munkavégzés gépi felgyorsítását. A géppel támogatott kódolás fő előnye a kézi kódolással szemben a nagyobb konzisztencia és áttekinthetőség. Mivel a kutató célirányosan a gazdaságpolitikát kereste a programokban, ezért az ezekkel kapcsolatos nehézségek csak kevésbé érintették munkáját. Mindemellett a gépi támogatású kódolási módszerek kiegészítő funkciói sem lehettek túlzott mértékben a segítségére (például nem vizsgálta az egyes szövegrészek kapcsolatait). A kézi kódolás legfőbb előnye a szöveg mélyebb értelmezésének lehetősége, melyet itt a kutatónak alaposan ki kellett és ki is tudott aknázni. Mivel a nehézségek elsősorban a módszer lassúságában, másrészt a megfelelő kutatók hiányában gyökereznek, ezek a hagyományos ellenvetések itt nem jelenthettek komolyabb problémát (persze ahogy megjegyeztük, a kódolást Artés nem egyedül végezte). Mindezek fényében megállapítható, hogy a kézi kódolás melletti döntés megfelelt a kutatási céloknak. Persze, ha a szerzők más kérdést tettek volna fel, így mondjuk több közpolitikai területet is megvizsgáltak volna, vagy éppen az összes induló párt programját elemezték volna, akkor valószínűsíthetően mérlegelték volna az alternatív megoldásokat a sokkal nagyobb és összetettebb korpusz feldolgozására.

I.2.2. példa

Második példánk Sebők Miklós és Berki Tamás (2016) tanulmánya. A szerzők vizsgálatuk során az 1991–2013 közötti évekre vonatkozó költségvetések és zárszámadások kiadási sorait kódolták közpolitikai tartalmuk szerint. Vizsgálatukban ezt gépi módszerrel végezték el. Érdemes átgondolnunk, hogy miért dönthettek ezen megoldás mellett, illetve milyen következményekkel járt volna egy másik lehetőség választása. A feladat végrehajtásához az összesen bő 108 000 kiadási előirányzatra vonatkozó költségvetési sort kellett elemezni. A lehetséges kategóriákat a magyar közpolitika dinamikájával foglalkozó kutatás (cap.tk.mta.hu) közpolitikai kódkönyve alapján állították össze, mely 21 fő- és 219 altémakörből áll. Az I.2.4. ábra egy példát mutat, a „románok magyarországi országos önkormányzata által fenntartott intézmények támogatása" címen szereplő költségvetési sort, melyet a 2-es főcsoportba (jogok) és a 201-es alcsoportba (etnikai kisebbségek, etnikai diszkrimináció és rasszizmus) soroltak a szerzők.

I.2.4. ábra – Egy költségvetési sor közpolitikai kódolása



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	id	year	major	subtopic	description	fiscal_value	chapter_title	budget_title	budget_subtitle	budget_article	budget_item	appropriation_group	headline_appropriation	chapter_number
32234	2006_01_17_04_00_00_00_00	2006	2	201		1	országgyű	országos l	németek mag					1
32235	2006_01_17_05_00_00_00_00	2006	2	201		1	országgyű	országos l	románok mag					1
32236	2006_01_17_06_00_00_00_00	2006	2	201		1	országgyű	országos l	országos cigá					1
32237	2006_01_17_07_00_00_00_00	2006	2	201		1	országgyű	országos l	lengyel országg					1
32238	2006_01_17_09_00_00_00_00	2006	2	201		1	országgyű	országos l	szlovák országg					1
32239	2006_01_17_10_00_00_00_00	2006	2	201		1	országgyű	országos l	szlovén országg					1
32240	2006_01_17_11_00_00_00_00	2006	2	201		1	országgyű	országos l	szerb országgy					1
32241	2006_01_17_12_00_00_00_00	2006	2	201		1	országgyű	országos l	ruszin országg					1
38976	2005_01_00_00_00_00_00_00	2005				0	országgyű							1
38977	2005_01_01_00_00_00_00_00	2005				0	országgyű	országgyű						1
38978	2005_01_01_01_00_00_00_00	2005	20	2011		1	országgyű	országgyű	országgyűlés					1
38979	2005_01_01_01_00_00_01_00	2005	20	2011		1	országgyű	országgyű	országgyűlés			működési		1
38980	2005_01_01_01_00_00_01_01	2005	20	2011		1	országgyű	országgyű	országgyűlés			működési	személyi juttatások	1
38981	2005_01_01_01_00_00_01_02	2005	20	2011		1	országgyű	országgyű	országgyűlés			működési	munkaadókat terhelő já	1
38982	2005_01_01_01_00_00_01_03	2005	20	2011		1	országgyű	országgyű	országgyűlés			működési	dologi kiadások	1
38983	2005_01_01_01_00_00_01_05	2005	20	2011		1	országgyű	országgyű	országgyűlés			működési	egyéb működési célú tán	1

Mivel a precízen megfogalmazott költségvetésekben költői képek és más géppel nehezen kezelhető nyelvi elemek nem jelennek meg, ezért a kutatóknak nem volt szükségük a személyes értelmezés előnyeire. Nem kellett különösebben bonyolult, mélyebb összefüggéseket felismerniük, ráadásul a költségvetés gyakorlatilag kifejezések listája, mely nem hordoz magában a szavak egyszerű jelentésén túli tartalmat. Összességében számukra a kézi kódolásnak elsődlegesen a hátrányai, s nem az előnyei jelentkeztek volna. Ezek közül is a döntő jelentőségű a kézi kódolási módszer nehézkessége, lassúsága és hatalmas energiaigénye volt: nem nagyon lehetséges néhány tucat sornál többet lekódolni óránként a kódkönyv alapján. Mivel az adatbázis százezernél is több sorból állt, így (a kettős kódolás mellett) ez sok ezer munkaórát jelentett volna kettejüknek, ami szinte automatikusan kizárhatta számukra a kézi kódolás lehetőségét.

Ehhez hasonló okokból vethették el a géppel támogatott kutatás lehetőségét is. A munkamódszer sebessége nem növekedett volna jelentősen, továbbra is a kutatást ellehetetlenítő erőforrásigénnyel kellett volna szembenézniük. Ráadásul a költségvetések precizitása értelmetlenné tette azokat az előnyöket, melyek ezekben az eljárásokban rejlenek. Ha eleve tudták, milyen csoportokba sorolhatják az egyes tételeket, mi értelme lett volna a csoportokat külön kigyűjteni? A kódkönyv ráadásul lehetővé tette a konzisztens munkavégzést, így a módszer előnyei nem jelentkezhettek. Mivel nem volt szükségük ezen programok kiegészítő funkcióira, vagy azokat könnyen pótolhatták, nem lett volna értelme géppel támogatott módszerekhez fordulniuk. A gépi kódolás számára ráadásul megfelelőek voltak az előfeltételek: a költségvetések szikár szövege homogén, így osztályozási módszerekkel jól értelmezhető. A kézi kódolásnál

nagyságrendekkel gyorsabb, ami elkerülhetővé tette annak hatalmas munkaterhét. A módszer konzisztenciája szintén előnyként jöhetett számításba a döntéshozás során. Mindent összevetve az ilyen feladatok végrehajtására a gépi kódolás optimális megoldás.

Ellenőrző kérdések

- Milyen módszertannal lenne érdemes kódolni Szabó Dezső Egész látóhatár című, költői képekben, példabeszédekben és gúnyban bővelkedő esszékötetének fő tematikus elemeit? Milyen előnyei és hátrányai lennének ennek a módszernek?
- Hogyan lenne érdemes lekódolnunk tematikájuk szerint a rendszerváltás óta az Országgyűlésben benyújtott írásbeli kérdéseket? Milyen előnyei és hátrányai lennének ennek a módszernek?
- Hogyan lenne érdemes lekódolnunk egy pártvezető elmúlt öt évben elhangzott évértékelő beszédeit? Milyen előnyei és hátrányai lennének ennek a módszernek?
- Tegyük fel, hogy valamiért összekeveredett egy nagy méretű reformkori és egy hasonlóan sok elemből álló rendszerváltás korabeli politikai írásokat tartalmazó korpusz! A három módszertani eljárás közül hogyan lenne érdemes szétválogatnunk őket?
- Melyik módszerhez forduljunk, ha ki akarjuk gyűjteni az elmúlt tíz év miniszteri és kormányrendeletei közül azokat, melyek valamely hazai nemzetiség helyzetéről rendelkeztek? Milyen előnyei és hátrányai lennének ennek a módszernek?

Szószedet

Magyar	Angol
Ábrázoló eszköz	Diagramming tool
Adatprofil-gyűjtemény	Foundation for data profiles
Csoportos gépi kódolás	Ensemble coding
Csoportosítás, klaszterezés	Clustering
Dokumentumrendszer	Document system
Egyszerű visszakeresési eszköz	Simple retrieval tool
Együttes előfordulási eszköz	Co-occurance tool
Emlékeztetőrendszer	Memo system
Érvényesség	Validity
Gyűjtőeszköz	Gathering tool
Kategóriarendszer	Category system
Osztályozás, klasszifikáció	Classification
Szakaszcímke	Label for section
Számítógéppel támogatott kvalitatív adatelemzés	Computer-Assisted Qualitative Data Analysis
Vizuális segítség	Visual aid

Ajánlott irodalom

A nyelv megértésének filozófiai és ismeretelméleti hátteréhez kapcsolódva hasznos olvasmány Ambrus (2007), Creswell (2013), illetve Farkas – Kelemen (2007). A kézi kódolás lehetséges alkalmazására nyújt példát: Artés (2011), Seale (1999), Soós Gábor – Körösényi András (2013), Soós (2015). A géppel támogatott kódolás alkalmazását tárgyalja Bhomwick (2006), White, Judd és Poliandri (2012). A gépi kódolás alkalmazásához jelent segítséget: Grimmer – Stewart (2013), Maietta (2008), illetve Schwartz – Ungar (2015).

II. SZÖVEGBÁNYÁSZATI FELADATOK: INFORMÁCIÓ-VISSZAKERESÉS ÉS -KINYERÉS

Balázs Ágnes

II.1. FOGALMI ALAPOK ÉS A SZÓZSÁK MÓDSZER

A fejezetben az információ-visszakeresés, illetve az információkinyerés egyik legalapvetőbb eljárása, a szózsák modell kerül bemutatásra, melynek segítségével megvizsgálhatjuk egyes szavak és kifejezések gyakoriságát egy adott korpuszon belül. A modell legegyszerűbb formájában leginkább homogén témájú, illetve alanyú szövegek vizsgálatára alkalmas, s nem veszi figyelembe a szavak sorrendjét, kapcsolatát, csupán gyakoriságát. A fejezetben bemutatásra kerül a szövegelőkészítés folyamata, a szózsák modell főbb alkalmazási lehetőségei, típusfeladatai, előnyei és hátrányai, valamint hazai és külföldi alkalmazási példái.

A *szózsák* (avagy vektortérmodell) az *információ-visszakeresés* egyik legalapvetőbb modellje, amely segítségével egyes szavak gyakoriságát vizsgálhatjuk meg egy adott korpuszon belül (Tikk – Kovács, 2007: 33). E módszer legszűkebben értelmezett formájában nem veszi figyelembe a szavak sorrendjét, kontextusát (Russel – Norvig, 2005: 742–744). Tágabb értelmében ugyanakkor kapcsolódik hozzá a kifejezések kinyerésének területe, ahol már a szavak sorrendje is számít, ám ez még nem jelent egyet a névelem-felismeréssel (ld. II.2. fejezet, ill. Baldwin et al., 2003: 89–96). Láthatjuk tehát, hogy a szózsák modellt komplexebb formájában *információkinyerésre* is lehet használni (alább bővebben kitérünk az információ-visszakeresés és az információkinyerés közötti különbségre).

E fejezetben bemutatjuk a szózsák megoldás főbb alkalmazási lehetőségeit, ismertetjük erősségeit és gyengeségeit. E feladat elvégzéséhez szükséges kitérni a *szövegelőkészítés* folyamatához kapcsolódó olyan műveletekre is, mint a *szótövezés/toldaléklevágás,* vagy a *tiltólistás szavak* beállítása. Mindezek után néhány példa segítségével ismertetjük az eljárás nemzetközi és magyar társadalom-, illetve politikatudományi kutatásokban való alkalmazási lehetőségeit.

A klasszikus feladat

Tikk és Kovács (2007: 33) úgy hivatkoznak a szózsák modellre, mint az információkinyerés területén széleskörűen elterjedt eszközre (ld. még Kovács, 2007: 122). Az *információ-visszakeresést* arra használjuk, hogy a már strukturált korpuszból visszakeressük a számunkra releváns információt (Russel – Norvig, 2005: 742; Vázsonyi – Tikk, 2007: 63). Az *információkinyerés* alternatív megoldása segítségével azonban már képesek lehetünk a kifejezések közötti kapcsolatok elemzésére, tendenciák és minták felismerésére és az információk összekapcsolása révén új információk létrehozására. Azaz a segítségével strukturálatlan szövegekből is előállíthatunk strukturált információkat (Hearst, 1999: 3–4; Szarvas – Farkas, 2007: 81).

II.1.1. táblázat – Az információ-visszakeresés és -kinyerés összehasonlítása

Információ-visszakeresés	Információkinyerés
A felhasználó a releváns *dokumentumokat* kapja vissza, s maga dolgozza fel azokat	A felhasználó a releváns dokumentumból kinyert *tényeket* használja, s nem saját magának kell feldolgoznia a visszakapott releváns dokumentumokat
Egyszerűbb (kevesebb háttérismeretet igényel)	Bonyolultabb (nagyobb szakértői tudás szükséges hozzá)
Általános eljárások alkalmazhatók a megoldására	Általában szövegtartomány-függő (azaz *domain*függő)
Tetszőleges lekérdezések kezelésére alkalmas	Rögzített típusú elemek (fontos szövegrészek, információk, tények) kigyűjtésére alkalmas
Gyors, kevésbé pontatlan	Lassú, gyakran pontatlan
Kevésbé hatékony, mint a kinyerés	Hatékonyabb, mint a visszakeresés

Forrás: Szarvas – Farkas (2007: 84–85) nyomán készített táblázat

A szózsák modell legszűkebb értelmében a szavak sorrendjét, kontextusát képtelen figyelembe venni, csupán az egyes szavak gyakoriságára, előfordulási számára kaphatunk választ segítségével (Tikk – Kovács 2007: 33, Tikk, 2007c: 122). A módszer alkalmazásához szükség van *a szöveg előkészítésére* a célból, hogy valós eredményt kapjunk. A szövegelőkészítés körében tárgyalni kell a különböző típusú toldaléklevágási módszereket, amelyek segítségével azonos szavak ragozott, különböző alakú formáit is ugyanazon szóként ismeri fel a program.

A toldaléklevágás egyik módszere a *szótőképzés,* melynek során a program levágja a szavak általa toldalékként felismert részeit. Előfordulhat azonban, hogy mindezek során téves adatot kapunk. A „zokni" szó esetében például elképzelhető az, hogy a program főnévi igenévként ismeri fel, s „zok" szótőt képez belőle. A toldaléklevágás másik módszere a *lemmatizálás*, ami már szó-

táralapú, és a kutató által a szótárba korábban felvitt adatok alapján ismeri fel az egyes szótöveket. A lemmatizálás a nyelvészetben használt fogalom, és a szó lemmájának, azaz normalizált vagy szótári alakjának megkeresését szolgálja (Tikk – Kovács, 2007: 41–55). Ez az eljárás azt feltételezi, hogy az állomány feldolgozásának megkezdését megelőzően a kutató készít egy szótárat, amely tartalmazza a különböző lehetséges szóalakokat és toldalékokat.

A szótövezés és a lemmatizálás közötti legfőbb különbség az, hogy míg utóbbi során rendszerint értelmes szóalak keletkezik, szótövezés esetén már előfordulhat, hogy a szó csonkolása miatt nem értelmes szótári alakot kapunk eredményül. A munkát, munkám stb. esetében a lemma és a szótő egybeesik, a ló, lovak, lovát esetében a lemma a ló, ám a használt szótövező függvényében a szótő már változik (ló vagy lo). Mindkét módszer esetén döntenünk kell róla, hogy a képzőket is levágjuk, vagy csupán a jeleket és a ragokat (i. m.: 41–42). A szövegelőkészítéshez tartozó fontos eszköz még a tiltólistás szavak (stopszavak) kialakítása. Ennek segítségével elérhetjük, hogy a számunkra nem releváns szavakat, például kötőszavakat, töltelékszavakat ne vegye figyelembe a program, s az elemzésünk szempontjából leggyakoribb érdemi szavak kerüljenek kiválogatásra. A stopszólista előállítására rendszerint a dokumentumgyűjtemény manuálisan is ellenőrzött szógyakorisági adatai alapján kerül sor (i. m.: 40–41).

A szózsák módszer előnye, hogy gyorsan, egyszerűen kiemeli az adott szövegben fontos (azaz a leggyakoribb) szavakat, ugyanakkor alkalmazása nem igényel különösebb szakértelmet, a számítógéppel támogatott kvalitatív adatelemzésben vagy a kvantitatív szövegelemzésben való magas szintű jártasságot. Ezzel együtt, részben egyszerűségéből fakadóan természetesen hátrányokkal is rendelkezik. Ilyen hátrány például, hogy egyes szavak jelentése különböző kontextusokban lehet negatív vagy pozitív is. Például nem mindegy, hogy az állam által újonnan vásárolt légvédelmi szirénák túl hangosak, vagy a frissen felépített repülőtér, illetve hogy egy smirgli a durva, vagy egy parlamenti felszólalás, esetleg a kormányt illető kritika. Az azonos alakú szavak is problémát jelentenek (pl. ég, vár, terem, fűz, nyom). Mivel a szavak sorrendjét sem kezeli ez a módszer, csupán a gyakoriságát, lényegében azonos jelentésűként fog szerepelni a következő két mondat: „Bajnai Gordon utódja a miniszterelnöki székben Orbán Viktor lett", illetve „Orbán Viktor utódja a miniszterelnöki székben Bajnai Gordon lett". Hátránya még, hogy nem tudja, a szöveg mely szavai kapcsolódnak egymáshoz, mi mire vonatkozik. Nem ismeri fel továbbá a hétköznapi nyelvhasználatot, így például problémát jelenthet, ha helyesírási hibák, elgépelések találhatók a szövegben.

Típusfeladatok

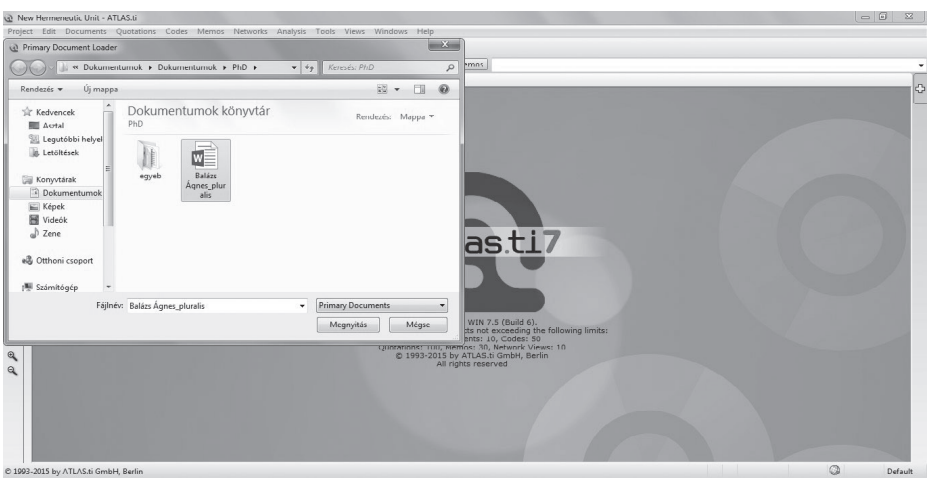
A szózsák homogén témájú vagy alanyú szövegek vizsgálatára alkalmas leginkább (Joachims, 2001: 129). Ilyen lehet például, ha meg kívánjuk vizsgálni, hogy mely szavak fordulnak elő leggyakrabban a „Vlagyimir Putyin"-nal kapcsolatos egyes szövegekben. Vizsgálhatunk továbbá egy-egy konkrét témához kötődő parlamenti felszólalásokat is ezzel a módszerrel, vagy kideríthetjük, hogy egy parlamenti képviselőnek vagy politikai gondolkodónak melyek a leggyakrabban használt szavai. A következőkben három kutatási feladatot láthatunk, amely során alkalmazható a szózsák modell. Az első esetben kiderül, hogyan találhatunk könnyen kulcsszavakat egy általunk publikálni kívánt tanulmányhoz, míg a második esetben viszonylag nagy tömegű, ám homogén témájú országgyűlési felszólalások leggyakoribb szavai kerülnek kinyerésre. A harmadik keretben a szózsák modell a valóságban megtörtént gyakorlati alkalmazására láthatunk példát.

II.1.1. példa

Tegyük fel, hogy egy általunk megírt tanulmányhoz kulcsszavakat kell megadnunk! A feladat szózsák módszerrel történő megoldását egy, a választójog egyenlőségéhez kapcsolódó írás kapcsán mutatjuk be (Balázs, 2015). A feladat megoldásához számítógépes támogatást (ld. I.1. és I.2. fejezet), az ATLAS.ti kvalitatív adatelemző program windowsos változatát használjuk. Első lépésként kiválasztjuk a vizsgálni kívánt dokumentumot (Documents → New → Add Documents – II.1.1–II.1.2. ábra).

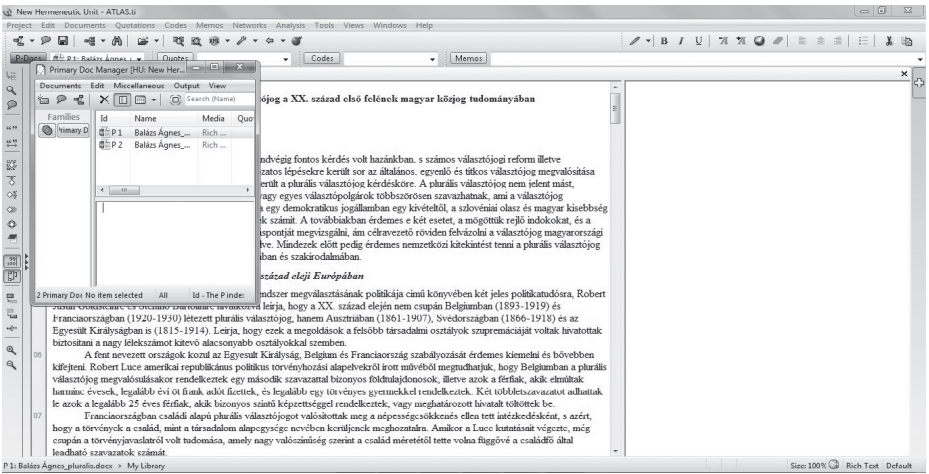
II.1.1. ábra – A vizsgálandó dokumentum hozzáadása

II.1.2. ábra – A vizsgálandó dokumentum kiválasztása

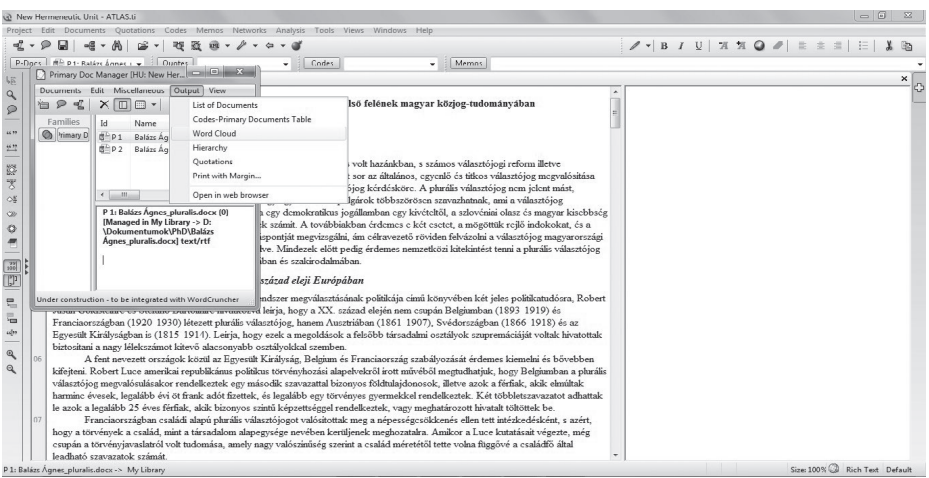


Ezt követően kiválasztjuk a bal felső sarokban lévő P-Docs gombot, majd a felső sorból kiválasztjuk és az alsó sorba húzzuk a vizsgálni kívánt dokumentumot, majd az Outputra kattintva kiválasztjuk a szófelhőt (Word Cloud – II.1.3–II.1.4. ábra).

II.1.3. ábra – Az aktuálisan vizsgálandó dokumentum kiválasztása

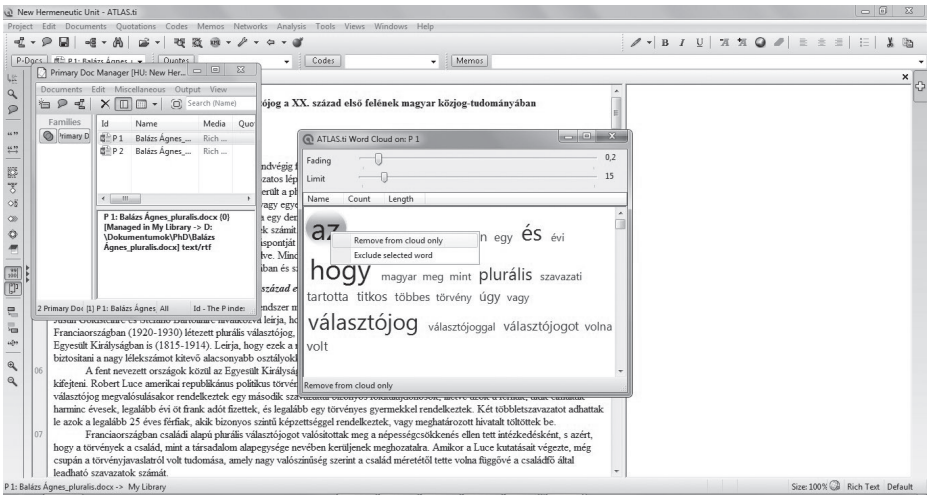


II.1.4. ábra – A szófelhő kimenetként való választása



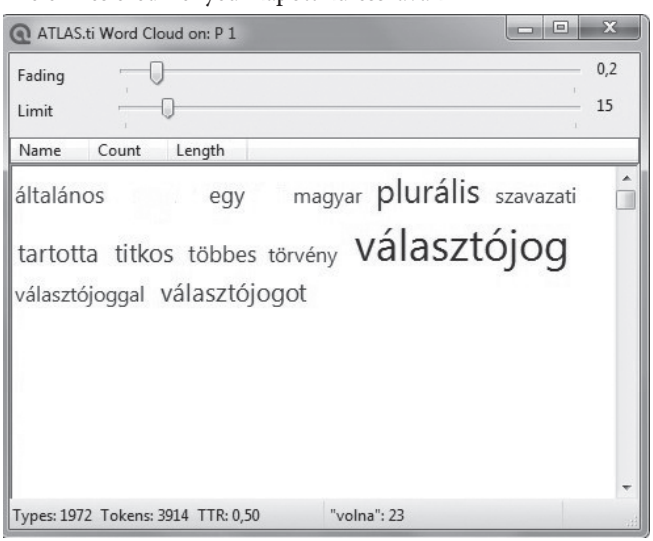
A létrejött szófelhőből tiltólistára tehetünk egyes szavakat, csak a szófelhőből törölve, vagy teljességgel kizárva azokat a vizsgálatból, emellett beállíthatjuk, hogy legalább hányszor kell szerepelnie az adott szónak a szövegben, hogy megjelenjen a szófelhőben. Mi jelen feladatnál az áttekinthetőség és a kellő mennyiségű releváns szó megjelenése érdekében tizenöt felbukkanást állítottunk be (II.1.5. ábra).

II.1.5. ábra – A nemkívánatos szavak eltávolítása



Miután eltávolítottuk a nemkívánatos szavakat (pl. egy, és, évi, hogy, meg, mint, vagy, úgy, volna, volt), készen is áll a szófelhő, melynek szavai megfelelő kulcsszavaknak tűnnek a tanulmányhoz, vagy legalábbis megfelelően leszűkítik a lehetséges kulcsszavak körét (II.1.6. ábra).

II.1.6. ábra – Az elemzés eredményéül kapott kulcsszavak

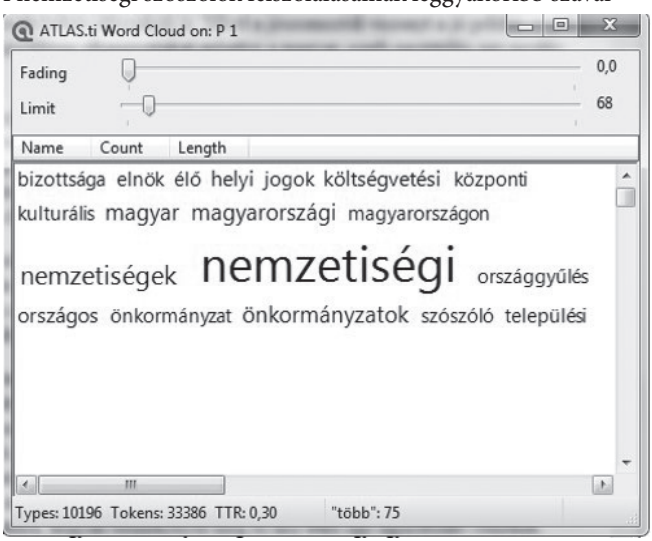


Hasonlóan hasznos alkalmazási terület lehet például, ha tudománytörténeti kutatásaink során nagy mennyiségű folyóirathoz kapcsolódóan kell kulcsszavakat kinyernünk, s a folyóiratok elérhetőek digitalizált formában.

II.1.2. példa

Tegyük fel, hogy a nemzetiségek országgyűlési képviseletének kérdését vizsgáljuk, és kíváncsiak vagyunk a nemzetiségi szószólók plenáris üléseken való szóhasználatára. Az ATLAS.ti program segítségével e felszólalásokból is könnyedén kinyerhetjük a szószólók által leggyakrabban használt szavakat. Elsőként létre kell hoznunk egy korpuszt az összes eddigi felszólalás szövegéből az Országgyűlés honlapja (www.parlament.hu) alapján, majd az előző feladathoz hasonlóan elvégezhetjük az elemzést.

II.1.7. ábra – A nemzetiségi szószólók felszólalásainak leggyakoribb szavai



Mivel a nemzetiségi szószólók a plenáris üléseken csupán a Házbizottság megítélése szerint a nemzetiségeket érintő kérdések kapcsán szólalhatnak fel,[1] talán nem meglepő, hogy a „nemzetiségi" és a „nemzetiségek" szavak bukkannak fel a leggyakrabban felszólalásaikban. A szabályozás ismeretében pedig a kapott szófelhő alapján képesek vagyunk következtetni arra, hogy a nemzetiségek bizottsága, a nemzetiségi jogok, a nemzetiségek költségvetési támogatása, a nem-

[1] 2012. évi XXXVI. törvény az Országgyűlésről 29. § (2).

zetiségek kulturális élete és a nemzetiségi önkormányzatok a plenáris üléseken leggyakrabban felmerülő kérdések e beszédek alapján.

Még ha messzemenő következtetéseket nem is vonhatunk le csupán ezen adatok alapján, mindenképpen könnyebbé teszik kutatásunk további irányainak meghatározását, hipotéziseink megfogalmazását. Ha arra lennénk kíváncsiak, hogy a nemzetiségi szószólók milyen személyeket, intézményeket vagy szervezeteket említenek meg leggyakrabban felszólalásaikban, akkor a szózsák modell már nem jelentene kielégítő megoldást, a névelem-felismerés módszeréhez kellene folyamodnunk (ld. II.2. fejezet).

II.1.3. példa

A Fidesz által 2011-ben kezdeményezett, alkotmányozással kapcsolatos „nemzeti konzultáció" is alkalmas alapanyagot kínál a szózsák megoldáshoz. Az akció során a párt közlése szerint a válaszadók 45%-a által kitöltött egyéni vélemények részre adott válaszok 125 leggyakoribb kulcsszavát egy szófelhőben prezentálták.[2] Mint az előbbi példánál is láthattuk, ez a módszer elsősorban kutatásunk kezdetén, a téma előzetes feltérképezése során és hipotéziseink megfogalmazásakor nyújthat segítséget, miközben a hipotézis igazolására már egyéb módszereket kell alkalmaznunk. E szófelhőnél például láthatjuk, hogy a legnagyobb méretű „Alkotmány", a „magyar" és a „Magyarország", valamint a valamivel kisebb méretű „jog" voltak a leggyakoribb szavak, és sejthetjük, hogy a „halálbüntetés", a „védelem" és a „választójog" kifejezések is igen gyakran felbukkantak a válaszokban, de erre nem alapozhatjuk állításainkat, ezt más módszerek alkalmazása segítségével tudjuk csupán igazolni. További módszertani nehézséget jelent, hogy az ábra segítségével nem tudjuk egyértelműen meghatározni, hogy a „védelem" szó mire vonatkozik az egyes válaszokban (például honvédelemre, a közrend védelmére, „munkahelyvédelemre" vagy alkotmányvédelemre).

Nemzetközi és magyar politikatudományi alkalmazások

Fejezetünk zárásaként bemutatjuk a módszer néhány nemzetközi, illetve magyar politikatudományi és egyéb alkalmazását. Egy 2007-es tanulmány elkészítése és az azt megelőző kutatások során Airoldi (2003) és kollégái (Airoldi

[2] Állampolgári kérdőív az Alaptörvényről, 2011. április 3. http://static.fidesz.hu/download/156/A_Nemzeti_Konzultacios_Testulet_kerdoivenek_eredmenyei_2156.pdf (Letöltés ideje: 2015. október 31.)

et al., 2007) a szózsák modellt használták Ronald Reagan rádióbeszédei szerzője kilétének kiderítésére. Reagan 1975 és 1979 között több mint ezer beszédet mondott a rádióban elnöki kampánya során. E beszédek közül hatszáz bizonyítottan Reagan saját maga által írt szöveg volt, s a kutatók szerették volna kideríteni a többi szöveg szerzőjének kilétét. A szózsák modell mellett kifejezések kinyerését is alkalmazva megállapították, hogy szövegei nagy részét maga Reagan írta, és nem alkalmazott szövegírót.

Laver és társai (2003) közpolitikai álláspontok különböző politikai szövegekből való kiszűrése során alkalmazták a szózsák modellt, méghozzá az egyes német pártok álláspontjainak feltárása érdekében, nyelvfüggetlen technikával, a parlamenti felszólalásokat is megvizsgálva. A kutatás során bizonytalansági méréseket is végeztek.

Cheryl Schonhardt-Bailey (2005) pedig George W. Bush és John Kerry nemzetbiztonsági és belbiztonsági témájú beszédeinek elemzésére használta fel a módszert. Megállapította, hogy míg Kerry esetében az iraki háborút illető kritika, a veteránok, a Nemzetbiztonsági Minisztérium és a nukleáris fenyegetettség csökkentése voltak a leggyakrabban megjelenő tartalmi elemek, Bush esetében a terrorizmus elleni harc és a fejlődő országok gazdasági növekedése jelentették a központi témákat. A szerző a kutatás során kapcsolatelemzést is végzett.

Magyarországon a 2010. évi Számítógépes Nyelvészeti Konferencián a szózsák modell több hasznosítási módját is bemutatták. Az egyik ilyen megoldás a többnyelvű online szövegek elemzésére létrehozott OpinHu rendszer, amely internetes hírportálokon, blogokon és közösségi oldalakon fellelhető szövegek tartalomelemzésére hivatott (Miháltz, 2010: 14). A szózsák modellt a véleményelemzés során a magyar, kínai és arab nyelv esetében használják az érzelmek felismerésére, feltételezve, hogy ha egy kulcsszó és egy érzelmet kifejező szó együtt fordul elő egy mondatban, egy érzelmi célpontra irányul (ld. III.2. fejezet). Az angol nyelvű szövegek esetében kulcsszó-előfordulási statisztikák létrehozása is elvégezhető (i. m.: 17).

Végezetül egy intézménynek küldött panaszlevelek megbízható tartalmi kivonatolására és az arra való válaszadásra képes rendszer kifejlesztése során is alkalmazták a szózsák algoritmusát, ám a kutatók felhívták a figyelmet, hogy a szózsák modell önmagában nem elegendő a feladat megoldására (Bártházi – Héder, 2010).

Ellenőrző kérdések

- Milyen típusú feladatok megoldására alkalmas a szózsák modell? Fogalmazzon meg egy kutatási problémát, amelyet a szózsák eljárással kielégítően meg lehet oldani!
- Milyen esetekben vezethet ez a módszer hibás eredményekhez?
- Milyen műveletek tartoznak a szövegelőkészítés folyamatához, és miért van rájuk szükség?
- Milyen problémák merülhetnek fel a szövegelőkészítés folyamán?
- Keressen példát a szózsák modell magyar nyelvű kutatásokban való megfelelő vagy erre alkalmatlan alkalmazási lehetőségére!
- Milyen módon prezentálható a szózsák modell használatával létrehozott eredmény?
- Alkalmas-e a módszer szóképekkel és alakzatokkal teletűzdelt politikai beszédek elemzésére?

Szószedet

Magyar	Angol
Információkinyerés	Information extraction
Információ-visszakeresés	Information retrieval
Kifejezések kinyerése	Extraction of expressions
Lemmatizálás	Lemmatization
Szótőképzés, szótövezés	Stemming
Szózsák (vektortérmodell)	Bag of words
Szövegelőkészítés, -előfeldolgozás	Preprocessing
Szövegtartomány (tématerület)	Domain
Tiltólistás szavak (stopszavak)	Stop words

Ajánlott irodalom

A mesterséges intelligenciáról, és azon belül a szózsák modellről további információk találhatók Russel és Norvig (2005) könyvében. Ha a szövegbányászatban szeretnénk jobban elmélyülni, a Tikk Domonkos (2007) szerkesztésében megjelent magyar nyelvű könyv kiváló segítséget nyújt számunkra. Ha a módszer konkrét alkalmazására szeretnénk további példákat keresni, érdemes a Magyar Számítógépes Nyelvészeti Konferencia kiadványaihoz fordulni (Tanács és Vincze, 2010, 2011). Ha kitekintést szeretnénk nyerni multimédiatartalmak szózsák modell segítségével történő intelligens feldolgozására, Li és társai (2011), Li (2012) Zhu és társai (2013) és Dimitrovski és társai (2016) írásai nyújthatnak segítséget.

Balázs Ágnes – Sebők Miklós

II.2. NÉVELEM-FELISMERÉS

A fejezet bemutatja a *névelem-felismerést*, mint az egyik legfontosabb szövegbányászati feladatot. Segítségével kinyerhetők egy adott korpuszon belül előforduló névelemek, s ezen belül a tulajdonnevek (személynevek, helyek, szervezetek és egyéb tulajdonnevek). A fejezetben meghatározzuk a névelem-felismeréshez kapcsolódó legfontosabb fogalmakat, valamint a módszer típusfeladatait és buktatóit. Ezt követően három konkrét példán mutatjuk be a módszer gyakorlati használatát, majd a hazai és nemzetközi alkalmazásokból ismertetünk néhányat.

A *névelem-felismerés* (NER) az egyik legfontosabb speciális szövegbányászati feladat. Legegyszerűbb formájában az *információ-visszakeresés* területéhez tartozik, komplexebb megoldásai ugyanakkor már a *szövegbányászathoz* tartoznak (ld. II.1. fejezet). *Információ-visszakeresés* esetén célunk az, hogy a már strukturált korpuszból visszakeressük a számunkra releváns információt (Russel – Norvig, 2005: 742; Vázsonyi – Tikk, 2007: 63). A szövegbányászat esetén már lehetőség van a kifejezések közötti kapcsolatok elemzésére, tendenciák és minták felismerésére, és az információk összekapcsolása révén új információk létrehozására (Hearst, 1999: 3–4; Szarvas – Farkas, 2007: 81).

A névelem-felismerés módszere az 1990-es években született meg, lényege, hogy egy program felismeri a korpuszban felbukkanó tulajdonneveket, azokat kigyűjti, és típusonként (pl. földrajzi név, márkanév, jogi személy stb.) csoportosítja (ennyiben egy hibrid információkinyerési és kategorizálási feladatként is tekinthetünk rá). A standardizált megoldások lehetővé teszik akár telefonszámok vagy időpontok kigyűjtését is, ennyiben tehát túlmutatnak a tulajdonnevek körén. A NER legáltalánosabb annotációs sémáit a Dokumentummegértési Konferenciák (MUC) és a Természetesnyelv-tanulási Konferenciák (CoNNL) fejlesztették ki (Jiang, 2012: 15–16). A fejezetben bemutatjuk e kutatási irány alapfogalmait és alkalmazásait.

A névelem-felismerés

A névelem-felismerés az információkinyerés egyik legalapvetőbb formája, amely igen fontos a szövegek összetettebb módszerek alkalmazására való előkészítésében (i. m.: 15–16). E módszer kapcsán fontos a *névelem* és a *tulajdonnév* fogalmának megkülönböztetése. Névelem minden olyan szóalaksorozat (*token*sorozat),[1] amely a világ valamely egyedi létezőjére utal. A tulajdonnév tehát csupán a névelem fogalmának egy részhalmaza, hiszen a tulajdonnév mellett névelemnek tekinthető például valamely azonosító, telefonszám, pénznem vagy e-mail cím is (i. m.: 15–16; Szarvas – Farkas, 2007: 91). Ezzel együtt a névelem-felismerés leggyakrabban a személynevek, helyek, szervezetek és egyéb tulajdonnevek felismerésére irányul, így a továbbiakban a két fogalmat szinonimaként használjuk.

A fogalmi alapokhoz tisztázni kell a strukturált és a strukturálatlan szöveg fogalmát is. A strukturálatlan szöveg alatt a hétköznapi értelemben vett szövegeket értjük. A strukturált szövegek olyan szövegek, amelyek valamilyen szempont szerint már feldolgozásra kerültek. A strukturált szöveg rendelkezik *metaadatokkal,* azaz az adatállományhoz kapcsolódó adatokkal, például kódokkal, *címkékkel*. Míg az *adatbányászat*i és az információ-visszakeresési módszereket összefüggések strukturált adatokból való kinyerésére használjuk, a *szövegbányászat* eszközeit strukturálatlan szöveges halmazok feldolgozásánál alkalmazzuk (Tikk, 2006: 344–346; Tikk, 2007a: 20–21). Ez utóbbi viszont külön *szövegelőkészítést* igényel, melynek módszerei közé tartozik a *tiltólistás szavak* alkalmazása, a *lemmatizálás* és a *toldaléklevágás* (Hu – Liu, 2012: 388–389 – minderről bővebben ld. II.1. fejezet).

A következő lépések már a természetesnyelv-feldolgozás területére vezetnek (amely szinte elválaszthatatlan az információkinyerés fogalmától). Mint az előző fejezetben láttuk, ahhoz, hogy a természetes nyelvből adatokat kapjunk, fel kell dolgoznunk a vizsgálandó szöveget, strukturált adatállományokat kell létrehoznunk (Aggarwal – Zhai, 2012: 3, Markov – Larose, 2007: 13). A webes információkinyerő rendszerek *(wrapperek)* alkalmazására azért van szükség, mert a honlapok gyakran tartalmaznak olyan strukturált vagy félig strukturált szövegeket, mint a táblázatok vagy a felsorolások, amelyek sokkal inkább állnak HTML vagy egyéb kódokat tartalmazó elemekből, mint természetes nyelvi elemekből (Jiang, 2012: 14–15).

[1] „*Token*nek nevezzük egy karaktersorozat konkrét dokumentumbeli előfordulását, míg típusnak hívjuk az azonos karaktersorozatot tartalmazó tokenek osztályát. A típusok összessége alapján állítjuk össze […] a szótárat…" (Tikk – Kovács, 2007: 39).

Az információkinyerésen belül a névelem-felismerés mellett a *kapcsolatbányászat* említhető fontos részterületként. Kapcsolatbányászatról akkor beszélhetünk, amikor egy adott szövegben található elemek között a szemantikus, azaz jelentésbeli kapcsolatok is felfedezésre és jellemzésre kerülnek (Jiang, 2012: 22).[2] Különösen fontos itt az elemzéshez felhasználandó szótár megfelelő kialakítása, hiszen a nem megfelelően kialakított szótár sokat torzíthat az eredményeken. Problémát jelenthet például, ha a különböző *névelemosztályok* között átfedés található (például valaki vezetékneve egyúttal földrajzi név is, ld. pl. Balaton Károly), vagy ha egy névelemen belül egy másik is található (pl. Budapest Főváros Kormányhivatala – Tikk et al., 2006: 29–30).

A névelem-felismerés körében két alapvető módszer alkalmazására van lehetőség. Az egyik a szabályalapú módszer, amikor előre megadott adatok alapján kerül kinyerésre az információ, (ilyen szabály lehet a mondatközi nagybetű mint a tulajdonnév kezdete). Mivel ezek a szabályok ütközhetnek egymással, szükséges közöttük hierarchiát felállítani. A másik módszer a statisztikai tanulás, amikor a gép alkot szabályokat a kutató előzetes mintakódolása alapján (Jiang, 2012: 16–22 – minderről bővebben ld. III.1. és III.3. fejezet).

Végezetül egy adott névelem-felismerő rendszer által produkált eredmények *hitelességének* ellenőrzésére többféle mutató létezik. A pontosság (megbízhatóság) azt mutatja meg, hogy az adott korpuszban a rendszer által felismert névelemek közül mekkora a helyes megoldások aránya. A *felidézés (fedés, teljesség)* pedig azt jelenti, hogy a találatok között hány darab lelhető fel az összes releváns dokumentum közül, mekkora az adott korpuszban talált névelemek aránya a szöveg többi részéhez képest. Ezek segítségével kiszámítható azok harmonikus közepét jelentő mutató, a széleskörűen használt F-mérték (Tjong Kim Sang – De Meulder, 2003: 143–144; Hobbs et al., 1991: 2; Vázsonyi – Tikk, 2007: 69–70).

Egy tipikus alkalmazás

A névelem-felismerés tipikus alkalmazásai az egyes névelemjellemzők vagy *tématerületek (szövegtartományok, domainek)* sajátosságaihoz kapcsolódnak. Ennek oka az, hogy a korábbi kutatások szerint például egy földrajzinév-kigyűjtésre kialakított technika csak korlátozottan hasznos eredményeket hoz például intézménynevek kapcsán. Igen fejlett névelem-felismerési módszereket alkalmaznak napjainkban például az orvostudomány, a honvédelem és a nem-

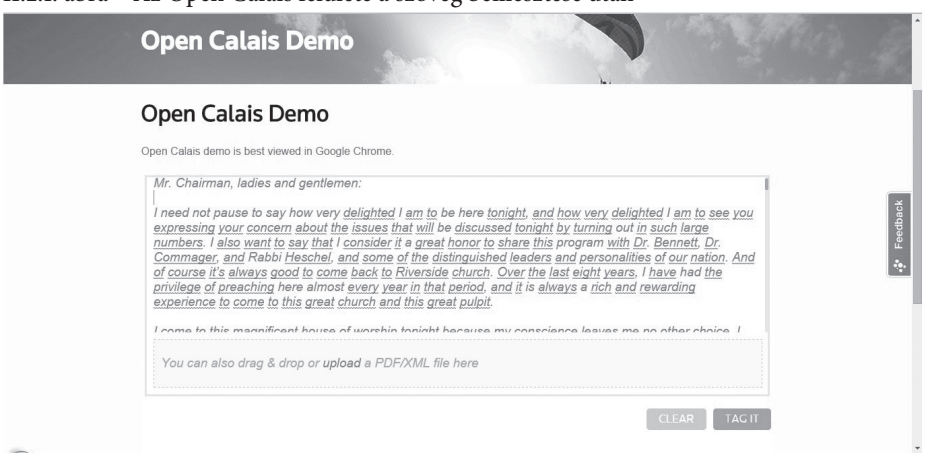
[2] A szemantika vagy jelentéstan legszűkebb értelmében a nyelv különböző összetevőinek (szavak, kifejezések) jelentésével foglalkozó, a nyelvészeten belüli tudományterület (Munk, 2014: 312). Szintaktikusan helyesnek kell lennie a mondatnak ahhoz, hogy szemantikájáról beszélhessünk (Bach, 2005: 20).

zetbiztonság (hírszerzés, felderítés, terrorizmusmegelőzés) terén (Neri et al., 2011: 393).

A következőkben egy konkrét politikatudományi alkalmazási lehetőséget mutatunk be. A feladat Martin Luther King, a vietnami háborút ellenző és a társadalmi igazságosságról szóló, 1967. április 4-i beszédének[3] vizsgálata a szemantikai információk strukturálatlan webes szövegekből való kinyerésére létrehozott Open Calais[4] online program segítségével.

Elsőként be kell illeszteni a vizsgált szöveget, majd a program egyetlen kattintásra elvégzi az elemzést (ld. II.2.1. ábra).

II.2.1. ábra – Az Open Calais felülete a szöveg beillesztése után



Mint látjuk (II.2.2–II.2.4. ábra) a program képes felismerni témaköröket, városokat, kontinenseket, országokat, ipari kifejezéseket, filmeket, szervezeteket, személyneveket, pozíciókat, államokat és tartományokat, régiókat, illetve különböző eseményeket és tényeket, akár idézeteket is. Azonban a felismerés során hibák is léphetnek fel.

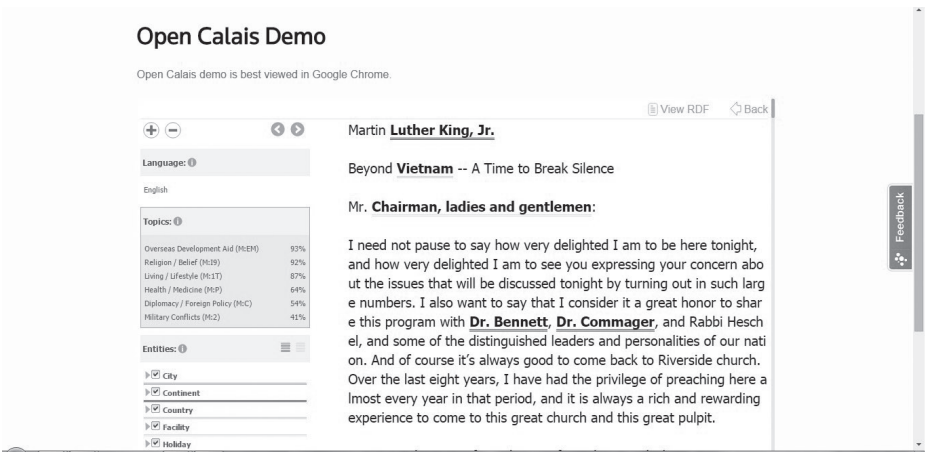
A II.2.2. ábrán láthatjuk például, hogy a program szerint igen jelentős mértékben vallási témájú szövegről van szó, illetve hogy volt olyan névelem, például Herschel rabbi neve, amelyet nem ismert fel. További hiba, hogy az applikáció a francia nemzetközösséget ünnepnapnak tekintette, hogy a kollektív megoldást az ipari kifejezésekhez sorolta be, illetve hogy a Dien Bien Phu nevet filmként ismerte fel (II.2.3. ábra). Láthatjuk továbbá, hogy Martin Luther

[3] Martin Luther King, Jr. Beyond Vietnam – A Time to Break Silence http://www.americanrhetoric.com/speeches/mlkatimetobreaksilence.htm (Letöltés ideje: 2015. október 30.)

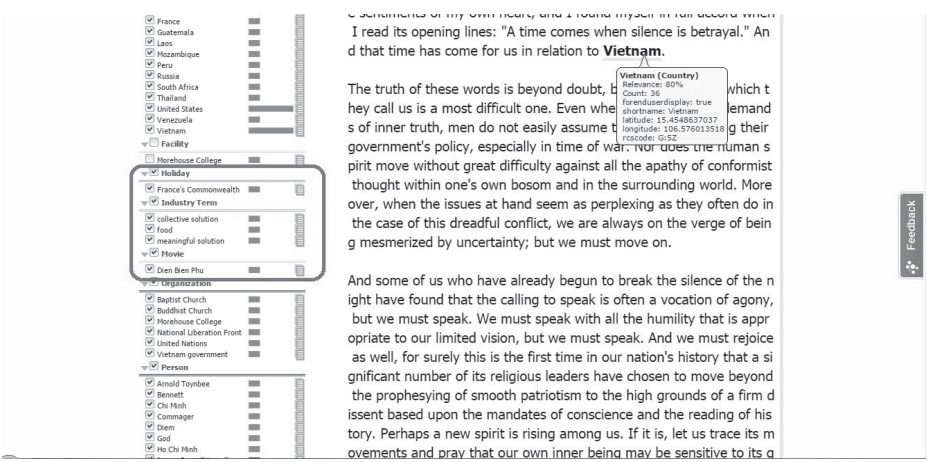
[4] Thomson Reuters – Open Calais Demo http://www.opencalais.com/opencalais-demo/

King vezetéknevét egy alkalommal pozícióként (királyként) ismerte fel, a Genfi Egyezményt pedig Genfnek tekintette (II.2.4. ábra).

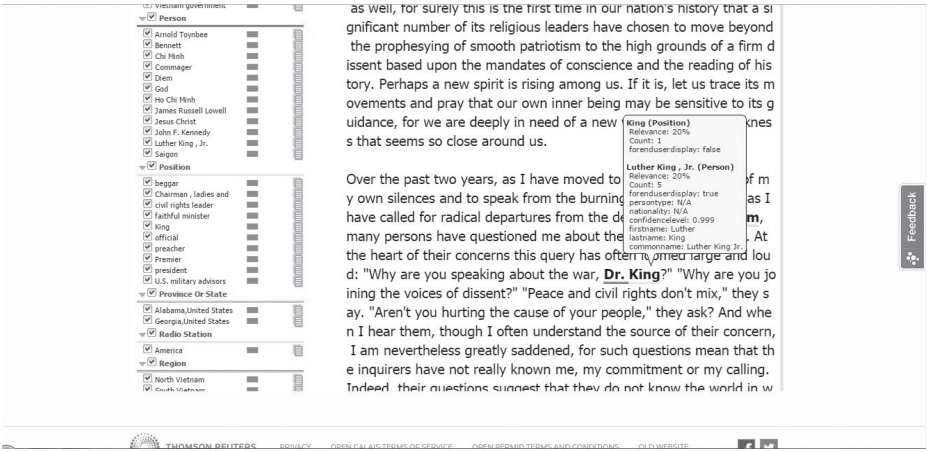
II.2.2. ábra – Elemzés az Open Calais programmal



II.2.3. ábra – Félreértelmezett kifejezések a szövegben

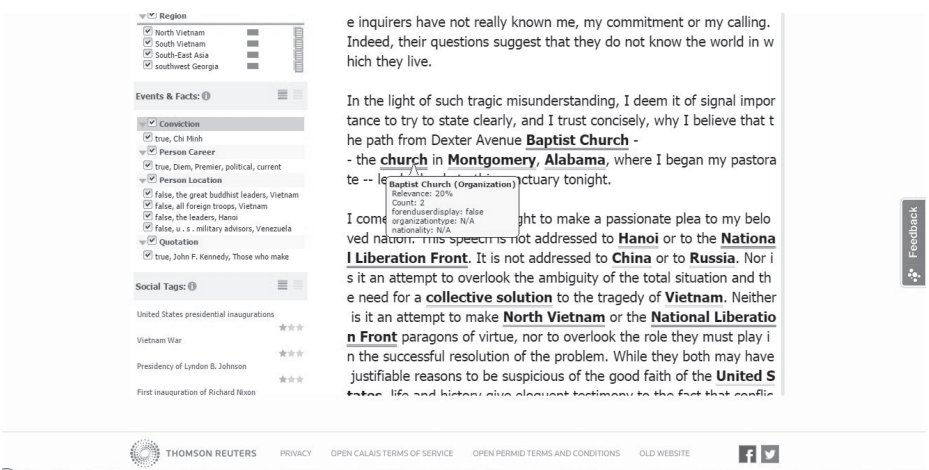


II.2.4. ábra – Viszonyok felismerése és elemzési hiba

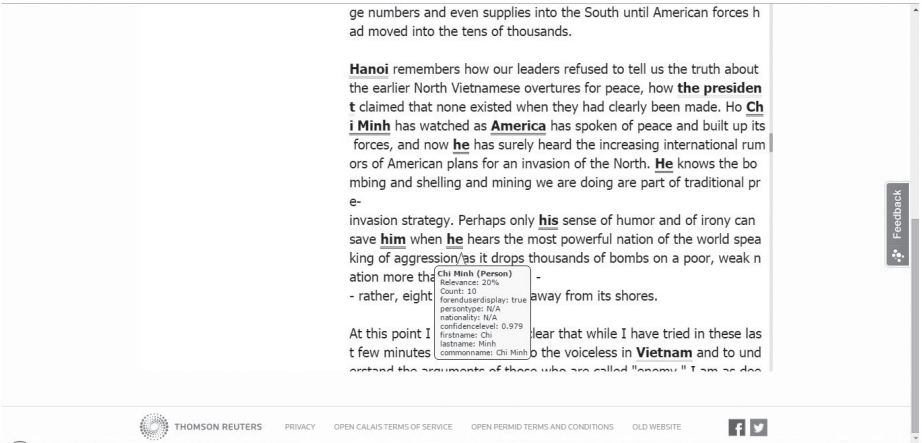


A program képes összekapcsolni a kifejezéseket, felfedezni a viszonyokat, így felismerni azt is, ha később csupán egy személyes névmással vagy rövidebb kifejezéssel utalunk egy személyre vagy szervezetre (II.2.5–II.2.6. ábra).

II.2.5. ábra – Egyes szavak, kifejezések közötti kapcsolatok felismerése 1.



II.2.6. ábra – Egyes szavak, kifejezések közötti kapcsolatok felismerése 2.



Láthatjuk tehát, hogy a névelem-felismerés jelentős mértékben képes segíteni a kutatómunkát, ám azt is, hogy az eredményeket óvatosan kell kezelni. További probléma, hogy mindeddig kevés magyar nyelvű szótár, illetve névelem-felismerő rendszer készült, amelyekre alább hozunk példákat.

Nemzetközi és hazai politikatudományi kutatások és egyéb alkalmazások

A politikatudományban gyakori alkalmazás az egyes közösségi oldalakon történő interakciók vizsgálata, ám például mikroblogok adatait egyéb politikatudományi kutatásokban is felhasználták. Számos tanulmány született Twitter-bejegyzések kapcsán (ld. pl. Nebhi, 2012). Tumasjan és társai (2010) azt vizsgálták, hogy a 2009. szeptemberi németországi választások eredményei előre jelezhetők-e Twitter-bejegyzések alapján. Az LIWC szövegelemző program segítségével több mint 100 000 bejegyzésben vizsgálták meg, hogy említésre kerül-e benne valamely német politikus vagy politikai párt. Arra jutottak, hogy a választás közeledtével növekedett a bejegyzések száma, s hogy a Twitter mára a politikai viták egyik színterévé vált. A bejegyzések vizsgálata során hasonló eredményre jutottak egyes pártok támogatottsága tekintetében, mint a közvélemény-kutatások. A névelem-felismerés az ilyen kutatások mellett alkalmas társadalmi hálózatok, így terroristacsoportok felderítésére is (ld. pl. Diesner – Carley, 2005).

Magyarországon egy érdekes alkalmazás a Szervezett Bűnözés Elleni Koordinációs Központ és a Szegedi Tudományegyetem közös projektjében elkészült

magyar nyelvű, bűnügyi névelem-felismerő rendszer, amelyről Molnár és társai (2010a; 2010b) írtak, illetve tartottak előadást a 2010. évi Magyar Számítógépes Nyelvészeti Konferencián. Ez a program egy, a Szegedi Tudományegyetemen korábban már kifejlesztett rendszer továbbfejlesztett változata. A rendszer megalkotásának célja a rendőrségi zárójelentések gyors feldolgozásának, a lényeges adatok kinyerésének és statisztikai adatok előállításának biztosítása volt. Ezen alkalmazás különlegessége, hogy nem a szokásos négy névelemosztályt (földrajzi név, személynév, szervezetnév, egyéb tulajdonnév), hanem összesen tizenhárom szemantikai osztályt különböztettek meg elkészítése során (például irányítószám, város, kerület, utca, házszám). Mivel a projekt során a névelemek egyes típusainak (például vezetéknév és keresztnév) megkülönböztetésére is szükség volt, kétszintű predikciós módszert alkalmaztak.

Nehézséget okozott a különböző névelemosztályok közötti gyakori átfedés (például egy adott szó településnév és vezetéknév is lehet). A *tanítóhalmazt* és a *teszthalmazt* kétszáz anonimizált dokumentumból állították össze, így a személynevek hiánya miatt a teszteredmények nem pontosak. Bár a valós adatokon készített modell egyes névelemosztályok tekintetében jobb eredményt mutatott, tesztelése összességében elmaradt az anonimizált dokumentumokon végzett vizsgálattól. A rendszer ugyanakkor lehetőséget nyújt az egyes név- és címadatok kiszűrésére (Molnár et al., 2010a: 366–369; Molnár et al., 2010b). Az ilyen üzleti és kormányzati alkalmazások egyre elterjedtebbek: Solt Illés és társai (2010) névelem-felismerést alkalmaztak általános egészségügyi problémák, gyógykezelések és vizsgálatok kórházi zárójelentésekben való azonosítására.

II.2.1. példa

A névelem-felismerésre érdekes példát nyújtanak az 1991-ben és 1992-ben megrendezett Dokumentummegértési Konferenciák (MUC-3 és MUC-4) témái. Mindkettő a Latin-Amerikában folytatott terroristacselekményekre vonatkozó adatok kinyerését tűzte ki célul. A projektben különböző, Latin-Amerikában elkövetett terroristacselekményekre vonatkozó hírösszefoglalókból hozták létre a mintegy ezerháromszáz szövegből álló korpuszt. Mindezt a hat modullal rendelkező, úgynevezett TACITUS rendszerrel végezték el, lehetővé téve a releváns események automatikus felfedezését és különválogatását (Hobbs et al., 1991). E két konferencia célja – csakúgy, mint a többié – az volt, hogy sor kerülhessen az információkinyerés kapcsán a szerzett tapasztalatok cseréjére és a terület további fejlesztésére, s ezzel továbbá a kormányzati igények kielégítésére.[5]

[5] Message Understanding Conference Proceedings http://www-nlpir.nist.gov/related_projects/muc/proceedings/proceedings_index.html (Letöltés ideje: 2015. december 3.)

Így jelentős eredménynek volt tekinthető, hogy a MUC-4 során kinyert adatok validitása magasabb szintű, a felidézés, a pontosság és az F-mérték is jobb, sokkal konzisztensebb az eredmény (Sundheim, 1992: 9–21).[6]

II.2.2. példa

A Twitter NewsCloud Alkalmazás segítségével megfigyelhetjük, hogy egy adott napszakban, tízperces bontásban egy adott hírfolyamban milyen gyakran és hányszor kerülnek említésre egyes névelemek (személyek, helyek, szervezetek).[7] Fehér Katalin írásából tájékozódhatunk a GeoX Kft. és a Zetema Ltd. által közösen kifejlesztett internetes tartalomelemző rendszer egy konkrét alkalmazásáról. Steve Ballmer, a Microsoft akkori vezérigazgatója 2013. augusztus 23-án jelentette be a cég éléről egy éven belül történő távozását. Az OpinHU webes véleményelemző rendszer segítségével megfigyelésre kerültek a Twitteren ezzel kapcsolatosan megjelenő aktivitások. Ennek keretében Steve Ballmer és a Microsoft említését vizsgálták 2013. augusztus 1-je és 29-e között, a Twitter NewsCloud alkalmazással. Ebből kiderült, hogy az első pár tíz percben a két tulajdonnév együtt jelent meg a hírfolyamban, a Microsoft említése ugrásszerűen növekedett, ahogyan Steve Ballmeré is, akinek a neve említésre sem került az esemény előtt. A hír hirtelen felfutása után az említések száma zuhanásszerűen le is csökkent.[8] Az OpinHu internetes tartalomelemző rendszer nyelvtechnológiai hátteréről bővebben Miháltz (2010) írásából tájékozódhatunk. Kitűnően alkalmazhatjuk tehát a névelem-felismerést akkor, ha egy meghatározott ügy közösségi médiában vagy mikroblogolás során történő felbukkanását szeretnénk vizsgálni.

II.2.3. példa

Érdekes alkalmazás az Európai Unió, az Európai Unió Belügyi Főigazgatósága közreműködésével kifejlesztett, az Európai Unió határvédelmi ügynökségéhez, a Frontexhez köthető nyílt forrású felderítő rendszer, amely a határvédelemhez köthetően képes azonosítani személyek és szervezetek neveit, helyeket, kapcso-

[6] A MUC-3 és MUC-4 adatbázisai ingyenesen hozzáférhetők: MUC Data Sets http://www-nlpir.nist.gov/related_projects/muc/muc_data/muc_data_index.html (Letöltés ideje: 2015. december 3.)

[7] Twitter NewsCloud alkalmazás https://sites.google.com/a/geox.hu/opinhu/elemzesi-eszkoezoek/twitter-newscloud-alkalmazas (Letöltés ideje: 2015. október 29.)

[8] Fehér Katalin: Microsoft – Steve Ballmer lemondása a Twitteren https://sites.google.com/a/geox.hu/opinhu/hirek-ujdonsagok/microsoft-steveballmerlemondasaatwitteren (Letöltés ideje: 2015. október 29.)

latfelvételi adatokat (pl. telefonszám, e-mail cím), hitelkártyaszámokat, továbbá adóazonosítókat. A rendszer az adatok összegyűjtésén és csoportosításán túl képes azok elemzésére, kapcsolatbányászatra is. A rendszer csak zárt körben elérhető, a végfelhasználók az Európai Unió egyes tagállamainak hatóságai, rendőrségei, határőrségei. A segítségével felhalmozott tudás a korábbi ismeretekkel kombinálva segíti elő a felderítést.[9]

Egy másik, a határőrizethez kapcsolódó alkalmazás a Frontex valós idejű híresemény-feldolgozó keretrendszere, amely képes nyolc nyelven, valós időben kigyűjteni az online hírekből és egyéb nyílt forrásokból (pl. mikroblogokból) a határőrizeti szempontból releváns strukturált információkat, szintén megkönnyítve a felderítést. Ilyen, lényeges információk lehetnek az illegális határátlépési ügyek és az ehhez kapcsolódó olyan bűncselekmények, mint az embercsempészet és emberkereskedelem, továbbá a válsághelyzetek (erőszakos események, tüntetések, fegyveres konfliktusok, humanitárius vagy természeti katasztrófák, illetve fertőző betegségek). A rendszer nem csupán az információk kinyerésére alkalmas, hanem annak egy térképen való elhelyezésére is képes. Segítségével nem csupán az illegális határátlépéshez kapcsolódó felderítés válik könnyebbé, hanem segítséget nyújthat a migrációs nyomás várható megnövekedésének előrejelzésére és az arra való felkészülésben is.[10]

Ellenőrző kérdések

- Mi a különbség a névelem és a tulajdonnév között?
- Mi a különbség a névelem-felismerés és az osztályozási probléma között?
- Melyek a névelem-felismerés előnyei a szózsák modellhez képest?
- Találjon ki egy feladatot, amelyet csupán a névelem-felismerés segítségével lehet megoldani, a szózsák modell segítségével nem!
- Milyen politikatudományi kérdés megoldására alkalmazná a névelem-felismerés módszerét?
- Tegyük fel, hogy az Országgyűlési Naplóból ki szeretnénk gyűjteni a személyneveket és a földrajzi neveket! Ez esetben milyen módszertani problémák merülhetnek fel a névelem-felismerés alkalmazása során?

[9] EMM Open Source Intelligence Suite http://btn.frontex.europa.eu/system/files/private/resources/tools/emm-osint-suite_productinfo.pdf (Letöltés ideje: 2015. október 30.)

[10] Frontex Real-time News Event Extraction http://btn.frontex.europa.eu/projects/internal/frontex-real-time-news-event-extraction (Letöltés ideje: 2015. december 3.)

Szószedet

Magyar	Angol
Címke	Label
Dokumentummegértési konferencia	Message Understanding Conference (MUC)
Felidézés, fedés, teljesség	Recall (score)
F-mérték	F1 score
Hitelesség	Accuracy
Információkinyerés	Information extraction
Kapcsolatbányászat	Relation extraction
Névelem-felismerés	Named entity recognition
Pontosság, megbízhatóság	Precision (score)
Szóalak	Token
Szövegelőkészítés, -előfeldolgozás	Preprocessing
Tématerület (szövegtartomány)	Domain
Természetesnyelv-feldolgozás	Natural language processing
Természetesnyelv-tanulási konferencia	Conference on Natural Language Learning (CoNLL)
Webes információkinyerő rendszer	Wrapper

Ajánlott irodalom

A névelem-felismeréssel kapcsolatos ismeretek mélyítéséhez, bővítéséhez az alábbi irodalmakat ajánljuk: Marrero et al. (2013); Russel – Norvig (2005); Tikk (2007); Tikk et al. (2005); Móra – Farkas (2010); Gosztolya – Tóth (2010).

III. SZÖVEGBÁNYÁSZATI FELADATOK: A SZÖVEGEK GÉPI KÓDOLÁSA

Molnár Csaba

III.1. OSZTÁLYOZÁS (KLASSZIFIKÁCIÓ)

A fejezetben a fejlett gépi szövegbányászati feladatok egy típusával, a szövegelemek ismert csoportokba való besorolását elvégző osztályozással ismerkedünk meg. A későbbi, konkrét eljárásokat részletező fejezeteket megelőzően bemutatjuk az ilyen feladatok két legelterjedtebb megoldási módszerét: a szótáralapú megoldásokat és a felügyelt tanulási eljárásokat, ezek előnyeit és hátrányait, illetve alkalmazhatóságát. Az eljárások lépéseinek ismertetése mellett kitérünk az eredmények érvényességének problémájára is.

Kutatásaink során gyakran szembesülhetünk azzal a problémával, hogy nagy mennyiségű szöveg közül kell kiválogatnunk a számunkra fontosakat. Kutatásunk során szükségünk lehet arra is, hogy a kutatott forrásszövegek egyes elemeit valamilyen szempont alapján szétválogassuk. E fejezetben az ilyen *osztályozási* problémák gépi megoldásait tekintjük át. Az idegen és magyar nyelvű szakkönyvek terminológiája korántsem egységes e technikák kapcsán. A *szövegosztályozás* (Tikk, 2007c: 102) az angol text classification (TC) megfelelője, amit néha *kategorizálásnak* (Abonyi, 2006: 397), *klasszifikációnak* vagy *csoportba sorolás*nak (Bodon, 2010: 113) is fordítanak.

További szóhasználati nehézséget okoz, hogy az irodalomban elterjedt a „kemény", illetve a „puha" osztályozásra való utalás. Előbbi a boole-i értelemben vett kategorikus, az adott halmazba tartozást vagy nem tartozást jelentő megkülönböztetést jelenti, míg utóbbi skálaszerű értelemmel bír: az eredmény azt jelenti, milyen mértékben sorolható egy dokumentum az adott csoportba (Wajeed – Adilakshmi, 2009: 121). Az osztályozási módszerek fő eltérése a kézi és géppel támogatott kódolástól abban áll, hogy itt nem vizsgálunk meg egyesével minden besorolandó szöveget, hanem azokat valamilyen automatizmus segítségével osztályozzuk.

A dokumentumok rendezése mindezek fényében két fő eljárásban végezhető el, melyek közül az egyik deduktív, a másik induktív logikát érvényesít. E két szempont alapján a dokumentumrendezés logikáját és eljárásait foglalja össze a III.1.1. táblázat.

III.1.1. táblázat – A dokumentumrendezés logikái és eljárásai

Rendezési logika / Eljárás	Boole-i bináris besorolás (beletartozik a kategóriába: igen/nem)	Valószínűségalapú besorolás
Deduktív (előre megadott kategóriák)	Szótáralapú	Felügyelt tanulás
Induktív (előre ismeretlen kategóriák)	-	Felügyelet nélküli tanulás

Az osztályozási megközelítés alkalmazása során már korábban kialakított, létező csoportokba soroljuk be a vizsgált elemeket (deduktív logika). Ilyen feladat lehet például a kormányról szóló újságcikkek csoportosítása azok pozitív vagy negatív tartalma szerint. A vizsgálat lényege abban áll, hogy már annak megkezdése előtt tudjuk, hogy milyen lehetséges kategóriákba kerülhetnek be a korpuszt alkotó szövegek (a fenti példánál maradva pozitív, semleges vagy negatív tartalmúak) vagy ezek elemei. A program feladata minden egyes szöveg esetében a megfelelő csoport kiválasztása lesz. Szemben a csoportosítási feladattal, a csoportok az osztályozási feladatnál előre adottak, és nem változnak a folyamat során (az ilyen deduktív eljárásokról részletesebben a IV.1. fejezetben írunk).

A klasszikus feladat

A rendezési logika (deduktív vagy induktív) kettőssége ugyanakkor nem jelenti azt, hogy az osztályozási feladatok ne lennének visszavezethetőek egy közös problémára. Rendelkezésünkre áll egy *tanítóhalmaz* (training set), melynek minden egyes eleméhez egy osztályozó értéket rendelünk egy több diszkrét értékből összeállított listából. A tanítóhalmaz adatai arra szolgálnak, hogy kialakítsunk egy osztályozási modellt, amely a tanítóhalmaz elemei tulajdonságaihoz rendeli hozzá a különböző értékeket. A modell feladata ezek alapján előre jelezni más elemek osztálytagságát. A problémának létezik „*kemény*" és „*puha*" változata is. Míg előbbi kategorikusan besorol minden elemet egy-egy osztályba, utóbbi csak az osztály-hovatartozás valószínűségét határozza meg (Aggarwal – Zhai, 2012b: 163–164).

E kutatási döntés a besorolás jellemzőiről (kemény, illetve puha osztályozás) egyben két tágan értelmezett eljárástípust eredményez (szótáralapú, illetve felügyelt tanulási megoldás, lásd lentebb). Ami közös bennük, hogy úgy kívánjuk egy korpusz elemeit szétosztani kategóriák között, hogy a csoportok közös is-

mérveit már kutatásunk elején ismerjük (szemben a csoportosítással és a hozzá kapcsolódó felügyelet nélküli tanulási technikákkal). A legegyszerűbb esetben a vizsgált korpusz elemei egyértelműen és teljes körűen besorolhatóak az általunk kialakított kategóriákba.

Vegyük azt a feladatot, amikor a parlamentben elhangzott képviselői kérdéseket próbáljuk közpolitikai témakörök szerint csoportosítani! Itt szembesülhetünk az egyszintű és a hierarchikus kategóriák problémájával. Előbbi csoportba sorolhatóak azok az eljárások, ahol olyan kategóriákat hozunk létre, melyek tagjai nem tartoznak más, közös gyűjtőcsoportba, míg utóbbi esetében a csoportok között alá-fölé rendeltségi viszony állhat fenn. (Qi – Davison, 2009: 12:3) Mivel olyan kategóriákat kell kidolgozni, melyekbe minden interpelláció besorolható, felmerül a kérdés, hogy rendezzük el ezeket a felszólalásokat. Több olyan közpolitikai terület van, melyek vagy nagyon kis elemszámot foglalnak magukban (például postaügyek), vagy beilleszthetőek más kategóriákba is (például a közoktatás az oktatásügybe). Ebben az esetben akkor járunk el helyesen, és teremtünk egyértelmű kategóriákat, ha hierarchikus kategóriákat alkotunk. Amennyiben sikerült kialakítanunk a majdani csoportokat, megkezdődhet az osztályozás eljárása.

Tipikus megoldások

A dokumentumok és ezek elemeinek ismert csoportokba rendezésére két fő módszer létezik: a *szótáralapú* (dictionary-based) és a *felügyelt tanulási* (supervised learning) megoldás. Előbbi lényege abban áll, hogy a meglévő szövegeket egy korábban elkészített szószedet elemeinek felbukkanása alapján rendezzük kategóriákba, míg utóbbiaknál egy program elemez egy általunk korábban elkészített tanítóhalmazt, és az abban felismerni vélt minták alapján osztja be a kategóriákba a teljes korpusz elemeit.

Szótáralapú megoldások

Az ismert csoportokba rendezési feladatok első megoldási módszerét szótáralapú megoldásoknak nevezzük. Az ilyen eljárás minden esetben egy *szószedet* (ld. Cioffi-Revilla, 2014: 61) elkészítésével veszi kezdetét. Az összes előre meghatározott csoport esetében összegyűjtjük azon szavak listáját, melyekről feltételezzük, hogy az adott kategóriába sorolható szövegekben felbukkannak. Például amennyiben a külpolitika témaköréhez akarunk összeállítani egy szószedetet, érdemes felvennünk hozzá a nagykövet, a nagykövetség, a konzul, a konzulátus, az attasé, a diplomata, a diplomácia szavakat és más hasonlóakat.

Ha elkészítettük a szószedetet, az általunk alkalmazott program megvizsgálja a korpusz elemeit, és a bennük felbukkanó szavak alapján valamelyik kategóriába besorolja azokat (Grimmer – Stewart, 2013: 8–9 – az osztályozási eljárás pontos végrehajtásának részleteibe a IV.1. fejezet nyújt bevezetést).

A szótáralapú megoldások két fő csoportját a kategóriák száma alapján különböztetjük meg: ilyen a kétkategóriás és a többkategóriás besorolás. Előbbire jó példa az a kutatási feladat, amikor arra vagyunk kíváncsiak, hogy egy parlamenti kérdésben megemlítik-e Magyarország valamely települését, vagy sem. A szótár ekkor minden településnevet tartalmaz minden toldalékolható formában. A program azt keresi meg, hogy a szövegekben ezen szavak közül felbukkan-e valamelyik. Ha igen, a településnevet tartalmazó, ha nem, a településnevet nem tartalmazó csoportba kerül a szöveg. A többkategóriás besorolásra a fent már említett közpolitikai témakör szerinti kategorizálás egy kézenfekvő példa. Ekkor minden meghatározott kategória esetében felsoroljuk annak fontos szavait (például hadsereg, légierő, honvédség, sorkötelesség, tartalékos stb. a honvédelem kategóriájában).

A szótáralapú megoldásoknak viszonylagos egyszerűségük ellenére vannak hátulütői. A munkafolyamatban komoly problémát jelent a szószedet összeállításának nehézsége. Érdemi feladat előzetesen összegyűjteni az olyan szavakat, melyek *csak az egyik* kategória esetében bukkanhatnak fel, és nem rendelkeznek olyan jelentéssel, mely egy másik csoportra is illik. Hasonlóképpen nehéz minden egy témához tartozó fontos kifejezést előzetesen meghatározni. A feladatot tovább nehezíti a kategóriák számának növekedése, illetve az a tény, hogy a program nem értelmezi a szövegkörnyezetet, egyszerűen csak szavakat keres benne. Így aztán nem ismeri fel a gúnyt, az iróniát, a metaforákat és az egyéb szóképeket sem. Nem tud átvitt értelemben gondolkodni, így számtalan esetben félreértheti a szöveget (például Orbán Viktor „csapataink harcban állnak" kijelentését a szótáralapú módszerek honvédelemhez sorolnák be, miközben a valóságban az Európai Unió költségvetési tárgyalásaival kapcsolatban mondta).

A másik probléma a szavak szövegkörnyezetétől függő jelentésében rejlik. Ha egy szöveg pozitív vagy negatív tartalmára vagyunk kíváncsiak (azaz ebbe a két osztályba szeretnénk besorolni őket), nem mindegy, hogy a „hangos" szó a hangszóró vagy a gépkocsi leírásában szerepel (Liu – Zhang, 2012: 429). A módszernek szintén problémát jelent, hogy egyszerű változatában nem ismer fel kifejezéseket, így egy Horn Gyula keresztnevét említő tételt könnyen besorolhat a településneveket tartalmazó felszólalások közé. Hasonló problémát jelentenek az azonos alakú szavak: az új *követ* kinevezése, az a tény, hogy egy személyt valaki *követ,* esetleg hogy hol bányásszák az útépítéshez felhasznált *követ,* a módszer eszközeivel megkülönböztethetetlen.

A szótáralapú eljárás harmadik módszertani korlátja nehézkes validálásában rejlik. Az elemzés érvényességének egyik ellenőrzési lehetősége a korpuszon belüli *validációs halmaz* kiválasztása. Ekkor miután a program kategorizálta a szövegeket, egy véletlen mintavétellel kiválasztott csoportot kézzel lekódolunk, ezzel ellenőrizzük, hogy megfelelő találati arányt sikerült-e elérnünk (Grimmer – Stewart, 2013: 13). A második lehetőség a szótár validálása, azaz többszöri átgondolása, ellenőrzése (i. m., 2013: 9). A harmadik lehetőség az *előrejelző validálás,* azaz annak ellenőrzése, hogy az eredmény megfelel-e várakozásainknak (i. m., 2013: 21). Így például, ha az 1990 előtt elfogadott törvényeket vizsgáljuk, feltételeznünk kell, hogy a világháborúk előtt és alatt az átlagnál több honvédelemmel kapcsolatos törvényt fogadott el az Országgyűlés. Negyedikként a *konvergens validálás* eszközéhez is nyúlhatunk: ez a módszer azt jelenti, hogy ugyanazt a korpuszt több lehetséges módszerrel is lekódoljuk, majd az eredményeket összevetjük (i. m., 2013: 24). Mindezen eljárásokat a gyakorlati kutatások során keverni is lehet.

Felügyelt tanulási megoldások

Az ismert csoportokba sorolás másik módját a felügyelt tanulási eljárások jelentik. A módszer lényege, hogy a korpusz egyik szövegcsoportját kézzel kategóriákba soroljuk, majd a program ez alapján sorolja be a korpusz teljes anyagát a különböző csoportokba (Grimmer – Stewart, 2013: 9). Ez az általános eljárás a gyakorlatban három szakaszra tagolódik. A legelső feladatunk az úgynevezett tanítóhalmaz elkészítése. A tanítóhalmaz egy, a korpuszból vett szövegminta. A véletlen mintavétellel kiválasztott szövegeket kézzel kódoljuk. A második szakasz akkor veszi kezdetét, ha a tanítóhalmaz elkészült: ekkor egy kiválasztott felügyelt tanulási programmal a gép a korpusz többi elemét is besorolja valamely kategóriába (a lehetséges eljárásokat a IV.1. fejezetben részletezzük). A munkafolyamat harmadik, lezáró szakasza a kapott eredmények érvényességének vizsgálata. Ennek módja lehet a korábban leírt validálási csoport kiválasztása, az előrejelző validálás, illetve a konvergens validálás. Az eltérő munkamenet a szótáralapú megoldásokhoz képest új validálási lehetőséget teremt: a *keresztvalidálást.* E művelet során a tanítóhalmazt csoportokra osztjuk, és azokkal újra és újra elvégezzük a felügyelt tanulásos kódolást (i. m., 2013: 13).

A felügyelt tanulási módszerek legnagyobb előnye, hogy a korpusz részletének szavait felhasználva működik, így nem szembesülünk a szótárkészítés nehézségeivel. Emellett az így lekódolt korpusz sokkal könnyebben validálható statisztikai módszerekkel, mint a kézi vagy szótáralapú kódolás eredménye.

Hátránya elsősorban a jelentős előerőt igénylő tanítóhalmaz-készítésben és gépi tanulási eljárás komplexitásában rejlik.

III.1.1. példa

A szótáralapú megoldások egyik legkézenfekvőbb alkalmazási területe a politikai kommunikáció vizsgálata. Egy szerzőhármas (Klebanov et al., 2008) Margaret Thatcher és Tony Blair beszédeit hasonlította össze a szótáralapú megoldások használatával. A kutatók arra voltak kíváncsiak, hogy milyen fogalomcsoportok (például politikára, munkára, földrajzi névre és mozgásra utaló kifejezés) bukkannak fel a szövegekben. Tanulmányuk alapvetően kísérleti jellegű, a különböző számítógépes szövegelemzési módszerek összehasonlítására szolgál, melynek egyik fontos következtetése, hogy összehasonlító elemzésekre kiválóan alkalmasak az ilyen, szótárakon alapuló megoldások.

III.1.2. példa

Napjainkban a számítógépes adatelemzés gyakorlatilag kimeríthetetlen kutatási anyagát szolgáltatja az internet. Egy koreai szerzőpáros (Chung – Noh, 2002) kísérletezve a szótáralapú megoldásokkal a Usenet-hálózat hírgyűjtője által szemlézett honlapokon fellelhető szövegeket próbálta meg csoportosítani, kiválogatva közülük a közgazdaságtannal kapcsolatosakat. Ezt a feladatot szótáralapú megoldással rendkívül magas, a validáltan helyes besoroláshoz képest 96%-os pontossággal teljesítették.

III.1.3. példa

Az osztályozási eljárásokat nemcsak szövegek, hanem multimédiás tartalmak esetében is lehet alkalmazni, melyre egy természettudományi példát hozunk. A zebradánió valószínűleg az akvarisztika minden rajongója számára ismerős halfajta. Az elmúlt időszakban esztétikai értékén kívül újabb hasznos tulajdonsága miatt kezdték tenyészteni: a vegyi anyagok hatása e halak egyedfejlődésére emlékeztet az emlősök, így az emberek esetében is észlelhető hatásokra, ráadásul igen jól megfigyelhető és jelentős elváltozást váltanak ki bennük ezek a szerek. Ennek köszönhetően a gyógyszerészet és a toxikológia előszeretettel használja fel azokat kísérleteihez. A kutatók hagyományosan kézi kódolással végzik kísérleteiket, azaz a különböző anyagok halivadékokra gyakorolt hatásának elemzését. Fényképet készítenek a háromnapos halivadékokról, majd azokat egyesével

elemzik. Egy kutatócsoport (Jeanray et al., 2015) ugyanezt a feladatot kézi kódolás helyett felügyelt tanulási módszerrel oldotta meg, amelynek eredményeképp megbízható, 90–100% közötti egyezésre sikerült jutni a kézi kódoláshoz képest.

Ellenőrző kérdések

- „Puha" vagy pedig „kemény" osztályozáshoz érdemesebb fordulnunk, ha arra vagyunk kíváncsiak, hogy egy újságcikkhalmaz elemei jónak vagy rossznak találnak egy kormányzati intézkedést? Miért?
- Fogalmazzon meg egy példát saját kutatási területén belül, ahol szótáralapú megoldást érdemes használni! Miért nem lenne itt megfelelő a felügyelt tanulási eljárás?
- Az osztályozás módszeréhez milyen validálási eljárások használhatóak fel?
- Milyen mértékben lennének ezek a módszerek alkalmasak egy általunk nem ismert ország belpolitikájához kapcsolódó szövegek közpolitikai témakörönkénti osztályozására?
- Hogyan validálható az a kutatás, melynek eredménye szerint az interpellációkban legsűrűbben emlegetett települések között szerepel a Heves megyei Balaton és a Hajdú-Bihar megyei Polgár?

Szószedet

Magyar	Angol
Előrejelző validitás (érvényesség)	Predicitive validity
Előrejelző validálás	Predictive validation
Keresztvaliditás	Cross-validity
Keresztvalidálás	Cross-validation
Klasszifikáció, osztályozás	Classification
Konvergens validitás	Convergent validity
Konvergens validálás	Convergent validation
Szószedet	Hash table
Szótáralapú megoldások	Dictionary-based methods
Szövegosztályozás	Text classification
Tanítóhalmaz	Training-set
Validációs halmaz	Validation set

Ajánlott irodalom

Az osztályozási feladat kapcsán áttekintést nyújt Abonyi (2006); Bodon (2010); Grimmer és Stewart (2013); valamint Qi és Davison (2009). A szótáralapú megoldásokba Aggarwal – Zhai (2012d); illetve Cioffi-Revilla (2014) nyújt bevezetést. A felügyelt tanulási megoldások alapjai kapcsán jó indulószöveg: Wajeed – Adilakshmi (2009). A példákban felbukkanó tanulmányok forrása: Chung – Noh (2002); Jeanray – Marée – Pruvot – Stern – Geurst – Wehenkel – Muller (2015); Klebanov – Diermeier – Beigman (2008).

Szabó Gabriella

III.2. VÉLEMÉNYELEMZÉS MINT SPECIÁLIS OSZTÁLYOZÁSI FELADAT

A fejezet bemutatja a véleményelemzés (sentiment analysis) eljárását, segít megérteni a módszer hátterében álló előfeltételezéseket, a véleményelemzést végző kutatások elméleti beágyazottságát, kérdésfelvetését, valamint operacionalizálási logikáját. Az olvasó képet kap a megközelítés teherbíró képességéről, erősségeiről és korlátairól. A fejezet lépésről lépésre végighalad az előkészítés, az érzelmi viszonyulások osztályozása, a szótárépítés és az indikátorszavak felismerése során elvégzendő feladatokon. Rámutatunk azokra az eljárás során felbukkanó dilemmákra, melyek feloldása kutatói döntést igényel. A fejezet végén feltüntetett példák jelzik a véleményelemzés által feltárható eredmények használhatóságát és továbbgondolásának lehetőségeit.

A *véleményelemzés* a módszeres szövegosztályozás (ld. III.1. és IV.1. fejezet) egyik lehetséges útja. Egy olyan részben automatizált információkereső vizsgálat, mely egy bizonyos verbális kommunikációnak a tárgyalt témához való érzelmi viszonyulását tárja fel.

Az érzelmi viszonyulások szövegelemzésére vállalkozó kutató a nemzetközi szakirodalom tanulmányozásakor egyfajta terminológiai kavalkádjával találja magát szembe. Magyarul nincs is bevett fordítása az *opinion mining, subjectivity analysis, review mining, appraisal extraction, affective computing* módszereknek, melyek tulajdonképpen ugyanazt az asztalt táncolják körül. A Pang – Lee (2008: 5–6) szerzőpáros hosszan értekezik az elnevezések tudománytörténeti aspektusairól, végső konklúziójuk azonban az, hogy a nemzetközi tudományos közösség mára leginkább a sentiment analysis és az opinion mining fogalmakat használja, azokat is jobbára szinonimaként. Bár a „szentimentelemzés" is használatos (Szabó, 2015), a módszer hazai alkalmazásában a számítógépes nyelvészek vállalnak úttörő feladatot, ezért úgy illik, hogy az általuk bevezetett véleményelemzés szókapcsolat-fordítást használjuk itt is (Tanács – Vincze, 2010).

Milyen tudáshoz segít hozzá a véleményelemzés? Röviden: megtudhatjuk, mi van az emberek fejében. Részletesebben: a fejezet segít megérteni a véle-

ményelemzés hátterében álló előfeltételezéseket, a véleményelemzést használó kutatások kérdésfelvetését, módszerét és operacionalizálási logikáját. Az olvasó a társadalomtudomány szemszögéből nézve képet kap a véleményelemzés teherbíró képességéről, erősségéről és korlátairól. Lépésről lépésre végighaladunk azon az úton, amit a kutató is bejár az előkészítéstől az adatok összeállításán át, azok korlátainak felismeréséig. A fejezet főszövege és a közölt példák igyekeznek világossá tenni a véleményelemzés által feltárható eredmények használhatóságát és továbbgondolási lehetőségeit.

A társadalomtudományi kutatások mögött gyakran annak a vágya áll, hogy közelebb jussunk a megszólaló gondolataihoz, kiderítsük, miként vélekedik egy bizonyos témáról, személyről, eseményről, helyről vagy termékről. Az ilyen vizsgálatok kiindulópontja szerint a véleményekből következtethetünk az emberek egyéni viselkedésére: ha azt mondom, *„a blézer színe szép és elegáns a szabása"*, akkor nagy valószínűség szerint szívesen megvásárolnám. Feltételezik azt is, hogy a valakivel vagy valamivel kapcsolatban megfogalmazott pozitív érzelmi megnyilvánulások erős kötődést jelölnek: ha azt írom, *„Angelina Jolie remek munkát végez Afrikában"*, akkor feltételezhetően megnő az esélye annak is, hogy adományt küldök a színésznő által támogatott segélyszervezetnek.

A széles nyilvánosság számára elérhetővé tett érzelmekre alapuló vélemények ráadásul gyakran a többiek viselkedését is befolyásolják. A megközelítés tehát arra épül, hogy az érzelmi kötődés bármilyen egyéb kalkulációnál, befolyásnál fontosabb lehet az egyén és a csoport viselkedésében, és fordítva: a vélemények megváltozása az érzelmi kötődés lazulását vagy felbomlását mutatja, mely végzetes lehet az adott termékre, személyre vagy országra nézvést. Ezért érdekli vállalatok, kormányok és kampánytanácsadók sorát az állampolgárok és fogyasztók érzelmi viszonyulásaira, szubjektív benyomásokra utaló véleménye, és ezért monitorozzák a közösségi médiát, a közösségi értékelő, illetve ajánló weboldalakat (mint például a Rotten Tomatoes vagy a magyarul is elérhető Tripadvisor).

Biztosra vehető, hogy az érzelmek az internetrobbanás előtt is fontos szerepet játszottak egy-egy személyes vagy kollektív döntés meghozatalában. Az új millennium azonban fordulatot hozott: egymást erősítő elméleti és empirikus, alap- és alkalmazott kutatások fedezték fel az érzelmeket, és általuk igyekeztek jobban megérteni a fogyasztási, üzleti, közéleti és politikai viselkedést (vö. Neuman et al., 2007; Kiss, 2013). A nagymintás adatokra építő empirikus vizsgálatok lendületét a webkettő[1] elterjedéséhez köthetjük. Az emberek jó része láthatóan szívesen osztja meg érzelmeit a nyilvánossággal. A blogok, fórumok, csetszobák, tartalommegosztó oldalak, Twitter, Facebook és egyéb

[1] A webkettő vagy web 2.0 elnevezéssel az internetes kultúra közösségre, kapcsolattartásra, megosztásra és a felhasználók által generált tartalmakra koncentráló fejleményeit szokás illetni (vö. Szűts, 2012).

közösségi médiafelületek pedig óriási és viszonylag könnyen feltárható kincsesbányát ajándékoznak a kutatóknak. Számtalan kérdésben beavatkozásmentes módszerrel (Babbie, 2003: 351), a kutatótól függetlenül előállt adatok (Silverman, 2007: 37–60) alapján megtudható, hogy ki mit érez egy-egy termék, személy, helyszín vagy politikai ügy kapcsán.

A klasszikus feladat

Az érzelmek feltárásában és kategorizálásában egy újabb állomás a magyar szaknyelvben véleményelemzésként meghonosuló *sentiment analysis*. Teoretikus háttere a pszichológia érzelemelmélete, annak is egyik leágazása, az ún. értékeléselmélet. E paradigma gazdag irodalmából számunkra elegendő annyi, hogy központi állítása szerint az érzelmek kiváltásáért és megkülönböztetéséért egy adott személy, esemény vagy termék értékelése a felelős (vö. Roseman – Smith, 2001). A véleményelemzés azonban megfordítva veszi fel a szálat, azaz az értékelésekből következtet a személy, a termék, az esemény, helyszín vagy bármi más iránti egyéni érzésekre. Célja a többnyire szövegekben megjelenő és abból kinyerhető érzelmek, értékelések, álláspontok, ítéletek, benyomások *semleges, pozitív vagy negatív* kifejezéseinek azonosítása és feltárása. Jellemző alkalmazási területe a marketing, pontosabban a márkaépítés *(branding),* de a katonai-polgári hírszerzés is felfedezte a véleményelemzésben rejlő lehetőségeket (Pang – Lee, 2008).

Politikatudományi alapkutatásokban az áttörés ugyanakkor még várat magára. A Twitteren megjelenő hangulatok és politikai preferenciák összefüggéseit feltáró néhány exploratív vizsgálatot leszámítva (Bakliwal et al., 2013; O'Connor et al., 2010) alig találunk példát a sentiment analysis használatára. Pedig nagyon is alkalmas lenne különösen a zártabb politikai közösségek (ld. pl. radikális jobboldal) működésének feltérképezésére, vezetők imázsának politikai preferenciákra gyakorolt hatásának vizsgálatára, vagy egy-egy szakpolitika körül kialakult állampolgári diskurzusok feltárására. Általánosságban pedig minden olyan helyzet leírására, ahol az informális, intézményeken túli, nem szabályozott, de politikai relevanciával bíró kommunikációnak, az érzelmeknek, az impressziónak, a szubjektív értékeléseknek kiemelt fontosságot tulajdonítunk.

Az érzelmi viszonyulások feltérképezéséhez hagyományosan két úton juthattunk el: szótáralapú megközelítéssel vagy gépi tanulási algoritmusok (ld. IV.1. és IV.2. fejezet) segítségével. Legújabban pedig e kettő ötvözetével.

A véleményelemzés a szövegeket három szinten osztályozza. A legáltalánosabb a dokumentumszintű osztályozás, amikor egy hosszabb szövegegység egészéről állapítjuk meg annak érzelmi viszonyulását. Aprólékosabb a mondat-

szintű osztályozás, mely ugyanezt teszi, de itt a vizsgálat alapegysége a mondat. A mondatszintű osztályozás esetében két feladatunk van: megállapítani, hogy van-e értékelés, vélemény vagy érzelem az adott szövegrészben, és ha igen, akkor milyen azok érzelmi tartalma.

A legrészletesebb adatokat akkor nyerjük, amikor az elemzés alapegységei annak a témának a különböző aspektusai, melyekre az érzelmek vonatkoznak. Előfordulhat ugyanis, hogy a megszólaló részletes véleményt fejt ki, és az adott terméket, személyt, eseményt vagy helyszínt több szemponttól tárgyalja (Wang – Liu, 2015: 1–2). Egy megszólaló dicsérheti egy szálloda személyzetének udvariasságát, de egy másik mondatban az ágyi poloskák okozta kellemetlen élményeiről is beszámolhat. A szálloda mint téma két különböző aspektusa tehát a személyzet viselkedése és a matracok tisztasága, melyekre ellentétes irányú vélemények és érzelmek vonatkozhatnak. A társadalomtudományi vizsgálatok többsége a dokumentum vagy a mondat szintjén elemez, ezért a későbbi fejezeteink is ehhez alkalmazkodva tárgyalják a véleményelemzés lépéseit.

Gyakorlati alkalmazás

E szakaszban áttekintjük a véleményelemzés lépéseit, rámutatunk egyes módszertani dilemmákra és ezek lehetséges megoldására. A megértést elősegítendő a komplikációkra konkrét példákat is hozunk.

Előkészületek

Az első munkafázis az előkészületeké, melyben legalább három dolgot kell elvégezni. Egyrészt készíteni kell egy olyan korpuszt, egy mesterfájlt, mely nem tartalmaz mást, csak az összes kódolandó tartalmat. Törlendők a HTML elemek, a beágyazott képek és videók (kivéve, ha ezek is elemzendő anyagok), dátumok, nevek, a weboldalak keretei stb. A korpusz kialakításakor célszerű leválogatni az értékeléseket, érzelmi viszonyulást nem tartalmazó elemeket is. Ha egy adott téma médiatálalását vizsgáljuk, akkor valószínűleg véleményközlő cikkekre fókuszálunk, így az újságok hír- és tényanyagaira nincs szükségünk. A válogatást végezhetjük manuálisan, de adott esetben felügyelt gépi tanulással működő algoritmusok is alkalmazhatók (Yu – Hatzivassiloglou, 2003; Pang – Lee, 2002; bővebben: IV.1. fejezet). A korpusz összeállítását nehezíthetik archiválási hiányosságok (a források részben vagy egészben nem kerültek rögzítésre), technikai problémák (a forrásokat analóg eszközökkel rögzítették), a papíralapú gyűjtések digitalizáltságának alacsony mértéke vagy a hanganya-

gok szöveges átiratának hiánya. Az is előfordulhat, hogy az adatforrások magas költségtérítés ellenében használhatók.

A teljes korpusz használatának alternatívája a mintavétel, vagyis a korpuszból reprezentatív részeket vehetünk ki, és a továbbiakban ezt a válogatást tekintjük az elemzendő anyagok gyűjteményének. A mintavétel nem kötelező, de kezelhetetlenül nagy vagy áttekinthetetlen korpusz esetén hasznos lehet. Amennyiben a mintavétel mellett döntünk, érdemes azt az előkészületi szakaszban megtenni. Hogy mi kerül a reprezentatív mintába, nyilvánvalóan meg kell indokolni, és ez a kutatási céloktól, a kutatási kérdéstől függ.

Végül a szövegek osztályozásához szükséges kategóriák definiálását is előkészítjük. Az érzelmek elemzéséhez szükséges kategórialista a kemény típusú klasszifikációhoz tartozik (ld. III.1. fejezet), kialakítása esetünkben tulajdonképpen egyfajta szótárépítés. Az induktív megközelítés szerint egy jól használható és érvényes szótár összeállításához nélkülözhetetlen a kutatási anyag valamilyen szintű áttekintése. Nincs egyértelmű szabály arra, hogy a korpusz mekkora részét érdemes ilyenkor megvizsgálni. A véleményelemzés esetében is támaszkodhatunk a kvalitatív tartalomelemzés általános elvére, mely előkutatásra a teljes elemzendő anyag 10 és 50 százalékát javasolja (Mayring, 2014: 80). Az előkészítés során tehát a kiválaszott korpuszrészek segítségével meghatározzuk azokat a szavakat és kifejezéseket, amiket a későbbiekben automatizált módon megfeleltetünk az érzelmi viszonyulásoknak. Minél nagyobb részt vonunk be az előkutatásba, annál inkább bízhatunk benne, hogy a szótár jó közelítéssel lefedi a vizsgálandó anyagot, mindezzel pedig az elemzés érvényességét javítjuk.

Az érzelmi viszonyulások osztályozása és a szótárépítés

A második fázis alighanem a kutatás leginkább idő-, munka- és szakértelemigényes része, ezért érdemes hosszabb időt rászánni. Ekkor történik ugyanis az érzelmi viszonyulások osztályozása, mely a kutatás sarokpontja. Tanácsos tapasztalt kutatóra bízni a munka elvégzését. Az elemzések többnyire bipolárisak, azaz pozitív vagy negatív érzelmi viszonyulásokra koncentrálnak. A polarizálás elemzési szempontból előnyös, szembe kell nézni azonban azzal, hogy a bináris értékelés valószínűleg egyszerűbbnek és széttartóbbnak mutatja az értékeléseket és érzelmi viszonyulásokat, mint ahogyan azok valójában vannak.

A poláris elemzés nem tudja kezelni az érzelmi viszonyulások fokozatait sem. Más érzelmi elköteleződésre utal, ha egy politikus Facebook-kommentelői közül az egyik azt írja, hogy *„ebben a kérdésben igaza van”*, míg a másik ezt *„Imádlak! Te vagy a király!”*. Ennek áthidalására történnek kísérletek, a pozitív-negatív viszonyulás a pszichológia értékeléselméletének bevonásával fino-

mítható (Bloom, 2011), illetve egy *véleményerősség-mérő szótár* kialakítására is történtek lépések (Taboada et al., 2011). Egyelőre azonban a véleményelemzések többsége pozitív-negatív orientációkban gondolkozik. Nem véletlen, hogy a módszert leginkább az internetes közösségi értékelő oldalak tartalmainak elemzésére használják: itt ugyanis többnyire világossá teszi az értékelő, hogy jó vagy rossz benyomásai vannak.

Az érzelmi viszonyulások mérése a megszólalók által használt kifejezések, szavak, nyelvi fordulatok kódolásával történik. A kódolást egy szótár segíti, amit az előkutatásra elkülönített anyag áttekintésekor készítünk el. A szótár listázza az érzelmi viszonyulásoknak megfelelő szavakat, kifejezéseket, nyelvi fordulatokat, vagyis az úgynevezett indikátorszavakat. Az elemzéseket nagyban segítik az érzelmek kifejezésével kapcsolatos szavak listái, mint például a Harvard Psychosociological Dictionary, WorldNet-Affect (Strapparava – Valitutti, 2004) vagy Affective Lexicon (Ortony et al., 1987), melyek angol nyelvű szövegek esetében alkalmazhatók. A szótárépítés feladata részben a kutató által elvégzett kivonatolással (manuális annotáció), részben számítógépes segítséggel végezhető el. A kombináció legegyszerűbb módja a kutató által kiválasztott indikátorszavak megszámolásának automatizálása. A számítógépes nyelvészet ennél szofisztikáltabb közreműködésre ajánlja a *szófaji egyértelműsítést* és a *lemmatizálást*. Előbbi azért hasznos, mert a szövegkorpuszban található szavakat általános lexikai jelentésük és kontextusuk alapján megjelöli és felcímkézi, utóbbi pedig a szavak szótári alakját meghatározva egymáshoz rendeli ugyanannak a szónak a toldalékokkal ellátott alakjait megfelelő algoritmusok segítségével.

A szótárépítés során azonban egy társadalomtudományi fókuszú vizsgálat több problémával is szembetalálkozhat. Ezek közül az egyik legtipikusabb az úgynevezett kontextusérzékenység: az elméleti nyelvészet által ismert jelenség, hogy adott kontextustól függően ugyanaz a szó más érzelmi viszonyulást jelezhet. Egyik példája ennek az irónia, melyet automatizált módon kis eséllyel lehet megfelelően kategorizálni.

Az intuitíve pozitív érzelmi viszonyulást sugalló szavak, mint például az *„okos"* egy csapásra pejoratívvá válik, ha valaki az *„okoska"* szót használja. Másik példája a polaritásváltás: a *„jó"* kifejezés többnyire kedvező megítélést sejtet, ám a *„jó nagy marha vagy"* értékelést aligha szánta pozitívnak a megszólaló. És fordítva: a szleng gyakran transzformál eredetileg negatív jelzőt pozitívvá, ahogyan az a *„brutális"*, az *„iszonyú"*, a *„pokoli"* és a *„durva"* kifejezésekkel megesik (vö. Szabó, 2015). Az internetes szleng további specialitása a kevertnyelvűség: a *„lúzer"*, *„badass"*, *„fukk"* és ehhez hasonlók bizonyos kommunikációs közösségek kutatása esetében elkerülhetetlenek, automatikus azonosításuk azonban nehézségekbe ütközhet.

Az is megeshet, hogy bizonyos szavak egyes szubkultúrákban vagy kommunikációs közösségekben más érzelmi viszonyulásra utalnak, mint ahogyan azt a kutató gondolja: így járhatunk például a *„nacionalista"* vagy a *„liberális"* jelzővel. Nem utolsósorban az érzelmi viszonyulások indikátorszavait hajlamosak vagyunk jelzőkkel azonosítani, ami súlyos adatvesztést eredményezhet. Ha szótárunk csak jelzőket tartalmaz, az elemzés vak lesz olyan megszólalásokra, mint például az alábbi komment, amely egy politikus tevékenységét véleményezi: *„Ne is törődjetek vele!"*

Amennyiben a fenti problémák bármelyike is komolyan veszélyezteti az elemzés érvényességét, mindenképpen javasolt beépíteni ellenőrző és korrekciós mechanizmusokat (mint például a manuális annotálást; Szabó, 2015). Ez utóbbi esetében meg kell bizonyosodni az annotálók közötti egyezés mértékéről. Ehhez az annotálóknak ugyanazt a korpuszt kell feldolgozni, és az így nyert adatokat kell összehasonlítani, melyre a például a Cohen-féle Kappa együttható alkalmazható.

Az érzelmek automatikus fellelése

Noha ez a szakasz az elemzés leginkább automatizálható fázisa, a kutatói döntések itt is fontos szerepet játszanak. A véleményelemzés során döntést kell hozni arról, hogy a megszólaló voltaképpen mivel kapcsolatban beszél benyomásairól, érzéseiről, véleményéről. Operacionalizálva a kérdés így hangzik: ki, miről, mit és milyen indikátorszavakat használva kommunikál? Az elemzés alapegységeit tehát az alábbi módon érdemes meghatározni és definiálni: a kommunikáló személy *(tulajdonos)*, a téma *(téma vagy célpont)*, a mondanivaló *(állítás)* és a pozitív – negatív – semleges érzelmi viszonyulásokra utaló indikátorszavak (érzelem – mind kapcsán ld. Kim – Hovy, 2004: 1–2). A módszer célja az érzelem automatikus azonosítása, majd annak összekapcsolása a *témával*. Kim és Hovy kísérletében az a modell mutatkozott a legpontosabbnak, mely a szókészletek polaritása alapján döntött. A magyar nyelv esetében azonban úgy tűnik, hogy az érzelmek automatizált felismerése és *témához* kapcsolása leginkább a szózsák algoritmussal kivitelezhető (ld. Miháltz, 2010, ill. II.1. fejezet). Itt az algoritmus a *téma* és az érzelem szavainak együttes előfordulását keresi és mutatja meg.

Ebben a kutatási szakaszban két nagyobb problémát kell alaposan megfontolni. Az egyik a *téma* minél pontosabb meghatározása körül jelentkezik. Egy konkrét termék, például egy mobiltelefon új modelljével kapcsolatos fogyasztói értékelések esetében sem elég a készülék nevének megadása, szükség lehet a köznyelvben elterjedt elnevezések integrálására is (például összecsukható,

széthajtható, butatelefon stb.). Egy politikus esetében pedig egyenesen kétséges, hogy előre megadható az összes lehetséges névelem.

Tegyük fel, hogy a korábbi magyar miniszterelnök Gyurcsány Ferenc aktuális megítélését akarjuk vizsgálni. Kutatásunk *témája* és keresőszava tehát *„Gyurcsány Ferenc"*, és minden Gyurcsány Ferenc nevét tartalmazó online kommentet meg akarunk vizsgálni. Ügyelnünk kell arra, nehogy az olyan kommentek, melyek a *Fletó, Feri* és egyéb ráaggatott csúfneveket használják, kiessenek a látókörünkből. Az automatizálást szintén megnehezíti a kommunikáció *tulajdonosa* által történő elírások, nyelvtani hibák, rövidítések. A *„Gyurcsány Ferenc"* keresőszóval az elütést tartalmazó *„Guyrcsány Ferenc"* verziót tartalmazó állítás könnyen kieshet az adatbázisból. Az elírásokkal és rövidítésekkel járó adatvesztés valószínűleg elfogadható mennyiségű, de egynéhányra rábukkanhatunk az ún. *fuzzy kereséssel,* mely az eredeti keresőszóhoz hasonló találatokat mutatja meg. A *témára* vonatkozó keresőszavak listáját pedig az előkutatás alkalmával bővíthetjük.

A második probléma a szózsák előállítása mögötti feltételezésből fakad. Amennyiben érzelmi viszonyulásra utaló kifejezés(ek) és a *téma* keresőszava együtt fordulnak elő egy mondatban, akkor az algoritmus úgy értelmezi, hogy az érzelem arra irányul. Tehát, amennyiben keresőszavunk egy politikus neve, amely névvel egy mondatban megtalálható a *„hazug"* szó, a szózsák a *„hazug"*-ot az adott politikusra vonatkoztatja.

Összességében elmondható, hogy a véleményelemzés a szövegek és a jövőben multimédiás tartalmak vizsgálatának egyik ígéretes útja. Fontosságát kiemeli, hogy remekül illeszkedik az egyéni és kollektív döntésekben újra felfedezett érzelmi dimenziók megismerésére és működésének leírására irányuló empirikus kutatói ambíciókhoz. Nemzetközi és hazai társadalomtudományi alkalmazása ugyanakkor egyelőre még korlátozott. A meglevő elméleti és módszertani dilemmák feloldásához, modellek teszteléséhez és kísérletezésekhez szociológusok, politológusok, antropológusok és számítógépes nyelvészek együttműködése szükséges.

III.2.1. példa

A *Predicting Movie Success and Academy Awards through Sentiment and Social Network Analysis* című tanulmányban összefoglalt elemzés a módszer egyik legtipikusabb alkalmazását mutatja (Krauss et al., 2008). A kutatás egy bizonyos fogyasztói csoport egy-egy termékkel kapcsolatos vélekedéseit kívánja felderíteni, majd azokat összefüggésbe hozza a termék sikerével. Jelen esetben a véleményelemzés a mozifilmekről szóló online diskurzusokra vonatkozik. A szerzők érve szerint a filmek értékelésében, sőt szakmai és közönségsikerének megjós-

lásában a telefonos megkereséseknél, a fókuszcsoportoknál és a véleményvezérekkel történt interjúknál hatékonyabb és pontosabb a *„web buzz”*, vagyis az internetes kommunikációs zsongás automatizált feltérképezése. Kimutatták, hogy az IMDb online közösségi filmértékelő fórumán zajló beszélgetésben egy-egy film fogadtatása és értékelése korrelál az alkotás Oscar-díjra történő jelölésével.

III.2.2. példa

Po-Ya Angela Wang (2013) tanulmánya jó példa a véleményelemzés és a kvalitatív tartalomelemzés közös kutatási tervbe foglalására. A projekt a számítógépes nyelvészet és a kommunikációkutatás érzelmek meghatározásával kapcsolatos egyik legnagyobb dilemmáját veti fel: a gúny és az irónia empirikus megragadásának lehetőségeit. Wang kétlépcsős eljárása először a kvantitatív megközelítést alkalmazza. Korábbi vizsgálatok által kidolgozott szótár alapján Twitter-üzeneteket vizsgált, melyeket ironikus és szarkasztikus elemeket tartalmazó csoportokra osztott. Megfigyelése szerint a szarkasztikus bejegyzésekben a megszólalók többnyire a szótár alapján pozitívnak kategorizált érzelmeket alkalmaznak gúnyos, ellenséges és támadó módon. Eredményeire támaszkodva kvalitatív módon is megvizsgálta az üzeneteket, mégpedig a szarkazmus által megtámadottak azonosítására, a megszólaló szándékaira, illetve az irónia és a szarkazmus közötti átfedésre koncentrálva.

III.2.3. példa

Alexander Hogenboom és szerzőtársai (2013) *Exploiting Emoticons in Sentiment Analysis* című tanulmányukban a véleményelemzés szövegközpontúságától elmozdulva az ún. emotikonokat vizsgálták. Az internetes interakciókban az emotikonok pótolják a személyközi kommunikáció nonverbális jeleit: a száj- és a szemmozgást, a mimikát és a testbeszédet. Bár a karakterek kombinálásával újabb és újabb jelentések születnek, az emotikonokat főként lelkiállapotok és hangulatok ábrázolására használják. A hangulatjelek többnyire egy arcot jelenítenek meg, és a felhasznált jelek variálásával számos érzelmi állapotot vagy érzelmi viszonyulásra utaló cselekedetet fejeznek ki. Hogenboom és munkatársai esettanulmányukban holland nyelvű Twitter-üzeneteket és webes fórumokat vizsgáltak. Először kvalitatív lingvisztikai elemzést folytattak a szöveges tartalmak és az emotikonok kapcsolatáról, majd ennek alapján mintázatokat különítettek el. Eredményeik szerint a hangulatjelek emelhetik a szöveg intenzitását, megváltoztathatják a szöveges tartalom által indikált érzelmi viszonyulást, illetve egy semleges szöveges közlésről egyértelművé tehetik a megszólalók véleményét. Az automatizált elemzés a kvali-

tatív kutatás alapján megszületett emotikon-lexikont alkalmazta, amit a szöveges érzelmekkel vetettek össze. Hogenboom és kollégái bebizonyították, hogy az emotikonok segítségével az interneten fellelhető érzelmi viszonyulások nagyobb pontossággal tárhatók fel, mint a pusztán szövegalapú osztályozások során.

Nemzetközi és hazai politikatudományi alkalmazások

A véleményelemzéssel foglalkozó kutatások jelentős része a számítógépes nyelvészet számára releváns témákat, illetve módszertani problémákat boncolgat. Akad azonban néhány tanulmány, mely politikatudományi kérdéseket válaszol meg a véleményelemzés alkalmazásával. A közösségi médiában megjelenő politikai tartalmak véleményelemzését gyakran szavazói magatartás előrejelzésére (Tjong Kim Sang – Bos, 2012, Franch, 2013), illetve politikusok népszerűségének és elismertségének kimutatására (O'Connor et al., 2010) alkalmazzák. Az angol, holland és francia mintákon bemutatott szavazói magatartás-vizsgálatok konklúziója szerint az internetes kommunikációk véleményelemzése pontosabb prognózist adott az általános parlamenti választások kimenetelére vonatkozóan, mint a hagyományos lekérdezésen alapuló nagymintás, reprezentatív survey-kutatások. A politikusok megítélésével kapcsolatos amerikai vizsgálat eredménye pedig szignifikáns korrelációt mutatott a telefonos vagy személyes megkereséssel felvett adatokkal.

Bár korai még temetni a nagymintás reprezentatív és személyes lekérdezést alkalmazó politikai viselkedést és választói magatartást célzó kutatásokat, hiszen a véleményelemzéssel kapott eredményeket további tesztelésnek és tudományos értékelésnek kell még alávetni, a módszer figyelemre méltó a politikatudomány számára. Látni kell azonban, hogy olyan országokban, mint Magyarország, ahol a közösségi médiahasználat generációs szegmentáltsága figyelhető meg, a Facebookra és Twitterre alapozott politikai viselkedést kutató munkák erőteljes korlátokba ütköznek.

A sentiment analysis bár alapvetően nyelvi, szövegalapú vizsgálódás, de könnyen és ésszerűen összekapcsolható más módszerekkel. Leginkább a *hálózatelemzés* kínál kapcsolódási pontokat, hiszen segítségével különböző érzelmi hálók rajzolhatók fel, nyomon követhető az érzelmek és vélekedések terjedése, mérhetővé válik a kellemes vagy kellemetlen érzéseket megfogalmazó személyek hatóköre.

Egy másik lehetőség a *netnográfiával* (Kozinets, 2015) való összekapcsolás. Az online közösségek ilyen antropológiai leírása ugyanis sokat segíthet a véleményelemzéshez szükséges kategorizálás és szótárépítés érvényességének erősítésében. Az érzelmi állapotokról persze a verbalitáson túli világból is nyer-

hetünk információkat. A vizuális és multimédiás kommunikáció (emotikonok, reakciógifek, mémek) terjedése arra ösztönöznek, hogy a véleményelemzés eszköztára a szövegen kívüli modalitásokra is vonatkozzék. A nem szöveges elemzések azonban még kísérleti fázisban vannak, így a fejezet is inkább a textuális – azon belül is inkább a társadalomtudományi – vizsgálatok logikáját mutatja be. Van azonban, ami egyértelműen kívül esik a módszer hatókörén: ilyen a pulzusszám-emelkedés, vércukor-ingadozás, verejtékezés, mimika, testbeszéd stb.

Magyarországon *A véleményváltozás azonosítása politikai témájú közösségi médiában megjelenő szövegekben* című tanulmány (Pólya et al., 2015) egy interdiszciplináris (pszichológiai-nyelvtudományi) együttműködésben megvalósuló kutatás egyik eredménye. A vizsgálat a közösségi médiában megjelenő politikai vélemények változásainak automatikus felismerésére vállalkozik. A pártok Facebook-posztjai és a hozzájuk érkezett kommentekre fókuszáló gyűjtés eredményeképpen 141 825 poszt és ezekhez kapcsolódóan 1 939 356 komment került be a szövegkorpuszba. A kutatás a manuális és gépi kódolás, illetve adatelemzés módozatait ötvözte. Pólya Tibor és kollégái öt modult különítettek el, ezek: individualizmus-kollektivizmus, optimizmus-pesszimizmus, közösségiség és ágencia, érzelmi polaritás és politikai szereplők. A véleményelemzés módszerét az érzelmi polaritás és a politikai szereplők elemzésekor alkalmazták. Legfontosabb megállapításuk szerint minél kisebb egy párt népszerűsége, annál erőteljesebben jelentkezik a Facebook-profiljukon más pártok leértékelése, ami a kevésbé támogatott párt identitásának védelmét és a csoportkohéziót szolgálhatja.

Ellenőrző kérdések

- Milyen hatással lehet egy-egy megszólalás kontextusa az érzelmi viszonyulás kódolására?
- Milyen módszerek egészíthetik ki a véleményelemzést? Milyen célokra javasolja a kiegészítő módszerek alkalmazását?
- Melyek a véleményelemzés eljárásának legfontosabb módszertani korlátai?
- Milyen problémákkal szembesülhet egy társadalomtudományi orientációjú kutatás a véleményelemzéshez szükséges korpusz kialakításakor?
- Hogyan lehet áthidalni a polarizált elemzés hiányosságait?
- Melyek a véleményelemzés elemzési alapegységei? Mondjon konkrét példát is mindegyikre, és indokolja meg, milyen esetben választandó egyik vagy másik alapegységként!
- Hogyan lehet az adatállományban megjelenő elírásokat kezelni?
- Készítsen jegyzéket az indikátorszavakról egy online hírhez érkezett kommentek véleményelemzéséhez!

Szószedet

Magyar	Angol
(Pozitív vagy negatív) érzelmi viszonyulás	Sentimental orientation
Elmosódó keresés	Fuzzy search
Értékeléselmélet	Appraisal theory
Érzelmek osztályozása (rendszerint pozitív és negatív csoportba sorolása)	Sentiment classification, classification of sentimental orientations
Hálózatelemzés	Network analysis
Két kategóriát alkalmazó osztályozás	Binary classification
Lemmatizálás (a lemma [szó] alaprészének algoritmikus definíciója)	Lemmatisation
Netnográfia (az internetes közösségek entográfiai leírása)	Netnography
Szófaji egyértelműsítés	POS tagging
Véleményelemzés, szentimentelemzés	Sentiment analysis
Véleményerősség-mérő szótármódszer	Strength-oriented lexical methods

Ajánlott irodalom

A véleményelemzés alapvetéseiről ld. Pang – Lee (2008). Érdekes, véleményelemzést alkalmazó politikatudományi kutatásokkal találkozhatunk Bakliwal et al. (2013); Johnson et al. (2012); Mullen – Malouf (2006); Ceron et al. (2013); Franch (2013) műveiben.

Zorigt Burtejin

III.3. CSOPORTOSÍTÁS (KLASZTEREZÉS)

A fejezetben a dokumentumrendezési feladatok egy másik lehetséges változatát, a csoportosítást (más néven klaszterezést) mutatjuk be. Ez a módszer azokban az esetekben hasznos, amikor nem állnak a kutató rendelkezésére előzetesen ismert csoportok, amelyek szerint a dokumentumot rendezni tudná, vagy maga a kutató szeretne új csoportokat létrehozni. A csoportosítás során a dokumentumokból olyan különálló csoportokat hozunk létre, amelynek tagjai valamilyen szempontból hasonlítanak egymásra. A csoportosítás legfőbb célja az, hogy az egy csoportba kerülők minél inkább hasonlítsanak egymásra, miközben a különböző csoportba kerülők minél inkább eltérjenek egymástól. A fejezetben bemutatásra kerül, hogy a csoportosítás miben tér el a korábban bemutatott osztályozási feladattól, miért és milyen esetekben érdemes ehhez az eljáráshoz fordulnunk. Röviden ismertetjük továbbá a társadalomtudomány területén is használt olyan leggyakoribb csoportosítási eljárásokat, mint a teljesen automatizált eljárások és a számítógéppel támogatott klaszterezés.

Osztályozáson a III.1. fejezetben egy szövegkorpusz elemeinek előzetesen kialakított és a kutató által ismert kategóriákba való rendezését értettük. A kutatók ebben az esetben előre tisztában vannak azokkal a csoportjellemzőkkel, melyek a sokaság rendezésének alapjául szolgálnak. E deduktív megközelítéssel szemben határozhatjuk meg a csoportosítási problémát, melyet gyakran klaszterezésnek is neveznek (Tikk, 2007b: 145 – a kettőt szinonimaként használjuk a továbbiakban). A megközelítés alapja, hogy a dokumentumrendezéshez előzetesen nem határozunk meg olyan csoportokat, amelyekbe a vizsgált elemeket (szövegeket/dokumentumokat) kívánjuk besorolni.

Ehhez kapcsolódva *felügyelet nélküli tanulásnak* nevezzük azokat a gépi módszereket, amelyekben a szövegek csoportokba rendezéséhez nem tudjuk előre a *csoportcímkéket* (ld. IV.2. fejezet). A felügyelet nélküli tanulás a dokumentum tulajdonságait és a modell becsléseit felhasználva hoz létre különböző kategóriákat, melyekhez később hozzárendeli a szöveget (Grimmer és Stewart, 2013: 15). A felügyelet nélküli tanulás tehát a szöveg tulajdonságaiból „tanul”

anélkül, hogy előre meghatározott csoportokat ismerne. A klaszterezési eljárás során a csoportokat induktívan – számítógép segítségével vagy anélkül – alakítjuk ki. Az osztályozás és a csoportosítás közötti legfontosabb különbség tehát az, hogy az utóbbi esetében nincsen ismert „címkékkel" ellátott kategóriarendszer vagy olyan minta, mint az osztályozás esetében a *tanítókörnyezet,* amiből tanulva a modellt fel lehet építeni (Tikk, 2007b: 145). A politikatudományban a felügyelt tanulást és a felügyelet nélküli tanulási módszertant gyakran egymás vetélytársának tekintik (ld. pl. Hillard, Purpura és Wilkerson, 2008; Quinn és szerzőtársai, 2010). Helyesebb viszont, ha a felügyelt és a nem felügyelt tanulást más-más célokat szolgáló eljárásoknak tekintjük, amelyek olykor egymás kiegészítésére is szolgálhatnak.

Fontos ugyanis hangsúlyozni, hogy az osztályozás deduktív logikája bizonyos dokumentumok, illetve kutatási stratégiák esetében nem alkalmazható. Vannak olyan esetek, amikor egyszerűen nem tudunk előre definiált („ismert") kategóriákat a vizsgált adatokra vetíteni. Ennek több oka is lehet. Tegyük fel, azt vizsgáljuk, hogy milyen törvényeket nyújtottak be 1990 előtt a magyar Országgyűlésben, ugyanakkor kevésbé vagyunk tisztában azzal, hogy ebben az időszakban milyen témák uralhatták a politikai közbeszédet. Ugyan előzetesen felállíthatnánk különféle csoportokat egy általunk írt közpolitikai szótár kapcsán, azonban félő, hogy sok olyan téma kimaradna ismereteink hiányosságai miatt, ami az adott időszakban tematizálhatta a politikai napirendet. Másfelől a probléma abból is eredhet, hogy már kevés számú adat esetében is sokféle csoportosítási lehetőség merül fel, hiszen az osztályozás egyik alaptulajdonsága, hogy olyan csoportokat hoz létre, amelyek egyszerre kizáróak és mindent magukba foglalóak. Ez a feltétel rengeteg lehetőséget kínál az osztályozásra, ami miatt szinte lehetetlennek tűnik teljes körűen felismerni az összes lehetséges, egyben kutatási értelemben hasznos kategorizálási módot.

A csoportosítási technikák képesek kezelni az ilyen kutatási problémákat, az ördög ugyanakkor a kutatásszervezés részleteiben van. A politikatudományban rengeteg szövegalapú adat áll a rendelkezésünkre, azonban ezeknek a feldolgozása idő- és pénzbefektetést, és nem utolsósorban humánerőforrást (emberi erőforrást) is igényel. Humán kódolók ugyan a feladat bizonyos részét el tudnák végezni, azonban a kiképzésük, betaníttatásuk sok erőforrást igényel, miközben a nagy mennyiségű adat feldolgozása és a dokumentumok rendezése a humán kódolók segítségével is többéves, évtizedes munka lehet. A gépi klaszterezés azáltal, hogy különböző matematikai formulákat és modelleket használ, képes csökkenteni ezeket a költségeket, miközben nem helyettesíti az emberi erőforrást, hiszen a kutatás helyes megtervezésében és az eredmények validálásában a kutató szerepe továbbra is kulcsfontosságú marad.

E fejezetben két olyan csoportosítási megoldást ismertetünk, melyeket Grimmer és Stewart (2013) összefoglalója alapján gyakran alkalmaznak a po-

litikatudomány területén. Az első a *teljesen automatizált klaszterezési algoritmus (FAC),* amely egyidejűleg becsüli meg a csoportokat, és rendezi ezekbe a kategóriákba a szövegeket. A FAC alkalmazása során nehézséget jelenthet az, hogy nem lehet előre megbecsülni, hogy valamelyik módszer hasznos vagy „jó" klaszterezést eredményez-e. Emiatt különösen fontos a klaszterezési eredmények validálása (erre visszatérünk). A második kategorizálási megoldást *számítógéppel támogatott klaszterezésnek (CAC)* nevezik (ld. pl. Grimmer – King, 2011), mellyel több ezer potenciális klasztert tudunk létrehozni viszonylag költséghatékonyan. A két stratégia közötti fő különbséget az adja, hogy míg a FAC egy teljesen automatizált módszer, addig a CAC a felügyelet nélküli tanuláson túl a csoportosítás bizonyos szakaszaiban a kutató részéről további elemzéseket igényel ahhoz, hogy az összes dokumentumot csoportokba rendezze. A CAC igényli a kutató aktívabb részvételét, ugyanakkor több lehetséges klaszterezési eredményt mutat.

A klaszterezés

Mielőtt rátérünk a szövegek és dokumentumok klaszterezésének bemutatására, érdemes definiálni a „klasszikus" klaszterezés fogalmát. A klaszterezést sokféle célból szokták használni egy cég vásárlóinak szegmentációjától komplex adatbázisok belső szerkezetének vizuális ábrázolásáig. Felmerülhet a kérdés, hogy miért érdemes ehhez az eljáráshoz fordulni? A válasz röviden az, hogy olyan esetekben, amikor nem az egyes elemekről, hanem a csoport jellemzőiről szeretnénk többet megtudni. A vásárlószegmentáció példájánál maradva a cégeket elsősorban nem az egyes emberek tulajdonságai foglalkoztatják, hanem az, hogy az egy csoportba tartozók milyen közös vagy hasonló jellemzőkkel bírnak. Ez a szövegek klaszterezésénél általánosságban azt jelenti, hogy nem egy-egy szöveg jellemzőire vagyunk kíváncsiak, hanem arra, hogy a szövegek egy-egy csoportja milyen hasonlóságokkal bír.

A második definíciós lépésben érdemes megkülönböztetni a klasszikus klaszterezést a klaszterezéstől mint szövegbányászati technikától. Ez utóbbi abban különbözik a klasszikus klaszterezéstől, hogy az adatforrás itt egy dokumentum, vagyis valamilyen szövegalapú forrás. Ezért a szövegalapú adatforrást át kell alakítani számokká ahhoz, hogy tudjunk vele matematikai műveleteket (vagyis klaszterezést) végezni (a fejezet további részében klaszterezés vagy csoportosítás alatt annak dokumentumokra vonatkozó részterületét értjük). A több dokumentumból álló korpuszok esetében a gépi klaszterelemzés különösen eredményes és költséghatékony lehet, mivel egy nagy korpusz vizsgálata sok erőforrást igényel (Grimmer – Stewart, 2013: 1).

Egy kutatásban a klaszterezés lehet önálló elemzési módszer, ugyanakkor lehet egy elemzés kiegészítő eljárása is. Ez azt jelenti, hogy a csoportosítás amellett, hogy egy egyedülálló eljárás (tehát egész cikket lehet egy ilyen eljárás eredményei alapján írni), egy kutatás során lehet első vagy második (harmadik stb.) elemzési stratégia is. Célját tekintve többféle szövegbányászati feladatra lehet használni (Tikk, 2007b: 147). Egyfelől segítheti az adataink vagy éppen a dokumentum rendszerezését. A szövegek hierarchikus rendezése koherens kategóriákba nagyon hasznos lehet például akkor, amikor a szöveget szisztematikusan szeretnénk böngészni.

Másfelől a klaszterezési technikák összefoglalót adnak, és kiváló betekintést nyújtanak az egész korpuszba. Az egyik legegyszerűbb erre alkalmas klaszterezési módszert szóklaszterezésnek hívják. A szóklaszterezés arra utal, amikor a szövegben előforduló szavak gyakoriságán alapszik a klaszterezés (később részletesen szó lesz a klaszterek további típusairól). Harmadrészt a klaszterezés az osztályozás pontosságát ellenőrizni tudja, illetve fordítva is igaz ez. Ugyanis az osztályozás is képes arra, hogy ellenőrizze a klaszterezés eredményét, amiről a fejezet későbbi részében még szó lesz.

A klaszterezés lépései a gyakorlatban

A következőkben bemutatjuk a klaszterezés főbb lépéseit, melyek egy lehetséges sorrendje a következő (Grimmer – Stewart, 2013: 3–6): az adathalmaz létrehozása; a szövegek számokká alakítása; a döntés a csoportosítás elvéről; a klaszterek számának megadása; az eredmények validálása. Nézzük ezeket sorban!

Az adathalmaz létrehozása

A klaszterezés során először meg kell határozni az adathalmazt, melynek csoportjait azonosítani szeretnénk. Az, hogy milyen politikatudományi dokumentumokat vizsgálhatunk klaszterelemzéssel, szinte csak a kutatótól, illetve a kutatási kérdésétől, témájától függ. Egy kampánnyal kapcsolatos kutatás fókuszában például pártprogramok, választási programok vagy újságcikkek vizsgálata állhat, míg a parlamenti viselkedéssel foglalkozó kutatókat inkább a törvényjavasatok, interpellációk vagy parlamenti kérdések érdekelhetik.

A klaszterezésnél a vizsgált szövegtartomány egyaránt lehet dokumentum, mondat vagy kifejezés. Ahhoz, hogy a szöveggel dolgozni tudjunk, nagyon fontos, hogy a dokumentum egyszerű szövegformátumban legyen. Ezt a megfelelő *szövegszerkesztő* (text editor – nem pedig word processor, mint amilyen pl. a Microsoft Word) programokkal viszonylag könnyedén elő lehet állítani.

Nehézséget az okozhat, ha bizonyos szövegeket nem tudunk máshogy beszerezni, csak úgy, ha papírváltozatukat beszkenneljük. Ugyanakkor erre az esetre is egyre több program áll rendelkezésre, amivel a szkennelt anyagokat át tudjuk alakítani olyan szövegekké, amelyeket a számítógép fel tud ismerni (ld. még *optikai karakterfelismerés,* ill. Grimmer – Stewart, 2013: 6).

Ehhez hasonló problémát okozhat az is, ha a szövegtípus, amelyet csoportosítani szeretnénk, nagyon kis terjedelmű (például tweetek vagy mondatok). A kis terjedelmű szövegek azért jelentenek módszertani kihívást, mert a számítógéppel támogatott eljárások bizonyos szómennyiséget igényelnek ahhoz, hogy megbízható eredményeket hozzanak létre (i. m.: 6). Ráadásul a szövegek adatokká (számokká) alakítása során az adathalmaz veszít terjedelméből, és ez a tweetek vagy mondatok esetében kevés kiinduló adatot jelent. A klaszterezés ezt jobban képes kezelni, mint a felügyelt tanulási módszerek, azonban ebben az esetben is igaz, hogy a megbízhatóság szempontjából előnyösebb, ha nagyobb terjedelmű szövegkorpusszal dolgozunk.

A szövegek számokká alakítása

A szövegek (kvantitatív) adattá, azaz számokká való átalakítására azért van szükség, hogy matematikai, illetve statisztikai műveleteket tudjunk végezni velük. Ennél a folyamatnál minden olyan tartalmat el kell távolítani a szövegből, amelyeknek a szöveg tartalmára nézve nincsen hozáadott értéke, így nem erősíti a műveletek eredményességét. Ebben segítséget nyújtanak különböző ingyen elérhető szofterek *szöveg-előkészítő* funkciói, tehát nem nekünk kell manuálisan végezni. Grimmer és Stewart (2013: 6–7) szerint ugyan leírható, hogy mit érdemes eltávolítani, azonban minden eset egyedi, ezért e lista csak kiindulópontot jelenthet a gyakorlati kutatások számára. Mindezek fényében a szövegek számokká alakításának lépései a következők.

Az első szakaszban átrendezzük a dokumentumban a szavak sorrendjét. Így egy szózsákszerű (ld. II. 1. fejezet) formátumot kapunk, mert feltételezzük, hogy a szavak sorrendje nem befolyásolja a szöveg értelmezését. Ez elsőleg némileg furának tűnhet, azonban több kutatás is igazolta, hogy a szavak sorrendje érdemben nem befolyásolja azt, hogy a szöveg egészéről milyen általános képet kapunk (i. m.: 4). Ha erős feltételezésünk van azt illetően, hogy bizonyos szavak együttes megjelenése vagy sorrendje számít, rendszerint beállíthatjuk a felhasznált szoftverben, hogy vegye figyelembe a szavak sorrendjét.

A második szakasz a szókincs egyszerűsítése a szótövek levágásával. Ehhez a számítógépes nyelvészet először megkeresi a szótöveket, majd leválasztja a toldalékokat a szótőtől különböző algoritmusokkal. Ezt a nyelvészek lemmatizálásnak nevezik (ld. II. 1. fejezet). Grimmer és Stewart (2013: 6) szerint a fő kü-

lönbség a kettő között az, hogy a lemmatizálás szótárakat használ, és a szöveg kontextusát is figyelembe veszi a szavak szótári alakjának a megállapításához. Azonban ez a művelet az angol nyelvben lényegesen egyszerűbb, mint a magyar nyelvben, nem beszélve arról, hogy lényegesen csökkentheti a korpusz információtartalmát. Ennek megfelelően törekedni kell arra, hogy a lecsupaszított adathalmaz hordozza azokat az információkat, amelyek a tartalom értelmezése szempontjából fontosak lehetnek számunkra.

A harmadik szakasz a szógyakoriság normalizálása. A hatékony klaszterezési folyamat és általában az adatbányászati technikák minőségének javítása érdekében meg kell szüntetni a szöveg tulajdonságainak ún. zajosságát azáltal, hogy eltávolítjuk az írásjeleket, nagybetűket és a gyakori szavakat (Grimmer – Stewart, 2013: 7). Zajosság alatt értünk minden olyan szövegelemet, amely elsődlegesen nem jelentést közvetít, hanem inkább nyelvtani funkciója van a szövegben. A dokumentumokban gyakran előforduló olyan szavak, mint például az „a", „az", „és" nem tesznek hozzá a szöveg értelmezéséhez, ezért csökkenteni kell ezeknek a kifejezéseknek a súlyát valamilyen súlyozási módszerrel (erről ld. Aggarwal – Zhai, 2012c: 81–82). Ezért az ilyen „zajos" vagy „tiltólistás" szavakat már a klaszterezés előtt ki lehet venni a korpuszból, az ún. tiltólistás szavak eltávolítása parancs használatával (Aggarwal – Zhai, 2012c: 81–82).

Eltávolíthatjuk azokat a szavakat is, amelyek nagyon ritkán szerepelnek a szövegben. Erre azért van szükség, mert feltételezhetően az ilyen szavak sem tesznek hozzá a szöveg tartalmának értelmezéséhez (Grimmer – Stewart, 2013: 7). Előfordulhat az is, hogy e szavak helyesírási vagy elgépelési hibák. Az olyan „zajos" szövegek, amelyek az internetről származnak (weboldalak tartalmai, blogról vagy közösségi oldalakról származó bejegyzések), nagyobb valószínűséggel tartalmaznak ilyen kifejezéseket. Végezetül érdemes mérlegelni, hogy a kutatás céljához milyen előkészítési eljárások a leghasznosabbak. Amikor egy szövegnek inkább a stílusára vagyunk kíváncsiak, és nem a tartalmára, a funkciószavaknak és a tiltószavaknak fontos szerepe lehet, ezért nem szabad eltávolítani őket.

Döntés a csoportosítás elvéről

A csoportosítás során az objektumoknak (elemeknek, megfigyelési egységeknek vagy esetünkben: a dokumentumoknak) olyan tulajdonságait keressük, amelyek között hasonlóságokat fedezhetünk fel. A klaszterezés nyomán az adatok automatikusan kialakított csoportokba kerülnek e hasonlóságok, illetve különbségek alapján: az egymáshoz hasonló objektumok egy csoportba, az egymástól különböző objektumok pedig különböző csoportokba kerülnek.

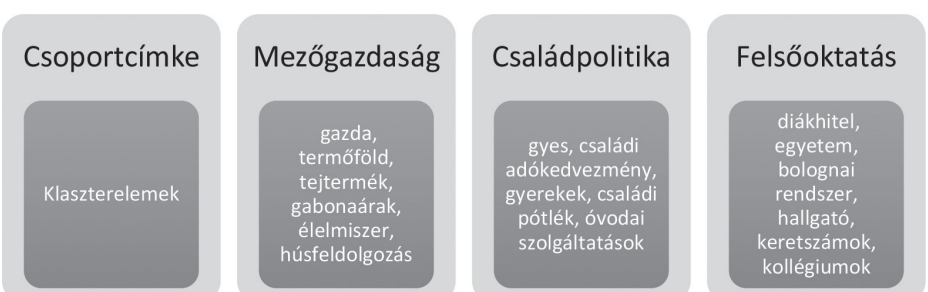
A klaszterezés egyik fontos lépése a csoportosítási elv meghatározása, melynél több tényezőre is figyelni kell. Egyrészt a feladat megoldása jellemzően számításigényes, a csoportosítási feladatok megoldása során bevett a számítógépek használata. Másrészt egy csoportosítási feladat megoldására számtalan lehetőség áll rendelkezésre, ami miatt fontos a folyamat automatizálása, objektivizálása, vagyis valamilyen matematikai formula segítségének az igénybevétele.

Egy hétköznapi példával élve, a nappalinkban lévő polcokon a könyveket különböző rendszerek szerint helyezhetjük el. Valaki ezt a feladatot úgy végezné el, hogy a könyveket a szerzők nevei szerinti ábécésorrendben csoportosítaná, más inkább tematikus rendbe tenné a könyveket. Ahogy e példában is, a dokumentumok csoportokba rendezése általában valamilyen tulajdonságuk (például nyelv, téma, stílus) alapján történik (Tikk, 2007b: 146).

Az eredmények felcímkézése és validálása

A dokumentumok csoportosítása utáni fontos mozzanat a csoportok *kategóriacímkékkel való ellátása*. A felcímkézést kétféle módon végezhetjük. Egyrészt a klaszterezés által kiadott csoportokban lévő szövegekből vehetünk mintát, majd a minta alapján az egyes csoportokba tartozó néhány szöveget elolvasva, a kulcsszavak alapján címkékkel láthatjuk el a dokumentumot. Másrészt statisztikai módszer segítségét is igénybe vehetjük, amellyel azonosíthatjuk a leggyakoribb szavakat, ami megkönnyíti a csoport felcímkézését. A III.3.1. ábra azt illusztrálja, hogy a különböző csoportokhoz tartozó kulcsszavak segítségével hogyan tudjuk a csoportokat felcímkézni egy közpolitikai tárgyú korpusz esetében.

III.3.1. ábra – A különböző csoportok felcímkézése



A klaszterezési feladatok típusától függetlenül fontos, hogy a klaszterezés eredményét némi szkepszissel értékeljük, mivel előzetesen nehéz tudni, hogy milyen klaszterekre számíthatunk, és azok mennyire tekinthetőek érvényesnek

(Grimmer – Stewart, 2013: 15). Ezért nélkülözhetetlen az eredmények validálása a csoportosítás esetében is. Ha már rendelkezésünkre állnak a felcímkézett csoportok, legalább három szempontból meg tudjuk vizsgálni a klaszterelemzés validitását: a szemantikai, az előrejelző (prediktív) és a konvergens érvényesség tekintetében (i. m.: 21). A szemantikai érvényesség azt jelenti, hogy a klaszterezés milyen mértékben volt képes valóban homogén és közben egymástól eltérő csoportokat létrehozni. Ezt nem egyszerű tökéletesen ellenőrizni, azonban vannak olyan megoldások, amelyekkel lehetséges a klaszterezés minőségét vizsgálni. Hasonlóan ahhoz, ahogy a különböző csoportokat felcímkéztük, itt is szükségünk lesz az emberi erőforrásra.

A szemantikai érvényesség ellenőrzésének egyik legegyszerűbb formája az, hogy a kódoló mintát vesz a klaszterezés által létrehozott csoportokból: kiválaszt egy csoportot, abból kiemel két dokumentumot, és megnézi, hogy ez a két dokumentum mennyiben tekinthető hasonlónak (Grimmer – Stewart, 2013: 21). Ezt követően vesz egy másik csoportból is két dokumentumot, és ezeket is összehasonlítja egymással. Végül a két különböző csoportba került dokumentumot hasonlítja össze egymással, és megvizsgálja, hogy azok valóban különböznek-e egymástól. Ezt a műveletet többször el kell végezni, majd megállapítható, hogy a szemantikai validálás eredménye milyen mértékben egyezik a klaszterezés eredményével. Ez határozza meg a klaszterezés minőségét.

A prediktív érvényesség ezzel szemben azt jelenti, hogy a módszer menynyire pontosan tudta a különböző csoportokat előre jelezni, és a jövőben várhatóan képes-e ugyanígy előre jelezni. Így például ha a miniszterek parlamenti felszólalásait vizsgáljuk, és a klaszterezés létrehozott különböző csoportokat, amelyekbe a felszólalásokat rendszerezte, akkor azt az eredményt kaphatjuk, hogy a miniszterek a saját területükhöz tartozó témákban szólalnak fel a legtöbbet. Ezt látva megfogalmazhatjuk hipotézisünket, mely szerint a miniszterek azokban a témákban szoktak felszólalni, amelyek a miniszteri területükhöz tartozik. A teszteredmények alapján felállított hipotézisünket úgy tudjuk ellenőrizni, hogy megvizsgálunk későbbi felszólalásokat is a miniszterektől, és megnézzük, hogy valóban az adott témákat részesítették-e előnyben.

Végezetül a konvergens érvényesség arra utal, hogy a klaszterezés során nyert eredmények milyen mértékben egyeznek egy másik módszerrel nyert eredményekkel. A konvergens érvényesség ellenőrzése során a felügyelet nélküli tanulási módszer és a felügyelt módszer egymás kiegészítő volta különösen hangsúlyt kap. Ehhez először is az szükséges, hogy a felügyelet nélküli módszer (azaz a klaszterelemzés) létrehozza a saját rendszere alapján kialakított csoportokat. Miután a klaszterelemzés felkínálja a lehetséges csoportokat (kategóriákat), az osztályozás mint felügyelt tanulási mechanizmus segítségét vehetjük igénybe ahhoz, hogy validáljuk a klaszterezés eredményeit, és azokat még álta-

lánosabb érvényűvé tegyük. Végső soron tehát a felügyelt tanulás által elnyert eredményekkel hasonlítjuk össze a klaszterezés eredményeit.

Klaszterezési technikák

A felügyelet nélküli tanulás kapcsán két nagyobb csoportosítási típusról beszélhetünk (Grimmer – King, 2011; Grimmer – Stewart, 2013). Az egyik legygyakrabban használt módszer a fentebb már említett teljesen automatizált klaszterezés (FAC). Ez egy erősen modellfüggő eljárás, mely sajátosságát nem is lehet teljesen megszüntetni. Vannak azonban olyan eszközök, amelyek további információk hozzáadásával rugalmasabbá teszik a modelleket.

A FAC eljárások alapvetően további két nagyobb csoportra oszthatóak aszerint, hogy a klaszterek egymásba ágyazottak vagy nem egymásba ágyazottak. A felosztó vagy másnéven *particionáló módszerek* alatt az adathalmaz olyan felbontását értjük, amikor a létrejövő csoportok nem egymásba ágyazottak. A particionáló módszerek így a dokumentumok egy lehetséges felosztását adják meg, ahol a csoportok függetlenek egymástól, vagyis nem egymásba ágyazottak (Tikk, 2007b: 146). Ezzel szemben a *hierarchikus klaszterezési eljárásoknál* a klaszterek egymásba vannak ágyazva, azaz a klaszthereknek lehetnek további alklasztereik. Ezt a technikát gyakran írják le a fa metaforával (ahogy a kódolásban is), hiszen az így létrejövő klaszterelemzés strukturája egy fához hasonlít, amelynek az ágai jelentik az alklaszthereket.

A második fő csoportosítási módszer a számítógéppel támogatott klaszterezés (CAC). Ez az eljárás nem egy konkrét típust takar, hanem arról van szó, hogy a számítógép segítségével ugyanazon az adathalmazon többféle klaszterezési eljárást tesztelünk. Ez a megoldás lehetővé teszi a kutatók számára, hogy hatékonyan keressenek a sok potenciális csoportosítási módszer között, ameddig nem azonosítják a számukra leginkább érdekes és hasznos csoportosítási lehetőségeket.

Fontos hangúlyozni, hogy számtalan klaszterezési eljárás létezik, és a kutatási kérdéseinktől függ, hogy melyik módszert alkalmazzuk. Terjedelmi okokból ezek közül csak néhányat mutatunk be, amelyek politikatudományi alkalmazás céljából hasznosak lehetnek. Ugyan némileg időigényesnek tekinthető, de a CAC módszerére igaz, hogy új csoportosítási rendszereket is elő tud állítani (ellentétben a felügyelt tanulási eljárásokkal, amelyek nem járulnak hozzá új kódolási rendszerek vagy tipológiák létrejöttéhez). Grimmer és Stewart (2013: 15) mindezek kapcsán a következő megközelítést javasolja: első lépésben a modellekre kell bízni a klaszterek és a témák azonosítását (kiválasztását), második lépésben pedig a kutatónak kell saját ítélőképessége alapján kiválasztani a végső

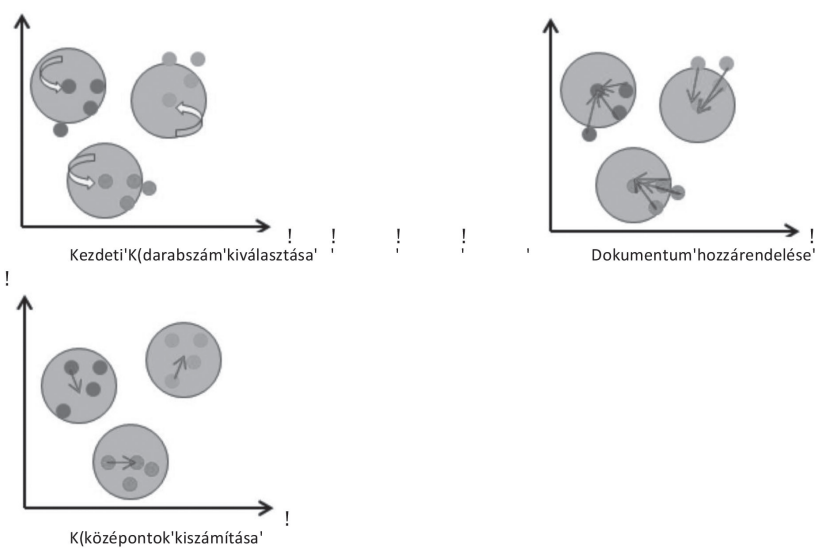
modellt, figyelembe véve a klaszterezési technikák és az így létrejövő csoportok közötti minőségbeli különbségeket.

A következőkben röviden ismertetjük e két fő eljárás elméleti hátterét. Bár a statisztikában kevésbé járatos olvasó számára ezek első látásra ijesztőnek tűnhetnek, érdemes visszaidézni, hogy a gyakorlati kutatás során e műveleteket az adatelemző program végzi el helyettünk. Ezzel együtt érdemes tisztában lenni a számítógépes folyamat elméleti hátterével.

Teljesen automatizált klaszterezés (FAC)

Az egyik legismertebb, s a társadalomtudományi kutatásokban talán legszélesebb körben használt FAC-típusú módszer a *K-közép klaszterezés,* a particionáló módszerek egyik esete. E technika alkalmazása során először is meg kell határozni az igényelt klaszterek darabszámát, ami ebben az esetben (K) darabszámot jelent (Tikk, 2007b: 148). A (K) darabszám kiválasztása tetszőleges, nagyban függ a klaszterezni kívánt dokumentum tulajdonságától, nagyságától. Így nem könnyű optimális darabszámot mondani, de a hangsúly azon van, hogy minden egyes elemet (adatot) úgy rendezzük különböző klaszterekbe, hogy az még áttekinthető maradjon. Egyes alkalmazások az áttekinthetőség miatt korlátot is szabnak a csoportok számára (i. m.: 148). A III.3.2. ábra áttekinti az eljárás lépéseit.

III.3.2. ábra – A K-közép eljárás



Forrás: Kunwar, S. (2013, online dokumentum)

A módszer a klasztereket úgy hozza létre, hogy a szöveg értékeinek egymástól vett távolságát veszi alapul. Az algoritmus induláskor véletlenszerűen kiválaszt K darab középpontot a klaszterek középpontjaként, majd a minta minden egyes elemét hozzárendeli a legközelebbi középponthoz. A K-közép klaszterezésnél a klaszterek középpontját a centroid, vagyis a klaszterekhez tartozó pontok átlaga jelenti. Így tehát egy pontot ahhoz a klaszterhez társítunk, amelynek a centroidja a legközelebb van hozzá.

A pontok a folyamat során az optimális hely (a megfelelő centroid) felé „vándorolnak". Ezután annyiszor ismételjük meg a műveletet, ahányszor szükséges. Az iteráció addig működik, amíg a középpontok viszonylag stabilak nem maradnak, és a klaszterezés nem éri el a konvergenciát (vagyis amikor már csak kevés pont helyzete változik). Ez általában nem jelent többet néhány ismétlésnél (Aggarwal – Zhai, 2012c: 93–94). A K-közép klaszterezésnél végeredményben azt kell látnunk, hogy egy klaszterbeli pont közelebb van a saját klaszterjének bármelyik pontjához, mint egy másik klaszter valamelyik pontjához.

A K-közép módszer alkalmazásának nyilvánvalóan vannak korlátai. Az egyik hátránya az, hogy érzékeny a kezdetben kijelölt klaszterszámra. Ezen úgy tudunk segíteni, ha egyrészt többször lefuttatjuk a klaszterelemzést, másrészt több klasztert keresünk, és a végén összevonjuk a különböző klasztereket. Ezenkívül is léteznek olyan technikák, például a félig felügyelt módszerek, amelyekkel kiküszöbölhetjük ezt a problémát. Ezek nagyon hasznosak a különböző zajok kezelésében, különösen a szövegdokumentumok kategorizálásánál.

Egy másik FAC-típusú feladatmegoldást tesz lehetővé az ún. *rejtett Dirichlet-allokáció* (LDA). E hierarchikus valószínűségi modell egy korpusz dokumentumait reprezentálja rögzített témák keverékeként (Bíró, 2009: 1). Az LDA az ún. topik- vagy *témamodellek* egyik típusa, és feltételezi, hogy minden dokumentumkorpusz témák keverékéből áll. Statisztikailag a csoportok (azaz a témák) a korpusz szókészletének valószínűségi függvényeinek (eloszlásának) tekinthetőek.

Az LDA egy olyan csoportosítási eljárás, amely a dokumentumokhoz ún. téma- (azaz topik-) szavakat rendel. A modell a témák megbecsléséhez a szavak együttes megjelenését vizsgálja a dokumentum egészében. Az LDA algoritmusnak is meg kell adni a keresett klaszterek számát, ami ebben az esetben a topikok számát jelenti. Ezt követően a dokumentumhalmazban szereplő szavak eloszlása alapján azonosítja a topikszavakat, ellentétben az információkinyerő eljárások egy részével, amelyek a dokumentumban előforduló szavak gyakoriságán alapulnak.

Az elemzés során a modell először azonosítja a témákat a dokumentumokból, majd listázza az egyes topikokhoz tartozó szavakat. Vegyünk például egy olyan törvényjavaslatot, amely a kórházak finanszírozásáról szól. Ebben a korpuszban várhatóan találkozhatunk a következő szavakkal: egészség, kórház,

finanszírozás, költségvetés. Az ilyen kulcsszavak eloszlása rajzolja ki a topikokat. Az LDA módszer előnye, hogy az azonosított témák olyan szavakat is tartalmazhatnak, amelyek elsőre nem feltétlen jutottak volna a humán kódoló eszébe, így vélhetően – ha osztályozással próbáltuk volna megoldani ezt a feladatot – nagy eséllyel kimaradtak volna. Ez a példa jól illusztrálja azt a korábbi megjegyzést is, mely szerint a felügyelt tanulási módszer nem alkalmas arra, hogy új csoportosítási lehetőségeket kínáljon a kutató számára.

Számítógéppel támogatott klaszterezés (CAC)

Miközben az automatikus klaszterezés egy relatíve egyszerű eszközt ad a kutató kezébe, nem könnyű egy *adott dokumentum vagy korpusz kapcsán* olyan elveket azonosítani, amelyek feltétlenül „jó" (azaz érvényes és megbízható) klaszterezéshez vezetnek. Grimmer és Stewart (2013: 19) szerint sokkal könnyebb azt megállapítani, hogy a csoportosítás valamelyik típusa hasznosnak bizonyul-e egy adott korpusz esetében. A CAC tulajdonképpen ezt a (FAC esetében fennálló) problémát próbálja áthidalni.

A számítógéppel támogatott klaszterezés egy olyan módszer, amely lehetővé teszi, hogy sokféle csoportosítási lehetőség között tudjunk válogatni és keresgélni. A kutatás során több klaszterezési eljárás működését is megvizsgáljuk, majd a kutató kiválasztja ezek közül a legmegfelelőbbet. A CAC alkalmazása során először is különböző FAC-alapú technikákkal vizsgáljuk a szöveghalmazt, melyek eltérő csoportosítási lehetőségeket kínálnak fel. Ezután kiválaszthatjuk azokat a klaszterezési eljárásokat, amelyek hasonló módon csoportosították a dokumentumokat. A számítógéppel támogatott klaszterezés legfőbb előnye így az, hogy változatos particionálási módszert kínál, ezért a kutatónak többféle lehetősége van az adatok csoportosítására. Másfelől nagy terhet is ró a kutatóra azáltal, hogy neki kell azonosítania a leghasznosabb algoritmusokat. Ez azt is jelenti, hogy a módszer sikeressége vagy eredményessége nagyban függ a kutatók kitartásától, azaz a számítógép használata nem csökkenti a kutatók szerepét.

A CAC és a FAC módszer is igényli a klaszterek számának megadását a modellben. A K-közép modell esetében a klaszterek számát (K), az LDA-ban pedig a témák számát kell kiválasztani. A számítógéppel támogatott módszerben ezen túl meg kell határozni a klaszterek számát a végső csoportosítás során is. A felügyelet nélküli tanulási módszerek alkalmazása esetében így a klaszterek számának a meghatározása jelenti az egyik legnagyobb kihívást.

III.3.1. példa

A klaszterezést gyakran használják információk kinyerésére webalapú adathalmazokból. Tegyük fel, hogy érdekel minket egy politikai párt facebookos követőinek az összetétele. Hány ismerősük van? Hol laknak? Milyen a párkapcsolati állapotuk? Milyen dolgokat szeretnek csinálni? Stephen Wolfram kutatása ugyan nem politikai jellegű, de jól bemutatja, hogy a Facebookon lévő adattartalmakból milyen információkat lehet kinyerni a felhasználók kapcsán.[1] A kutatásból kiderül, hogy egy átlagos felhasználónak hány ismerőse van, mely életkorban szerzi legtöbb ismerősét, és különböző életkoraiban milyen életkorú barátai vannak (nem beszélve párkapcsolati állapotáról, barátai számáról, földrajzi helyzetéről és így tovább).

III.3.2. példa

Vegyünk egy konkrét példát az LDA módszer alkalmazására: egy interpelláció elemzését (ld. III.3.3. ábra).[2]

III.3.3. ábra – Példa egy interpelláció elemzésére

PÜSKI ANDRÁS (MDF): Köszönöm a szót. Tisztelt Elnök Asszony! Tisztelt Miniszter Úr! Tisztelt Képviselőtársaim! (1) Néhány héttel ezelőtt a Gazdasági Versenyhivatal az M3-as, M7-es, M70-es autópálya-szakaszok építésével összefüggő versenyfelügyeleti eljárásban minden idők eddigi legmagasabb, 7 milliárd forintos bírságát szabta ki a pályázaton nyertes cégekkel szemben. (2) Azt is megállapították, hogy a pályázók versenyjogot sértő kartellbe tömörülése 18,5 milliárd forinttal drágította meg a beruházást. Azt is mondhatjuk tehát, hogy ennyivel károsították meg a költségvetést, ennyi pénzt húztak ki az adófizető polgárok zsebéből. Ez a pályázat a 2002-es kormányváltást követően került kiírásra, tehát miniszter úr döntése alapján állhatott elő ez a helyzet. (3) A Medgyessy Péter miniszterelnök úr személye körüli bizalmi válság elindítója pedig éppen az ön tervezett leváltása volt. A miniszterét védelmébe vevő SZDSZ többször is elmondta, hogy nem kívánnak minisztert beáldozni a kormányváltás során.Tisztelt Miniszter Úr! (4) Joggal vetődik fel tehát a kérdés: a Gazdasági Versenyhivatal autópályacégeket elmarasztaló döntése miatt került volna sor az ön leváltására? (5) Mindezek után pedig vajon az ön SZDSZ által támogatott politikai és gazdasági sérthetetlensége vezetett Medgyessy Péter bukásához? (6) Medgyessy Péter a közszolgálati televízióban elhangzott nyilatkozatában azt sejtette, többet tud az ügyekről, arra utalt, ezek a zavaros, korrupciós ügyek több szereplősek annál, mint amit az emberek látnak. (7) Miniszter úr, kérdezem tehát, mi az összefüggés az autópálya-építések körül kialakult zavaros, korrupciógyanús ügyek és a miniszterelnök bukását eredményező bizalmi válság között? (Taps az ellenzéki pártok padsoraiban.)

[1] Forrás: Stephen Wolfram: Data Science of Facebook World, April 24, 2013. Elérhető: http://blog.stephenwolfram.com/2013/04/data-science-of-the-facebook-world/

[2] A hangsúly az algoritmus működésének a bemutatásán van, ezért nem lényeges az interpellációról szóló információ. A bemutatott interpellációk száma sem releváns.

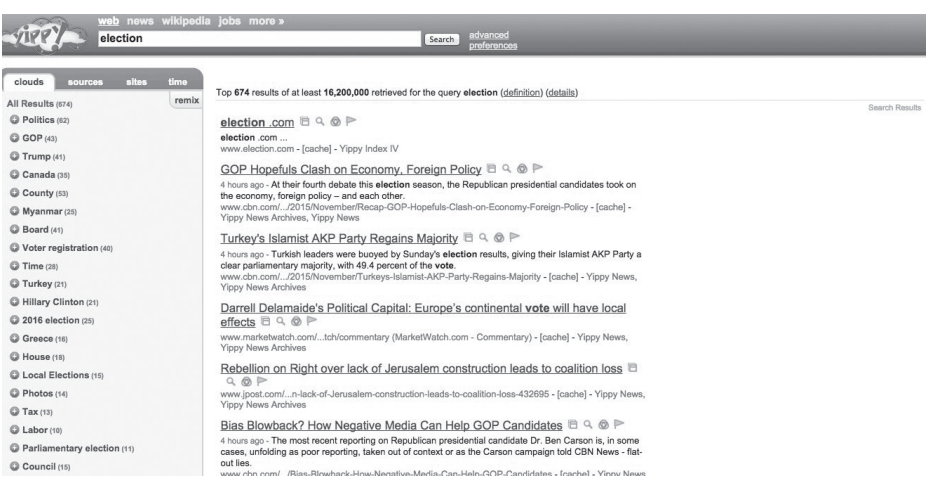
A példa jobb érzékeltetése érdekében vegyük figyelembe a számmal jelölt mondatokat a szövegből. A kiemelések a szövegben megjelenő különböző témákat jelölik: azokat a szavakat, amelyek az autópályához, versenyhez kapcsolódnak, illetve azokat, amelyek Medgyessy Péter személyéhez köthetők.

Az LDA algoritmusnak meg kell adni, hogy hány topikot azonosítson. Tegyük fel, hogy ebben az esetben két topikot keresünk. Az 1-es és a 2-es mondatokat az algoritmus az A topikhoz rendelné. A 3-as, az 5-ös és a 6-os mondatot ezzel szemben a B témához csoportosítaná. A 4-es és a 7-es mondat esetében ugyanakkor nem egyértelmű a besorolás, mivel a szavak előfordulása alapján 50-50% százalék az esélye az egyik vagy a másik topikhoz való tartozásnak. Végül az algoritmus listázza a topikszavakat a témák szerint: A (autópálya, versenyjog, beruházás, autópályacég, autópálya-építés stb.) és B (bizalmi válság, leváltás, bukás, korrupciós ügyek, korrupciógyanús, miniszterelnök-bukás stb.). Az LDA alkalmazása nyomán tehát a dokumentumban lévő szavak eloszlása alapján kapjuk meg a dokumentum témacsoportjait.

III.3.3. példa

A Yippi (korábban Clusty – http://yippy.com/) metakereső jó példáját adja a klaszterelemzés gyakorlati felhasználásának. A metakereső nem más, mint egy olyan keresőmotor, ami a felhasználó kéréseit továbbítja más keresőknek vagy adatbázisoknak, és az így visszakapott találatokat összegzi. A Yippy pedig egy olyan speciális metakereső, amely a keresés eredményeit úgy tálalja, hogy azokat különböző klaszterekbe rendezi.

III.3.4. ábra – Egy Yippy oldalon készült keresés eredménye



A keresőbe beírhatjuk az általunk keresett kifejezést, mely a III.3.4. ábra esetében a „választás" angol nyelvű alakja volt. A Yippy a keresési eredményeket különböző témák szerint rendezi úgy, hogy a hasonló elemeket egy témába csoportosítja. Így egymástól némiképp különböző csoportok („mappák") jönnek létre. A bal oldali sávban láthatjuk, hogy a 674 kiemelt találatból a „választás" kifejezésre milyen csoportokat alkotott a program: politika (*politics)*, Republikánus Párt (GOP), Trump, Kanada, megye (county), Mianmar stb. Ezeket a mappákat tovább lehet bontogatni, és például a Republikánus Párt mappa a következő „almappákat" vagy klasztereket tartalmazza: Trump, republikánus jelöltek, republikánus vita stb. A keresőmotor egyik legnagyobb előnye, hogy a keresési eredmények alapján foglalkozhatunk csak a számunkra releváns csoportokkal. A weboldal felső részében pedig beállíthatjuk, hogy a teljes weben vagy például csak a hírek vagy álláshirdetések között keressen (a példa a teljes webre vonatkozott).

Nemzetközi politikatudományi alkalmazások

Grimmer (2010) egy felügyelet nélküli tanulási eljárást használt annak a megvizsgálásához, hogy az USA Kongresszusának tagjai hogyan kommunikálnak a választókkal. A klaszterelemzéshez 24 000 olyan sajtóanyagot gyűjtött össze 2007-ből, amelyekben szenátorok beszélnek a választókhoz. Ezeket a sajtóanyagokat olyan ideális eszköznek vagy forrásnak tekintette, amelyek jól képesek illusztrálni azt, hogy a szenátorok hogyan magyarázzák el washingtoni munkájukat a választóiknak.

Quinn és társai (2010) szintén az amerikai Kongresszust, s ezen belül is a szenátus törvényalkotási napirendjét, az ott felmerülő témákat vizsgálták. Ez a tanulmány a szövegadatok egyik jellegzetes klaszterezési eljárását, a topikmodellezés hasznát illusztrálja. Az elemzéshez a szenátorok törvényalkotási tevékenységét vették alapul, és az 1997 és 2004 között elhangzott felszólalásokat elemezték. Az adatbázis 118 000 beszédet tartalmazott, az így létrejött korpusz pedig 70 millió szó hosszú volt.

A kutatók azt feltételezték, hogy a szenátusban mindennap különböző témák merülnek fel. A rejtett Dirichlet-allokáció alkalmazásával minden egyes beszédet különböző topikokhoz csoportosítottak úgy, hogy az egyes szavak eloszlása rajzolta ki az egyes témákat. Végül az egy nap során elhangzott összes beszédet egyetlen témához rendelték. Az LDA módszerrel tehát először azonosították az eltérő témákat, másodszor az egyes kulcsszavakat, amelyek a témákat meghatározták, harmadszor pedig megbecsülték a témák hierarchikus szerkezetét. Összefoglalva tehát ezzel a módszerrel megtudták azt, hogy milyen témák merülnek fel az egyes napokon, és azok milyen intenzitással jelennek meg a szenátus napirendjén.

Ellenőrző kérdések

- Mi a fő különbség az osztályozási és a csoportosítási eljárás között?
- Milyen kutatási terv esetében célszerű a felügyelt helyett felügyelet nélküli tanulási módszertant alkalmazni?
- Mi a különbség a teljesen automatizált és a számítógéppel támogatott klaszterezés között?
- Hogyan validálhatjuk a klaszterezés során kapott eredményeinket?

Szószedet

Magyar	Angol
Címke	Label
Csoportosítás (klaszterezés)	Clustering
Felcímkézés	Labeling
Felügyelet nélküli tanulás	Unsupervised methods
Felügyelt tanulás	Supervised learning
Félig felügyelt tanulás	Semi-supervised learning/ methods
Hierarchikus módszerek	Hierarchical clustering
K-közép algoritmus	K-means algorithm
Konvergens érvényesség	Convergent validity
Látens Dirichlet-allokáció	Latent dirichlet allocation (LDA)
Lemmatizálás (szótövek keresése)	Lemmatization
Particionáló (felosztó) módszerek	Partitional clustering
Prediktív (előrejelző) érvényesség	Predictive validity
Számítógéppel támogatott klaszterezés	Computer assisted clustering (CAC)
Szemantikai érvényesség	Semantic validity
Szózsák	Bag of words
Szótőképzés, szótövezés	Stemming
Szövegelőkészítés	Pre-processing
Tanítóhalmaz	Traing-set
Teljesen automatizált klaszterezés	Fully automated clustering (FAC)
Tiltólistás szavak eltávolítása	Stop-word removal
Topikmodellek	Topic modeling methods
Validitás (érvényesség)	Validity

Ajánlott irodalom

A klaszterezésbe való általános betekintésért ld. Aggarwal, C. C. – Zhao. Y – Yu, P. S. (2012); Hopkins – King (2010); Jain, – Murty – Flynn (1999) és Manning – Raghavan – Schütze (2008). A szövegek számokká alakításához nyújt segítséget: Hopkins – King (2010b); Manning – Schütze (1999). Ha a klaszterezési típusokról szeretnénk többet megtudni Blei – Ng – Jordan (2003); valamint Croft (1977) műve ajánlott. A klaszterezés eredményeinek validálásához nyújt segítséget Adcock – Collier (2001); illetve Quinn et al. (2010).

IV. A SZÖVEGBÁNYÁSZATI MÓDSZEREK KUTATÁSI ALKALMAZÁSA

Kubik Bálint György

IV.1. OSZTÁLYOZÁS: FELÜGYELT TANULÁSI MÓDSZEREK

A fejezet bevezetőt nyújt a felügyelt tanulási módszerek alapvető logikájába. A szöveg fő ívét egy az USA Kongresszusában benyújtott törvényjavaslat-szövegekre épülő gyakorlati példa adja. A módszertant egy R programozási nyelven írt forráskód értelmezésével illusztráljuk, melynek során bemutatjuk a vonatkozó felügyelt tanulási modellt, ennek becslését, illetve röviden értékeljük a modell eredményeit is.

A *felügyelt tanulás* elvére épülő módszereket *induktív tanulási megoldásoknak* vagy *induktív osztályozásnak* is nevezhetjük. A felügyelt tanulási alapú módszerek feltételezik, hogy rendelkezünk osztálycímkékkel, tehát tudjuk a megfigyeléseink legalább egy részéről, hogy azok pontosan milyen csoportba tartoznak (Liu, 2007). A kategóriabesorolás (legalább részleges) ismerete jelentős különbséget jelent a nem felügyelt tanulási módszerekkel szemben (ezeket a könyv IV.2. fejezete tárgyalja).

A felügyelt tanulás *statisztikai modellezés,* melynek során a megfigyelési egységek valamilyen tulajdonságai alapján „magyarázzuk" azt, hogy az adott elem milyen osztályba tartozik. Az alternatív elnevezések jól érzékeltetik, hogy milyen módszerrel is állunk szemben: induktív módon a megfigyeléseink (jelen esetben a szövegkorpuszunk adatai) felől haladunk az általánosítások (a szöveg adataiban fellelhető mintázatok) felé. E folyamat célja az, hogy új elemeket tudjunk beilleszteni a már rendelkezésre álló osztálytagolásba. A függvényt, mely bizonyos szövegtulajdonságok (pl. bizonyos kifejezések valamilyen súlyú előfordulása) alapján megmondja számunkra, hogy az adott szöveg milyen csoportba is tartozik, *klasszifikációs* vagy *predikciós modellnek* hívjuk (i. m.: 56).

A felügyelt tanulási alapú módszerek igen sokfélék lehetnek, és megértésük érdemi statisztikai előismereteket igényel, így a különböző technikák részletes ismertetésére e fejezetben nem vállalkozunk. Egy általános bevezetőt adunk ugyanakkor a felügyelt tanulási módszertanba, amit a politikatudományhoz szorosan kapcsolódó gyakorlati példák segítségével illusztrálunk. Ennek keretében lépésről lépésre elemzünk egy törvényhozási korpuszt, illetve röviden kitekintünk a módszer egyes nemzetközi politikatudományi felhasználásaira is.

Célunk így elsődlegesen az, hogy az olvasó tisztába kerüljön a felügyelt tanulási módszerek logikájával, illetve el tudjon végezni egy többkategóriás klasszifikációs elemzést az R programozási nyelv felhasználásával (a forráskódok lefuttatásához az *RStudio*[1] nevű program nyílt forráskódú verzióját ajánljuk).

A kutatási probléma és megoldása

A modern törvényhozások nagy mennyiségű törvényjavaslatot „termelnek" munkájuk során. A hatalmas mennyiségű irat feldolgozása komoly nehézség elé állítja a képviselők mellett a politikatudósokat is. Amennyiben nem rendelkezünk jelentős erőforrásokkal (például egy, a javaslatszövegeket elolvasni és csoportosítani képes kutatócsoporttal), hogy manuálisan azonosítsuk például a különféle közpolitikai területek megjelenését az egyes szövegekben, akkor a kvantitatív szövegelemzés felügyelt tanulási ága hasznos módszertani választás lehet. A „felügyelt" melléknév arra utal, hogy a felügyelt tanulási módszerek esetében ismernünk kell a megfigyeléseink legalább egy részének osztálybesorolását.

Amennyiben e feltétel adott, úgy a törvényszövegek közpolitikai tartalmát meg tudjuk jósolni még akkor is, ha nem tudtuk előzetesen és „biztosan" (tehát más politikatudományi szakemberekkel konszenzust kialakítva) az összes javaslat közpolitikai kódját meghatározni. Ehhez ugyanakkor tudnunk kell legalább a törvényjavaslatok egy részének közpolitikai tartalmát. Az alábbiakban ezen általános elvet mutatjuk be egy konkrét kutatási projekt és a hozzá tartozó forráskód hat lépésben történő elemzésével.

1. lépés: az adatok betöltése

Vegyük az Egyesült Államok 113. Kongresszusának adatait, és oldjuk meg az előzőekben leírt problémát felügyelt tanulási módszerek segítségével! Használjuk fel a *Congressional Bills Project*[2] által ingyenesen rendelkezésre bocsátott adatbázist erre a célra. Az adatbázis tartalmazza az egyes javaslatok címeit, illetve közpolitikai tartalomra utaló változókat is. A közpolitikai csoportosítást

[1] A program telepítéséhez kövessük a https://www.rstudio.com/products/rstudio/download/ weboldal utasításait. A teljes, megszakítások nélküli forráskód elérhető a kötet honlapján (qta.tk.mta.hu).

[2] A Congressional Bills Project egy, az amerikai törvényhozás törvényalkotási adatainak gyűjtésével, illetve azok elemzésével foglalkozó, politikatudományi projekt. A letölthető adatbázis itt található: http://congressionalbills.org/download.html. A közvetlen letöltési link a következő: http://congressionalbills.org/billfiles/bills93-113.zip

a *Policy Agendas Project*[3] végezte el, melynek során kézi kódolók látták el az egyes javaslatokat egy fő *(major topic)* és egy másodlagos *(subtopic)* közpolitikai kóddal.

Elemzésünket kezdjük azzal, hogy importáljuk a szükséges R *csomagokat*. Ezek a következők: „xgboost", „tm", „readr" és „wordcloud". Ha az alábbiak valamelyike nem áll rendelkezésünkre, akkor telepítsük őket az *install.packages* paranccsal (pl. *install.packages[„xgboost"]*).

Következő lépésként töltsük be az adatbázist az R-be (az elérési cím értelemszerű módosítása után), illetve szűrjük ki azokat a javaslatokat, melyek egyrészt nem rendelkeznek fő közpolitikai kóddal (a *Major* kód esetükben 99), illetve amelyek nem a 113. Kongresszus alatt keletkeztek. Oldjuk meg egy műveletben azt is, hogy csak az egyedi azonosító és a számunkra releváns változók *(Major, Title)* maradjanak bent az adatbázisban.

```
congressData <- read_delim(„bills93-113.txt", delim="\t")
congressData <- congressData [congressData$Cong=="113" & congressData$Major!
="99",c(„id","Major","Title")]
```

Az egyszerűség kedvéért szabaduljunk meg továbbá a több közpolitikai kóddal rendelkező törvényszövegektől az alábbi paranccsal:

```
congressData <- congressData[duplicated(congressData$Title)==FALSE,]
```

Vessünk egy pillantást az adatbázisra az alábbi parancs lefuttatásával. A IV.1.1. ábrának megfelelő adatsort fogunk látni.

```
View(congressData)
```

[3] A Policy Agendas Project a nemzetközi Comparative Agendas Project (http://www.policyagendas.org/) részeként törvényhozási szövegeket, illetve médiacikkeket elemez annak érdekében, hogy meghatározza azok közpolitikai tartalmát. A projekt magyar kutatócsoportjáról a http://cap.tk.mta.hu-n lehet tájékozódni.

IV.1.1. ábra – Az adatbázis megjelenítése az RStudióban

	id	Major	Title
218687	1000001	1	To amend the Internal Revenue Code of 1986 to pro...
218688	1000002	8	To remove Federal Government obstacles to the pro...
218689	1000003	8	To approve the construction, operation, and mainten...
218690	1000004	1	To make revisions to Federal law to improve the con...
218691	1000005	6	To support State and local accountability for public ...
218692	1000006	8	To provide for expedited approval of exportation of ...
218693	1000007	2	To prohibit taxpayer funded abortions.
218694	1000008	6	To amend the charter school program under the Elem...
218695	1000009	20	To reauthorize the Violence Against Women Act of 1...
218696	1000010	2	To modernize voter registration, promote access to ...
218697	1000011	9	To provide for comprehensive immigration reform an...
218698	1000012	20	To reform the financing of Congressional elections b...
218699	1000013	12	To provide for greater safety in the use of firearms.
218700	1000014	18	To provide for the exchange of information related t...
218701	1000015	2	To provide that human life shall be deemed to begin ...
218702	1000016	1	To require a full audit of the Board of Governors of th...
218703	1000017	1	To promote freedom, fairness, and economic opport...

Showing 1 to 18 of 8,583 entries

Ezen a ponton rendelkezésünkre áll minden információ, hogy a korábbiakban leírt politikatudományi problémát megoldjuk, hiszen – mint látni fogjuk – a javaslatcímek alapján az esetek döntő részében eldönthető, hogy az adott törvényjavaslat milyen közpolitikai tartalommal bír.

2. lépés: a szövegkorpusz tisztítása

Az elemzést megelőzően a szövegkorpuszt érdemes megtisztítani a fölösleges szavaktól, karakterektől. A szövegelőkészítés során gyakran eltávolítják az ún. *tiltólistás szavakat* (ld. II.1. fejezet), melyek nyelvtani funkciót látnak el, és nem adnak többletinformációt az elemzéshez. Javasolt továbbá megválni az írásjelektől, illetve kisbetűsíteni a szöveget (Grimmer – Stewart, 2013: 7).

A korábbiakban már telepített és importált *tm* elnevezésű R csomag megkönnyíti a szövegtisztítási műveleteket. Ehhez első lépésben alakítsuk át az adatbázisunk *Title* oszlopát a csomag számára értelmezhető *volatile corpus*, vagy *VCorpus* formátumba.

```
titleCorpus <- VCorpus(VectorSource(unique(congressData$Title)))
```

Kisbetűsítsük a szöveget, majd távolítsuk el az írásjeleket és a stopszavakat!

```
titleCorpus <- tm_map(titleCorpus, content_transformer(tolower))
titleCorpus <- tm_map(titleCorpus, removeWords, stopwords(„english”))
titleCorpus <- tm_map(titleCorpus, removePunctuation)
```

3. lépés: a szövegkorpusz elemezhető formába hozása

Adatainkat ebben a szakaszban először elemezhető formátumba kell hoznunk, mely többet takar az előző részben elvégzett tisztításoknál. A törvényjavaslatok címei jelenleg még szövegként vannak eltárolva, ami nem teszi lehetővé az elemzés végrehajtását. Alakítsuk hát át őket olyan formára, mely már értelmezhető a felügyelt tanulási algoritmusok számára. Ez az értelmezhető forma a *dokumentum-kifejezés mátrix* (DTM – a konkrét alkalmazástól függően lehet *kifejezés-dokumentum mátrix* is). A DTM-táblázatban hozzárendeljük az egyes szövegekben szereplő kifejezések számát a korpusz teljes dokumentumlistájához úgy, hogy az egyes cellákban az adott dokumentumban (sor) az adott szó (oszlop) gyakoriságának száma szerepel. A DTM kialakításával a szöveget matematikai számításokra alkalmas, numerikus formába hoztuk (Grimmer – Stewart, 2013: 7).

Mindezt illusztrálandó vegyünk egy példát a már kialakított korpuszunkból, válasszunk ki két javaslatcímet a 113. Kongresszus irományai közül:

1. cím: To prohibit taxpayer funded abortions.
2. cím: To restore safety to Americas schools.

A két mondat dokumentum-kifejezés mátrixa a következőképp fog kinézni:

IV.1.1. táblázat – Példa a dokumentum-kifejezés mátrixra

	abortions	americas	funded	prohibit	restore	safety	schools	taxpayer	to
1. cím	1	0	1	1	0	0	0	1	1
2. cím	0	1	0	0	1	1	1	0	2

Látható, hogy az egyes szavak előfordulási gyakoriságai jelentik itt a szöveg *adatreprezentációját* (ez egyben hasonlít a II.1. fejezetben tárgyalt *szózsák* megoldáshoz). A szakirodalomban találhatunk példákat több szóból álló kifejezések kezelésére is, azonban elmondható, hogy a bonyolultabb módszerek nem

eredményeznek jelentős javulást a gyakorlatban az egyes szavak egymástól független kezeléséhez képest (Grimmer – Stewart, 2013: 6).

Mindezek után alakítsuk át a 8583 javaslatcímből álló korpuszunkat egy teljes dokumentum-kifejezés mátrixszá, majd konvertáljuk az objektumot *data. matrix* formátumúvá, hogy a későbbiekben könnyebben tudjunk vele dolgozni.

```
titleDTMatrix <- DocumentTermMatrix(titleCorpus)
titleDTMatrix <- data.matrix(titleDTMatrix)
```

Vessünk egy pillantást a tisztított szövegkorpusz leggyakoribb szavaira! Ehhez vegyük a kialakított DTM oszlopainak (tehát az egyes kifejezések gyakoriságainak) összegét, majd az összegeket rendezzük csökkenő sorrendbe a *sort* paranccsal.

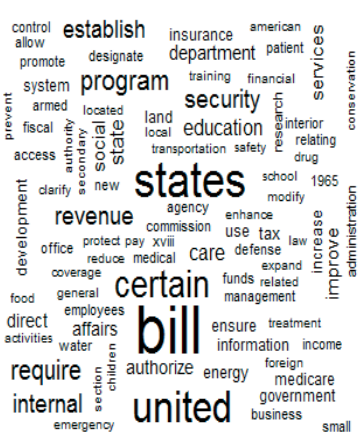
```
termFrequencies <- colSums(titleDTMatrix)
termFrequencies <- sort(termFrequencies, decreasing = TRUE)
```

Vizsgáljuk meg a kifejezésgyakoriságokat a szófelhővel! A *wordcloud* elnevezésű R kiegészítő megfelelő eszközt biztosít a művelethez.

```
wordcloud(names(termFrequencies), termFrequencies, min.freq = 100)
```

Ahogy a IV.1.2. ábrán is látható, a leginkább gyakori szavak között olyan kifejezéseket találunk, mint a *provide, program, veterans,* vagy *states*. Mint ahogy azt feltételezhettük, számos olyan elemet találhatunk a szófelhőnkben, melyek közpolitikai tartalmat hordoznak (*veterans, energy, investment* stb.).

IV.1.2. ábra – A 113. Kongresszus javaslatcímeinek szófelhője



A létrehozott dokumentum-kifejezés mátrix 8583 sort és 9183 oszlopot tartalmaz, ami a sorok és oszlopok számának szorzataként 78 817 689 elemet jelent. A mátrix 602,3 MB-ot foglal a számítógépünk memóriájában, tehát egy igen nagy méretű objektumról beszélhetünk. A korábbiak fényében érdemes lehet a legkevésbé vagy akár a leginkább gyakori kifejezések egy részétől megválni, hiszen azok valószínűleg kevés többletinformációt tartalmaznak számunkra (elemzésünket ugyanakkor a nagy méretű objektumban folytatjuk).

4. lépés: a minta kettéosztása tanulási és teszthalmazra

A felügyelt tanulási módszerek alkalmazásának alapvető logikája az, hogy a rendelkezésünkre álló (tehát csoportokba előzetesen besorolt) megfigyeléseket véletlenszerűen kettéosztjuk. Az egyik részen, az ún. *tanítóhalmazon* végezzük a modellillesztést, az adatbázis másik felén, az ún. *teszthalmazon* pedig értékeljük módszerünk eredményességét (Liu, 2007: 57).

Osszuk tehát ketté az adatainkat olyan módon, hogy egy véletlenszerűen kiválasztott, 4000 javaslatcímből álló tanítóhalmazon alkalmazzuk a felügyelt tanulási módszerünket, míg a maradék 4583 megfigyelést tartalmazó teszthalmazon a megoldás pontosságát értékeljük. Annak érdekében, hogy a véletlenszerű számításokon alapuló műveletek pontosan megismételhetőek legyenek, az előzőek előtt állítsuk be az ún. *seed* értékét.

```
set.seed(1)
train_ind<-sample(seq_len(nrow(titleDTMatrix)), size = 4000)
train <- titleDTMatrix[train_ind, ]
test<- titleDTMatrix[-train_ind,]
```

Ezt a felosztást az indokolja, hogy a teljes mintán belül végzett modellbecslés könnyen *túlillesztéshez* vezethet. Ez azt jelenti, hogy az algoritmus annyira „rátanul" a meglévő mintánkra, hogy a mintán kívüli, azaz új megfigyelések csoportosítása alacsony eredményességű lesz. A tanítóhalmaz mérete már lehetővé teszi az eredményes felügyelt tanulást, tehát az adatmennyiség növelésével nem érünk el jelentős javulást a felügyelt tanulási módszer teljesítményében (erre alább visszatérünk). Így a többi megfigyelés teszthalmazként való alkalmazása indokolt.

5. lépés: a felügyelt tanulás művelete

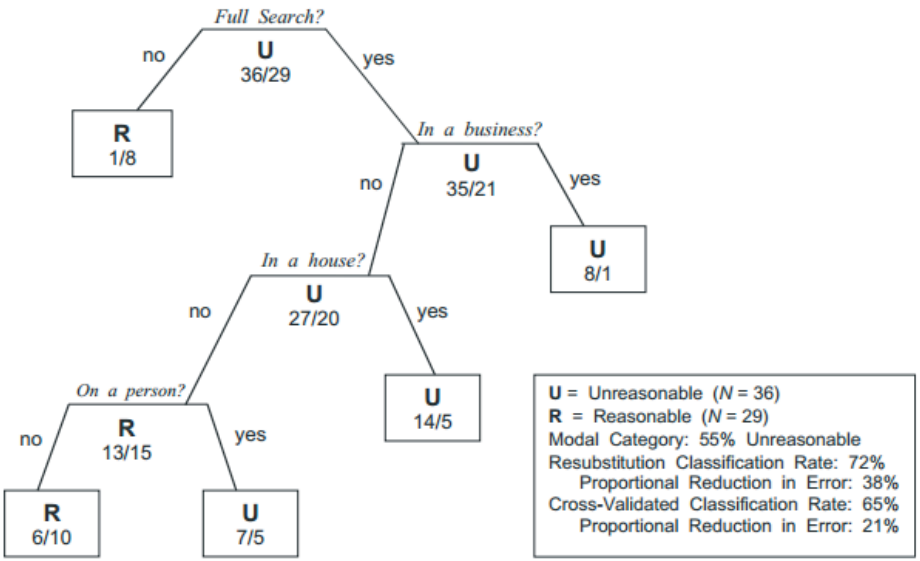
Az eddigiek során a javaslatcímeket elemezhető formátumba hoztuk, továbbá kettéosztottuk a mintánkat egy tanítóhalmazra, illetve egy teszthalmazra, így tehát már elvégezhetjük a felügyelt tanulási számításokat. A példában az egyszerűség kedvéért alkalmazzunk egyetlen algoritmust, az igen elterjedt és magas pontosságot elérni képes klasszifikációs döntési fák módszerét.

A döntési fák egy olyan módszer, mely képes az adatainkban fellelhető fő szabályosságok meghatározására. Az algoritmus lépésenként kisebb és kisebb darabokra osztja az adatmennyiséget aszerint, hogy melyik változók jelentik a legjobb elválasztási logikát, azaz hogy milyen tényezők osztják fel legjobban a megfigyeléseinket osztályokra (az osztályok jelen esetben az osztálytagságot jelentik). A döntési fák módszere igen jó osztályozási pontosságot eredményez a többi felügyelt tanulási módszerrel összehasonlítva is (Liu, 2007: 67, ill. IV.1.1. keret).

IV.1.1. példa

A döntési fák módszerének megértéséhez tekintsük át Kastellec (2010) cikkét, aki az amerikai Legfelsőbb Bíróság döntéseit elemezte döntési fák módszerével. A szerző a klasszifikációs fák politika-, illetve jogtudományi használhatósága mellett érvel, mondván a módszertan kapcsolatot teremt a megfigyelt tények (jelen esetben a bírósági döntések) és a tényekhez kapcsolódó információk mintázatai között. Kastellec a Legfelsőbb Bíróság 1962 és 1972 közötti házkutatási és lefoglalási ügyeiben hozott döntéseinek elemzésével rajzolta fel az alábbi döntési fát (ld. IV.1.3. ábra).

IV.1.3. ábra – Az USA Legfelsőbb Bírósága 1962 és 1972 közötti döntéseinek döntési fája



Forrás: Kastellec, 2010

Az első tényező, mely a bírósági döntéseket két csoportra osztja, a házkutatás mértéke volt, így meg is határozhatunk két ágat a döntési fánkban *(full search – yes; full search – no).* Amennyiben nem teljes házkutatásról volt szó, úgy a bírók nagy valószínűséggel helytállónak minősítették a hatóságok eljárását. Mivel ettől a ponttól nem haladt tovább az algoritmus (tehát a házkutatás mértéke már önmagában jól előre jelezte a bíróság döntését), ezt egy ún. *levélcsúcsnak* tekinthetjük. Amennyiben teljes házkutatásról volt szó *(full search – yes),* úgy már egy új döntési változó (érintett-e vállalkozás, vagy sem) bevezetésére van szükség *(in a business? – yes* és *in a business? – no).*

Ahol az águnk valamilyen döntési szabály esetén elágazik, azt döntési vagy *belső csúcsnak* tekintjük. Ha vállalkozással kapcsolatos volt az ügy, akkor a bíróság valószínűleg helytelennek minősítette az eljárást, ellenkező esetben pedig egy új döntési csúcshoz érünk. Látható, hogy hamarosan eljutunk a legalsó szintig, ahol már mindkét ág egy levélcsúcsként végződik, tehát a klasszifikációs algoritmusunk végzett az egyes döntések csoportosításával. Kastellec példája egy kis méretű döntési fa, mely elegendőnek bizonyult az adatban fellelhető fő szabályosságok azonosítására és általánosítások megfogalmazására. Fontos, hogy a döntési fák logikája alapján minden ügy csak és kizárólag egy ághoz tartozhat (Liu, 2007: 69).

A közpolitikai kódok javaslatcímekkel történő „megjóslását" is egy, a döntési fák logikájához igen hasonló módszerrel (*Gradient Boosted Decision Trees*,

GBDT) hajtjuk végre. A számítások elvégzéséhez az R *xgboost* nevű kiegészítőjét használjuk, mely a számításokat gyorsan és hatékonyan hajtja végre, továbbá nagy osztályozási pontosságot biztosít. A művelet a *Major* közpolitikai kódok kategorikus változóit fogja a korábban kialakított dokumentum-kifejezés mátrix értékei alapján megjósolni. Első lépésként alakítsuk át a *Major* változót kategorikussá, melynek numerikussá alakított értékeiből vonjunk ki egyet, hogy azt a felügyelt tanulási algoritmus képes legyen megfelelően kezelni (az *xgboost* R kiegészítő csak a 0-val kezdődő numerikus csoportváltozókat kezeli).

```
levels <- sort(unique(congressData$Major))
congressData$Major2 <- as.numeric(factor(congressData$Major, levels=levels))-1
```

Rendeljük továbbá az átalakított közpolitikai kódokat az eredetiekhez, hogy majd vissza tudjuk fejteni a modell által megjósolt értékekből a valós közpolitikai kódokat.

```
oldNew<-unique(congressData[,c(„Major”,”Major2”)])
```

A következő lépésben átalakítjuk a *train,* illetve a *test* mátrixainkat ún. *DMatrix* formátumú objektumokká, mely erőforrás-takarékos (a lehetőségekhez mérve kevés helyet foglal a számítógép memóriájában) és az XGBoost kiegészítő számára könnyen értelmezhető módon tárolja el az adatainkat. A parancs első argumentuma adja meg, hogy milyen adatbázis tartalmazza a felügyelt tanulási modell „magyarázó változóit” (jelen esetben az egyes szavak gyakoriságát a különféle törvénycímekben), a *label* argumentum értéke pedig azt, hogy mely értékek jelzik a csoport-hovatartozást (a mi esetünkben ezek a tanítóhalmaz, illetve a teszthalmaz soraihoz tartozó átalakított közpolitikai kódok).

```
dtrain <- xgb.DMatrix(train, label = congressData$Major2[train_ind])
dtest <- xgb.DMatrix(test, label = congressData$Major2[-train_ind])
```

Futtassuk le az így meghatározott felügyelt tanulási algoritmust a tanítóhalmaz értékein, és értékeljük a modellépítés eredményességét a teszthalmaz elemein! Az xgb.train parancs hajtja végre a felügyelt tanulási műveletet, és ehhez meg kell adnunk az alábbi főbb argumentumokat:

- A *num_class* a csoportok számát jelöli (esetünkben a fő közpolitikai kódok számát, azaz 20-at).
- Az *nrounds* azt adja meg, hogy az algoritmus maximum hány iterációig fusson. A mögöttes mechanizmus ismertetésére itt nincs hely, elég annyit megjegyeznünk, hogy az újabb és újabb iterációk során a felügyelt tanulási modellünk egyre jobban „rátanul” az adatainkra.

- Az *objective* értéke a lefuttatandó algoritmus típusát mondja meg (most egy többkategóriás klasszifikációról van szó, melyet a *multi:softmax* jelöl).
- Az *eval_metric* az általunk használni kívánt teljesítményindikátort definiálja (esetünkben ez a hibaszázalék, melyre alább visszatérünk).
- A *watchlist* megadja azokat az adatbázisokat, melyeken nyomon kívánjuk követni a teljesítménymutatónk alakulását (ez a tesztelési halmaz és a tanítóhalmaz).
- Az *early.stop.round* azt adja meg, hogy hány iteráció fusson le, mielőtt az algoritmus a teljesítményindikátor romlása miatt megállna.
- A *set.seed* argumentum a véletlenszám-generálást szabályozza annak érdekében, hogy pontosan megismételhetőek legyenek a számítások.

A lefuttatandó parancs az előbbiek fényében a következőképp néz ki (a parancs futtatása számítógépünk teljesítményétől függően igényel némi időt):

```
model <- xgb.train(data      = dtrain,
                   booster   = „gbtree”,
                   num_class = 20,
                   nrounds   = 100,
                   objective = „multi:softmax”,
                   eval_metric = „merror”,
                   watchlist = list(eval = dtest, train = dtrain),
                   early.stop.round = 10,
                   set.seed      = 1)
```

A megjelenő számok az algoritmus teljesítményét mérik (a korábban már megemlített hibaszázalékot), és az egyre kisebb értékek egyre „jobb” modellt jelölnek. A bal oldali oszlop a teszthalmaz elemein, a jobb oldali oszlop pedig a tanítóhalmaz elemein mért teljesítményt mutatják. Körülbelül 60 iterációt követően az algoritmus megáll, és a *model* nevű objektumban megtalálható a becsült modellünk, mely már alkalmas új törvénycímek klasszifikációjára.

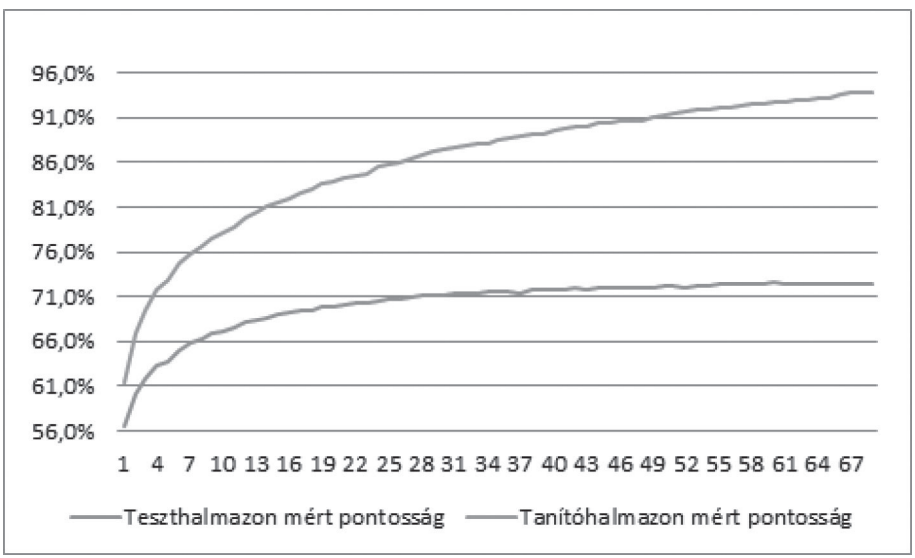
6. lépés: a modellfuttatás eredményei

A modell előrejelzési teljesítményének legalapvetőbb mutatószáma a *hibaszázalék*, mely a rosszul előre jelzett értékek arányát közli az összes értékre vetítve. A *pontosság* a hibaszázalék ellentéte, tehát a megfelelően besorolt megfigyeléseink aránya az összes megfigyelés körében (Liu, 2007: 81). Az algoritmusunk megáll a 60. iteráció körül, hiszen az azt követő futások során a teszthalmaz értékein mérve már nem tud jobb (tehát kb. 27,7%-nál alacsonyabb) hibaszáza-

lékot elérni. Ez azt jelenti, hogy a teszthalmaz törvénycímeinek mintegy 72%-át jól sorolta be a felügyelt tanulási modellünk, ami igen jó pontosságot jelent!

Érdemes azt is megvizsgálni, hogy miképp alakul a hibaszázalék a tanító-, illetve a teszthalmaz értékein. A IV.1.4. ábrán jól látható, hogy a tanítóhalmaz adatain és a teszthalmaz elemein mért pontosság az iterációk számának növekedésével egyre inkább elválik egymástól: a modellépítésre használt tanítóhalmaz elemein mért pontosság már a 13. futásnál meghaladta a 80%-ot, míg a teszthalmaz közpolitikai kódjait csak maximum kb. 72,3%-os pontossággal tudjuk a modellünkkel megbecsülni, ráadásul ez az érték a 60. iterációt követően romlani kezd.

IV.1.4. ábra – A modell pontosságának eltérése a tanító-, illetve a teszthalmazon



Szembekerültünk tehát a már röviden említett *túlillesztés* problémájával. A futásszám növekedésével egyre inkább „rátanulunk" a tanítóhalmazunk adataira, a halmazba tartozó törvénycímek közpolitikai kódjait egyre nagyobb és nagyobb pontossággal tudjuk megjósolni. A teszthalmaz elemeinek sikeres megjóslása azonban egy bizonyos pontosságon túl már nem lehetséges, és a teljesítmény az újabb és újabb iterációk során még romlik is.

A felügyelt tanulási modelljeink építése során a túlillesztés jelenségére igen fontos odafigyelnünk, hiszen számtalan tényező vezethet annak megjelenéséhez. Tisztában kell lennünk azzal, hogy a modellépítésre felhasznált adatokon (tehát a tanítóhalmazon) mért jó teljesítmény nem feltétlenül eredményezi azt, hogy az új adatokon (jelen esetben a teszthalmaz elemein) való becslés is ha-

sonlóan jó lesz. A tanulási, illetve a teszthalmazok elkülönítése az egyik széles körben alkalmazott megoldás a túlillesztés problémájára.

Következő lépésként azt ábrázoljuk, hogy a korpusz mely kifejezései szabták meg leginkább azt, hogy az adott törvényszöveg milyen fő közpolitikai kód alá tartozik! Az *xgboost* nevű csomag ezt a *fontossági mátrix* definiálásával könnyen lehetővé teszi. Első lépésként számítsuk ki ezt a mátrixot!

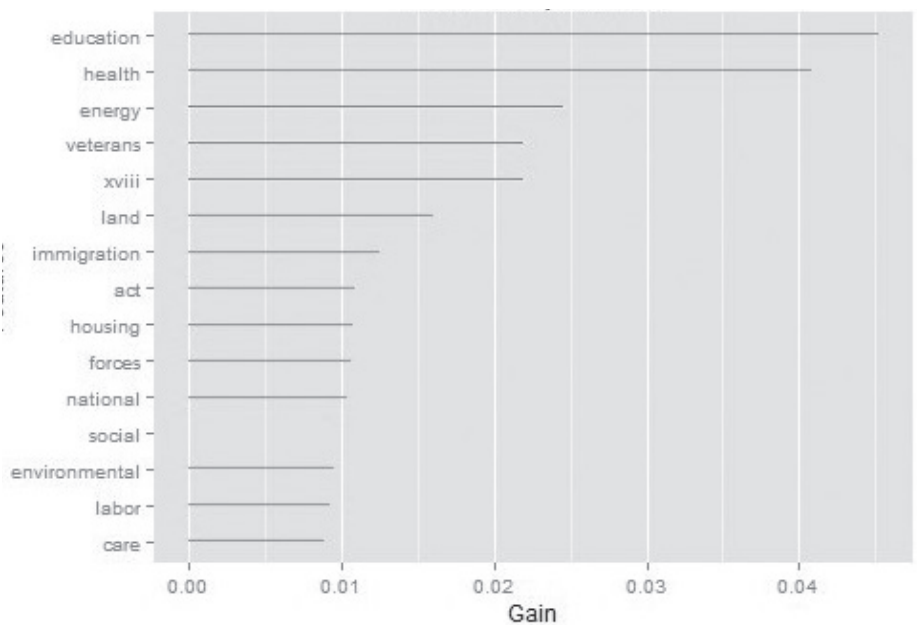
```
importance_matrix <- xgb.importance(colnames(train), model = model)
```

Második lépésként jelenítsük meg a 15 „legfontosabb" szót az alábbi paranccsal.

```
xgb.plot.importance(importance_matrix[1:15])
```

A parancs lefuttatásával létrejön a IV.1.5. ábra. Látható, hogy az olyan egyértelműen közpolitikai tartalmú szavak bizonyultak a „legfontosabbnak" a klasszifikáció során, mint az egészség, oktatás vagy a háborús veteránok.

IV.1.5. ábra – A korpusz klasszifikációjának „legfontosabb" szavai



Az eddigiekben áttekintett hat kutatási lépés során a felügyelt tanulási modellünk lefutott, és értékeltük annak teljesítményét a teszthalmaz értékein. A következő lépés egy politikatudományi kutatómunka során az lenne, hogy olyan törvénycímeket soroljunk be osztályokba, melyek nem találhatóak

meg sem a tanulási, sem a teszthalmaz értékei között. Ezt úgy tudjuk a legegyszerűbben elérni, ha a „szűz" elemeket hozzáadjuk a kezdeti korpuszhoz, majd a modellbecslést, illetve modellértékelést követően az új elemekre külön megbecsüljük a fő közpolitikai kódot.[4]

A gyakorlati alkalmazás egyes kérdései

Egyéb felügyelt tanulási módszerek

A példánkban alkalmazott döntési fák módszere jellemzően jó teljesítményt nyújt a többi felügyelt tanulási módszerrel összehasonlítva is (Liu, 2007: 67), de számos egyéb módszer is rendelkezésünkre áll. A választás során fontos, hogy az alkalmazni kívánt megoldásokat előzetesen áttanulmányozzuk, hiszen az algoritmusok lefuttatásához egyes esetekben számos bemeneti értéket kell megadnunk, és a rossz értékek választása elégtelen becsléshez vagy túlillesztéshez vezethet. Tikk (2007c: 111–134) alapján a következő főbb szövegosztályozási módszerek használatosak: döntési fák, a naiv Bayes-módszer, a legközelebbi szomszédokon alapuló osztályozók, a neurális hálózatok, valamint a különféle lineáris és nemlineáris regressziók.

Kiemelendő továbbá az ún. *osztályozó bizottság* megoldás (Tikk, 2007c: 133), mely többféle modell kimeneti eredményeit egyesíti a jobb tanulási teljesítmény érdekében. Gyakorlatilag arról van szó, hogy képessé válunk a különböző felügyelt tanulási megközelítések előnyeinek egyesítésére (Hastie et al., 2015: 605). A felügyelt tanulási modellek teljesítményének értékelésére a korábban bemutatott *pontosság* és *hibaszázalék* csak korlátozottan képes, sőt egyes esetekben (mint például kiegyensúlyozatlan csoportlétszámok esetén) egyenesen félrevezetőek lehetnek (Liu, 2007: 81).

A szövegkorpusz mérete

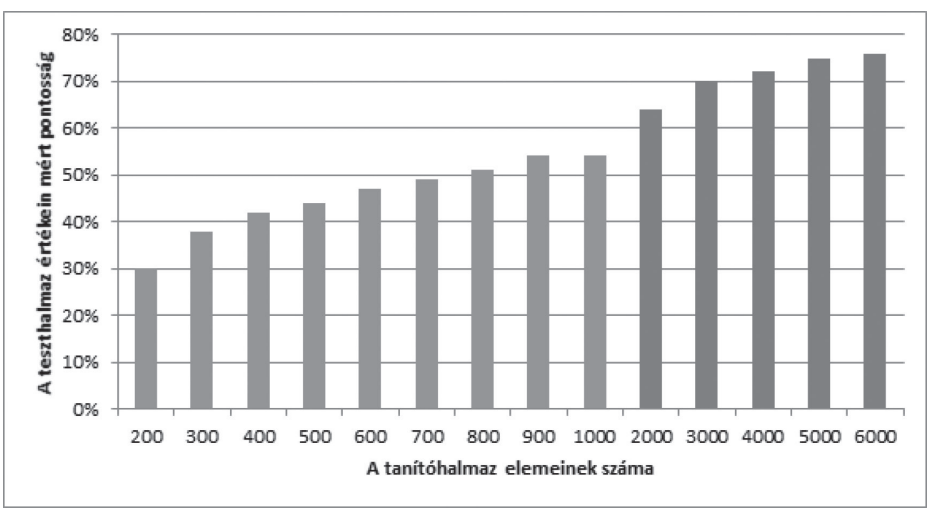
Fontos kérdés, hogy a szövegkorpusz milyen méretű mintájára lehet szükségünk annak érdekében, hogy minél jobb hatékonyságú klasszifikációt tudjunk végrehajtani. Ez egy nagyon fontos kérdés, hiszen a felügyelt tanulási módszerek feltételezik, hogy rendelkezünk legalább a megfigyelések egy részére osz-

[4] A könyv honlapján rendelkezésre áll a fejezet során taglalt forráskód kibővített verziója, melyben az ilyen szűz elemek csoportosítására is sor kerül.

tálytagsági értékekkel. A manuális (tehát emberek által történő) előzetes kódolás ráadásul igen költséges és időigényes (Hopkins – King, 2010: 229).

Elmondható, hogy az osztályozó megoldás pontossága nem nő egyenletesen az előzetesen bekódolt szövegek mennyiségével. Ez azt jelenti, hogy egy bizonyos szövegmennyiségen túl már nem érhető el jelentős javulás a felügyelt tanulási alapú módszer pontosságát illetően. Hopkins és King (2010) korpusza jóval ezer fölötti mennyiségű szöveget tartalmazott és ötszáz előzetesen csoportosított szövegen túl már nem tudtak jelentős javulást elérni az osztályozó mechanizmus pontosságát illetően. E probléma illusztrálása céljából lefuttattuk a korábbiakban bemutatott példamodellt különböző nagyságú tanítóhalmazokon. A IV.1.6. ábra jól mutatja, hogy a tanítóhalmaz elemszámának növekedésével párhuzamosan a teszthalmazon mért pontosság is fokozódik, azonban egyre csökkenő mértékben.

IV.1.6. ábra – A tanítóhalmaz elemeinek száma, illetve a teszthalmazon elért pontosság



Az osztályozási teljesítmény fokozása a korpusz átalakításával

A szövegkorpuszunk átalakításával, illetve a dokumentum-kifejezés mátrix megfelelő súlyozásával gyakran jobb teljesítményt érhetünk el (Feinerer et al., 2008: 10), és hasonlóan javulhat modellünk teljesítménye a szövegkorpusz szavainak szótövesítésével, illetve lemmatizálásával (ld. II.1. fejezet).

Míg a szótövesítés például az angol nyelvben eredményesen működik, a komplex magyar nyelv jelentős kihívás elé állítja a szövegelemzést végrehajtani kívánó

kutatót. Érdemes áttekinteni a *Snowball* projekt magyar szótövesítő megoldását vagy a különféle szóváltozatok mellett a szótöveket is tartalmazó szótárakat, mint amilyen például a *Hungarian National Corpus*.[5]

A korábbi példánál a szavak gyakoriságait szerepeltettük a modellépítésre használt dokumentum-kifejezés mátrixban, azonban a nyers gyakoriságok az eltérő méretű szövegek esetében természetesen problémát jelentenek. Ennek egyik oka, hogy a hosszú javaslatcímek szavainak nagyobb gyakorisága a címszöveg hosszával függ össze. A problémát megoldandó érdemes lehet az igen széles körben használt és igen eredményesen használható *IDF,* illetve *TF-IDF* reprezentációkat alkalmazni (Tikk, 2007a: 33–37). A *TF-IDF* megoldás gyakorlatilag egy súlyozási módszer, mely az egyes szavak „fontosságát" méri egy szövegkorpuszon belül, és gyakorta alkalmazzák információkinyerésre, illetve klasszifikációra (Robertson, 2014).

Szószedet

Magyar	Angol
Adatreprezentáció	Data representation
Ág (döntési fán)	Branch
Belső csúcs	Internal node (inner node, inode, branch node)
Dokumentum-kifejezés mátrix	Document-term matrix
Döntési fa	Decision tree
Felügyelt tanulás	Supervised learning
Fontossági mátrix	Importance matrix
Hibaszázalék	Error rate
Induktív osztályozás	Inductive classification
Induktív tanulás	Inductive learning
Inverz-dokumentumfrekvencia (IDF) reprezentáció	Inverse document frequency (IDF) representation

[5] Bővebben: http://snowball.tartarus.org/algorithms/hungarian/stemmer.html ill. http://corpus.nytud.hu/mnsz/index_hun.html

Magyar	Angol
Levélcsúcs (külső csúcs)	Letter node (external node, outer node, terminal node)
Osztályozó bizottság	Classifier committee, ensemble classifier
Pontosság	Precision
R csomagok	R-Packages
Statisztikai modellezés	Statistical modeling
Tanítóhalmaz	Training-set
Terminusfrekvencia és inverz-dokumentum-frekvencia (TF-IDF) reprezentáció	Term frequency and inverse document frequency (TDF-IDF) representation
Teszthalmaz	Testing set
Túlillesztés	Overfitting

Ellenőrző kérdések

- Milyen eredményességgel lenne képes a fejezetben felépített modell a 19. századi amerikai törvényjavaslatok közpolitikai tartalom szerinti besorolására?
- Miért indokolt a tanító- és a teszthalmaz véletlenszerűen történő szétválasztása? Milyen esetekben érdemes a véletlenszerű kiválasztástól eltérni?
- Egy kutató kizárólag a törvényjavaslatok szövege alapján szeretné megjósolni, hogy a köztársasági elnök az adott javaslatot aláírja-e, vagy sem. Megfelelő a felügyelt tanulás modellkerete a jelenség vizsgálatára? Lehetnek olyan tényezők, melyek nem derülnek ki a törvényszöveg alapján, és ha igen, melyek lehetnek azok?
- A magyar Országgyűlés képviselői felszólalásaiból milyen kategóriák hozhatóak létre, és azok a felszólalásszövegekből milyen teljesítménnyel jósolhatóak meg?

Ajánlott irodalom

A felügyelt tanulási módszerekben való mélyebb elmélyüléshez az olvasó figyelmébe ajánljuk Bing Liu (2007) könyvét, a technikák statisztikai és matematikai alapjainak elsajátításához pedig Trevor Hastie, Robert Tibshirani és Jerome Friedman (2005) könyvét érdemes áttanulmányozni. A módszerek számítógépes alkalmazásához további segítséget az interneten fellelhető számtalan ingyenes forrás mellett a tankönyv honlapján (qta.tk.mta.hu) is kaphatunk, ahol teljes forráskódok is rendelkezésre állnak a különféle kvantitatív szövegelemzési módszerek alkalmazásához.

Kubik Bálint György – Vancsó Anna

IV.2. CSOPORTOSÍTÁS: FELÜGYELET NÉLKÜLI TANULÁSI MÓDSZEREK

A fejezetben bevezetést nyújtunk a felügyelet nélküli gépi tanulás eljárásába egy gyakorlati politikatudományi kutatási probléma megoldásával. A fejezet vázát egy Python nyelven írt forráskód bemutatása adja, mely egy magyar nyelvű szövegkorpusz, a napimigrans.hu weboldal 168 cikkének előkészítését és elemzését hajtja végre. A programkód tisztázása mellett kitérünk a felügyelet nélküli tanulás alapfogalmainak definiálására, illetve bemutatjuk a szövegelőkészítést, illetve -elemzést segítő legfontosabb eljárásokat.

Szemben az előző fejezetben tárgyalt felügyelt tanulási módszerekkel, a felügyelet nélküli tanulást eredendően nem szövegek elemzésére fejlesztették ki: hagyományosan a kognitív tudományok (kognitív nyelvészet és idegtudományok), valamint az orvosi kutatások területén használják (Rojas 1996, Ghahramani, 2004). Újabban ugyanakkor egyre többet találkozni e megoldáscsoporttal a szövegbányászati technikák között is, például szinonimák, névelemek vagy éppen témacsoportok felismerésének tárgyalása kapcsán (Aggarwal – Zhai, 2012).

A felügyelet nélküli tanulást akkor alkalmazzuk, amikor nem rendelkezünk semmilyen a priori kategóriarendszerrel, címkékkel az adatok struktúráját vagy jellemzőit illetően, és a létrehozandó csoportok számáról sincsen előzetes tudásunk; tehát nincs olyan információnk, amiből tanulva a modellt fel lehetne építeni (Tikk, 2007b). Így célunk, hogy hagyjuk az adatokat magukért beszélni, és saját, szabad szemmel nem látható, de statisztikailag jól megragadható jellemzőik alapján kategorizáljuk őket (feltételezve, hogy nem csak „zaj" található a szövegekben). A felügyelet nélküli tanulási módszerek esetében a felügyelt technikákkal szemben nincs szükség tanítóhalmazra, tehát akármilyen szövegkorpuszra alkalmazható mindenféle előzetes manuális erőfeszítés nélkül (Aggarwal – Zhai, 2012: 5).

A strukturálást sokféle tulajdonság mentén végre lehet hajtani attól függően, hogy mi a kutatási kérdés, illetve milyen típusú eredményeket szeretnénk kapni. A kulcskérdés az, hogy mit határozunk meg a keresendő entitásként: mondatokat, szavakat, párbeszédeket stb. A szöveges dokumentumok elemzése

kapcsán a leggyakoribb a téma szerinti kategorizáció, amit a szakirodalom összegző technikának is nevez (mivel nagy méretű szövegkorpuszt vagyunk képesek jellemezni anélkül, hogy bármilyen háttér-információval rendelkeznénk annak tartalmával, szerkezetével kapcsolatban).

Az összegzést végző felügyelet nélküli tanulási algoritmus – a szövegek megfelelő formátumra történő alakítása után (szótövesítés, tiltólistás szavak stb.) – a dokumentumokat külön-külön jellemzi a szavak gyakorisága, a leggyakoribb szókapcsolatok, valamint a szavak távolsága alapján. A dokumentumok statisztikai jellemzőit a bennük megjelenő szavak mint legkisebb entitások alapján definiálja, majd ezen statisztikai jellemzők segítségével a dokumentumokat csoportokba/klaszterekbe rendezi. Az eljárás pontos módja minden esetben a választott algoritmus logikájától függ.

Ami minden ilyen eljárásban közös, az az adathalmaz statisztikai módszerekkel történő strukturálásának egy módja, a dimenziócsökkentés, avagy a látens dimenzió feltárása. Ennek különböző módszereivel – ld. főkomponens-elemzés, faktoranalízis, klaszterezés – a statisztika tudományága foglalkozik. A felügyelet nélküli tanulás szövegkorpuszra történő alkalmazása során ezek közül leggyakrabban a klaszteranalízist használjuk – a fejezet gyakorlati példája során is egy ilyen csoportosítási eljárást mutatunk be.

A klaszterezési eljárás

Egy szövegalapú adatbázis klaszterezésénél abból a hipotézisből indulunk ki, hogy a szóhasználatukban hasonló dokumentumok nagy valószínűséggel témájuk szerint is hasonlóak lesznek (Tikk 2007), a tartalmi hasonlóság pedig együtt jár bizonyos szavak, kifejezések használatával. A szövegkorpusz elemzésekor tehát az a cél, hogy a meglévő dokumentumok közötti strukturális elemek feltárásával minél homogénebb – tehát témában hasonló – klasztereket hozzunk létre úgy, hogy a különböző klaszterek minél jobban eltérjenek egymástól.

A klaszterezés módszerét szokás két csoportra bontani: a hierarchikus és a nem hierarchikus klaszterezésre. Jelen fejezetben a második, azaz a particionáló módszert vesszük górcső alá.[1] Ez az eljárás elsősorban akkor használatos, amikor nagyon sok adat áll rendelkezésünkre – ilyen a legtöbb szöveges adatbázis –, ami feltételezhetően egy számunkra nem látható struktúrával rendelkezik. Emellett megkülönböztetünk puha és szigorú klaszterezést is; az elsőt témamodellezésnek is nevezik, ahol valószínűségi modellt alkalmazva minden egyes dokumentumnak egy adott p valószínűsége van arra, hogy egy klaszterbe

[1] A statisztika magyar nyelvezetében ezt jellemzően nem hierarchikus klaszterezésnek nevezik, azonban ennek egyéb változatai is vannak.

kerüljön; ezzel szemben a szigorú klaszterezésnél minden dokumentum csak egy adott klaszterhez tartozhat, figyelmen kívül hagyva egy másik csoportba való kerülés valószínűségét.

Nézzük előzetes illusztrációként az eljárás egy gyakorlati politikatudományi alkalmazását! A közösségi média erősödésével óriási mennyiségű strukturálatlan adathalmaz áll a rendelkezésünkre,[2] amit a megfelelő módszerrel értelmezhetővé és a hétköznapi életben is felhasználhatóvá lehet alakítani. Ezt tűzték ki célul a melbourne-i Monash Egyetem kutatói, akik terrorcselekmények Twitteren történő megjelenését, felhasználók általi feldolgozását vizsgálták ún. *félig felügyelt tanulási módszerrel* (Cheong-Lee, 2011).[3]

A megközelítés lényege, hogy egy ismeretlen adathalmaz elemzésekor segítségül hívnak már címkézett, kategorizált adatokat – természetesen azonos témában és hasonló csatornán létrehozott adatokat –, melyek segíthetnek az újak értelmezésében és a gépi tanulás pontosításában, továbbfejlesztésében. A Twitteren fellelhető források elemzése azért is az egyik leggyakoribb téma az adatbányászat területén, mert a rendszer maga már sokféleképpen kategorizálja az adatokat – a helyre, időre, felhasználó tulajdonságaira vonatkozó információk mellett a kommunikáció módjának osztályozása is rendelkezésünkre áll (tweet, retweet, mention), míg a hashtag funkció használata a témák detektálását is segíti.

A kutatók az adatok elemzéséhez fel is használták ezt a segítséget, tehát az adatok begyűjtése után azok csoportosítása is megtörtént a különböző, előzőekben felsorolt szempontok szerint. A terrorcselekmény témakörben tehát a legfontosabb szempontok a releváns hashtagek megtalálása és klaszterezése, valamint a kommunikáció módja voltak. Ezek után véleményelemzést is végeztek, aminek során klaszterezték a bejegyzésekben megjelenő érzelemkifejezéseket.

Ehhez a folyamathoz egy saját szótárt hoztak létre már korábban elvégzett kutatások alapján,[4] mellyel egyben a nem felügyelt tanulási módszerrel történő véleményelemzés eredményeit is ellenőrizhették. Az elvégzett elemzés eredményeképpen tisztább képet kaptak arról, hogy milyen mintázatokat követhet a terrorcselekmények közösségi médiában történő feldolgozása. Az említett kutatás módszertana jóval komplexebb az itt bemutatandó gyakorlati problémánál, azonban jól mutatja a felügyelet nélküli tanulási módszer hasznosságát nagy mennyiségű adathalmaz strukturálásában.

[2] 2016-os statisztikák szerint több mint 1,5 milliárd aktív Facebook- és 320 millió aktív Twitter-felhasználó van (statista.com).

[3] Fontos megjegyezni, hogy csak szöveges tartalmat vizsgáltak, tehát a képek és linkek megjelenését nem vizsgálták külön.

[4] A 2001. szeptember 11-i terrortámadáshoz kapcsolódó cikkek, bejegyzések elemzésével ld. Clark (2009).

A gyakorlati példa

Gyakorlati példánkban a napimigrans.hu nevű bevándorlóellenes weboldal 168 cikkének szövegét elemezzük a felügyelet nélküli tanulási módszer segítségével.[5] A radikális szervezetek politikatudományi relevanciája nem igényel hosszabb indoklást: az Európa-szerte előtérbe kerülő szélsőséges pártok gyakran jelentős választói támogatást élveznek, és a fősodor tematizációjától eltérő módon szerkesztett, alternatív hírforrásokat tartanak fenn. Bár a napimigrans.hu weboldal nem köthető össze egyértelműen egyetlen hazai radikális szervezettel sem, a szélsőséges tartalmú információforrások elemzéséhez jól felhasználható felügyelet nélküli tanulási módszerekre példával szolgál. A magyar nyelvű szövegkorpusz elemzése kapcsán továbbá számos, sajátosan a magyar nyelvvel kapcsolatos gyakorlati problémára és annak megoldására tudunk kitérni.

A felügyelt tanulási módszerekről szóló fejezettel ellentétben most nem az R programozási nyelvet használjuk, hanem az ugyancsak nagy népszerűségnek örvendő Pythont (ennek 2.7-es verzióját). A forráskód sikeres futtatásához az olvasó figyelmébe ajánljuk az Anaconda[6] elnevezésű eszköztárt, mely számos, többek között tudományos számításokhoz használható Python kiegészítőt (ún. *library*t) tartalmaz, és munkánkat a könnyen használható, átlátható kódstrukturálást és látványos vizualizációt lehetővé tevő IPython fejlesztési környezettel segíti.

A probléma megoldása

A megoldandó politikatudományi probléma a következő: mintázatokat kell azonosítanunk a napimigrans.hu weboldal cikkei, pontosabban ennek szövegei között. Politikatudósként kíváncsiak vagyunk arra, hogy a médium híreit milyen témák dominálják, milyen kontextusban jelenik meg a célkeresztbe állított migránsok csoportja, illetve egyes országok és politikai vezetők milyen konnotációval tűnnek fel az egyes oldalakon.

Mivel a cikkek egyenkénti alapos átolvasása és kategorizálása meghaladhatja erőforrásainkat, a felügyelet nélküli tanulási módszercsoporthoz fordulunk a célból, hogy egy ún. *topik- (avagy téma-) modellt* alkossunk. Egy témamodellben a szövegkorpuszunk elemeinek szavaihoz valószínűségi értékeket rendelünk, ahol a témára leginkább jellemző kifejezések kapják a legnagyobb va-

[5] A tankönyv weboldalán (qta.tk.mta.hu) az alábbiakban részletezett programkód mellett hozzáférhető a cikkszövegek automatizált összegyűjtését lehetővé tevő Python program is.

[6] Az Anaconda letöltéséhez és telepítéséhez kövessük a következő weboldal utasításait: https://www.continuum.io/downloads.

lószínűséget. A témák és a klaszterek ebben az igen általános keretrendszerben analóg fogalmaknak tekinthetőek (Aggarwal – Zhai, 2012: 5).

A témák azonosítása céljára a III.3. fejezetben bemutatott K-közép klaszterezés módszerét alkalmazzuk, mely egy adattípustól függetlenül általánosan hasznosítható megoldás. A módszer ezáltal hasznosítható szöveges adatok elemzésére is, ugyanakkor Aggarwal és szerzőtársai (2012c: 79) óvatosságra intenek: a szöveges adat olyan tulajdonságai, mint a magas dimenzionalitás, a *szétszórtság* vagy az egyes szövegek méretének eltérése gyakran szükségessé teszi a szövegspecifikus algoritmusok alkalmazását. Gyakorlati példánkban ezzel együtt a K elemű klaszterezés technikájára hagyatkozunk, mivel – alább látni fogjuk – a módszer sikeresen azonosít egyes fontosabb témákat a napimigrans.hu cikkeinek szövegeiben.

A szövegkorpusz betöltése

Első lépésként töltsük le a napimigrans.hu vizsgálandó cikkeit a tankönyv GitHub oldaláról. A letöltéshez szükségünk lesz a Pandas nevű Python könyvtárra, így azt első lépésként importáljuk be, és használjuk az elterjedt pd rövidítést a Pandas jelölésére a továbbiakban.

```
import pandas as pd
```

Az ún. CSV-formátumú fájl (vesszővel tagolt adatfájl) értékei tabulátorkarakterekkel vannak elszeparálva, és az adatsor elemei ún. UTF-8 kódolásúak. Az UTF-8 kódolás tartalmazza a magyar nyelv ékezetes karaktereit, tehát indokolt a használata magyar nyelvű korpuszok kezelésénél. Nevezzük el továbbá a tábla két változóját a munka megkönnyítése érdekében. Az adatbázis betöltését követően jelenítsük meg az adatbázist.

```
url = „http://qta.tk.mta.hu/uploads/files/napimigrans_corpus.csv”
corpus_df = pd.read_csv(url, sep=”\t”, header=None, encoding=”utf-8”)
corpus_df.columns = [„url”, „text”]
corpus_df.head(10)
```

A fenti parancsok lefuttatása után az alábbi adatbázist fogjuk látni.

IV.2.1. ábra – Az adatbázis képe (részlet)

```
corpus_df
```

	url	text
0	http://napimigrans.com/egykori-szexrabszolgakb...	2016-02-11 Több száz, egykor az ISIS fogságába...
1	http://napimigrans.com/egy-muszlim-rab-megprob...	2016-02-12 Washington – A 25 éves muszlim fogv...
2	http://napimigrans.com/magyarorszag-fele-szori...	2016-02-21 Horvátország nem engedi tovább a Sz...
3	http://napimigrans.com/vajon-mikor-hal-meg-az-...	2016-02-24 A népvándorlás nem más, mint egy ol...
4	http://napimigrans.com/a-parizsi-gyilkosok-sem...	2016-02-16 MILOS ZEMAN A KORÁNRA HIVATKOZIK – ...
5	http://napimigrans.com/verfurdot-rendezett-az-...	2016-02-09 Több mint 300 embert végeztek ki az...
6	http://napimigrans.com/video-oriasi-tuntetes-a...	2016-02-21 Minnesota – Úgy látszik az USA-ban ...
7	http://napimigrans.com/video-220-milliot-kolte...	2016-02-24 Még 2014-ben szó volt róla, hogy né...
8	http://napimigrans.com/hogyan-romboljuk-le-a-h...	2016-02-21 Az értékek viszonylagosak. Más kult...
9	http://napimigrans.com/europa-megeroszakolasa-...	2016-02-18 Meghökkentő címlappal jelent meg a ...
10	http://napimigrans.com/robbanooveket-es-robban...	2016-02-10 A biztonságiak 34 embert állítottak...
11	http://napimigrans.com/titkosszolgalati-szaker...	2016-02-16 A kérdés az vajon kik állnak a csap...
12	http://napimigrans.com/video-megvertek-a-foldo...	2016-02-17 London – A rendőrség közzé tette az...

A szövegkorpusz előkészítése

A következő lépésben alkalmazzuk a korábbiakban már említett eljárásokat, hogy elemezhetővé tegyük a cikkek szövegeit. A szöveg előkészítésének célja nem más, mint az elemzési probléma leegyszerűsítése oly módon, hogy az igen hasonló tulajdonságokkal rendelkező szavakat azonosként kezeljük, továbbá a számunkra szükségtelen kifejezésektől megváljunk (Lucas – Nielsen – Roberts – Stewart – Storer – Tingley, 2014).

Végezzük tehát el szövegkorpuszunk elemzésre való előkészítését! Első lépésként importáljuk be az átalakításhoz szükséges Python könyvtárakat, melyek közül mindhárom az ún. *Natural Language Toolkit* (röviden NLTK) platform[7] ernyője alá tartozik. Az NLTK az egyik legnépszerűbb megoldás a különféle szövegek feldolgozásához, mivel használatát egyszerű utasítások és alapos online útmutatók segítik.

```
import nltk
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem.snowball import SnowballStemmer
import unicodedata
```

Definiáljuk a szótövesítést végző objektumot *(stemmer)* a magyar nyelvre, továbbá töltsük le az NLTK-hoz kapcsolódó, tiltólistás szavakat tartalmazó fájlokat.

[7] További információ az NLTK honlapján: http://www.nltk.org/

```
stemmer = SnowballStemmer(„hungarian")
nltk.download(„stopwords")
nltk.download(,punkt')
```

Munkánk megkönnyítése érdekében hozzunk létre egy függvényt, mely a számára értékként (argumentumként) megadott szöveget kisbetűsíti, eltávolítja az írásjeleket, kiszűri a tiltólistás szavakat, majd záró lépésként elvégzi a szótövesítést is. Itt már látszik, hogy UTF-8 kódolású szövegkorpuszunk problémát okoz a *tokenizálás* ezen műveletében. A probléma pontos technikai leírása meghaladja a jelen fejezet kereteit, ezért vegyük adottnak a következő függvényt:

```
def strip_punctuation(text):
  punctutation_cats = set([,Pc', ,Pd', ,Ps', ,Pe', ,Pi', ,Pf', ,Po'])
  return ,'.join(x for x in text
        if unicodedata.category(x) not in punctutation_cats)
```

A fenti függvény röviden összefoglalva összeveti az általa kapott szöveglánc elemeit az írásjelekhez tartozó Unicode-kategóriákkal. Ez a függvény végzi tehát az írásjelek eltávolítását.

Az írásjeleket eltávolító függvény birtokában már megalkothatjuk az egyes cikkek szövegét tokenizáló függvényt, úgy, hogy kimeneti értéke az előkészített szöveg szavainak listája legyen.

```
def tokenize_article(text):
  text = text.lower()
  text = strip_punctuation(text)
  tokenized_text = word_tokenize(text)
  tokenized_text = [word for word in tokenized_text
        if word not in stopwords.words(,hungarian')]
  tokenized_text = map(stemmer.stem,tokenized_text)
  return tokenized_text
```

Próbaképp alkalmazzuk az előbb létrehozott függvényt korpuszunk első mondatára, illetve (mivel az előbb megadott függvény szavak listáját adja vissza) vonjuk össze a lista elemeit szóközzel elválasztott karakterlánccá.

```
print ' '.join(tokenize_article(corpus_df[„text"][0]))
```

Az átalakított mondat első néhány szava a következő alakot ölti:

> 2016 02 11 száz egy isis fogság szexrabszolg szenvedő nő csatlakozot ir egység nap hölgy nevez

Az eredeti (átalakítás előtti) mondat még a következőképp festett:

> 2016-02-11 Több száz, egykor az ISIS fogságában, szexrabszolgaként szenvedő nő csatlakozott ahhoz az iraki egységhez, mely a Nap Hölgyeinek nevezi magát

A fent részletesen tárgyalt előkészítő műveletek mellett fontos külön megemlíteni a szótövesítés lépését. A szótövesítés a hasonló jelentéstartalmú szavak azonos szótőhöz rendelésével csökkenti a szöveg egyedi kifejezéseinek számát, így a komplexitást is (Grimmer – Stewart, 2013). Gyakorlati példánkban az ún. Snowball[8] eljárást alkalmazzuk, mely az angolon túl számos egyéb (akár magyar) nyelvű szöveg szótövesítésére is megoldást kínál (Porter, 2011). Fontos ugyanakkor megemlíteni, hogy az agglutináló nyelvekben, ahol az egyes szavak igen sok különböző alakban előfordulhatnak, a lemmatizálás hasznosabb lehet a szótövezésnél (Lucas et al., 2014). Szövegkorpuszunk első cikkének szavai jól láthatóan kisbetűsítésre kerültek, az ékezeteket és a tiltólistás szavakat kiszűrtük, továbbá az egyes kifejezéseket szótövesítettük. A Snowball algoritmus jól láthatóan eltávolította az egyes szavak végződéseit (ragjait), így kevesebb egyedi szóval kell dolgoznunk.

A szöveges adatok elemezhető formába hozása

A felügyelt tanulási fejezet gyakorlati példájához hasonlóan a szövegkorpusz előkészítése még nem elegendő az elemzés elvégzéséhez, hiszen ahhoz először valamilyen módon létre kell hoznunk a szövegek *adatreprezentációját*. Az adatreprezentációt jelen esetben is a dokumentum-kifejezés mátrix jelenti, mely az egyes dokumentumokhoz hozzárendeli a különféle szavak előfordulásának mérőszámait (legegyszerűbb esetben az egyes kifejezések gyakoriságait).

Az adatelemzés műveletét jelentősen leegyszerűsítő és számos gépi tanulási módszert egyesítő *scikit-learn* nevű Python könyvtár[9] jelentős segítséget nyújt nekünk: a *CountVectorizer* nevű *osztály* egy lépésben elvégzi a tokenizációt és az egyes szavak előfordulási gyakoriságainak kinyerését, és ehhez mindössze a korpuszunk szövegeinek listájára és az előbbiekben kialakított segédfüggvényre van szükség.

[8] További információ a Snowball-projekt honlapján: http://snowball.tartarus.org/.

[9] További információ és számos ingyenes útmutató a Python könyvtár honlapján: http://scikit-learn.org/stable/.

```
from sklearn.feature_extraction.text import CountVectorizer
vectorizer = CountVectorizer(tokenizer=tokenize_article)
document_term_matrix =vectorizer.fit_transform(corpus_df[[„text"]])
```

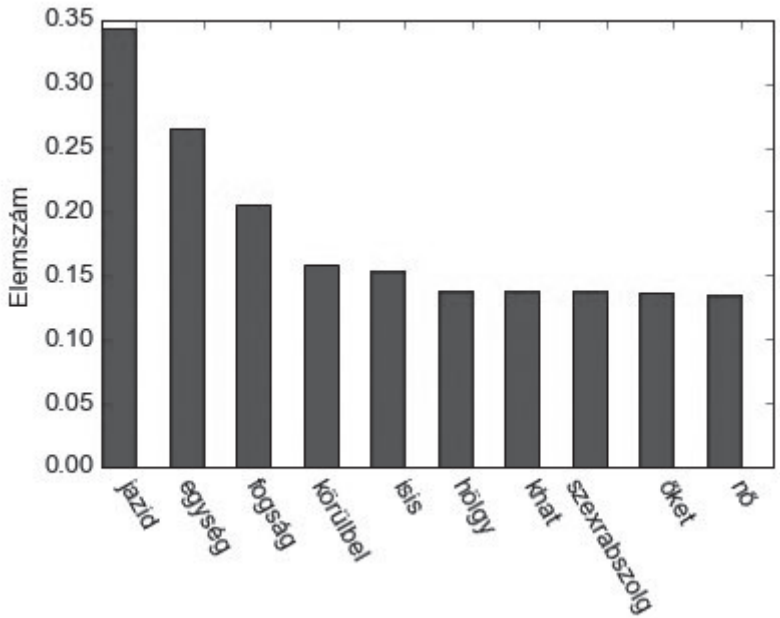
Az előbbi műveletekkel meg is kaptuk a dokumentum-kifejezés mátrixunkat, azonban az előző fejezet szógyakoriságokat tartalmazó mátrixával szemben most próbálkozzunk meg a korábbiakban már röviden megemlített szógyakorisági-inverz szógyakorisági (*term frequency-inverse term frequency, TF-IDF*) súlyozással is. A TF-IDF súlyozási eljárás egy statisztikai módszer az egyes (akár több szóból álló) kifejezések „fontosságának" mérésére a szövegkorpusz egyedi szövegeire vetítve. A szövegkorpusz szövegeiben egyenletesen gyakori kifejezéshez kisebb súly, míg a néhány szövegben gyakorta előforduló elemekhez nagyobb súly tartozik (Ali – Wang – Haddad, 2015). A scikit-learn a *TfidfTransformer* nevű osztály formájában a TF-IDF súlyozáshoz is egy egyszerű megoldást kínál:

```
from sklearn.feature_extraction.text import TfidfTransformer
transformer = TfidfTransformer()
tfidf_matrix = transformer.fit_transform(document_term_matrix)
```

A TF-IDF súlyozás demonstrálása érdekében nézzük meg, hogy szövegkorpuszunk első mondatában milyen szavak kapták a legmagasabb súlyt, tehát mely kifejezések számítanak a „legfontosabbnak" az algoritmus alapján.[10] Korábban már láttuk e cikk első mondatát, melyből következtethetünk a cikk témájára, de a cikk címe alapján már igazán egyértelművé válhat számunkra: *Egykori szexrabszolgákból összeállt női egység harcol az ISIS ellen Irakban*. Mint a cikkszövegben a tíz legnagyobb súlyt kapott szót mutató IV.2.2. ábrán látható, a magasabb súlyt kapó szavak alapvetően jól megragadják a cikk témáját: az ISIS magas értéket, tehát magas súlyt kapott, ahogy a nő és hölgy szavak, továbbá a fogsággal, harccal, védekezéssel kapcsolatos kifejezések is igen fontosnak számítanak. Elmondhatjuk tehát, hogy a TF-IDF súlyozás intuíciójával összhangban sikerült beazonosítani azokat a szavakat, melyek az első cikket leginkább megkülönböztetik a többi cikktől. Megjegyzendő továbbá, hogy a Snowball szótövesítő technika korlátairól ad tanúbizonyságot az a tény, hogy a módszer nem vonta össze a nő és a nők szavakat.

[10] A legmagasabb TF-IDF súlyt kapó kifejezések kinyerésének pontos módja megtalálható a fejezethez tartozó forráskódban, mely ugyancsak elérhető a tankönyv honlapján.

IV.2.2. ábra – A korpusz első cikkének legmagasabb TF-IDF súllyal rendelkező szavai



Rendelkezésünkre áll tehát a dokumentum-kifejezés mátrix, mely az egyes szavak gyakoriságait TF-IDF súlyozott formában tartalmazza. Ez már lehetővé teszi a felügyelet nélküli tanulási módszer alkalmazását, hiszen kaptunk egy mérőszámot arra nézve, hogy az egyes szövegeket milyen szavak teszik „egyedivé" a teljes korpuszon belül. Ennek megfelelően elkezdhetjük a „hasonlóan egyedi" szövegek klaszterekbe történő csoportosítását.

A szövegek hasonlóságának mérése: a koszinuszhasonlóság

Az elemzéshez szükség lesz az egyes szövegek közötti hasonlóság mérésére szolgáló mérőszámra. Egy ilyen, gyakran használt indikátor a *koszinuszhasonlóság* (Nenkova – McKeown: 60). A koszinuszhasonlóság olyan távolságértéket ad a kutató kezébe, melyet például az agglomeratív és hierarchikus klaszterezési módszerek könnyedén értelmeznek és kezelnek (Aggarwal – Zhai , 2012c: 90).

Számítsuk ki a koszinuszhasonlóságot a szövegkorpuszunk összes cikkszövegére! A *sklearn* Python könyvtár most is a segítségünkre lesz a *cosine_similarity* függvénnyel, mely mátrixformában megadja számunkra minden szöveg

minden más szöveggel szemben mért koszinuszhasonlóságát egy 0 és 1 közötti számmal (az egy a tökéletes megfelelést, a nulla a hasonlóság teljes hiányát jelöli).

```
from sklearn.metrics.pairwise import cosine_similarity
cos_sim_matrix = cosine_similarity(tfidf_matrix)
cos_sim_matrix
```

Az előbbi sorok lefuttatását követően a IV.2.3. ábrán látható listát látjuk.

IV.2.3. ábra – A szövegkorpusz koszinuszhasonlóságai

```
cos_sim_matrix

: array([[ 1.        ,  0.03794233,  0.00724466, ...,  0.0228782 ,
           0.02988523,  0.0152713 ],
         [ 0.03794233,  1.        ,  0.0140241 , ...,  0.00296434,
           0.02362931,  0.00465752],
         [ 0.00724466,  0.0140241 ,  1.        , ...,  0.00991861,
           0.01556769,  0.02255349],
         ...,
         [ 0.0228782 ,  0.00296434,  0.00991861, ...,  1.        ,
           0.00802531,  0.00257985],
         [ 0.02988523,  0.02362931,  0.01556769, ...,  0.00802531,
           1.        ,  0.02858622],
         [ 0.0152713 ,  0.00465752,  0.02255349, ...,  0.00257985,
           0.02858622,  1.        ]])
```

Az ábra a koszinuszhasonlósági mátrix egy részlete, hiszen a teljes mátrix túl nagy (168*168 elemű) ahhoz, hogy megjelenítsük. A mátrix megadja szövegkorpuszunk minden cikkpárja közötti koszinuszhasonlóság mértékét. A főátlóban (ahol a sor és az oszlop száma megegyezik) egyeseket látunk, mely azt jelenti, hogy a korpusz egyes szövegei önmaguknak tökéletesen megfelelnek. Az első cikk (a mátrix első sora) és a második cikkszöveg közötti hasonlóság mértéke 0.0379 (a mátrix második oszlopa). Vizsgáljuk meg, hogy az első cikk melyik másik cikkhez hasonlít a leginkább (önmagát nem számítva). Emlékeztetőképp az első cikk első néhány szava a következő:

> 2016-02-11 Több száz, egykor az ISIS fogságában, szexrabszolgaként szenvedő nő csatlakozott ahhoz az iraki egységhez, mely a Nap Hölgyeinek nevezi magát

Nyerjük ki a koszinuszhasonlósági mátrixból az első cikkhez leginkább hasonló szöveg sorszámát, a hasonlóság mértékét, illetve írassuk ki magát a cikkszöveget is! Ehhez importáljuk be a Python matematikai számítások elvégzésére

használt *numpy*[11] nevű könyvtárát, mellyel kezelhető a *numpy.ndarray* típusú koszinuszhasonlósági mátrixunk.

```
import numpy as np
print np.argmax(cos_sim_matrix[0,1:])
print cos_sim_matrix[0,1:][50]
print corpus_df[„text”][51]
```

E sorok lefuttatását követően megjelenő eredmény első sorából megtudjuk, hogy a hasonlósági mátrix első sorában (tehát az első cikk és az összes másik szöveg közötti hasonlósági érték esetében) az 50. elemnél a legnagyobb a koszinuszhasonlósági érték (0,11518). Megjelent továbbá maga a cikkszöveg is, melynek első mondata a következő:

> 2016-02-12 Néhány élelmiszer ára rendkívüli módon megemelkedett az utóbbi hetekben Deir Ez-Zour több, az Iszlám Állam által körbezárt és ostrom alatt tartott kerületének piacain.

A két cikkszöveg közötti viszonylag magas hasonlóság logikus, hiszen mindkét cikk az Iszlám Állammal és a szíriai–iraki háborúval kapcsolatos, ugyanakkor a 0,1 körüli érték emlékeztet minket a két cikk markáns különbségeire: az első cikk szövegében magas TF-IDF értéket kapó jazidi vagy nő szavak nem jelennek meg e másik cikkszövegben.

A felügyelet nélküli tanulási módszer alkalmazása

Az így előkészített korpuszt a K elemű klaszterezéssel mint felügyelet nélküli tanulási módszerrel elemezzük. A klaszterezésről bővebben szóltunk a III.3. fejezetben, így e helyt emlékeztetésképp csak annyit említünk, hogy az eljárás során egy megadott mennyiségű („K” darab) klaszterbe rendezzük megfigyeléseinket (jelen esetben a napimigrans.hu egyes cikkeit) olyan módon, hogy az egyes elemek (cikkszövegek) csak egy klaszterbe tartozzanak. A klaszterezésnél használatos távolságdefiníciót jelen esetben az egyes szövegek szavainak TF-IDF súlyozott mennyiségei alapján számolt euklideszi távolságok adják. A TF-IDF súlyozással megkaptuk, hogy az egyes cikkszövegeket mely szavak különböztetik meg leginkább az összes többi szövegtől, a klaszterezés során pedig a „hasonlóan egyedi” cikkszövegeket vonjuk össze.

[11] További információ a *numpy* könyvtár honlapján található: http://www.numpy.org.

A K elemű klaszterezést a *sklearn* Python library által biztosított *KMeans* osztállyal hajtsuk végre. Legyen 5 klaszterünk, és rögzítsük a *random_state* értékét 1-re, hogy a továbbiakban bemutatott eredmények pontosan megismételhetőek legyenek:

```
from sklearn.cluster import KMeans
km = KMeans(n_clusters=5, random_state=1)
km.fit(tfidf_matrix)
```

A fenti kódrészlet lefuttatásával végre is hajtottuk a K-közép klaszterezést, és az illesztett modell megtalálható a *model* nevű változóban. Egyszerűen kinyerhetőek a klasztertagságok is a szövegkorpusz összes cikkére, amit rögzítsünk az eredeti adatbázisban!

```
memberships = list(km.labels_)
corpus_df[„cluster_membership”] = memberships
corpus_df.head(10)
```

Ezt követően vessünk egy pillantást az eredménytáblára (IV.2.4. ábra).

IV.2.4. ábra – A klasztertagságokkal kiegészített adatbázis

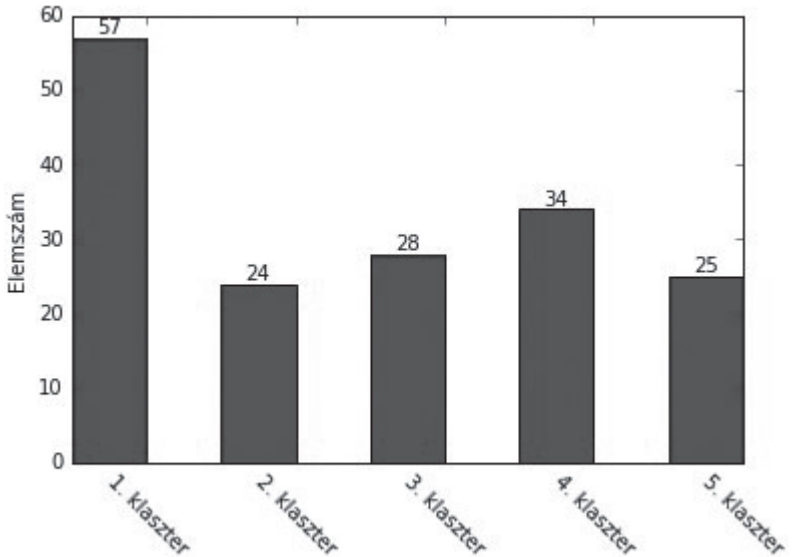
```
corpus_df
```

23]:

	url	text	cluster_membership
0	http://napimigrans.com/egykori-szexrabszolgakb...	2016-02-11 Több száz, egykor az ISIS fogságába...	3
1	http://napimigrans.com/egy-muszlim-rab-megprob...	2016-02-12 Washington – A 25 éves muszlim fogv...	0
2	http://napimigrans.com/magyarorszag-fele-szori...	2016-02-21 Horvátország nem engedi tovább a Sz...	0
3	http://napimigrans.com/vajon-mikor-hal-meg-az-...	2016-02-24 A népvándorlás nem más, mint egy ol...	0
4	http://napimigrans.com/a-parizsi-gyilkosok-sem...	2016-02-16 MILOS ZEMAN A KORÁNRA HIVATKOZIK – ...	0
5	http://napimigrans.com/verfurdot-rendezett-az-...	2016-02-09 Több mint 300 embert végeztek ki az...	1
6	http://napimigrans.com/video-oriasi-tuntetes-a...	2016-02-21 Minnesota – Úgy látszik az USA-ban ...	4
7	http://napimigrans.com/video-220-milliot-kolte...	2016-02-24 Még 2014-ben szó volt róla, hogy né...	1
8	http://napimigrans.com/hogyan-romboljuk-le-a-h...	2016-02-21 Az értékek viszonylagosak. Más kult...	3
9	http://napimigrans.com/europa-megeroszakolasa-...	2016-02-18 Meghökkentő címlappal jelent meg a ...	0
10	http://napimigrans.com/robbanooveket-es-robban...	2016-02-10 A biztonságiak 34 embert állítottak...	3
11	http://napimigrans.com/titkosszolgalati-szaker...	2016-02-16 A kérdés az vajon kik állnak a csap...	4
12	http://napimigrans.com/video-megvertek-a-foldo...	2016-02-17 London – A rendőrség közzé tette az...	4

Eredményeinkből látható a kialakult klaszterek elemszáma. Az 5 klaszter közül az első (a Python számozásában nulladik) csoport 57 cikkszöveget sűrít össze, mely mintegy kétszeresen meghaladja a többi, 20-30 közötti darab cikket egyesítő klaszterek elemszámát (ld. IV.2.5. ábra).

IV.2.5. ábra – Az egyes klaszterek elemszámai



Vizsgáljuk meg az egyes klasztereket olyan módon, hogy kinyerjük az adott klaszterre leginkább jellemző 30 kifejezést, és tegyünk kísérletet az egyes klaszterek témáinak azonosítására! Ezt a *klaszterközéppontok* vizsgálatával tehetjük meg. Első lépésként gyűjtsük ki a mátrix egyes oszlopaihoz tartozó szavak listáját *(feature names),* a rangsorolt klaszterközéppontokat, majd írassuk ki az adott klaszter középpontját leginkább meghatározó 30 kifejezést.

```
terms = vectorizer.get_feature_names()
order_centroids = km.cluster_centers_.argsort()[:,::-1]
for i in range(0,5):
  term_list = []
  for ind in order_centroids[i,:30]:
    term_list.append(terms[ind])
  print „„”.join(term_list)+”\n”
```

A kinyert szavak alapján próbáljuk meg értelmezni – azaz felcímkézni – az egyes klasztereket! Kezdjük a legnagyobb tagsággal rendelkező első klaszterrel. Itt a program a következő 30 szót azonosította:

európ, határ, is, migráns, ország, szerb, merkel, eu, görög, bevándorló, macedón, menekült, iszla, ha, ném, kancellár, görögország, 2016, nat, balkán, németország, unió, 02, ez, ember, magyar, uniós, áll, migrációs, osztra

A szavak arra engednek következtetni, hogy az első klaszter cikkei az európai migrációval, az Európai Unióval, Angela Merkel német kancellárral, illetve a migránsok útvonalán elhelyezkedő, illetve célállomását adó országokkal foglalkoznak.

A második, 26 cikket tartalmazó csoport legfontosabb kifejezései a következők:

tör, iszla, szír, áll, is, szíri, hadsereg, város, kur, katon, al, öngyilkos, törökország, szerd, eln, ész, terrortámadás, svéd, férf, merénylet, erdog, egység, rész, forrás, orosz, tovább, támadás, iszlamist, mondt, európ

A szavak áttekintése arra enged következtetni, hogy a vizsgált klaszter a közelkeleti eseményekkel foglalkozó szövegeket tartalmaz: szó esik Törökországról, Szíriáról, a szír polgárháborúról, merényletekről, továbbá az iszlámról és az iszlamistákról.

A harmadik klaszter 22 cikkének legfontosabb szavai a következőképp festenek:

bíróság, év, vádlott, ügy, is, rendőr, szeged, bűncselekmény, migráns, bűntett, szám, ügyészség, magyarország, férf, határsértő, ké, vádlot, éves, el, határzár, iszla, tegnap, mti, megrongálás, elkövető, fogt, lány, mansour, ország, 16

A kifejezések egy, a magyar határral, illetve az ott határsértést elkövető bevándorlókkal foglalkozó cikkcsoportra engednek következtetni.

A negyedik csoport szövegeiben azonosított 30 legfontosabb szó a következő:

férf, rendőrség, éves, is, ember, gyer, lány, ir, bűncselekmény, őket, részlet, tovább, kisfiú, szexuális, 2016, 02, nyom, eset, 10, menekült, törvény, év, ké, allah, egy, miat, muszl, hár, őrizet, az

Az azonosított szavak arra engednek következtetni, hogy a negyedik klaszterbe tartozó napimigrans.hu cikkek a migránsok által elkövetett, kiemelten a szexuális tartalmú bűncselekményekkel, illetve azok szankcióival foglalkoznak.

Az ötödik klaszter 30 leglényegesebb kifejezése a következő:

016, 02, 14, 22, közzétett, un, 21, február, 24, par, tv, ben, facebo, látsz, vaj, videó, tüntetés, idiót, gyerek, 16, megin, français, látható, merkel, nyom, bevándorló, of, métr, 2015, európ

Úgy tűnik, hogy a klaszter cikkeinek részletesebb vizsgálata szükséges, hiszen a szavak igen vegyes témákra engednek következtetni: feltűnik Angela Merkel neve, a français szó feltételezhetően a franciaországi terrorista merényletekre utalhatnak *(je suis français),* a többi kifejezés funkciója azonban nem egyértelmű.

Érvényességvizsgálat egy alternatív módszerrel

Eredményeink érvényességének megállapítása érdekében vessük össze a korábbiakban végrehajtott K-közép klaszterezés eredményeit a könyv III.3. fejezetében már említett *látens Dirichlet allokáció* (LDA) módszerével. Röviden összefoglalva az LDA valószínűségi modellje egy szövegkorpusz dokumentumait bizonyos témák keverékeként reprezentálja, és a kutató által meghatározott számú témát a témamodell révén maga az LDA határozza meg.

Az LDA művelete is végrehajtható a sklearn könyvtárral. Az LDA-t az eredeti, tehát a TF-IDF súlyozást megelőzően kapott dokumentum-kifejezés mátrixra illesztjük. A K-közép módszer segítségével 5 klasztert számítottunk, ezért azonosítsunk az LDA technikával is 5 témát.

```
from sklearn.decomposition import LatentDirichletAllocation
lda = LatentDirichletAllocation(n_topics = 5, random_state = 1)
lda.fit(document_term_matrix)
```

A modellbecslés lefutását követően a modellobjektumból kinyerhető az egyes témák kifejezéseloszlása a következő parancssal:

```
lda.components_
```

A klaszterezés műveletéhez hasonlóan írassuk ki az egyes témák 30 legfontosabb szavát!

```
for topic_idx, topic in enumerate(lda.components_):
  print „Topic #%d:" % (topic_idx + 1)
  print „ „.join([vectorizer.get_feature_names()[i]
        for i in topic.argsort()[:-30 - 1:-1]])
```

Az LDA által azonosított első téma legnagyobb súllyal bíró szavai a következők:

> is, határ, európ, migráns, ország, szerb, bevándorló, nyom, unió, ha, rendőr, menekült, merkel, fog, magyar, vég, útvonal, magyarország, mti, kerítés, németország, ót, törökország, illegális, macedón, határsértő, év, ember, február, tör

Úgy tűnik, hogy az első témába tartozó cikkek elsősorban a migráció európai útvonalával és az általa érintett európai országokkal foglalkoznak. Az első kifejezés valószínűsíthetőleg az ISIS szótövezett alakja (az „is" szó a tiltólistás szavak közé sorolandó, tehát a tokenizáció során eltávolítható), a csoport cikkei így valószínűleg párhuzamot vonnak a terrorszervezet és a migráció között. Első ránézésre a téma cikkei jelentős átfedést mutatnak a klaszterezés során kapott első számú csoport cikkeivel.

A kettes számú téma domináns szavai a következők:

> the, ahluwal, new, yor, of, an, gyilkosság, sykes, légitársaság, is, was, nélkül, at, nő, cutler, felvétel, utc, outs, punch, őket, university, közzétett, stop, kisfiú, fehér, fash, divatvilág, szikh, turbán, cnnnek

A téma szavai láthatóan Waris Ahluwalia, a szikh divattervező esetével kapcsolatosak. Ahluwaliát 2016 elején egy légitársaság nem engedte fel a gépére, mivel a divattervező a reptéri ellenőrzésnél nem volt hajlandó levenni a turbánját. A többi kifejezés önmagában nehezen értelmezhető.

Az LDA által becsült harmadik téma legfontosabb szavai a következők:

> kultúra, egyenlő, kultú, kultúr, valamenny, calais, is, muzulm, jelenleg, migráns, miat, feleség, őket, rendőrség, ezer, ha, bíráln, ember, hely, érték, vallás, gyerek, euró, tesz, kellen, dzsungel, nyugateuróp, 20160213, első, szervez

A kifejezésekből arra következtethetünk, hogy a témába tartozó cikkek a muszlim kultúrával és vallással kapcsolatosak. Előkerül továbbá Nyugat-Európa és a rendőrségi fellépés is (feltételezhetően a migránsokkal szembeni akciókról lehet szó).

A negyedik téma legfontosabb kifejezései a következők:

> is, iszla, hadsereg, város, férf, áll, ember, katon, tör, éves, szíri, európ, rész, terrorist, ország, tartomány, nap, ellen, forrás, év, egység, terrorista, nyom, rendőrség, szír, ha, erő, őket, mag, kur

A kategória jól láthatóan hasonlít a K-közép klaszterezés során kapott kettes számú klaszterre. A téma cikkeiben az ISIS-ról, a közel-keleti fegyveres konfliktusról (hangsúlyosan a szíriai polgárháborúról), illetve a kurdokról esik szó.

Az LDA témái közül az ötödik téma esetében a következő szavak dominálnak:

un, par, vádlott, français, szeged, franc, femm, à, közzétett, bíróság, ausztrál, racism, 20160224, február, segély, százale, hét, le, 2016, mun, antiblanc, dans, pas, ét, la, sos, en, dun, ké, 20160218

A számos francia kifejezés kapcsán a franciaországi terrormerényletek juthatnak eszünkbe, azonban Szeged város neve vagy a segély szó egy igen vegyes témára engednek következtetni.

Azonosítsuk mindezek után az egyes szövegek leginkább domináns témáit! Ehhez először vizsgáljuk meg, hogy miképp kapjuk meg a témák súlyát az egyes szövegekre nézve! Futtassuk le a következő műveletet!

```
lda_doc_topics = lda.fit_transform(document_term_matrix)
print lda_doc_topics
```

A korábban létrejött „lda" objektum felhasználásával transzformáljuk a dokumentum-kifejezés mátrixunkat (amin az LDA-becslést is végeztük). A transzformáció eredménye megjelenik előttünk: egy tömböt látunk, mely a szövegkorpusz mindegyik dokumentumára nézve megadja, hogy milyen súlyt kapott az öt, az LDA művelete által azonosított téma (a nagyobb értékek természetesen magasabb súlyt kapnak). Nincs más teendőnk a domináns téma azonosításához, mint hogy soronként (cikkenként) azonosítsuk a maximális értéketm és határozzuk meg az értékhez tartozó oszlop sorszámát (az oszlopok sorszámai megfelelnek az előzőekben taglalt témák számainak). A matematikai műveletekre kiválóan hasznosítható numpy Python könyvtárral ezt egyszerűen végre tudjuk hajtani.

```
dominant_topics = np.argmax(lda_matrix, axis=1)
```

A dominant_topics változó már mindegyik dokumentumra nézve tartalmazza az adott cikk domináns témáját (pontosabban annak sorszámát). Adjunk hozzá egy új oszlopot a kezdeti adatbázisunkhoz, mely ezeket a témaazonosítókat tartalmazza!

```
corpus_df[„dominant_topics"] = dominant_topics
corpus_df.head(10)
```

A fenti sorok lefuttatását követően megjelenik előttünk a domináns témák sorszámaival kiegészített adatbázis (IV.2.6.).

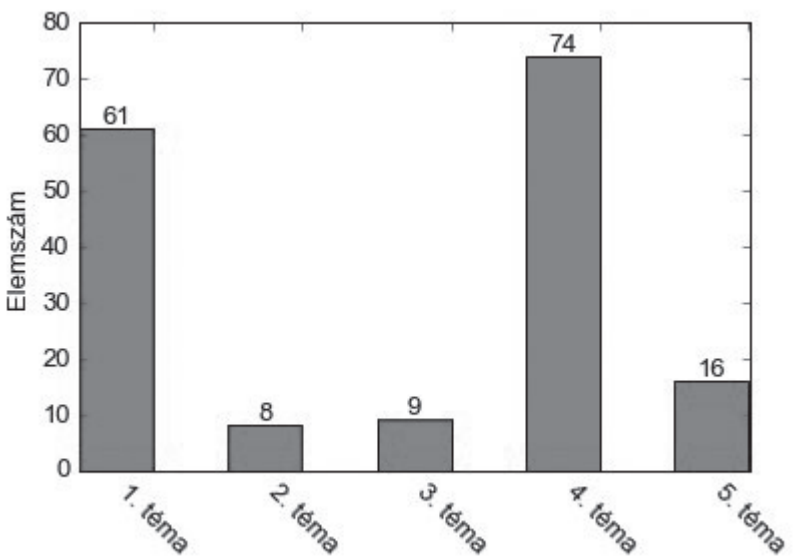
IV.2.6. ábra – A domináns LDA-témák sorszámaival kiegészített adatbázis

Out[138]:

	url	text	cluster_membership	dominant_topics
0	http://napimigrans.com/egykori-szexrabszolgakb...	2016-02-11 Több száz, egykor az ISIS fogságába...	3	3
1	http://napimigrans.com/egy-muszlim-rab-megprob...	2016-02-12 Washington – A 25 éves muszlim fogv...	3	3
2	http://napimigrans.com/magyarorszag-fele-szori...	2016-02-21 Horvátország nem engedi tovább a Sz...	0	0
3	http://napimigrans.com/vajon-mikor-hal-meg-az-...	2016-02-24 A népvándorlás nem más, mint egy ol...	0	0
4	http://napimigrans.com/a-parizsi-gyilkosok-sem...	2016-02-16 MILOS ZEMAN A KORÁNRA HIVATKOZIK – ...	0	0
5	http://napimigrans.com/verfurdot-rendezett-az-...	2016-02-09 Több mint 300 embert végeztek ki az...	1	3
6	http://napimigrans.com/video-oriasi-tuntetes-a...	2016-02-21 Minnesota – Úgy látszik az USA-ban ...	4	3
7	http://napimigrans.com/video-220-milliot-kolte...	2016-02-24 Még 2014-ben szó volt róla, hogy né...	1	3
8	http://napimigrans.com/hogyan-romboljuk-le-a-h...	2016-02-21 Az értékek viszonylagosak. Más kult...	3	2
9	http://napimigrans.com/europa-megeroszakolasa-...	2016-02-18 Meghökkentő címlappal jelent meg a ...	4	3

Vessünk egy pillantást a domináns témák gyakoriságaira (IV.2.7. ábra)!

IV.2.7. ábra – A domináns LDA-témák gyakoriságai



A K-közép klaszterezéshez hasonlóan az egyes csoportok elemszáma igen kiegyenlítetlen. Az első és a negyedik téma 60 fölötti gyakoriságához képest a többi téma nagyjából 10 cikket tömörít.

Eredményeink érvényességének megállapítása érdekében vessük össze a klaszterezés során kapott klasztertagságokat és a domináns témák azonosítóit! Ehhez

számítsunk kereszttáblát az adatbázis klasztertagságokat és a domináns témák sorszámait tartalmazó változók között!

```
cluster_topic_crosstab = pd.crosstab(corpus_df[„dominant_topics”], corpus_df[„cluster_membership”]).as_matrix()
cluster_topic_crosstab
```

A megjelenő kereszttábla sorszámai megfelelnek a domináns témák azonosítóinak, az oszlopok sorszámai pedig a klasztertagságokat jelentik (azaz az első sorban és első oszlopban megtalálható érték azokat az eseteket jelöli, ahol az adott dokumentum egyszerre kapott egyes sorszámot a klasztertagságok és a domináns témák tekintetében is). A kereszttábla celláiban lévő számokat áttekintve látható, hogy bizonyos klasztertagságok szorosan összefüggnek egyes domináns LDA-témákkal. Az első számú klaszterbe 43 olyan cikk tartozik, mely ugyancsak egyes sorszámú domináns témával rendelkezik, továbbá a négyes számú domináns téma dokumentumai nagy számban feleltethetők meg a kettes, hármas és négyes klaszterbe tartozó cikkeknek.

Számítsunk Khi-négyzet statisztikát a két kategorikus változó között, hogy mérhetővé tegyük a klasztertagságok és a domináns LDA-témák közötti kapcsolat szorosságát! Ehhez vegyük igénybe a scipy nevű Python könyvtárat!

```
from scipy.stats import chi2_contingency
chi2, p, dof, ex = chi2_contingency(cluster_topic_crosstab)
print chi2, p
```

Látható, hogy a Khi-négyzet statisztikaértéke magas (87,056), tehát a p-érték igen alacsony ($8{,}69*10^{-12}$). Az egyes cikkek klasztertagságai és a domináns LDA-témák között statisztikailag szignifikáns kapcsolatot mérhetünk, azaz a két módszer eredményei jelentősen átfednek. Az eredmény nem lehet különösebben meglepő, hiszen az egyes témák jellemző kifejezéseinek vizsgálatakor egyértelmű párhuzamokat vettünk észre a K-közép módszer során létrejött klaszterekekkel. A két felügyelet nélküli tanulási módszer eredményei ennyiben kölcsönösen validálták egymást, s így kutatási kérdésünkre megfelelő választ hozott elemzésünk.

Ellenőrző kérdések

- Milyen kutatási kérdés megválaszolására lehetne felhasználni a magyar Országgyűlés képviselőinek felügyelet nélküli tanulási módszerekkel vizsgált felszólalásait?
- Milyen előnyökkel jár a TF-IDF súlyozás a nyers szógyakoriságokkal szemben?
- Mennyire nehezíti meg a magyar nyelv a kutató dolgát a felügyelet nélküli tanulási módszerek alkalmazása során?
- Milyen módszertani problémákat kell megoldani annak érdekében, hogy egy felügyelet nélküli tanulási módszerrel azonosítsuk az index.hu adott napi címlapjának témáit?

Szószedet

Magyar	Angol
K-közép klaszterezés	K-means clustering
Klaszterközéppont	Cluster centroid
Klasztertagság	Cluster membership
Koszinuszhasonlóság	Cosine similarity
Python könyvtár	Python library
Szógyakorisági inverz	Term frequency-inverse
Szógyakorisági súlyozás	Term frequency weighting
Szótövesítés, szótövezés, szótőképzés	Stemming
Tokenizáció	Tokenization
Topikmodell, témamodell	Topic model

Ajánlott irodalom

A felügyelet nélküli tanulási algoritmusok statisztikai és matematikai alapjaiban való mélyebb elmélyüléshez az olvasó figyelmébe ajánljuk Hastie, Tibshirani és Friedman művét (2005). A szövegelemzési és szövegbányászati módszerek jobb megértéséhez érdemes áttekinteni Aggarwal és szerzőtársai (2012) könyvét. A technikák és mintaprogramkódok gyakorlati alkalmazásához ajánljuk a fejezetben említett Python könyvtárak *(nltk, sklearn)* honlapjának áttanulmányozását, illetve a tankönyv honlapján (http://qta.tk.mta.hu) fellelhető anyagokat.

V. KITEKINTÉS

Sebők Miklós

V.1. TOVÁBBI KUTATÁSI IRÁNYOK

A fejezet célja, hogy kitekintést nyújtson a kötetben bemutatott alapvető problémákon és technikákon túl a kvantitatív szövegelemzés tudományterületére. E fejezetben bemutatjuk azokat a legfontosabb kiterjesztési lehetőségeket, melyek egy természetes következő lépést jelentenek a kötet ismeretanyagát már elsajátító olvasó számára. Ennek keretében vázolunk olyan, a klasszikus szövegbányászati feladatokhoz tartozó, illetve ezen részben túlmutató feladatokat, mint a kivonatolás vagy a többnyelvű információkinyerés. Másodsorban tárgyaljuk a „szöveg mint adat"-megközelítés multimédiás és internetalapú kiterjesztéseit, valamint az ezekkel kapcsolatos új információbányászati technikákat. Kitérünk a számítógépes QTA elemzések olyan speciális problémáira, mint a programválasztás és a programok használata. Zárásképpen pedig a tágabb tudományági kontextust villantjuk fel a Big Data szociológiája és a számítógépes társadalomtudomány kapcsán.

A jelen kötet legfontosabb célja az volt, hogy bevezetést nyújtson a nemzetközi politikatudomány két – részben átfedést mutató – kurrens irányzatába, a kvalitatív adatok elemzésébe és a szövegbányászatba. E bevezető jelleg ugyanakkor szükségképpen szelekcióra késztette a szerkesztőt és a szerzőket. Feladatunk az volt, hogy bemutassuk a QTA elemzések alapfogalmait és leggyakoribb módszereit, tudván hogy tényleges kutatási alkalmazásukhoz a területtel ismerkedő olvasónak még további elmélyülésre lesz szüksége. Az, hogy röviden kitekintettünk a kutatási alkalmazás fontosabb lépéseire – így például az elméletalkotásra és a kutatási stratégiákra is –, egy tudatos döntés volt, még akkor is, ha tudtuk, hogy ezzel a technikák bemutatása vagy bizonyos speciális feladatok elvégzése csak korlátozott terjedelemben, egy-egy fejezetben, illetve néhány példa kapcsán szerepelhet csak a kéziratban.

Annyi bizonyos, hogy a későbbiek során mind az anyag bemutatásának elvontsági szintje, mind tematikus kiterjedtsége emelendő, még akkor is, ha a számítógépes nyelvészetben és szövegbányászati irodalomban bevett matematikai-informatikai apparátusra a társadalomtudósnak nincs feltétlenül szüksége. Abban is bízhatunk, hogy a hazai szakirodalom fejlődése és az itt

bemutatott technikákat alkalmazó magyar nyelvű cikkek gyarapodása megfelelően kiegészíti, s ha kell – amint az rendszerint a technológiai tárgyú könyvek osztályrésze – meghaladja majd e kötet tartalmát. Annak érdekében, hogy jelezzük, hogy milyen esetekben éreztük kényelmetlennek a fenti kompromisszumokat, e rövid fejezet kitekintés nyújt négy, a könyvben nem szereplő (vagy legalábbis megkerült), de fontos problémakörre.

1. A *klasszikus szövegbányászati* feladatok köre részben túlmutat a könyvben tárgyalt, s a névelem-felismeréssel, az osztályozással és a csoportosítással leírható legszűkebb problémaegyüttesen. A szövegek *kivonatolása,* illetve összefoglalása (Kovács, 2007: 166–175) egy ilyen, meglehetős technikai kihívásokkal terhes, de egyben nagy gyakorlati jelentőségű terület. A szövegbányászat egyik legalapvetőbb területét jelentő információkinyerés esetében a névelemek kiemelése mellett ugyanilyen fontos ezek *kapcsolatának feltárása* (Jing Jiang, 2012: 22–29).

Nem foglalkoztunk a kötetben az egyes természetes nyelvek speciális problémáival. Figyelemmel arra, hogy a magyar nyelv nem része a nagy világnyelvek nyelvcsaládjainak – s ettől nem függetlenül a magyar nyelvvel foglalkozó kutatócsoportok relatíve kis méretűek – az olyan többnyelvű feladatok és megoldásaik, mint a *tudástranszfer* (Pan et al., 2012: 223–238) és a *többnyelvű szövegbányászat* (Nie, 2012: 324) a haladó kutató számára különösen fontosak lehetnek. Részben ehhez kapcsolódik a *szó-dokumentum mátrix* leegyszerűsítésének problémaköre, mely szinte minden empirikus kutatás során felmerül. Az e célt is szolgáló *témamodellezési* megoldások a jelen könyv szövegének első kiterjesztései között szerepelhetnének (Crain et al., 2012: 129).

Egy másik és már a politikatudományban is alkalmazott terület a különböző osztályozási *algoritmusok együttes használata* (Zhang – Ma, 2012). A különböző logikát alkalmazó függvényekre épülő technikák eredményeinek súlyozott vagy *szavazásos összesítése* (i. m.: 35) egyfajta biztosítékot jelenthet az egyes megoldások speciális módszertani korlátai ellen. A fentieken túl számos olyan határterülete van a politikatudománynak a számítógépes nyelvészettel, mely korábban kivitelezhetetlennek tűnő kutatási stratégiák előtt nyit teret. Ezek közül kiemelkedik a nagy szöveges adatbázisok (törvények, parlamenti beszédek, médiatermékek tartalmának) elemzése. Az ilyen interdiszciplináris kutatások előtt álló lehetőségek túlzás nélkül korlátlannak tekinthetők (ezek kapcsán ld. az éves számítógépes nyelvészeti konferenciák előadásait, így pl. MSZNY, 2010).

2. A második nagy témaegyüttes, mely további feldolgozásra vár, egyik kiindulópontunkkal, a *szöveg mint adat* feltevésével kapcsolatos (Grimmer – Stewart, 2013; Laver et al., 2003). Egyrészt felmerülhet a megközelítés kiterjeszthetősége

olyan más kvalitatív adatforrásokra, mint amilyenek a képek vagy filmek. Más kutatások az adatforráshoz kapcsolódó elemzési módszereket használnak, melyek részben túlmutatnak a szöveg mint adat paradigma hagyományos kérdésfeltevésein és megoldásain.

E tekintetben talán a leggyorsabban fejlődő terület az internetes adatok feldolgozása. Az olyan valamilyen módon újszerű adatforrások, mint a multimédiás és közösségi médiás tartalmak, s általában a metaadatokkal rendelkező webes/web 2.0-ás dokumentumok sok szempontból előnyöket kínálnak a papíralapú vagy rosszabbul strukturált elektronikus adatforrásokkal szemben. Az osztályozási algoritmusok dolgát jelentősen könnyítik az ilyen többletinformációk, nem beszélve a webes információk tömeges beszerzésének relatíve alacsony költségeiről (vö. *web crawling*; a strukturált adatbázisok kinyerése webes információkinyerő rendszerek (*wrapper*ek) segítségével – Ferrara et al., 2014). E területnek részben egy sajátos kutatási területe és technológiája is kialakult (ld. pl. a *közösségi hálózatok elemzését* [Hogan, 2008] és a *webes adatok bányászatát* [Liu, 2011; Markov – Larose, 2007]).

3. Legfeljebb jelzésértékű módon foglalkoztunk a szövegben a számítógépes QTA olyan speciális problémáival, mint a szoftverek kiválasztása és célirányos használata a társadalomtudományi kutatásokban (ld. pl. Gilbert et al., 2014; Silver – Lewins, 2014). E tekintetben különbséget kell tenni a gépi támogatású és a gépi adatelemzés között. Előbbiek esetében a kutató személyét és szubjektív döntéseit felértékelő antropológiával és a szociológia egyes ágaival mutathat átfedést a politikatudós érdeklődése. A kézi annotálás változatos funkcióitól a fogalmak közötti hálózatok ábrázolásáig a kódolást támogató CAQDAs programok fontos szerepet játszhatnak a politikatudományi kutatás eszköztárában. A legelterjedtebb ilyen kereskedelmi szoftverek (ATLAS.ti; MAXQDA; Nvivo, QDA Miner – ld. Bazeley – Jackson, 2013; Friese, 2014) közül e kötetben némi ízelítőt adtunk az ATLAS.ti bizonyos funkciói kapcsán (egy ingyenes alternatíva az R programnyelvhez készített „RQDA" csomag). Optimális esetben ugyanakkor egy önálló monográfia foglalkozhatna az ilyen megoldások társadalomtudományi alkalmazásának lehetőségeivel és a szoftverfunkciók részleteivel.

Hasonlóan önálló kutatási területnek tekinthető mára az automatizált gépi szövegbányászat, legyen szó szótáralapú, felügyelt tanulási vagy felügyelet nélküli tanulási módszerekről. E tekintetben némileg nagyobb a választék, hiszen a kereskedelmi szoftverek (mint amilyen a WordStat) mellett például a szabad hozzáférésű R programnyelvhez is számos témánkba vágó kiegészítő modul érhető el (ld. pl. „tm" kiegészítőt – Feinerer, 2015). Bár a kötet számos példát hoz ezek alkalmazására, módszeres bevezetésre e tekintetben sem volt lehetőségünk. Márpedig a megfelelő szoftver, illetve applikáció kiválasztása egy olyan döntés, mely akár a kutatási eredményekre is kihatással lehet.

Általánosságban itt azzal a tanáccsal élhetünk, hogy egyszerűbb és mára sztenderdizált feladatok elvégzésére célszerű kereskedelmi szoftvereket alkalmazni (mint amilyen a Tikk [2007: 238] által is tárgyalt SPSS Clementine vagy a Provalis már említett WordStat szoftvere). Komplexebb műveletek elvégzésére és speciális kutatási igények kivitelezéséhez ugyanakkor már megéri „befektetni" valamely gyakran használt programnyelv megismerésébe. A kötetünket kísérő honlapon (qta.tk.mta.hu) bővebben is kitérünk az egyes szoftverek funkcióinak áttekintésére, illetve mintascripteket is közlünk az R és Python programnyelvben bizonyos alapműveletek kapcsán (illetve általánosságban ld. Ledolter, 2013).

4. Végezetül fontos megemlítenünk a szövegbányászat lehetséges tudományos és társadalmi hatását tárgyaló különböző kutatási irányokat, így például a Big Data szociológiáját (Csepeli – Dessewffy, 2015) és a *számítógépes társadalomtudomány* gyorsan fejlődő területét (Cioffi-Revilla, 2014). E még porózus körvonalakkal rendelkező két tudományág a hagyományos QTA-megközelítésnél sokkal közvetlenebb módon kapcsolódik a filozófiához és a természettudományokhoz, amennyiben a társadalmi komplexitás elméleti kérdéseivel is foglalkoznak (Miller – Page, 2007). Egyes szerzők (pl. Barabási-Albert, 2010) egyenesen az *emberi dinamika* új korszakáról értekeznek, melyben a jövő (mint az emberi viselkedések összessége) „kiszámíthatóvá" válik. A számítógépes társadalomtudományban e célt részben ágensalapú szimulációs eljárásokkal modellezik (Vág, 2006; Kovács – Takács, 2003).

Annyi talán már e rövid felsorolásból is látszik, hogy kötetünk legfeljebb egy első, kezdetleges lépés a fejlett QTA-alapelvek és technikák magyarországi társadalomtudományi alkalmazása terén. Reményeink szerint könyvünket tanulmányok és monografikus művek sora követi majd, melyek választ adnak az olvasó olyan kérdéseire is, melyeket nem tudtunk részletesebben tárgyalni.

Szószedet

Magyar	Angol
Ágensalapú modellezés/szimuláció	Agent-based modelling/simulation
Csoportok közötti csatornakövetés	Ensemble learning
Emberi dinamika	Human dynamics
Információbányászat	Information extraction
Kapcsolatbányászat	Relation extraction
Kifejezés-dokumentum mátrix	Term-document matrix
Osztályozó bizottság	Classifier committee
Összefoglalás, kivonatolás	Summarization
Robottal támogatott internetes keresés	Web crawling
Számítógépes társadalomtudomány	Computational social science
Társadalmihálózat-elemzés	Social network analysis
Topikmodellezés (témamodellezés)	Topic modelling
Többnyelvű szövegbányászat	Translingual mining
Tudástranszfer	Transfer learning (inductive transfer)
Webes adatbányászat	Web data mining
Webes információkinyerő rendszer	Wrapper

Ajánlott irodalom

A kötet tematikáján részben túlmutató területek megismeréséhez kiváló bevezetést nyújt magyar nyelven Tikk (2007), angol nyelven Aggarwal és Zhai (2012d). A további témák kapcsán ajánlott irodalmak: a kvalitatív adatelemzési szoftverek közötti választásról: Silver – Levins (2014); az internetbányászatról: Liu (2011); a számítógépes társadalomtudományról: Cioffi-Revilla (2014).

FELHASZNÁLT IRODALOM

Abonyi János (szerk.) (2006): *Adatbányászat a hatékonyság eszköze. Gyakorlati útmutató kezdőknek és haladóknak.* Budapest, ComputerBooks.

Abney, Steven (2007): *Semisupervised Learning for Computational Linguistics.* Computer Science and Data Analysis Series. Chapman and Hall CRC.

Adcock, Robert – Collier, David (2001): Measurement validity: A shared standard for qualitative and quantitative research. In: *American Political Science Review*, Vol. 95., No. 3. (September, 2001), 529–546.

Aggarwal, Charu C. – Zhai, Cheng Xiang (2012): An introduction to text mining. In: Aggrawal, Charu C. – Zhai, Cheng Xiang (eds.): *Mining Text Data*. New York, Springer, 1–10.

Aggarwal, Charu C. – Zhai, Cheng Xiang (2012b): A Survey of text classification algorythms. In: Aggrawal, Charu C. – Zhai, Cheng Xiang (eds.): *Mining Text Data*. New York, Springer, 163–222.

Aggarwal, Charu C – Zhaim Cheng Xiang (2012c): A survey of text clustering algorythms. In: In: Aggrawval, Charu C. – Zhai, Cheng Xiang (eds.): *Mining Text Data*. New York, Springer. 77-128.

Aggarwal, Charu C. – Zhai, Cheng Xiang (eds.) (2012d): *Mining Text Data.* New York, Springer.Aggarwal, Charu C. – Zhao, Yuchen – Yu, Philip S. (2012): *On Text Clustering with Side Information.* ICDE ,12 Proceedings of the 2012 IEEE 28th International Conference on Data Engineering. IEEE Computer Society Washington, DC, USA, 894–904.

Airoldi, Edoardo M. – Fienberg, Stephen E. – Skinner, Kiron K. (2007): Whose Ideas? Whose Words? Authorship of Ronald Reagan's Radio Addresses. In: *Political Science and Politics,* Vol. 40., No. 3. (Jul., 2007), 501–506.

Airoldi, Edoardo M. (2003) *Who wrote Ronald Reagan's Radio Addresses?* http://repository.cmu.edu/cgi/viewcontent.cgi?article=1162&context=statistics (Letöltés ideje: 2015. november 11.)

Albaugh, Quinn – Soroka, Stuart – Loewen, Peter J. – Sevenans, Julie – Walgrave, Stefan (2014): *Comparing and Combining Machine Learning and Dictionary-Based Approaches to Topic Coding.* Paper presented at the 7th annual Comparative Agendas Project (CAP) conference, Konstanz, June 12–14.

Ali, Chedi Bechikh – Wang, Rui – Haddad, Hatem (2015): A Two-Level Keyphrase Extraction Approach. In: Gelbukh, Alexander (Ed.): *Computational Linguistics and Intelligent Text Processing. 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Proceedings, Part II.* Springer International Publishing, 390–401.

Ambrus Gergely (2007): Analitikus filozófia. In: Boros Gábor (szerk.): *Filozófia,* Budapest, Akadémiai Kiadó, 1065–1145.

Artés, Joaquín (2011): Do Spanish politicians keep their promises? In: *Party Politics*, Vol. 19, No. 1, 143–159.

Babbie, Earl (1998): *A társadalomtudományi kutatás gyakorlata.* Budapest, Balassi Kiadó.

Babbie, Earl (2003): *A társadalomtudományi kutatás gyakorlata.* Budapest, Balassi Kiadó.

Babbie, Earl (2004): Kvalitatív adatelemzés. In: Babbie, Earl: *A társadalomtudományi kutatás gyakorlata.* Budapest, Balassi Kiadó, 413–437.

Babbie, Earl (2008): Paradigmák, elmélet és társadalomtudományi kutatás. In: Babbie, Earl: *A társadalomtudományi kutatás gyakorlata.* Hatodik, átdolgozott kiadás. Budapest, Balassi Kiadó, 46–77.

Bach Iván (2005): *Formális nyelvek*. Második, javított kiadás. Budapest, Typotex Kiadó.

Bakliwal, Akshat – Foster, Jennifer – van der Puil, Jennifer – O'Brien, Ron – Tounsi, Lamia – Hughes, Mark (2013): Sentiment Analysis of Political Tweets: Towards an Accurate Classifier. In*: Proceedings of the Workshop on Language in Social Media (LASM 2013)*. 49–58.

Balázs Ágnes (2015): Plurális választójog a XX. század első felének magyar közjog-tudományában. In: *Acta Humana. Emberi Jogi Közlemények.* Új folyam III. 2015/1, 9–23.

Baldwin, Timothy – Bannard, Colin – Tanaka, Takaaki – Widdows, Dominic (2003): An empirical model of multiword expression decomposability. In: *Proceedings of the ACL 2003 workshop on Multiword expressions, Association for Computational Linguistics*. 89–96.

Banerjee, Sanjoy (1986): Reproduction of Social Structures: An Artificial Intelligence Model. In: *The Journal of Conflict Resolution*. Vol. 30., No. (June 1986), 221–252.

Barabási Albert László (2010): *Villanások. A jövő kiszámítható.* Fordította: Kepes János. Budapest, Nyitott Könyvműhely.

Bártházi Eszter – Héder Mihály (2010): Panaszlevelek szerkezetének gépi felismerése. In: Tanács Attila – Vincze Veronika (szerk.): *VII. Magyar Számítógépes Nyelvészeti Konferencia.* Szeged, Szegedi Tudományegyetem, Informatikai Tanszékcsoport, 3–13.

Baruque, Bruno – Chorchado, Emilio (2010): *Fusion Methods for Unsupervised Learning Ensembles.* Berlin, Heidelberg, Springer.

Bauer, Martin W. – Bicquelet, Aude – Suerdem, Ahmed K. (eds.) (2014): Text analysis: an indroductory manifesto. In: *New Literary History.* Vol. 5., No. 1., 91–117.

Bazeley, Pat – Jackson, Kirsti (2013): *Qualitative Data Analysis with NVivo. Second Edition.* SAGE Publications.

Bazeley, Pat (2009): Analysing Qualitative Data: More Than Identifying Themes. In: *Malaysian Journal of Qualitative Research*, Vol. 2., No. 2., 6–22.

Bazeley, Pat (2013): Comparative analyses as a means of furthering analysis. In: Bazeley, Pat: *Qualitative Data Analysis: Practical strategies.* SAGE Publishing London. 254–281.

Berg, Bruce L. – Lune, Howard (2014): *Qualitative Research Methods for the Social Sciences. 8th Edition.* Harlow, Pearson Education Limited.

Berry, Michael J. A. – Linoff, Gordon S. (2004): *Data Mining Techniques. For Marketing, Sales, and Customer Support. Second Edition.* Indianapolis, Indiana, Wiley Publishing.

Bhomwick, Tanuka (2006): *Building an Exploratory Visual Analysis Tool for Qualitative Researchers.* AutoCarto2006, Vancouver, WA, June 26–28. http://www.geovista.psu.edu/publications/2006/Bhowmick_AutoCarto_QualRes_06.pdf (Letöltés ideje: 2016. február 29.)

Bíró Lajos (2009): *Dokumentum osztályozás rejtett Dirichlet-allokációval. Doktori Értekezés.* Budapest, Eötvös Loránd Tudományegyetem, Informatikai Kar, Információtudományi tanszék.

Blei, David M. – Ng, Andrew Y. – Jordan, Michael I. (2003): Latent Dirichlet Allocation. In: *Journal of Machine Learning Research*, Vol. 3. January. 993–1022.

Bloom, Kenneth (2011): *Sentiment Analysis based on Appraisal Theory and Functional Local Grammars.* Chicago, Illnois, Illinois Institute of Technology.

Boda Zsolt – Sebők Miklós (2015): Előszó: a Hungarian Comparative Agendas Project bemutatása. In. *Politikatudományi Szemle.* 2015/4. 33-40.

Bodon Ferenc (2010): *Adatbányászati algoritmusok.* http://www.cs.bme.hu/~bodon/magyar/adatbanyaszat/tanulmany/adatbanyaszat.pdf (Letöltés ideje: 2015. november 10.)

Bringer, Joy D. – Johnston, Lynne Halley – Brackenridge, Celia H. (2006): Using computer-assisted qualitative data analysis software to develop a grounded theory project. In: *Field Methods*, August 2006, Vol. 18, No. 3., 245–266.

Bryman, Alan (2005): Kvantitatív és kvalitatív módszerek összekapcsolása. Fordította: Erdődi Katalin. In: Letenyei László (szerk.): *Településkutatás.* Budapest, Ráció Kiadó, 371–391.

Ceron, Andrea – Curini, Luigi – Iacus, Stefano M. – Porro, Giuseppe (2013): Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. In: *New Media & Society,* Vol. 16., No. 2., 340–358.

Charmaz, Kathy (2006): *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis.* London, SAGE Publications.

Cheong, Marc – Lee, Vincent CS (2011): A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter. In: *Information Systems Frontiers*. Vol. 13., No. 1., 45–59.

Chung, Young Mee – Noh, Young-Hee (2002): Developing a specialized directory system by automatically classifying Web documents. In: *Journal of Information Science,* Vol. 29, No. 2., 117–126.

Cioffi-Revilla, Claudio (2014): *Introduction to Computational Social Science: Principles and Applications.* London, Springer.

Coffey, Amanda Jane – Atkinson, Paul A. (1996): Concepts and Coding. In: Coffey, Amanda Jane – Atkinson, Paul A. (1996): *Making Sense of Qualitative Data. Complementary Research Strategies.* Thousand Oaks, SAGE Publications, 26–45.

Coffey, Amanda Jane – Holbrook, Beverley – Atkinson, Paul (1996): Qualitative Data Analysis: Technologies and Representations. In: *Sociological Research Online,* Vol. 1, No. 1. http://www.socresonline.org.uk/1/1/4.html

Copeland, B. Jack – Proudfoot, Diane (2005): Turing and the computer. In: Copeland, B. Jack (ed.): *Alan Turing's Automatic Computing Engine. The Master Codebreaker's Struggle to Build the Modern Computer.* Oxford and New York, Oxford University Press, 107–148.

Creswell, John W. (2009): Research design: *Qualitative, Quantitative, and Mixed Methods Approaches. Third Edition.* Thousand Oaks, SAGE Publishing.

Creswell, John W. (2013): Research design: *Qualitative, Quantitative, and Mixed Methods Approaches. Fourth Edition.* Thousand Oaks, SAGE Publishing.

Croft, W. Bruce (1977): Clustering large files of documents using the single-link method. In: *Journal of the American Society of Information Science,* Vol. 28., No. 6., 341–344.

Csepeli György – Dessewffy Tibor (2015): Big Data, mint a szociológiai megismerés új paradigmája. In: *Szellem és tudomány: a Miskolci Egyetem Szociológiai Intézetének folyóirata,* 6. évf. 2015/1–2. 173–188.

David-Tabibi, Omid E. – Van Den Herik, H. Jaap – Koppel, Moshe – Netanyahu, Nathan S. (2009): Simulating human grandmasters: evolution and coevolution of evaluation functions. In: *GECCO ,09 Proceedings ofthe 11th Annual conference on Genetic and evolutionary computation*, 1483–1490.

Dey, Ian (1993): *Qualitative data analysis, A user-friendly guide for social scientists.* London, Routledge.

Diesner, Jana – Carley, Kathleen M. (2005): Revealing Social Structure from Texts: Meta-Matrix Text Analysis as a novel method for Network Text Analysis. In: Narayanan, V. K – Amstrong, Deborah (eds.): *Causal Mapping for Research in Information Technology.* London, Idea Group Publishing, 81–108.

Dimitrovski, Ivica – Kocev, Dragi – Loskovska, Suzana – Džeroski, Sašo (2016): Improving bag-of-visual-words image retrieval with predictive clustering trees. In: *Information Sciences.* (Feb., 2016), Vol. 329., 851–865.

Doukidis, Georgios I. (1987): An Anthology on the Homology of Simulation with Artificial Intelligence. In: *The Journal of the Operational Research Society*, Vol. 38., No. 8., Current Simulation Research (Aug., 1987), 701–712.

Doukidis, Georgios I. – Paul, Raul J. (1986): Experiences in automating the formulation of discrete event simulation models. In: Kerckhoffs, Eugene J. H. – Vanstenkiste, Ghislain C. – Zeigler, Bernard P. (eds.): *AI Applied to Simulation*. Simulation Series, Vol. 18., No. 1. San Diego, Calif., The Society for Computer Simulation.

Doukidis, Georgios I. – Paul, Raul J. (1987): ASPES–a skeletal Pascal expert system. In: Sol, Henk G. – Takkenberg, Cees A.Th – De Vries Robbé, Pieter F. (ed.): *Expert Systems and Artificial Intelligence in Decision Support Systems.* Proceedings of the Second Mini Euroconference, Lunteren, The Netherlands, 17–20 November 1985., 227–246.

Dudás László (2011): *Alkalmazott mesterséges intelligencia*. Miskolc, Miskolci Egyetem, Alkalmazott Informatikai Tanszék.

Duffy, Gavan – Tucker, Seth A. (1995): Political Science: Artificial Intelligence Applications. In: *Social Science Computer Review.* Vol. 13., No. 1. (April 1995), 1–20.

Ensmenger, Nathan (2012): Is chess the drosophila of artificial intelligence? A social history of an algorithm. In: *Social Studies of Science,* Vol. 42, No. 1. (February 2012), 5–30.

Esterberg, Kristin G. (2002): Qualitative Methods in Social Research. Boston, McGraw-Hill.

Ezzy, Douglas (2002): *Qualitative Analysis. Practice and Innovation*. London, Routledge.

Farkas Katalin – Kelemen János (2007): Nyelvfilozófia. In: Boros Gábor (szerk.): *Filozófia*. Budapest, Akadémia Kiadó, 1273–1288.

Feinerer, Ingo (2015): *Introduction to the tm. Package. Text Mining in R*. https://cran.rstudio.com/web/packages/tm/vignettes/tm.pdf (Letöltés ideje: 2015. december 4.)

Feldman, Ronen – Sanger, James (2007): *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. New York: Cambridge University Press.

Ferrara, Emilio – De Meo, Pasquale – Fiumara, Giacomo – Baumgartner, Robert (2014): *Web Data Extraction, Applications and Techniques: A Survey*. http://arxiv.org/pdf/1207.0246v4.pdf (Letöltés ideje: 2015. december 4.)

Filho, Dalson Britto – Figuerido – da Rocha, Enivaldo Carvalho – da Sliva Júnior, José Alexandre – Paranhos, Ranulfo – da Silva, Mariana Batista – Duarte, Bárbara Sofia Felix (2014): Cluster Analysis for Political Scientists. In: *Applied Mathematics,* 2014/5. 2408–2415.

Franklin, James (2005): The Elements of Statistical Learning: Data Mining, Inference and Prediction by Trevor Hastie, Robert Tbishirani and Jerome Friendman (Review). In: *The Mathematical Intelligencer,* Vol. 27., No. 2., 83–85.

Friedman, Jerome H. (1998): Data Mining and Statistics: What's the connection? In: *Computing Science and Statistics*, Vol. 29., No.1., 3–9.

Friese, Susanne (2014): *Qualitative Data Analysis with ATLAS.ti. Second Edition*. London, Sage.

Furlong, Paul – Marsh, David (2010): A Skin, not a Sweater: Ontology and Epistemology in Political Science. In: Marsh, David – Stroker (eds.): *Theory and Methods in Political Science. Third Edition*. London, Palgrave Macmillan, 184–210.

Franch, Fabio (2013): (Wisdom of the Crowds)2: 2010 UK Election Prediction with Social Media. In: *Journal of Information Technology & Politics,* Vol. 10., No.1., 57–71.

Ghahramani, Zoubin (2004): Unsupervised learning. In Bosquet, Olivier – Luxburg, Ulrike von – Rätsch, Gunnar (eds): *Advanced lectures on machine learning*. ML Summer Schools 2003, Canberra, Australia, February 2–14, 2003, Tübingen, Germany, August 4–16, 2003, Revised Lectures. Berlin Heidelberg, Springer, 72–112.

Gibbs, Graham R. (2002): *Qualitative Data Analysis: Explorations with NVivo (Understanding Social Research)*. Buckingham, Open University Press.

Gibson, William J. – Brown, Andrew (2009): Introduction to qualitative data: analysis in context In: Gibson, William – Brown, Andrew: *Working with Qualitative Data*. London, Sage, 1–14.

Gilbert, Linda S. – Jackson, Kirsti – di Gregorio, Silvana (2014): Tools for Analyzing Qualitative Data: The History and Relevance of Using Qualitaive Data Analysis Software. In: Spector, J. Michael – Merill, M. David – Elen, Jan – Bishop, M.J. (eds): *Handbook of Research on Educational Communications and Technology. Fourth Edition*. New York, Springer, 221–236.

Glaser, Barney – Strauss, Anselm (1967): *The discovery of grounded theory*. Chicago, Aldine.

Glaser, Barney – Strauss Anselm (1999): *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago, Aldine Transaction.

Gosztolya Gábor – Tóth László (2010): Kulcsszókeresési kísérletek hangzó híranyagokon beszédhang alapú felismerési technikákkal. In: Tanács Attila – Vincze Veronika (szerk.): *VII. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged, Szegedi Tudományegyetem, Informatikai Tanszékcsoport, 224–235.

Grbich, Carol (2013): *Qualitative Data Analysis: An introduction*. London, Sage.

Grimmer, Justin (2010): A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. In: *Political Analysis,* Vol. 18., No. 1., 1–35.

Grimmer, Justin – King, Gary (2011): General purpose computer-assisted clustering and conceptualization. In: *Proceedings of the National Academy of Sciences,* Vol. 108., No. 7., 2643–2450.

Grimmer, Justin – Stewart, Brandon M. (2013): Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. In: *Political Analysis,* Vol. 21., No. 3., 1–31.

Hastie, Trevor – Tibshirani, Robert – Friedman, Jerome (2005): *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* New York, Springer.

Hearst, Marti A. (1999): Untangling Text Data Mining. In: *Proceedings of 37th Annual Meeting for Computational Linguistics,* New York, ACM Press, 3–10.

Henderson, Harry (2011): *Alan Turing. Computer Genius and Wartime Code Breaker.* Makers of Modern Science. New York, Chelsea House Publishers.

Héra Gábor – Ligeti György (2014): *Módszertan – A társadalmi jelenségek kutatása*. Budapest, Osiris Kiadó, 2014.

Hillard, Dustin – Purpura, Stephen – Wilkerson, John (2008): Computer-Assisted Topic Classification for Mixed-Methods Social Science Research. In: *Journal of Information Technology & Politics*. 31–46. http://faculty.

washington.edu/jwilker/TopicClassification.pdf (Letöltés ideje: 2016. október 5.)

Hobbs, Jerry L. – Appelt, Douglas E. – Bear, John S. Tyson, Mabry – Magerman, David (1991): *The TACITUS System: The MUC-3 Experience.* Artifical Intelligence Center. SRI International. October 1991. https://www.sri.com/sites/default/files/uploads/publications/pdf/463.pdf (Letöltés ideje: 2015. október 30.)

Van den Hoonard, Deborah K. – van den Hoonard, Will C. (2008): Data analysis. In: Given, Lisa M. (ed.): *The Sage Encyclopedia of Qualitative Data Analysis.* Volumes 1&2. Los Angeles, CA, SAGE Publishing, 186–188.

Hogan, Bernie (2008): Analyzing Social Networks via the Internet. In: Fielding, Nigel – Lee, Raymond M. – Blank, Grant (eds.): *The SAGE Handbook of Online Research Methods.* London, SAGE Publications, 141–160.

Hogenboom, Alexander – Bal, Daniella – Frasincar, Flavius – Bal, Malissa – De Jong, Franciska – Kaymak, Uzay (2013): *Exploiting emoticons in sentiment analysis. In: Proceedings of the 28th Annual ACM Symposium on Applied Computing, SAC 2013,* 18–22 Mar 2013, Lisbon, Portugal, 703–710.

Hopkins, Daniel J. – King, Gary (2010): A Method of Automated Nonparametric Content Analysis for Social Science. In: *American Journal of Political Science,* Vol. 54., No.1., 229–247.

Hopkins, Daniel – King, Gary (2010b): Extracting systematic social science meaning from text. http://gking.harvard.edu/files/gking/files/wordstlk-high.pdf (Letöltés ideje: 2016. október 3.)

Huang, Te-Ming – Kecman, Vojislav – Kopriva, Ivica (2006): *Kernel Based Algorithms for Mining Huge Data Sets: Supervised, Semi-supervised, and Unsupervised Learning.* Varsaw, Springer.

Huang, Thomas S. – Nijholt, Anton – Pantic, Maja – Pentland, Alex (2006b): *Artifical Intelligence for Human Computing: ICMI 2006 and IJCAI 2007 International Workshops,* Canada, Banff.

Hudson, Valerie M. (ed.) (1991): Artificial Intelligence and International Politics. Boulder, CO., WestView Press.

Hu, Xia – Liu, Huan (2012): Text Analytics in Social Media. In: Aggarwal, Charu C. – Zhai, ChengXiang (eds.): *Mining Text Data.* New York, Springer, 385–414.

Jain, A. K. – Murty, M. N. – Flynn, P. J. (1999): Data clustering: A review. In: *ACM Computing Surveys,* Vol. 31., No. 3., 264–323.

Jeanray, Nathalie – Marée, Raphaël – Pruvot, Benoist – Stern, Oliver – Geurst, Pierre – Wehenkel, Louis – Muller, Marc (2015): *Phenotype Classification of Zebrafish Embryos by Supervised Learning,* PLOS ONE 10 (1).

Jiang, Jing (2012): Information extraction from text. In: Aggarwal, Charu C. – Zhai, ChengXiang (eds.): *Mining Text Data*. New York, Springer, 11–41.

Joachims, Thorsten (2001): A Statistical Learning Model of Text Classification for Support Vector Machines. In: *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, USA, 128–136.

Johnson, Cristopher – Shukla, Parul – Shukla, Shilpa (2012): *On Classifying the Political Sentiment of Tweets*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.229.3927&rep=rep1&type=pdf (Letöltés ideje: 2016. március 1.)

Jurka, Timothy P. – Collingwood, Loren – Boydstun, Amber E. – Grossman, Emiliano – van Atteveldt, Wouter (2011) RTextTools: A supervised learning package for text classification. In: *The R. Journal,* Vol. 5., No. 1., 6–12.

Kastellec, Jonathan P. (2010): The Statistical Analysis of Judicial Decisions and Legal Rules with Classification Trees. In: *Journal of Empirical Legal Studies,* Vol. 7., Issue 2, June 2010, 202–230.

Kim, Soo-Min – Hovy, Eduard (2004): *Determining the Sentiment of Opinions*. Proceedings of the COLING conference, Geneva, 2004. http://www.isi.edu/natural-language/people/hovy/papers/04Coling-opinion-valences.pdf (Letöltés ideje: 2015. október 2.)

Kiss Balázs (2013): Érzelmek és politikatudomány. In: *Politikatudományi Szemle,* XXII. évfolyam, 2013/3., 7–28.

Klebanov, Beata Beigman – Diermeier, Daniel – Beigman Eyal (2008): Automatic Annotation of Semantic Fields for Political Science Research In: *Journal of Information Technology & Politics,* Vol. 5., No. 1., 95–120.

Kovács Balázs – Takács Károly (2003): Szimuláció a társadalomtudományban. In: *Szociológiai Szemle,* 2003/3., 27–49.

Kovács László (2007): Tartalomkeresés webdokumentumokban. In: Tikk Domonkos (szerk.): *Szövegbányászat*. Budapest, Typotex Kiadó, 176–216.

Kozinets, Robert (2015): *Netnography: Redefined. Second Edition*. London, SAGE Publications.

Kunwar, Samir (2013): *Text Documents Clustering using K-Means Algorithm*. http://www.codeproject.com/Articles/439890/Text-Documents-Clustering-using-K-Means-Algorithm (Letöltés ideje: 2016. február 29.)

Krauss, Jonas – Nann, Stefan – Simon, Daniel – Fischbach, Kai – Gloor, Peter (2008): Predicting Movie Success and Academy Awards through Sentiment and Social Network Analysis. In: *ECIS 2008 Proceedings. Paper 116*. http://www.ickn.org/documents/Oscar_ECIS_Final_v1.3.pdf (Letöltés ideje: 2016. március 1.)

Lánczi András (2005): A politika alapfogalmai. In: Gallai Sándor – Török Gábor (szerk.): *Politika és politikatudomány.* Budapest, Aula Kiadó, 29–30.

Laver, Michael – Benoit, Kenneth – Gary, John (2003): Extracting Policy Positions from Political Texts Using Words as Data. In: *The American Political Science Review,* Vol. 97., No. 2. (May, 2003), 311–331.

Ledolter, Johannes (2013): *Data mining and Business Analytics with R.* Hoboken, New Yersey, John Wiley & Sons Inc.

Li, Lei (2012): A novel violent videos classification scheme based on the bag of audio words feature. In: *International Journal of Computational Intelligence and Applications,* Vol. 11., No. 2., 1–21.

Li, Lei – Mei, Tao – In-So, Kweon – Xian-Sheng, Hua (2011): Contextual Bag-of-Words for Visual Categorization. In: *IEEE Transactions on Circuits & Systems for Video Technology.* 04/01/2011, Vol. 21., Issue 4, 381–392.

Liu, Bing (2007): *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Heidelberg, Springer Science & Business Media.

Liu, Bing (2011): *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Heidelberg, Springer Science & Business Media.

Liu, Bing – Zhang, Lei (2012): A survey of opinion mining and sentiment analysis. In: Aggrawval, Charu C. – Zhai, Cheng Xiang (eds.): *Mining Text Data*. New York, Springer, 415-463.

Lucas, Christopher – Nielsen, Richard A. – Roberts, Margaret E. – Stewart, Brandon M. –Storer, Alex – Tingley, Dustin (2015): Computer-assisted text analysis for comparative politics. In: *Political Analysis*, mpu019.

MacQueen, J. B. (1967): Some Methods for classification and Analysis of Multivariate Observations In: *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability,* Berkeley, University of California Press, Vol. 1., 281–297.

Manning, Christopher – Raghavan, Prabhakar – Schütze, Hinrich (2008): *Introduction to information retrieval.* Cambridge, Cambridge University Press.

Manning, Christopher D. – Schütze, Hinrich (1999): *Foundations of Statistical Natural Language Processing. Second Printing.* Cambridge, Massachusets, London, England, Massachusetts Institute of Technology.

Markov, Zdravko – Larose, Daniel T. (2007): *Data Mining the Web. Uncovering Patterns in Web Content, Structure and Usage*. New Jersey, John Wiley and Sons Inc.

Marrero, Mónica – Urbano, Julián – Sánchez-Cuadrado, Sonja – Morato, Jorge – Gómez-Berbís, Juan Miguel (2013): Named Entity Recognition: Fallacies, Challenges and Opportunities. In: *Journal of Computer Standards and Interfaces,* Vol. 35., No. 5., 482–489.

Mayring, Philipp (2014): *Qualitative content analysis: theoretical foundation, basic procedures and software solution*. SSOAR. Open Access Repository. http://nbn-resolving.de/urn:nbn:de:0168-ssoar-395173 (Letöltés ideje: 2016. március 1.)

Maietta, Raymond C. (2008): Computer Assited Data Analysis. In: Given, Lisa M. (ed.): *The SAGE Encyclopedia of Qualitative Research Methods*. Thousand Oaks, SAGE Publications.

Marks, David F. – Yardley, Lucy (2004): *Research Methods for Clinical and Health Psychology*. London, SAGE Publications.

Mefford, Dwain (1990): Case-Based Reasoning, Legal Reasoning and the Study of Politics. In: *Political Behavior*, Vol. 12., No. 2., Cognition and Political Action (Jun., 1990), 125–158.

Meyer, David – Hornik, Kurt – Feinerer, Ingo (2008): Text mining infrastructure. In: *R. Journal of Statistical Software*. Vol. 25., No.5., 1–54.

Miháltz Márton (2010): OpinHu: online szövegek többnyelvű véleményelemzése. In: Tanács Attila – Vincze Veronika (szerk.): *VII. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged, Szegedi Tudományegyetem, Informatikai Tanszékcsoport, 14–23.

Miles, Matthew B. – Huberman, A. Michael (1994): Early Steps in Analysis. In: Miles, Matthew B. – Huberman, A. Michael: *Qualitative Data Analysis: An Expanded Sourcebook. Second Edition*. Thousand Oaks, SAGE Publications, 55–76.

Miles, Matthew B. – Huberman, A. Michael (1994): *Qualitative Data Analysis: An Expanded Sourcebook. Second Edition*. Thousand Oaks, SAGE Publications.

Miles, Matthew B. – Huberman, A. Michael (2013): *Qualitative Data Analysis: A Methods Sourcebook. Third Edition*. SAGE Publications, Thousand Oaks.

Miles, Matthew B. (1979): Qualitative data as an attractive nuisance: The problem of analysis. In: *Administrative Science Quarterly,* Vol. 24., No.4., 590–601.

Miller, John H. – Page, Scott E. (2007): *Complex Adaptive Systems: An Introduction to Compunational Models of Social Life*. Princeton and Woodstock, Princeton University Press.

Molnár Gábor József – Kojedzinszky Tamás – Farkas Richárd (2010a) Bűnügyi névelem-felismerés. In: Tanács Attila – Vincze Veronika (szerk.): *VII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2010)*. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged. pp. 366-370

Molnár Gábor József – Kojedzinszky Tamás – Farkas Richárd (2010b): *Bűnügyi névelem-felismerés*. http://www.textrend.org/publications/msznyk.pdf (Letöltés ideje: 2015. november 11.)

Montoya, Mark S. (2012): A Brief Survey of Chess AI: Strategy and Implementation. University of New Mexico. CS 427 – Fall 2012. www.cs.unm.edu/~pdevineni/papers/Montoya.pdf (Letöltés ideje: 2016. március 20.)

Móra György – Farkas Richárd (2010): Szótáralapú névelem-felismerés szóhatárainak javítása gépi tanulási módszerrel. In: Tanács Attila – Vincze Veronika (szerk.): *VII. Magyar Számítógépes Nyelvészeti Konferencia. MSZNY2010.* Szeged, Szegedi Tudományegyetem, Informatikai Tanszékcsoport, 317–324.

Mullen, Tony – Malouf, Robert (2006): A Preliminary Investigation into Sentiment Analysis of Informal Political Discourse. In: *Proceedings of the AAAI Workshop on Analysis of Weblogs*. http://www-rohan.sdsu.edu/~malouf/pubs/aaai-politics.pdf (Letöltés ideje: 2016. március 1.)

Munk Sándor: Szemantika az informatikában. In: *Hadmérnök,* IX. évfolyam, 2014/2., 311–331.

Nagel, Stuart S. (1988): Updating Microcomputers and Public Policy Analysis. In: *Public Productivity Review,* Vol. 11., No. 3. (Spring, 1988), 117–122.

Nagel, Stuart S. (1994): Decision-Aiding Software and Super-Optimum Solutions. In. Nagel, Stuart (ed.): Encyclopedia of Policy Studies. Second Edition, Revised and Expanded. New York, Basel, Marcel Dekker Inc., 49–68.

Nebhi, Kamel (2012): Ontology-based information extraction from Twitter. In: *Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data,* COLING 2012, Mumbai, December 2012., 17–22.

Negnevitsky, Michael (2005): *Artificial Intelligence. A Guide to Intelligent Systems.* Second Edition. Harlow, Addison-Wesley.

Nenkova, Ani – Mc Keown, Kathleen (2012): A survey of text summarization techniques. In: Aggrawal, Charu C. – Zhai, Cheng Xiang (eds.): *Mining Text Data.* New York, Springer, 43-76.

Neri, Federico – Aliprandi, Carlo – Camillo, Furio (2011): Mining the Web to Monitor Political Consensus. In: Wiil, Uffe Kock (ed.): *Counterterrorism and Open Source Intelligence.* Wien, Springer-Verlag, 391–412.

Neuman, W. Russel – Marcus, George E. – Crigler, Ann N. –MacKuen, Michael (eds.) (2007): *The Affect Effect. Dynamics of Emotion in Political Thinking and Behavior.* Chicago, The University of Chicago Press.

Newman, Maxwell Herman (1955): Alan Mathison Turing. 1912-1954. In: *Biographical Memoirs of Fellows of the Royal Society*, Vol. 1. (Nov., 1955), 253–263.

O'Connor Brian – Balasubramanyan, Ramnath – Routledge, Bryan R. – Smith, Noah A. (2010): From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In: *Proceedings of the International AAAI Conference on Weblogs and Social Media*. http://www.cs.cmu.edu/~nasmith/papers/oconnor+balasubramanyan+routledge+smith.icwsm10.pdf (Letöltés ideje: 2016. március 1.)

Ortony, Andrew – Clore, Gerald – Foss, Mark A. (1987): *The Referential Structure of the Affective Lexicon.* Cognitive Science 11, 341–364.

Pang, Bo – Lee, Lillian (2008): Opinion mining and sentiment analysis. In: Foundations and Trends. In: *Information Retrieval,* Vol. 2., No. 1-2., 1–135.

Pataki Máté (2011): Fordítási plágiumok keresése. In: Tanács Attila – Vincze Veronika (szerk.): *VIII. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged, Szegedi Tudományegyetem, Informatikai Tanszékcsoport, 24–34.

Paul, Ray J. – Doukidis, Georgios I. (1986): Further Developments in the Use of Artificial Intelligence Techniques Which Formulate Simulation Problems. In: *The Journal of the Operational Research Society,* Vol. 37., No. 8. (Aug., 1986), 787–810.

Polányi Károly (1992): *A hallgatólagos következtetés logikája*. Előadás. http://nyitottegyetem.phil-inst.hu/tudfil/ktar/forr_ed/polanyi.htm (Letöltés ideje: 2015. szeptember 10.)

Pólya Tibor – Csertő István – Fülöp Éva – Kővágó Pál – Miháltz Márton – Váradi Tamás (2015): A véleményváltozás azonosítása politikai témájú közösségi médiában megjelenő szövegekben. In: Tanács Attila – Varga Viktor – Vincze Veronika (szerk.): *XI. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged, Szegedi Tudományegyetem, Informatikai Tanszékcsoport, 198–209.

Porter, Martin F. (2001) *Snowball: A language for stemming algorithms.* Retrieved from: http://snowball.tartarus.org/texts/introduction.html (Letöltés ideje: 2016. október 5.)

Prior, Lindsay F. (2008): Document Analysis. In: Given, Lisa M. (ed.): The Sage Encyclopedia of Qualitative Data Analysis. Volumes 1&2. Los Angeles, CA., SAGE Publishing, 230–231.

Quinn, Kevin M. – Monroe, Burt L. – Colaresi, Michael – Crespin, Michael, H. – Radev, Dragomir R. (2010): How to analyze political attention with minimal assumptions and costs. In: *American Journal of Political Science,* Vol. 54., No. 1., 209–228.

Qi, Xiaoguang – Davison, Brian D. (2009): Web Page Classification: Features and Algorithms. In: *ACM Computing Surveys,* Vol. 41., No. 2., Article 12 (February 2009), 31 pages DOI = 10.1145/1459352.1459357 http://doi.acm.org/10.1145/1459352.1459357 (Letöltés ideje: 2016. március 2.)

Roberts, Carl W. (2000): A Conceptual Framework for Quantitative Text Analysis. On Joining Probabilities and Substantive Inferences about Tests. In: *Quality & Quantity.* Vol. 34., No. 3., 259–274.

Robertson, Stephen (2004): Understanding inverse document frequency: on theoretical arguments for IDF. In: *Journal of Documentation,* Vol. 60. ,No. 5., 503–520.

Rojas, Raul (1996): Unsupervised Learning and Clustering Aglorithms In: Rojas, Raul: Neural Networks. A Systematic Introduction. Berlin, Springer, 99–121.

Roseman, Ira J. – Smith, Craig A. (2001): Appraisal theory: Overview, assumptions, varieties, controversies. In: Scherer, Klaus R. – Schorr, Angela – Johnstone, Tom (eds.): *Appraisal processes in emotion: Theory, methods, research.* Series in affective science. New York, Oxford University Press, 3–19.

Russel, Stuart – Norvig, Peter (eds.) (2010): *Artificial Intelligence. A modern approach. Third Edition.* Englewood Cliffs, Prentice Hall.

Russel, Stuart – Norvig, Peter (2005): *Mesterséges intelligencia modern megközelítésben. Második, átdolgozott, bővített kiadás.* Budapest, Panem Könyvkiadó.

Rácz József (2006): *Kvalitatív drogkutatások. Kvalitatív kutatások budapesti droghasználók között.* Budapest, L'Harmattan.

Sandelowski, Margarete (1995): Qualitative Analysis: What It Is and How to Begin. In: *Research in Nursing & Health,* Vol. 18., No. 4., 371–375.

Sanders, David (2010): Behavioural Analysis. In: Marsh, David – Stroker, Gerry (eds.): *Theory and Methods in Political Science.* London, Palgrave Macmillan, 24–41.

Schreiber, James B. (2008): Data. In: Given, Lisa M. (ed.): The Sage Encyclopedia of Qualitative Data Analysis. Volumes 1&2. Los Angeles, CA., SAGE Publishing, 195–196.

Schonhardt-Bailey, Cheryl (2005): Measuring Ideas More Effectively: An Analysis of Bush and Kerry's National Security Speeches. In: *Political Science and Politics,* Vol. 38., No. 4., 701–711.

Schreiber, James B. (2008): Data. In: Given, Lisa M. (ed.): The Sage Encyclopedia of Qualitative Data Analysis. Volumes 1&2. Los Angeles, CA., SAGE Publishing, 195–196.

Schwartz, H. Andrew – Ungar, Lyle H. (2015): Data-Driven Content Analysis of Social Media: A Sytematic Overview of Automated Methods. In: *The Annals of the American Academy of Political and Social Science,* May 2015, Vol. 659., No. 1., 78–94.

Seale, Clive (1999): „Quality in Qualitative Research". In: *Qualitative INquiry,* Vol. 5., No. 4., 465–478.

Sebők, Miklós – Berki, Tamás (2016): Incrementalism and Punctuated Equilibrium in Hungarian Budgeting (1991-2013). In: Journal of Public Budgeting, Accounting and Financial Management. (megjelenés alatt)

Seidel, John V. (1998): Appendix E: Qualitative Data Analyisis. In: *The Ethnograph v5 Manual,* http://www.qualisresearch.com/DownLoads/qda.pdf (Letöltés ideje: 2016. március 1.)

Shannon, Claude (1950): Programming a Computer for Playing Chess. In: Philosophical Magazine. Ser. 7, Vol. 41., No. 314. (March 1950) http://archive.computerhistory.org/projects/chess/related_materials/text/2-0%20and%202-1.Programming_a_computer_for_playing_chess.shannon/2-0%20and%202-1.Programming_a_computer_for_playing_chess.shannon.062303002.pdf (Letöltés ideje: 2016. március 29.)

Silver, Christina – Lewins, Ann (2014): *Using Software in Qualitative Research: A Step-by-Step Guide.* London, Sage.

Silverman, David (2007): *A Very Short, Fairly Interesting and Reasonably Cheap Book about Qualitative Research.* London, SAGE Publishing.

Simon, Herbert A. – Dantzig, George B. – Hogarth, Robin – Plott, Charles R. – Raiffa, Howard – Schelling, Thomas C. – Shepsle, Kenneth A. – Thaler, Richard – Tversky, Amos – Winter, Sidney (1987): Decision Making and Problem Solving. In: *Interfaces*. Vol. 17., No. 5. (Sep. - Oct., 1987), 11–31.

Sim, Janchuan – Acree, Brice D. L. – Gross, Justin H. – Smith, Noah A. (2013): Measuring Ideological Proportions in Political Speeches. In: *Conference on Empirical Methods in Natural Language Processing* (EMNLP 2013). Oct, 2013. Seattle, WA. http://homes.cs.washington.edu/~nasmith/papers/sim+acree+gross+smith.emnlp13.pdf (Letöltés Ideje: 2016. március 29.)

Slapin, Jonathan B. – Proksch, Sven-Oliver (2008): A Scaling Model for Estimating Time-Series Party Positions from Texts. In: American Journal of Political Science. Vol. 52., No. 3., (July 2008.), 705–722.

Slapin, Jonathan B. – Proksch, Sven-Oliver (2014): Words as Data: Content Analysis in Legislative Studies. In: Martin, Shane – Saalfeld, Thomas – Strøm, Kaare W. (eds.*): The Oxford Handbook of Legislative Studies*. Oxford, Oxford University Press, 127–145.

Solt Illés – Szidarovszky P. Ferenc – Tikk Domonkos (2010): Kontextualizált névelem-felismerés és relációkinyerés kórházi zárójelentésekben. In: Tanács Attila – Vincze Veronika (szerk.): *VII. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged, Szegedi Tudományegyetem, Informatikai Tanszékcsoport, 35–46.

Soós Gábor – Körösényi András (szerk.) (2013): *Azt tették, amit mondtak? Választási ígéretek és teljesülésük, 2002-2006*. Budapest, MTA Társadalomtudományi Kutatóközpont Politikatudományi Intézet.

Soós Gábor (szerk.) (2015): Ígéret, felhatalmazás, teljesítés. Választási programok és kormányzati megvalósulásuk, 1998-2010. Budapest, MTA Társadalomtudományi Kutatóközpont Politikatudományi Intézet.

Spencer, Liz – Ritchie, Jane – O'Connor, William – Barnard, Matt (2014): Analysis: Principles and Processes. In: Ritchie, Jane – Lewis, Jane – McNaughton Nicholls, Carol – Ormston, Rachel (eds.): *Qualitative Research Practice: A Guide for Social Science Students and Researchers*. London, SAGE Publishing, 269–293.

Steen, Lynn Arthut (1975): Computer Chess: Mind vs. Machine. In: *Science News*, Vol. 108., No. 22. (Nov. 29, 1975), 345; 350.

Stevenson, William (1990): *Titkos háború*. Budapest, I.P.C.

Stoker, Gerry – Marsh, David (2010): Introduction. In: Marsh, David – Stroker Gerry (eds.): *Theory and Methods in Political Science. Third Edition*. London, Palgrave Macmillan, 9–10.

Strapparava, Carlo – Valitutti, Alessandro (2004): WordNet-Affect: an Affective Extension of WordNet. In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-2004), Lisbon, Portugal. European Language Resources Association (ELRA)*. 1083–1086.

Sundheim, Beth M. (1992): Overview of the fourth message understanding evaluation and conference. In: *Proceeding MUC4 '92 Proceedings of 4th conference on Message understanding. Association for Compunational Liguistics, Stroudsburg, PA, USA*. 3–21.

Szabó Martina Katalin (2015): A polaritásváltás problémája a szentimentelemzés szempontjából. In: Váradi Tamás (szerk.): *Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből. IX. Alkalmazott Nyelvészeti Doktoranduszkonferencia*. Budapest, 2015. 02. 06. Budapest, MTA Nyelvtudományi Intézet, 51–60.

Szarvas György – Farkas Richárd (2007): Információkinyerés. In: Tikk Domonkos (szerk.): *Szövegbányászat*. Budapest, Typotex Kiadó, 81–101.

Szűts Zoltán (2012): A web 2.0 kommunikációelméleti kérdései. In: *Jelkép*. 2012/1–4. http://communicatio.hu/jelkep/2012/1_4/szuts_zoltan.htm (Letöltés ideje: 2015. november 10.)

Taboada, Maite – Brooke, Julian – Tofiloski, Milian – Voll, Kimberly – Stede, Manfred (2011): Lexicon-based methods for sentiment analysis. In: *Computational Linguistics,* Vol. 37., No. 2., 267–307.

Tanács Attila – Vincze Veronika (szerk.) (2011): *VIII. Magyar Számítógépes Nyelvészeti Konferencia MSZNY 2011.* Szeged, Szegedi Tudományegyetem, Informatikai Tanszékcsoport.

Tanács Attila – Vincze Veronika (szerk.) (2010): *VIII. Magyar Számítógépes Nyelvészeti Konferencia MSZNY 2011.* Szeged, Szegedi Tudományegyetem, Informatikai Tanszékcsoport.

Thomas, David R. (2006): A General Inductive Approach for Analyzing Qualitative Evaluation Data. In: *American Journal of Evaluation,* Vol. 27., No. 2., 237–246.

Tikk Domonkos (2007a): Bevezetés. In: Tikk Domonkos (szerk.): *Szövegbányászat.* Budapest, Typotex Kiadó, 20–24.

Tikk Domonkos (2007b): Csoportosítás. In: Tikk Domonkos (szerk.): *Szövegbányászat.* Budapest, Typotex Kiadó, 145–165.

Tikk Domonkos (2007c): Osztályozás. In: Tikk Domonkos (szerk.): *Szövegbányászat.* Budapest, Typotex Kiadó, 102–144.

Tikk Domonkos (2006): Szövegbányászat. In: Abonyi János (szerk.): *Adatbányászat a hatékonyság eszköze. Útmutató kezdőknek és haladóknak.* Budapest, Computer Books Kiadói Kft., 343–365.

Tikk Domonkos – Kardkovács Zsolt Tivadar – Magyar Gábor – Szidarovszky Ferenc P. (2006): Szótári névelemek felismerése és morfológiai annotálása. In: *Híradástechnika,* LXI. évfolyam, 2006/1., 29–34.

Tikk Domonkos – Kovács László (2007): Előfeldolgozás, modellalkotás, reprezentáció. In: Tikk Domonkos (szerk.): *Szövegbányászat.* Budapest, Typotex Kiadó, 25–62.

Tikk Domonkos – Szidarovszky Ferenc P – Kardkovács Zsolt Tivadar – Magyar Gábor (2005): Ismert névelemek felismerése és morfológiai annotálása szabad szövegben. In: Alexin Zoltán – Csendes Dóra (szerk.): *III. Magyar Számítógépes Nyelvészeti Konferencia: MSZNY 2005.* Szeged, Szegedi Tudományegyetem, Informatikai Tanszékcsoport, 190–199.

Tjong Kim Sang, Erik. – Bos Johan (2012): Predicting the 2011 Dutch Senate Election Results with Twitter. In: *Proceedings of SASN 2012, the EACL 2012 Workshop on Semantic Analysis in Social Networks, Avignon, France, 2012.* http://ifarm.nl/erikt/papers/sasn2012.pdf (Letöltés ideje: 2016. március 1.)

Tjong Kim Sang, Erik – De Meulder, Fien (2003): Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition In: *Proceeding CONLL ,03 Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003* - Volume 4., 142–147.

Trappl, Robert (ed.) (2006): *Programming for Peace. Computer-Aided Methods for International Conflict Resolution and Prevention.* Advances in Group Decison and Negotiation. Vol. 2., Dordrecht, Springer.

Tumasjan, Andranik – Sprenger, Timm O. – Sandner, Philipp G. – Welpe, Isabell M. (2010): Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In: *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media.* 178–185.

Turing, Alan Mathison (1946): Proposed electronic calculator (National Physical Laboratory report, 1946). In: Carpenter, Brian E. – Doran, Robert W. (eds.): *A.M. Turing's ACE Report of 1946 and Other Papers*. Cambridge, MA: MIT Press.

Turing, Alan Mathison (1950): Computing machinery and intelligence. In: *Mind, New Series*, Vol. 59., No. 236. (Oct., 1950), 433–460.

Turing, Alan Mathison – Strachey, Christopher – Bates, M. Audrey – Bowden, Bertram Vivian (1953): Digital computers applied to games. In: Bowden, Bertram Vivian (ed.): *Faster than Thought.* London, Pitman, 286–310.

Vág András (2006): Multiágens modellek a társadalomtudományokban. In: *Statisztikai Szemle*. 84. évfolyam, I. szám, 25–52.

Vázsonyi Miklós – Tikk Domonkos (2007): Az információ-visszakeresés alapjai. In: Tikk Domonkos (szerk.): *Szövegbányászat.* Budapest, Typotex Kiadó, 63–79.

Wang, Bo – Liu, Min (2015): *Deep Learning for Aspect-Based Sentiment Analysis.* Stanford Research Report. http://cs224d.stanford.edu/reports/WangBo.pdf (Letöltés ideje: 2015. december 18.)

Wang, Po-Ya Angela (2013): #Irony or #Sarcasm-A Quantitative and Qualitative Study Based on Twitter. In: *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation.* http://aclweb.org/anthology/Y/Y13/Y13-1035.pdf (Letöltés ideje: 2016. március 1.)

Wang, John (ed.) (2005): *Encyclopedia of data warehousing and mining.* Hershey és New York, Information Science Reference

Wajeed, Mohammed Abdul – Adilakshmi, Thondepu (2009): Text Classification Using Machine Learning. In: *Journal of Theoretical and Applied Information Technology,* Vol. 7., No. 2., 119–123.

White, Michael J. – Judd, Maya D. – Poliandri, Simone (2012): Illumination with a Dim Bulb? What Do Social Scientists Learn by Employing Qualitative Data Analysis Software in the Service of Multimethod Designs? In: *Sociological Methodology,* Vol. 42., No. 1., 43–76.

Wolfram, Stephen (2013): *Data Science of Facebook World*, April 24, 2013. http://blog.stephenwolfram.com/2013/04/data-science-of-the-facebook-world/ (Letöltés ideje: 2016. március 1.)

Wooldridge, Jeffrey M. (2012): *Introductory econometrics: A modern approach. Fifth Edition. (US Edition).* Mason, Cengage Learning.

Yoffe, Hélène – Yardley, Lucy (2004): Content and Thematic Analysis. In: Marks, David F. – Yardley, Lucy: *Research Methods for Clinical and Health Psychology.* London, SAGE Publications, 56–68.

Yu, Hong – Hatzivassiloglou, Vasileios (2003): Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-03).* 129–136.

Zhang, Yan – Wildemuth, Barbara M. (2009): Qualitative Analysis of Content. In: Wildemuth Barbara M. (ed.): *Applications of Social Research Methods to Questions in Information and Library Science.* Westport, Libraries Unlimited, 308–319.

Zhu, Xiaojin – Goldberg, Andrew B. – Brachman, Ronald – Dietterich, Thomas (2009): *Introduction to Semi-Supervised Learning.* Synthesis Lectures on Artificial Intelligence and Machine Learning. n.a., Morgan and Claypool Publishers.

Zhu, Lei – Jin, Hai – Zheng, Ran –Feng, Xiaowen (2013): Weighting scheme for image retrieval based on bag-of-visual-words. In: *IET Image Processing.* Sep 2014, Vol. 8., Issue 9, 509–518.

FÜGGELÉK

A KÖTETET KIEGÉSZÍTŐ HONLAP BEMUTATÁSA

A kötetet a http://qta.tk.mta.hu címen elérhető oldal egészíti ki, amelyen további információk találhatók a könyv anyagához. Mivel kötetünket aktív kutatóknak-oktatóknak és felsőoktatási hallgatóknak egyaránt ajánljuk, szerkezetében keverednek a kézikönyv és a tankönyv elemei. Az oktatásban való közvetlen alkalmazást segíteni hivatottak a fejezetek végén megadott vizsgakérdések mellett a kötet honlapján szereplő további tartalmi egységek.

Forrás: http://qta.tk.mta.hu/a-konyvrol

Ilyen a könyv fejezeteinek kivonata, valamint a kötethez készült munkafüzet, melynek segítségével a gyakorlatban is elsajátíthatjuk a tanult módszereket. Ez gyakorló feladatokat (megoldásokkal), illetve az egyes feladatokra alkalmazható scripteket is tartalmaz. A honlapon szereplő linkekről elérhetőek olyan szabad hozzáférésű szövegelemző programok is, amelyekkel alkalmazott kvantitatív szövegelemzési kutatásokat végezhetünk. Ehhez olyan magyar, illetve angol nyelvű szövegrepozitóriumokat is bemutatunk, melyek megfelelő adatbázisokat kínálnak a QTA-eljárások alkalmazására. Végezetül a honlap egy magyar–angol szószedettel is segítséget nyújt a szakkifejezések közötti eligazodásban.

TÁRGYMUTATÓ

TÁBLÁZATOK JEGYZÉKE

ÁBRÁK JEGYZÉKE

A SZERZŐKRŐL

Balázs Ágnes – Közigazgatási menedzser. Az MTA Társadalomtudományi Kutatóközpontjának kutatási asszisztense, a Nemzeti Közszolgálati Egyetem Közigazgatás-tudományi Doktori Iskolájának doktoranduszhallgatója.

Kubik Bálint – Közgazdász, politológus. Az MTA Társadalomtudományi Kutatóközpontjának kutatási asszisztense, a Starschema Kft. munkatársa.

Molnár Csaba – Politológus. Az MTA Társadalomtudományi Kutatóközpontjának tudományos segédmunkatársa, a Budapesti Corvinus Egyetem Politikatudományi Doktori Iskolájának hallgatója.

Sebők Miklós – Közgazdász, politológus. Az MTA Társadalomtudományi Kutatóközpontjának tudományos munkatársa, az MTA TK PTI Kormányzás és Közpolitika Osztályának vezetője.

Szabó Gabriella – Politológus. Az MTA Társadalomtudományi Kutatóközpontjának tudományos munkatársa.

Vancsó Anna – Szociológus. Az MTA Társadalomtudományi Kutatóközpontjának tudományos segédmunkatársa, a Budapesti Corvinus Egyetem Szociológia Doktori Iskolájának doktorjelöltje.

Zágoni Bella – Politológus-közgazdász. Az MTA Társadalomtudományi Kutatóközpontjának tudományos segédmunkatársa.

Zorigt Burtejin – Politológus. Az MTA Társadalomtudományi Kutatóközpontjának tudományos segédmunkatársa.

SUMMARY

The aim of this edited volume is to provide an introduction to quantitative text analysis and text-mining, this cutting-edge methodology that is now widely used in international political science. Based on the paradigm of *text as data*, the book offers a beginner's guide to the analysis of qualitative data, on the automated or computer-assisted coding of large-scale text corpora as well as on core text mining methods.

The Introduction presents the book's subject and structure, along with some brief remarks on the creative process leading to its publication. Chapter 1 offers a general treatment of the analysis of qualitative data and texts and positions quantitative text analysis and text mining in the field of social sciences. It also highlights the innovative aspects of text mining within the methodological subfield of documentary and content analysis and evaluates the relative merits of hand-coding, computer-assisted coding and automated coding.

Chapter 2 provides an introduction to basic concepts of qualitative and quantitative text analysis as applied for the social sciences. Basic research methodologies, such as the bag-of-words method and named entity recognition are presented with ample practical illustrations from real-life political science research projects. Chapter 3 introduces more sophisticated text mining methods, including classification, sentiment analysis and clustering.

Chapter 4 applies these general approaches to selected research problems in political science. It presents sample scripts in R and Python in a step towards the empirical application of the theoretical models and tools outlined in the previous chapters. The basic methods of supervised and unsupervised machine learning are discussed in more detail. Chapter 5 concludes by considering avenues for further research and the general context of quantitative text analysis and text mining in the social sciences: the sociology of Big Data and computational social science.

L'Harmattan France
5-7 rue de l'Ecole Polytechnique
75005 Paris
T.: 33.1.40.46.79.20
Email: diffusion.harmattan@wanadoo.fr

L'Harmattan Italia SRL
Via Degli Artisti 15, 37
10124 Torino
T.: (39) 011.817.13.88 / (39) 348.39.89.198
Email: harmattan.italia@agora.it

Korrektor: Keszthelyi-Kiss Judit
Borítóterv: Kára László
Nyomdai előkészítés: Kovácsné Daróczi Annamária
A nyomdai munkákat a Robinco Kft. végezte, felelős vezető Kecskeméthy Péter.