Szociológiai Szemle 30(1): 44–62 .

Az automatizált szövegelemzés perspektívája a társadalomtudományokban

Németh Renáta[1] – Katona Eszter Rita[2] – Kmetty Zoltán
nemeth.renata@tatk.elte.hu, katona.eszter@tatk.elte.hu , kmetty.zoltan@tatk.elte.hu

Beérkezés: 2019.02.19.
Átdolgozott változat beérkezése: 2020.02.18.
Elfogadás: 2020.03.23.

Összefoglaló: Cikkünkben a „Big Data" paradigma térnyerésével párhuzamosan rohamosan terjedő természetesnyelv-feldolgozási (NLP) módszereket tekintjük át. Bemutatjuk a társadalomkutatási szempontból leginkább perspektivikus eszközöket, a hozzájuk illeszthető társadalomkutatási kérdéseket és azokat a technikai-módszertani jellegzetességeket, amelyek a klasszikus kvantitatív kutatáshoz képest az NLP specifikumát jellemzik. Ezek a módszerek lényegesen túllépnek a szógyakoriság-elemzésen alapuló klasszikus kvantitatív szövegelemzésen, és a gépi tanulási paradigmán alapuló modellezési logikájuk gyökeresen eltér a magyarázatot / oksági hatás kimutatását elérni kívánó klasszikus társadalomkutatási logikától. Célunk, hogy ebbe az itthon még kevéssé intézményesült területbe betekintést engedve inspirációt nyújtsunk a hazai társadalomkutatók számára, mert meggyőződésünk szerint a szövegbányászat néhány éven belül standard eszköze lesz a nemzetközi alkalmazott társadalomkutatásnak.

Kulcsszavak: kvantitatív szövegelemzés, természetes nyelvfeldolgozás, szövegbányászat, számítógépes szövegelemzés

Bevezetés

A kvantitatív társadalomkutatás Durkheim óta adathalmazokba rendezett empirikus adatokkal foglalkozik. Ezek az adatok jellemzően jól strukturáltak és numerikusak. Igaz ez még az olyan speciális kutatási területen is, mint a kapcsolathálózat-elemzés, ahol a „hagyományostól" nagyon eltérő adatmodellekben kell gondolkodnunk. Azonban a számokon túl is van élet (sőt mindig is volt). Egyre nagyobb mennyiségben érhetők el (kutatási célokra is) olyan tartalmak, amelyek szövegekből épülnek fel. A korábban inkább csak professzionális szereplők (például újságírók, tudományos/szépirodalmi szerzők) által előállított nyilvános szövegek mellett az internet terjedésével párhuzamosan megjelent a laikus felhasználók által előállított tartalom is. Honlapszövegek, blogok, kommentek, a közösségi médiában megjelenő posztok jelentik ma a digitálisan előállított tartalom jelentős részét. Ezek a szövegek

1 A szerző a tanulmány elkészítésekor a Felsőoktatási Intézményi Kiválósági Program támogatásában részesült.
2 A szerző a tanulmány elkészítésekor az Emberi Erőforrások Minisztériuma Új Nemzeti Kiválóság Programjának támogatásában részesült.

bár strukturálatlanok (vagyis nem illeszkednek előre definiált adatbázisba), de az internet előtti időkhöz képest elképzelhetetlen mennyiségben tartalmaznak adatokat az emberek véleményéről, preferenciáiról, attitűdjeiről, sőt cselekvéseiről is (Evans–Acaves 2016).

A szövegek társadalomtudományi vizsgálata természetesen nem új kutatási paradigma, bár használata korábban inkább a kvalitatív kutatásokra korlátozódott. A szisztematikus (kvantitatív) szövegelemzés a tömegmédia elemzésének céljával indult el a két világháború között, és folytatódott a világháború után is (pl. Berelson–Lazarsfeld 1948). A kvantitatív elemzés egy tipikus használata volt, amikor kvalitatív módon azonosított „kódok" megjelenését kvantifikálták a szövegekben (lásd pl. Bales 1950). Ezekben a korai elemzésekben a kódokon kívül nyers szövegelemek (pl. releváns szavak) gyakoriságát és metaadatokat (pl. szerző) is felhasználtak a stilisztikai/szemantikai mintázatok feltárására. Az 1960-as évektől kezdve számítógépek is támogatták ezt a munkát (Hayes 1960).

Miner és szerzőtársai (2012) alapján a kvantitatív szövegelemzés utóbbi tizenöt évben tapasztalható drámai felfutásához két tényezőnek kellett együttesen megjelenni. Egyrészről szükség volt digitális/digitalizált szövegekre, másrészről olyan számítógépes kapacitásra, ami képes volt ezeket a szövegeket feldolgozni. Kiindulópontnak a 80-as évek végét, a 90-es évek elejét tekinthetjük, ekkor jelennek meg az első olyan statisztikai eljárások (például a látens szemantikai indexelés), amelyek nagyobb szövegkorpuszokat voltak képesek komplex módon elemezni (Miner et al. 2012). Az igazi elterjedés azonban inkább a 2000 utáni évekhez köthető, amikor az egyre nagyobb mennyiségű digitális tartalom feldolgozására újabb és újabb módszerek jelentek meg (Liu 2015). A jelenleg is szemünk előtt zajló Big Data[3] forradalom természetesen ezt a területet is magával ragadta. E forradalom nagyságrendjének érzékeltetésére elég egyetlen számot idézni: 2020-ban már 3,8 milliárd ember használ legalább egy közösségimédia-platformot a földön (Digital 2020 reports). Természetesen ezt a kutatási területet is magával ragadta, sőt megfordíthatjuk a dolgot, a szövegbányászat egyfajta katalizátora a Big Data-elemzés jelentős felfutásának. Az automatizált szövegelemzés üzleti/ipari alkalmazásai gyorsan elérték az alkalmazó tudományokat, így pl. az irodalomtudományt és a kultúratudományt is, lásd a „távolról olvasás" fogalmát Moretti (2013) distant reading – close reading dichotómiáján alapulva.

Az újabb kutatási paradigmák esetében mindig felmerülhet, hogy mi újat tud adni a korábbi megközelítésekhez képest. Ez különösen igaz abban az esetben, ha olyan módszerekről, irányokról van szó, amelyek nem standard társadalomkutatói eszköztárat használnak. A Big Data társadalomtudományi használatáról már számos hazai tanulmány született (Csepeli 2015, Dessewffy–Láng 2015, Kmetty 2018). Tanulmányunkban

3 A „Big Data" kifejezést az utóbbi tizenöt évet tekintve előbb egy hatalmas felhajtás kísérte (egy szemléletes példa erre: Anderson 2008), majd egyfajta csalódás volt tapasztalható, éppen a laikusok túlfűtött várakozásai miatt (lásd „big data is dead" 124.000 Google Search találat, „big data projects fail" 16.000 találat). Digitális adatelemzői körökben ennek megfelelően ma már kevésbé használják; mi mégis a használata mellett döntöttünk ebben a cikkben, hogy az olvasók számára könnyebben azonosíthatóvá tegyük tárgyunkat.

amellett érvelünk, hogy a Big Data egy speciális ágának, az automatizált szöveganalitikának is egyre nagyobb szerepe lehet az empirikus társadalomkutatáson belül[4].

Érvelésünk egyik ága a klasszikus survey-kutatások egyre nehezedő környezetéből indul ki. A survey típusú kvantitatív adatgyűjtés válaszmegtagadásra visszavezethető problémái állandó részét képezik a téma szakértői diskurzusának. A survey beszűkülésével szemben a digitális adatok egyre nagyobb mennyiségben állnak rendelkezésre[5]. Ha a mostani tendenciák folytatódását várjuk, akkor vélhetően egyre olcsóbb lesz digitális adathoz hozzájutni, és egyre drágább lesz jó minőségű survey-adatot generálni. De legalább ugyanennyire fontos az is, hogy digitális adatokból olyan vélemények és attitűdök is kinyerhetők, amelyekhez nagyon nehéz hozzáférni survey-ekben vagy a téma kényessége, vagy nehéz operacionalizálhatósága miatt. Az online adatforrás és a survey természetesen nem tökéletes alternatívája egymásnak episztemológiai szempontból sem (Németh 2015). Nem ugyanaz, ha előítéletes attitűd feltárása céljával kérdezünk önbevallásra alapulva egyéneket, vagy ha az egyének online térben megvalósult gyűlöletbeszéde alapján vonunk le róluk következtetéseket (Barna–Knap 2019). A digitális adatforrások előnyeit azonban árnyalja, hogy ezek a szövegek – mivel elsősorban nem elemzési célokra születtek – zajosak és strukturálatlanok, és a survey-hez hasonlóan megbízhatósági és érvényességi kérdésekkel terheltek. Továbbá – mint ahogy azt cikkünk példáiból látni fogjuk – bár a szövegek tartalmi elemzésének vannak egyes részkérdéseket megválaszoló és a társadalomtudomány számára is inspiratív technikái, ezek a technikák korántsem reprodukálják a szövegek ember általi tartalmi megértését.

Tanulmányunk célja az új szöveganalitikai paradigma társadalomtudományi lehetőségeinek bemutatása, az e szempontból releváns főbb eszközeinek, ezek logikájának és a hozzájuk illeszthető társadalomkutatási kérdéseknek az ismertetése. Az automatizált szöveganalitika kurrens eszközeit azért tartjuk fontosnak legalább heurisztikájukat tekintve bemutatni, mert (1) azok lényegesen túllépnek a szógyakoriság-elemzésen alapuló klasszikus kvantitatív szövegelemzésen, továbbá (2) a gépi tanulási paradigmán alapuló modellezési logikájuk gyökeresen eltér a magyarázatot / oksági hatás kimutatását elérni kívánó klasszikus társadalomkutatási logikán. Célközönségünket azok a kvantitatív társadalomkutatók adják, akiknek nincsen mély előismeretük ezen a téren, de érdeklődnek a kvantitatív szövegelemzés iránt. A bemutatott módszerek megértésének megkönnyítésére igyekszünk az új módszereket a klasszikus kvantitatív módszerekhez kapcsolni. Áttekintésünk vállaltan nem technikai, és az egyes módszerek alkalmazásának támogatása is túlmutat tanulmányunkon, ugyanakkor a továbblépéshez ajánlunk jó kiindulópontokat. Az a célunk, hogy ebbe az itthon még kevéssé intézményesült területbe betekintést engedve inspirációt nyújtsunk a hazai társadalomkutatók számára.

4 Természetesen nem minden kvantitatív szövegelemzés Big Data-alapú, de az új módszerek megjelenését elsősorban a Big Data-felhasználások implikálják. A szövegeknek nemcsak a mennyisége okoz problémát, hanem a strukturálatlansága és a standardizálatlansága is. Ilyen szempontból a szöveg nagysága csaknem mellékes.

5 Bár kétségtelen, hogy az adathozzáférés ezen a területen sem mindig egyszerű, lásd a technológiai és adatvédelmi kérdéseket.

A természetesnyelv-feldolgozás

Az automatizált szövegelemzés abban különbözik a hagyományos adatelemzéstől, hogy strukturálatlan adatokon dolgozik. Így mielőtt a szövegek elemzése kapcsán használt fogalmakat tisztázzuk, fontos különbséget tenni a strukturált és a strukturálatlan adatok között.

A strukturált adatoknak azokat az adatokat nevezzük, amelyek hagyományosan sorokban és oszlopokban rögzítettek, könnyen kereshetők.

A félig strukturált adatok olyan tulajdonságokkal bírnak, amelyek megkönnyítik az elemzést, amelyek segítségével hierarchiába rendezhetők az információk. Ilyen például egy webáruház, amely minden termékről ismétlődő struktúrában tárol adatot. Ennek segítségével az adatokat gyorsan és könnyen egy strukturált adatbázisba rendezhetjük.

A strukturálatlan adatoknak egyáltalán nincsen adatbázis jellege, esetünkben lehetnek újságcikkek, blogbejegyzések, hosszabb szöveges dokumentumok. A tanulmányunk következő szakasza a strukturálatlan adatok feldolgozásával foglalkozik.

A természetesnyelv-feldolgozás (a magyar fordításra ez az írásmód terjedt el[6], angolul Natural Language Processing, a továbbiakban NLP) az informatika, a mesterségesintelligencia-kutatás és a nyelvészet határterülete. Természetes nyelvnek azokat a nyelveket nevezzük, amelyek spontán alakultak ki, amelyek nyelvtana nem az emberek közötti nyelvi kommunikáció természetes fejlődésének eredményeként alakult ki, ilyenek például a különböző nemzetek nyelvei. Ezzel szemben a mesterséges nyelvet az ember hozza létre, szabályai tudatosan tervezettek, ilyenek lehetnek a programozási nyelvek, de akár a morze is ide tartozhat. Az NLP lényege olyan módszerek kidolgozása, amelyek alkalmasak nagy mennyiségű, természetes nyelven előállt szöveg elemzésére, célzott információ kinyerésére és nyelvi tartalom generálására (Hirschberg–Manning 2015). Ide tartozik a beszédfelismerés vagy a szövegek szintaktikai feldolgozása is, de a társadalomtudományokat a fentiek szellemében elsősorban az írott nyelvi források szemantikai/tartalmi megközelítése érinti.

E részterületet több más, nem feltétlen szinonim elnevezéssel is illetik, nevezik például számítógépes nyelvészetnek (computational linguistics), automatizált szövegelemzésnek (automated text analytics), szövegbányászatnak (text mining). Ha az elnevezések közötti relációkat részletesebben megvizsgáljuk, akkor azt mondhatjuk, hogy az NLP gyakorlatorientált, eszközöket készít, a számítógépes nyelvészet pedig a módszer elméleti hátterét adja. Az NLP-t sokszor a mesterséges intelligencia (AI, Artificial Intelligence) aldiszciplínájának tekintik. Mivel sokszor kognitív folyamatok megértésére törekszik, egyes szerzők szerint a kognitív számítástechnika (Cognitive Science) diszciplínájaként is értelmezhető (Moreno–Redondo 2016). Moreno–Sandoval és Redondo (2016) szerint a különböző írott tartalmak elemzését szövegelemzésnek (text analytics) nevezzük, amely náluk a szövegbányászat (text mining) szinonimája. Szemléletükben maga a szövegelemzés vagy szövegbányászat

6 Lásd: https://www.inf.u-szeged.hu/algmi/kutatas/nlp. Többek között a Szegedi Egyetem tanszéke is ezt az írásmódot alkalmazza.

az NLP alkategóriája. Cikkük alapján az NLP eszköztárába tartozik a Text Analytics, ami a Natural Language Understanding, és a Text Mining egy másik elnevezése.

Ha különbséget szeretnénk tenni az automatizált szövegelemzés és a szövegbányászat között, azt mondhatnánk, hogy a szövegelemzés a számítógépes nyelvészethez tartozik, míg a szövegbányászat egy újabb tudományág, amely a statisztika, az adatbányászat és a gépi tanulás területéhez kapcsolódik szorosan. A szövegbányászat és az adatbányászat rokon területek, a fő különbség abban áll, hogy utóbbi strukturált adatokkal dolgozik, míg előbbi strukturálatlan vagy félig strukturált adatokkal. A szövegbányászat feladata új, korábban azonosítatlan információk feltárása különböző írásbeli forrásokból

A téma iránt érdeklődőknek ajánlható Aggarwal és Zhai sokat hivatkozott általános összefoglalója (2012), amely korrekt statisztikai alapokat is átad. A társadalomtudományokhoz szól Ignatow és Mihalcea kézikönyve (2016), amely a kvalitatív és a kvantitatív módszereket egyaránt tárgyalja, tankönyvszerű, alkalmazásorientált megközelítésben, áttekintést adva az aktuálisan elérhető adatforrásokról, programnyelvekről, szoftvercsomagokról, elemzési módszerekről[7]. Evans és Aceves (2016) kiváló tanulmánya a társadalomtudományok felől közelít, és a társadalomtudományi tudás létrehozásának kontextusában, a kvalitatív eszközökkel való együttműködés lehetőségében vizsgálja az NLP új módszereit. Cikkünkben mi is felhasználtuk ezeket a munkákat. Magyar nyelven a Tikk Domonkos (2007) által szerkesztett alapos, ám több mint tízéves összefoglaló technikai oldalról ajánlható, míg alkalmazói oldalról a Sebők Miklós által szerkesztett kötet (2016), amely politikatudományi alkalmazásokra koncentrál, de általában a társadalomtudomány számára is közvetlenül használható.

Meg kell jegyezni, hogy a tudományterület rendkívül gyorsan változó/fejlődő képet mutat. Ennek oka egyrészt a nyelvészet, a nyelvtechnológia, másrészt a gépi tanulás rohamos fejlődése. Ennek következménye például, hogy a *.hu* domain 2003-as állapotát tükröző, 18 millió letöltött oldalt tartalmazó, bárki által hozzáférhető Webkorpusz (http://mokk.bme.hu/resources/webcorpus/) minősége mára elmaradt a jelenlegi eszközeink által elérhető minőségtől, mert a letöltött nyers html-oldalak feldolgozásához egykor használt eszközök sokat fejlődtek[8]. A gépi tanulás gyors fejlődését mutatja, hogy a fent hivatkozott szöveganalitikai munkák egyike sem tartalmazza még a Mikolov és társai (2013) által fejlesztett és az utóbbi 3-4 évben elterjedt szóbeágyazási modelleket (lásd a következő fejezetekben), miközben ezek a modellek ma a legnépszerűbbek között vannak[9].

7 Ignatow az Észak-Texasi Egyetem szociológusa, az NLP szociológiai alkalmazásainak egyik ismert alakja.

8 Mivel a Webkorpusz készítői nem őrizték meg az eredeti html-oldalakat, csak a feldolgozott változatot, így utólag ez már nehezen javítható (Indig 2018).

9 A Google Scholaron 25.000 cikk hivatkozza a „word embedding" kifejezést, míg a múlt évezredben is már létező, klasszikusabb „topic model"-t csak 36.000 – igaz, felhasználási területeik nem feltétlenül egyeznek.

Az automatizált szövegelemzés sajátosságai

Előfeldolgozás

Ahogy a bevezetőben már utaltunk rá, a szöveges adatok automatizált feldolgozásának inputja nagyban különbözik a társadalomkutatás klasszikus, jól strukturált adatforrásaitól, mint a standard survey-ek vagy adminisztratív közigazgatási adatok statisztikai elemzésre közvetlenül használható adatbázisai. Röviden, a feladat specifikusságának és komplexitásának bemutatása, illetve a potenciális társadalomtudományi alkalmazások támogatása céljából a következőkben részletezzük ebből a strukturálatlanságból fakadó problémákat/feladatokat. Bár technikainak tűnhet, mi mégis fontosnak tartjuk ezt az ismertetést, mert éppen ebből ítélhető meg ez az új adatforrás mint empirikus bázis kutatási érvényessége – hasonlóan a survey-hez, itt is kutatói döntések sora szükséges, egy konceptualizációs és operacionalizációs folyamat eredménye az output.

A szöveges adatbázisok elérésére/összegyűjtésére itt nem térnénk ki, annyit említünk csak, hogy szöveges adatokat gyűjthetünk közvetlenül saját programmal, elérhetjük azokat tartalomgyűjtő szolgáltatókon keresztül, illetve léteznek szabadon elérhető, mások által gyűjtött és valamilyen szinten előfeldolgozott korpuszok is. Itt kell utalnunk az adatgyűjtés és -felhasználás adatvédelmi és adatbiztonsági oldalára is. A nyers szövegkorpusz összeállítása után az elemzés első lépése egy elemzésre alkalmas numerikus(!) adatbázis előállítása, ezt a lépést előfeldolgozásnak (preprocessing) nevezzük. Részben technikai, részben nyelvészeti megalapozású lépések ezek, köztük olyan feladatok vannak, mint a mondatok és a szavak azonosítása a szövegben (tokenization), a tartalmatlan szavak, például a névelők eltávolítása a szövegből (stop word removal), a szótövesítés (stemming, lemmatization), a szófajok és más nyelvészeti kategóriák azonosítása (part of speech tagging) és a tulajdonnevek vagy más névelemek felismerése (named entity recognition) stb.

Mindezekről a lépésekről lásd részletesen Ignatow és Mihalcea (2016) 5. fejezetét. Ennek az előfeldolgozási szakasznak a specifikálása kulcsfontosságú: a hibák javítása később gyakran nagyon költséges lehet, így megéri az elemzés előtt kezelni őket (Rehurek 2011). Az egyes lépések pontos tartalma és sorrendje erősen függ az adott alkalmazástól is, tehát egyedi kutatói döntés függvénye. Ritkán hangsúlyozott szempont, hogy ezek a döntések (mivel az egyes lépések kölcsönösen hatnak egymásra) befolyásolják a teljes elemzés eredményét (Denny–Spirling 2018).

A fenti lépésekhez használt megoldások nyilván nyelvfüggők, hiszen például egy magyar nyelvű szöveg szótöveinek azonosításához magyar lexikonra és magyar nyelvi szabályokra van szükség. Ezért az NLP-technológiák nem univerzálisak, az NLP adott területen történő alkalmazhatósága attól is függ, hogy az adott nyelvben milyen megoldásokat implementáltak már. Mivel a magyar nyelv agglutináló, elemzése meglehetősen nehéz, és a nyelvtechnológiai fejlesztő közösség is nagyságrendekkel kisebb, mint az angol nyelv esetében, ezért a programok még nem dolgoznak az angolhoz

hasonló pontossággal. Több párhuzamos fejlesztés folyik többféle programnyelven (R, Python, Java) párhuzamosan. A magyar számítógépes nyelvészet egyik meghatározó műhelye az MTA–SZTE Mesterséges Intelligencia Kutatócsoport és annak Nyelvtechnológiai Csoportja. Több fejlesztés mellett a fenti feladatok elvégzésére alkalmas *Magyarlánc* fűződik hozzájuk, amely több nyelvtechnológiai megoldás kiindulópontja. Az egységes és up-to-date megoldást keresőknek egy másik jó kiindulópont az „e-magyar" rendszer, a Magyar Tudományos Akadémia támogatásával és a Nyelvtudományi Intézet koordinálásával létrejött nyelvfeldolgozó rendszer (http://e-magyar.hu/hu/). De nem csak a tudományos műhelyekben készítenek nyelvfeldolgozó eljárásokat, erre jó példa az Orosz György által magyarra ültetett Python csomag, a Spacy (https://github.com/oroszgy/spacy-hungarian-models).

Az előfeldolgozás logikája szinte teljes mértékben megegyezik a survey-es adattisztítással. Ez egy szükséges (bár nem feltétlenül elégséges) lépés az elemzések elkezdése előtt. Korábban utaltunk rá, hogy az NLP módszerei még elmaradnak a szövegek teljes tartalmi megértésétől. Ezt jól alátámasztja, hogy a legnépszerűbb módszerek nyelvészeti szempontból a leegyszerűsített, ún. szózsákmodellt (bag of words) használják, vagyis nem veszik figyelembe a szavak szövegbeli sorrendjét, szintaktikai hierarchiáját, csupán szavak halmazaként kezelik a szövegeket (a többször előforduló szavakat többször véve). E szózsákmodellt használva azután például szóelőfordulási gyakoriságok segítségével állapítják meg egy-egy szöveg érzelmi töltetét, vagy tesznek spam-címkét az e-mailekre. Az e cikkben tárgyalt modellek közül egyedül a szóbeágyazási modellek lépnek túl ezen, az egyes kifejezések közvetlen mondatbeli környezetét is számon tartva.

Felügyelt vs. nem felügyelt módszerek

Az alábbiakban az NLP társadalomtudományi szempontból releváns módszereit kutatási kérdésük logikája szerint kategorizálva mutatjuk be.

Az NLP felhasználási területe eredendően az informatika határterületeire, a gépi fordításra/nyelvfeldolgozásra és az üzleti alkalmazásokra terjedt ki, ebből eredően a célfeladat leggyakrabban valamilyen predikció létrehozása (lásd például azt a kérdést, hogy spam kategóriába sorolható-e egy e-mail). A gépi tanulás legfontosabb célkitűzése ennek megfelelően olyan hatékony és robusztus algoritmusok előállítása, amelyekkel pontos előrejelzéseket tehetünk. Fontos hangsúlyozni, hogy ez a predikciós cél alapvetően különbözik a klasszikus társadalomtudományi elemzések céljától, hiszen utóbbi célja elsősorban a leírás, magyarázat vagy egyfajta oksági hatás kimutatása. Ez a különbség nem csupán interpretációs eltérés: a használt modellekre is kihat. A predikciós modellek egy része ugyanis nem is használható magyarázati séma alapjaként, mert egyfajta fekete dobozként működik: tudjuk, hogy jól prediktál, de nem tudjuk, hogy miért prediktál így vagy úgy – szemben például a társadalomtudományi magyarázati modellekben klasszikusan használt regresszióval, amely a regressziós együtthatók révén fogódzót kínál a predikció mellett

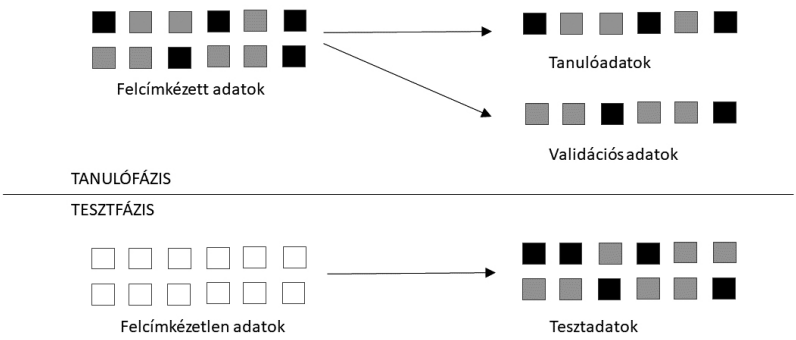
annak magyarázatára is. A predikció/magyarázat különbség a modellek alkalmazási logikájára is kihat, az előbbieknél a predikciós teljesítmény, az utóbbiaknál a modell statisztikai illeszkedése értékelendő. Az alábbiakban ezt részletezzük.

A predikciós modellt tekintve a rendelkezésre álló háttérinformáció szempontjából megkülönböztetünk felügyelt (supervised) és nem felügyelt (unsupervised) módszereket. Felügyelt esetben rendelkezünk bizonyos kategóriákkal, amelyeket előzetesen már ismerünk, és elemeink egy részéről tudjuk, hogy melyik kategóriába tartoznak (felcímkézett adathalmaz). A cél az, hogy új (felcímkézetlen) elemeket is be tudjunk illeszteni a kategóriarendszerbe, minél nagyobb pontossággal. Ha például szakértők egy újság valamennyi cikkét bekategorizálták témacsoportok szerint, akkor ezt az információt felhasználva megjósolhatjuk azt, hogy a be nem kategorizált cikkek melyik csoportba tartoznak. Ehhez valamely statisztikai eszköz segítségével létrehozunk egy előrejelző modellt, amely megpróbál a szakértők által hozott döntések mögötti nyelvi mintázatot találni, majd e mintázatokra alapozva automatizált módon hozzárendeli a témát a be nem kategorizált cikkekhez. Erre a modellezésre használhatjuk például a klasszikus módszertanból ismert logisztikus regressziót vagy döntési fákat, de akár valamilyen neurálisháló-alapú megközelítést is. Elsődleges fontosságú szempont, hogy tudnunk kell értékelni a modell előrejelző képességét azért, hogy (1) információnk legyen a jövőben várható predikciós képességéről, valamint (2) kiválaszthassuk a legjobb modellt.

A modell előrejelző képességének értékelése nehézségekbe ütközik, mivel az nem végezhető el egyszerűen az adott adatbázisunkra illesztett modell értékelésével, hiszen annak működése nem tükrözi hűen egy másik, külső adatbázison várható működését (nyilván azért, mert éppen az adott adatbázishoz leginkább illeszkedő modellként definiáltuk). Mivel általában nincs lehetőségünk egy külső adathalmazon tesztelni a modell sikerességét, ugyanezt a saját adatbázist kell használnunk értékelésre is, leggyakrabban az ún. keresztvalidálási logikával (ezt gyakran használják a klasszikus statisztikában máshol, például az orvostudományban, a diagnosztikus tesztek értékelésében). Ahogy azt az 1. ábra mutatja, a felcímkézett adathalmazt kettéosztjuk tanuló- és validációs adathalmazra (training/validation data set). A különböző modelleket a tanulómintán illesztjük (itt kapjuk meg például a logisztikus regresszió együtthatóit), majd a validációs mintán értékeljük a teljesítményt aszerint, hogy a korábban kapott együtthatókkal definiált logisztikus regressziós modell milyen pontossággal sorolja be helyesen az újságcikkeket. Ez a helyes besorolási arány a legegyszerűbb, pontosság (precision) elnevezésű teljesítményértékelési mutató. Ez a fázis az úgynevezett tanulófázis.

Ezután a következtetési (inference) fázisban a legjobbnak bizonyult modellt új újságcikkek besorolására használjuk anélkül, hogy ehhez humán szakértőt kellene igénybe vennünk – azaz fel nem címkézett újságcikkeket látunk el automatikusan kategóriacímkével.

1. ábra. *A felügyelt osztályozás logikája: a modellezés fázisai és adathalmazai (saját szerkesztés)*



A felügyelt módszerek közé tartozik a klasszikus társadalomkutatásban ismert minden regressziótípus (hiszen ismert a függő változó értéke), a döntési fa, a diszkriminanciaelemzés és az NLP-ben gyakran használt, de a klasszikus módszertanban nem szereplő Random Forest, Naive Bayes, Support Vector Machine (SVM) vagy a különböző neurális hálók is (e modellekről bővebben lásd Ignatow–Mihalcea 2016). Ezek különböző előnyökkel rendelkeznek, amelyek mindegyike a Big Data-feladatnak való megfelelőségre vezethető vissza: vannak köztük például nem paraméteresek, olyanok, amelyek nem igényelnek eloszlásfeltételeket, nem szorítják meg a függő és a független változók közötti függvénykapcsolatot, alkalmazhatók nagyon sok magyarázó változó esetén is.

A nem felügyelt módszerek esetén nincsen előzetes ismeretünk, nincsenek korábban felcímkézett eseteink. Klasszikus, a társadalomkutatók számára is ismert nem felügyelt módszer a klaszterelemzés: nincs előzetes információnk a klaszterek számáról, és egyetlen esetnek sem ismerjük a klaszter-hovatartozását. A nem felügyelt módszereket az NLP-ben hagyományosan többek között technikai célokra – például szinonimák felderítésére – használják, ám számos tartalmi modell is támaszkodik a módszerre, így például a később tárgyalt topikmodell is.

Evans és Aceves (2016) a felügyelt és nem felügyelt módszerek különbségét abban ragadja meg, hogy míg az előbbiek meglévő elméleteket vizsgálnak új adatokon, addig az utóbbiak ismeretlen mintázatok felderítését célozzák új elméletek létrehozása érdekében. Ez a nézőpont sok esetben helytálló, de nem általánosítható, hiszen például egy nem felügyelt klaszterelemzés már a releváns háttérváltozók kiválogatásánál igényel elméletet. Például a struktúrakutatásban a háttérváltozók kiválasztását az határozza meg, hogy elméletünk szerint milyen dimenziók mentén tagolódik a magyar társadalom, miközben a feltárt klaszterek nyilván újat adnak hozzá az elmélethez.

Potenciális kutatási kérdések

A továbbiakban az NLP társadalomtudományi szempontból releváns módszereit kutatási kérdésük szerint csoportosítva mutatjuk be. Ez a tárgyalásmód tehát a tartalomra fókuszál, míg az előző fejezet a kutatási logikára alapozott, így a két csoportosítás átfedi egymást: létezik például felügyelt és nem felügyelt szentimentelemzés.

Szentimentelemzés, emócióelemzés

Ezek a módszerek a szöveg szerzőjének egy tárggyal kapcsolatos álláspontját vizsgálják. Tipikusan a szerző attitűdjének, véleményének, értékelésének (szentimentelemzés) vagy a tárggyal kapcsolatos érzelmi megnyilvánulásának (emócióelemzés) megtalálását célozzák. A szentimentelemzés (másik, ritkábban használt magyar nyelvű kifejezéssel értékeléselemzés) a véleményeknek általában csak a polaritását (például negatív, pozitív, semleges) határozza meg, míg az emócióelemzés (lásd még érzelemelemzés) olyan alapérzelmeket különböztet meg, mint például a félelem, az undor, az öröm vagy a harag. Meghatározott számú kategóriába való besorolás a feladat, tehát egy osztályozási problémáról van szó.

A szentiment- és az emócióelemzés az üzleti szféra talán legnépszerűbb NLP-módszere, használják például a közösségi média hozzászólásaiban adott, a céggel vagy a cég bizonyos termékeivel, mozifilmekkel stb. kapcsolatos vélemények detektálására. Az utóbbi időben megjelentek társadalomtudományi alkalmazások is. Martins és szerzőtársai (2018) célja a gyűlöletbeszéd detektálása volt a közösségi média adataiban. Ez tipikusan felügyelt klasszifikációs feladat, vagyis a szakértők által korábban két kategóriába (gyűlöletbeszéd/nem gyűlöletbeszéd) sorolt szövegek alapján alkotnak osztályozó algoritmust. Martinsék újítása az volt, hogy a szövegek érzelmi töltetére vonatkozó információt is bevonták a modell magyarázó változói közé, ami így sokkal jobban teljesített.

Ezeken az elemzéseken belül is létezik felügyelt és nem felügyelt megoldás. Felügyelt esetre egyszerű példa, amikor a korábbi, újságcikk-besoroló példánk analógiájára megint felcímkézett adathalmazt hozunk létre úgy, hogy tesztalanyokat kérünk fel mozikritikáknak a három szentimentpolaritási típus egyikébe való besorolására, majd predikciós modellt hozunk létre a segítségükkel, amely a jövőben automatizáltan tudja monitorozni egy-egy új film fogadtatását.

A nem felügyelt elemzés legegyszerűbb megoldása a szótáralapú módszer. Ilyenkor az adott nyelvre érvényes szentiment-, illetve emóciószótárt használunk. A szótárak egy részében az egyes szavak polaritása van meghatározva, mint a 7 600 kifejezést tartalmazó Magyar Szentiment Lexikonban (Szabó 2014), ahol például a „megkukul" negatív, az „összefog" pozitív töltetű kifejezésként van besorolva. Más szótárakban a szavak szentimentje egy 0–100 pozitivitási skálán értékelt, ilyen lexikont használ pl. a hedonometer.org az USA Twitter-szentimentjének követésekor. A teljes szöveg szentimentjének/emóciójának meghatározása az azt alkotó szavak szótárból vett besorolásán alapul (bizonyos módszereknél a szavak környezetét, szófaját stb. is figyelembe véve), valamilyen aggregáló módszert alkalmazva.

A legújabb nem felügyelt megoldások mélytanulás-alapúak (deep learning), ilyen például dos Santos és Gatti szentimentelemzése (2014). Ennek előnye, hogy a feladat szempontjából releváns komplex mintázatok felismerését támogatja. Ezek a modellek már nemcsak a szavak szintjén vizsgálódnak, hanem a mondat szintjét is figyelembe veszik az elemzés során, ami például a tagadó szerkezetek felismerését támogatja. A kutatási gyakorlatban egyre elterjedtebbé válnak azok a megoldások is, amelyek terület- és időspecifikus szótárak építését támogatják. Jó példát nyújt erre Rice és Zorn (2019) tanulmánya. A szerzők munkájukban bemutatják, hogyan változott egyes szavak időbeli érzelmi töltete, vagy mennyiben különbözik ugyanannak a szónak az érzelmi töltete különböző területeken.

A látens tartalom elemzése

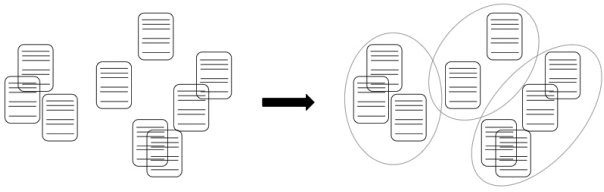
A tanulmány korábbi részeiben elsősorban klasszifikációs problémákat tárgyaltunk egyrészt népszerűségük miatt, másrészt azért, mert könnyen érthetőek, hiszen a kapcsolódó kutatási kérdés (besorolás) és módszer (például regresszió) a klasszikus kvantitatív társadalomkutatási módszerek használói számára is ismerős. A klasszifikáción túl egy másik általános, a társadalomkutatás számára fontos potenciált jelentő kutatási kérdés a szövegek látens tartalmának kinyerése. Ennek három legismertebb módszere a klaszterelemzés, a topikmodell és a szóbeágyazási modell.

Klaszterelemzés

A klaszterelemzés nem egyetlen modell, hanem osztályozási algoritmusok összessége, amelyeket a kutatási kérdés köt össze: bizonyos jellemzők szerinti hasonlóságukkülönbözőségük alapján sorolják be csoportokba a vizsgált egyedeket (itt: szövegeket). Jól ismert módszer a társadalomtudományokban is, így például a struktúrakutatásban hagyományosan klaszterelemzés segítségével detektálunk státuszcsoportjellegű rétegeket, leggyakrabban gazdasági-kulturális jellemzőik alapján. Nincsenek a priori ismert kategóriák, csoportok, sőt általában a csoportok száma sem ismert előzetesen. A csoportképzés magából az adatbázis sajátosságaiból indul ki, és a jellemzőik szerint hasonló szövegek kerülnek ugyanazon csoportba. A szövegek közötti hasonlóságot legegyszerűbb esetben az általuk tartalmazott szavak alapján (ismét szózsákmodellt alkalmazva) definiálhatjuk. Ennek megfelelően a klaszterelemzés nem felügyelt logikát követ (megjegyezzük, hogy a priori ismert csoportok esetén léteznek felügyelt klaszterezési eljárások is). Mivel társadalomtudós olvasóink között vélhetően sokan ismerik a klaszterelemzés vektortér-reprezentációját, röviden utalunk rá, hogy a szövegek klaszterezése is hasonlóan történik, egy olyan sokdimenziós vektortérben, ahol a tengelyek az egyes szavaknak felelnek meg, míg a vektortérbeli pontok a szövegeknek, amelyek elhelyezését az határozza meg, hogy az adott szó hányszor szerepel a szövegben. A szövegek/dokumentumok klaszterezése e vektortérben a távolságuk/közelségük által meghatározott csoportokon alapul

(lásd a 2. ábrát), a lehető leghomogénebb csoportok létrehozására törekszünk, amelyek egymástól a lehető legnagyobb távolságra vannak.

2. ábra. *Dokumentumok klaszterelemzése vektortérbeli elhelyezkedésük alapján*



A szövegbányászati alkalmazásokban a klaszterelemzés célja a szöveghalmazunkban rejlő struktúra felismerése, a dokumentumok csoportokba rendezése. Üzleti felhasználásai közül például az ajánlórendszerek említhetők, amikor az adott felhasználónak kiajánlott terméket a vele egy klaszterben levő felhasználók fogyasztási preferenciáira építjük.

Topikmodellek

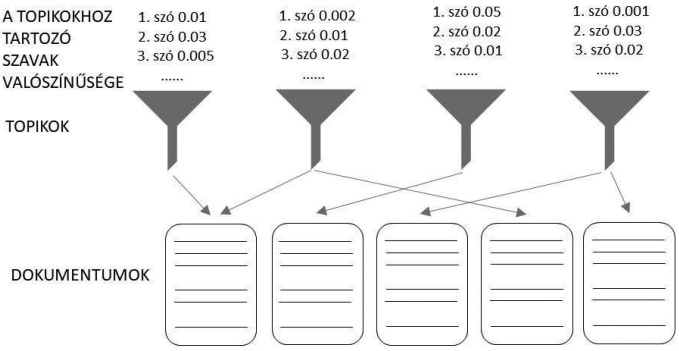
A topikmodellek (Blei–Lafferty 2009) olyan automatizált eljárások, amelyek célja dokumentumgyűjtemények (például valamely online média cikkei) témáinak azonosítása. A modell mögött intuitíve az a megfontolás áll, hogy létezik a témáknak egy véges halmaza, amelyben a témák statisztikailag az adott nyelv kifejezésein értelmezett valószínűségeloszlásként definiálhatók. Például a sportról szóló cikkekben a „győztes" szó előfordulásának nagyobb a valószínűsége, mint az „infláció" szóénak, míg a gazdasági témájú cikkekben ez éppen fordítva van. A dokumentumok néhány topik keverékeként azonosíthatók, például egy stadionépítésről szóló cikk 80%-ban gazdasági, míg 20%-ban sporttémát dolgozhat fel. Ezek a mögöttes topikok aztán a hozzájuk tartozó szó-valószínűségeloszláshoz igazodva generálják a dokumentumokat (lásd a 3. ábrát). Itt is szózsákmodellt alkalmazunk, a dokumentumoknak csak a szógyakoriság-eloszlását vizsgáljuk. Ez a megközelítés a survey-logikán belül leginkább a modellalapú klaszterezésekhez van közel, elsősorban a véges kevert modell illesztésekhez (más néven látens profilelemzés).

A topikok száma és tartalma a priori nem ismert, tehát ez is egy nem felügyelt módszer. Akárcsak a k-közép klaszterezés esetén, itt is a modell bemenő paramétere a topikok száma (K), és az optimális topikszám megválasztása többféle K mellett illesztett modell közül a „legjobb" modell[10] kiválasztásán alapul. A modell interpretációjában a K értéke mellett a topikok valószínűségét (népszerűségét), az egyes

10 A „legjobb" modell kiválasztása minden NLP-módszer, de leginkább a nem felügyelt módszerek esetén (amilyen a topikmodell is) nagy kihívás, és többfajta megoldási javaslat létezik.

topikokhoz tartozó szóeloszlást (például a legvalószínűbb tíz szó listáját), illetve a topikok „keveredési" hajlandóságát értelmezzük.

3. ábra. *Dokumentumok létrejötte a topikmodell feltevései szerint*



A topikmodellek az utóbbi években gyorsan fejlődtek. Blei, Ng és Jordan (2003) írt először a látens Dirichlet-allokációról (Latent Dirichlet Allocation, LDA), amely az egyik legismertebb topikmodellezési eljárás. Az elnevezése a modell azon matematikai feltevéséből ered, hogy a dokumentumonkénti topikeloszlás Dirichlet-eloszlást követ. Az eloszlást úgy hangolják, hogy a topikok keveredését minimalizálják. A topikok szöveggeneráló mechanizmusára vonatkozó előfeltevések mint megszorítások mellett a modellünkre bízzuk, hogy a szövegekben jellemző tartalmi struktúrákat találjanak. Jó példa erre Barna és Knap (2019) tanulmánya. A szerzők topikmodellel elemezték, hogy a kuruc.info „zsidó" szót tartalmazó cikkei tartalmilag tipikusan milyen témákra/topikokra bonthatók, így azonosították pl. a faji alapú, a vallási alapú vagy az összeesküvésre épülő topikokat.

A fenti modellnek több kiterjesztése létezik, ezek közé tartoznak például a korrelált topikmodellek (correlated topic models), amelyek a topikok szövegbeli együttes előfordulását modellezik (Blei–Lafferty 2007), a dinamikus topikmodellek (dynamic topic models), amelyek a topikok időbeli változását vizsgálják (Blei–Lafferty 2006), illetve a szerző-topikmodellek (author topic model), amelyek a szövegre vonatkozó metaadattal (például szerzőségi információval) egészítik ki az alapmodellt (Rosen-Zvi et al. 2008). A topikmodellezés a társadalomtudományban széles felhasználási spektrummal rendelkezhet. Elemezhetjük bármely, a digitális társadalmi térben megjelenő csoport (adott politikai platformhoz tartozók, adott betegségben szenvedők, adott tudományos folyóirat szerzői) megnyilvánulásainak tematizációját, a témák népszerűségváltozását, a témák tartalmának változását stb. A hagyományos kérdőíves módszert alkalmazó kutatóknak is segítségére lehet a kérdőív nyílt kérdéseire adott válaszok elemzésében.

Szóbeágyazási modellek

A szóbeágyazási modellek (word embedding models) a vizsgált korpusz látens szemantikai struktúrájának reprezentálására szolgáló, a gépi tanulásban elterjedt neurális hálókat használó módszerek. Néhány éves múltra tekintenek vissza, népszerűségük nagy ütemben nő. Többféle technikai megvalósításuk létezik (Mikolov és munkatársai [2013] word2vec modellje az első ezek közül), az alábbiakban intuitív lényegüket igyekszünk bemutatni.

A modell, leegyszerűsítve, a korpuszunk szavainak vektortér-reprezentációját adja, ahol a vektortérben egy vektor egy szónak felel meg (lásd a 4. ábra leegyszerűsített háromdimenziós terét), a szavak elhelyezkedését pedig a jelentésük határozza meg. Az egymáshoz közel eső szavaknak a jelentése is közel esik egymáshoz. Itt a szójelentés a szó használatával azonosított fogalom, konkrétabban a szavak mondatbeli előfordulásának szűk környezetét (általában a szót megelőző, illetve követő 3-10 szót) veszi figyelembe a modell. Aszerint kerül közel vagy távol egymástól két szó a vektortérben, hogy mennyire egyezik meg ez a környezet a korpuszunkban. A környezeten belül nem vizsgál sorrendiséget, egyszerű szózsákként kezeli azt. A vektortér dimenzióját, amely általában néhány száz körül van, az határozza meg, hogy a mögöttes neurálisháló-modell hány dimenzióban képes elég jól reprodukálni az eredeti használati környezeteket – vagyis tulajdonképpen egy dimenziócsökkentő eljárásról van szó, amely az eredeti komplex teret, ahol minden szó önálló dimenziót jelenít meg, egy kisebb, néhány száz dimenziós térbe ágyazza bele. A dimenzióknak, tengelyeknek nincs közvetlen interpretálhatósága.

A vektortérben két szó jelentési közelségét a nekik megfelelő vektorok által bezárt szög nagysága (pontosabban általában annak koszinusza) segítségével határozzák meg. Lásd a 3. ábrát: a „matek" és a „matematika" szó egészen kis szöget zár be, hiszen a használati szövegkörnyezetük csaknem megegyezik, és a „fizika" is közel van hozzájuk (közelebb a „matematika" szóhoz), hiszen sokszor szerepelnek hasonló környezetben. Fontos hangsúlyozni, hogy e használatalapú definícióból következően a vektortér nem a szavak jelentésének hasonlóságán, hanem a szavak jelentésének kapcsolatán alapszik. A „férfi" és a „nő" szó például nincs nagyon nagy távolságra egymástól, hiszen gyakran szerepelnek hasonló környezetben a mondatainkban. Ennek a „disztribúciós szemantikának" a lényege, hogy a szójelentést nem a szavak és a „valóság" elemeinek kapcsolatára alapozza, hanem azt a szavak használatával azonosítja; a megközelítés legkorábbi nyelvfilozófiai reprezentánsai Wittgenstein 1930-as évekbeli munkái.

A modell által létrehozott vektortér (Garg és munkatársai [2018] példáival) képes tehát szemantikai kapcsolatok megragadására (a foglalkozások neve például közel van egymáshoz), jól elkülöníthetők benne jellegzetes szócsoportok (a női jellegű foglalkozások, mint a „táncos" és a „háziasszony" elkülönülnek a férfiasaktól, mint az „ács" és a „mérnök"). Ennek a vektortérnek önmagában is van szociológiai relevanciája: Magu és Luo (2018) Twitter-szövegeken végzett gyűlöletbeszéd-kutatásá-

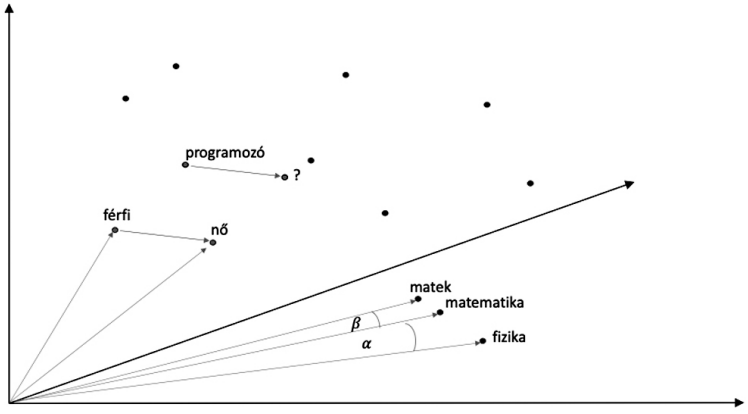
ban például előítélettel sújtott csoportokat (feketéket, zsidókat, mexikóiakat) jelölő eufemisztikus kódszavakat azonosított más, ismert kódszavakhoz való vektortérbeli közelségük alapján.

De a közelség-távolság értelmezésénél is továbbmehetünk. A vektortér jól teljesít analógiás teszteken is, például a főváros relációt egy ország és fővárosa különbségével megadva választ kaphatunk a „Mi Oroszország fővárosa?" kérdésre:

Ilyenkor a megfelelő vektortérelemeket tekintve Moszkva úgy adódik, hogy megkeressük az egyenlet fenti megoldásához legközelebbi vektort. Elég nagy (több százmillió szót tartalmazó) korpuszokon tanítva a modell jól működő vektorteret ad ilyen analógiás teszteken (vagyis az egyenlet jobb oldalához legközelebbi vektor valóban Moszkva lesz). A vektortér teljesítményét általában is ilyen analógiás teszteken értékelik, léteznek több száz kérdésből álló tesztbázisok, lásd például Kozlowski et al. 2018.

A fenti analógiák nem csak technikai példákon működtethetők: feltehető például a kérdés, hogy van-e az általunk vizsgált korpuszban a foglalkozásneveknek a nemi különbségeket reprezentáló nyelvhasználati különbsége. Ehhez például fel kell tennünk a kérdést, hogy mi az a foglalkozás, amit úgy kapunk a „programozóból", hogy ugyanolyan irányba és távolságra toljuk el, mint amilyen eltolással a „férfiból" a „nőt" kapjuk (lásd a 3. ábrát). Vajon például az „adminisztrátort" kapjuk-e? Vagyis vajon van-e a vizsgált korpuszban olyan nyelvhasználati jelleg, amely a programozókat inkább a férfiakhoz, az adminisztrátort inkább a nőkhöz kapcsolja? A társadalomtudós olvasó számára itt már vélhetően nyilvánvaló, hogy ezek a szóbeágyazási modellek figyelemre méltó kulturális-társadalmi tudást tartalmaznak. Garg és munkatársai (2018), valamint Kozlowski és munkatársai (2018) például amerikai korpuszokat vizsgálva több mint 100 éves intervallumon státuszalapú, társadalmi nemi szerepekkel kapcsolatos és kulturális trendeket volt képes detektálni a módszer használatával.

4. ábra. A szavak jelentésbeli hasonlósága, illetve jelentésanalógiák megjelenése a szóbeágyazási modell vektortér-reprezentációjában



Összefoglalás

A társadalomtudomány folyamatosan változik és fejlődik. Ez egyrészről új témák vizsgálatát és klasszikus témák újrafeldolgozását jelenti, de ugyanúgy új módszerek felfedezését és integrálását is a társadalomkutatási kánonba. Az elmúlt pár évtized gyökeresen megváltoztatta a társadalomkutatók által felhasználható módszertani és statisztikai apparátust. A 30-40 éve még csak a „kiválasztottak" által alkalmazott nagy számításigényű statisztikai számolások a 90-es években egyre inkább mindenki számára elérhetővé váltak az asztali számítógépek elterjedésével és a statisztikai szoftverek kommercializálódásával. A digitalizáció továbbgyűrűzése pedig tovább erősítette ezt a folyamatot, a társadalomtudományon belül (is) egyre inkább teret nyer(t) a Big Data paradigma. A szövegelemzés, a szövegbányászat elsősorban a növekvő digitalizáció miatt válik egyre fontosabb vizsgálati tereppé.

A tanulmányban arra koncentráltunk, hogy a megközelítés milyen lehetőségeket teremt a társadalomtudomány számára. Az írott szövegek fontos lenyomatai az emberek gondolkodásának. Ezért azokban a kutatási témákban, amelyekben elsősorban a (köz)gondolkodásnak, az emberi viselkedésnek a megértése a cél, nagy hasznot hozhat a kvantitatív szövegelemzés. Diszkrimináció, gyűlöletbeszéd, egyenlőtlenségek – mind-mind olyan témák, amelyek sokszor akár rejtetten, de megjelennek az írott kommunikációban is. De a kultúra-, a szabadidő-, a zeneszociológia is sokat profitálhat a szövegelemzésen alapuló kutatásokból. Nem véletlen, hogy a kultúraszociológia egyik vezető lapja, a *Poetics* már 2013-ban különszámot szentelt a topikmodellezésnek (lásd többek között McFarland et al. 2013). De az idődimenzió is megjeleníthető: a korábban már hivatkozott tanulmányok (Garg et al. 2018; Kozlowski et al. 2018) 100 év távlatában elemeznek társadalomtudományi tendenciákat igen nagy sikerrel. Az ehhez hasonló történeti vizsgálatok tere is kiszélesedőben van, hiszen nemcsak a „born-digital", hanem a digitalizált szövegtárak spektruma is folyamatosan nő. Hazai példa erre a Kádár-korszak vizsgálata a *Pártélet* című újságon keresztül (Szabó et al. 2019). Ezeket a munkákat támogatják az olyan projektek, mint a Google Books Library óriási volumenű vállalkozása, amely az 1500-as évektől digitalizálja az (elsősorban angol nyelvű) könyveket. Magyarországról az *Arcanum* folyóirattára említhető meg ennek kapcsán.

A szövegelemzési módszerek nem helyettesítői, hanem kiegészítői a klasszikus kvantitatív társadalomkutatási eszközöknek. A kérdőíves vizsgálatok precízen kidolgozott módszertana alapvetően megbízható és érvényes tudást nyújt. A kvantitatív szövegelemzés viszont képes lehet olyan kérdések megválaszolására is, amelyek a survey-ekből nem vagy csak nagyon nehezen és közvetve megválaszolhatók. Ezért mi a módszer jelentős társadalomtudományi felfutását várjuk a közeljövőben. Ugyanakkor – ahogy írásunk is mutatta – a módszer belépési küszöbe viszonylag magas, új módszertani/programozói tudást igényel. Erre vagy az empirikus tudás önképzés útján való megszerzése (és az új társadalomtudós-generációk számára a társadalomtudományi képzésekbe való integrálása), vagy interdiszciplináris kuta-

tócsoportok alakítása lehet a válasz. Tanulmányunkkal ezekhez a komoly tudományos befektetést igénylő döntésekhez próbáltunk érveket szolgáltatni.

Abstract: In our paper, we present an overview of Natural Language Processing (NLP) methods, which developed parallel with the spread of 'Big Data' paradigm. We present the most promising methods for social sciences, the specific research questions they can answer and the methodological features that distinguish them from classic quantitative methods. These methods go far beyond classic quantitative text analysis based on simple word frequencies. Their modelling logic arises from machine learning methods; hence, it is substantially differing from the classic social science logic that seeks for explanation and casual effects. Our goal is to inspire Hungarian social scientists by providing an insight into a less-institutionalized area, since we believe that at an international level, text mining will be a standard method for empirical social science research within a few years.

Keywords: quantitative text analysis, natural language processing, textmining, computational text analysis

Irodalom

Aggarwal, C. C. – Zhai, C. X. (eds.) (2012): *Mining Text Data*. New York: Springer.

Anderson, C. (2008): The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired*. https://www.wired.com/2008/06/pb-theory/. (letöltés: 2019. március 27.)

Bales, R. F. (1950): A set of categories for the analysis of small group interaction. *American Sociological Review*, 15(2): 257–263.

Barna, I. – Knap, Á. (2019): Antisemitism in Contemporary Hungary: Exploring Topics of Antisemitism in the Far-Right Media Using Natural Language Processing. *Theo Web – Academic Journal of Religious Education*, 18(1): 75–92.

Berelson, B. – Lazarsfeld, P. F. (1948): *The Analysis of Communication Content*. Oslo: Univ. Stud.

Blei, D. M. – Lafferty, J. D. (2006): Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*. New York: Association for Computing Machinery, 113–120.

Blei, D. M. – Lafferty, J. D. (2007): A correlated topic model of science. *The Annals of Applied Statistics*, 1(1): 17–35.

Blei, D. M. – Lafferty, J. D. (2009): Topic models. In Srivastava, A. N. – Sahami, M. (eds.) *Text mining: Classication, Clustering, and Applications*. London: Chapman & Hall/CRC Data Mining and Knowledge Discovery Series.

Blei, D. M. – Ng, A.Y. – Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3(4–5): 993–1022.

Csepeli Gy. (2015): A szociológia és a Big Data. *Replika*, 92–93(2015/3–4): 169–174.

Denny, M. J. – Spirling, A. (2018): Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. *Political Analysis*, 26(2). doi:10.1017/pan.2017.44

Dessewffy T. – Láng L. (2015): Big Data és a társadalomtudományok véletlen találkozása a műtőasztalon. *Replika*, 92–93 (2015/3–4): 155–168.

Digital 2020 reports, Hootsuite, wearesocial.com/digital-2020

Evans, J. A. – Aceves, P. (2016): Machine translation: mining text for social theory. *Annual Review of Sociology*, 42(1): 21–50.

Dos Santos, C. – Gatti M. (2014): Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*. Dublin: Dublin City University and Association for Computational Linguistics. https://www.aclweb.org/anthology/C14-1008 (letöltés: 2019. január 18.)

Garg, N. – Schiebinger, L. – Jurafsky, D. – Zou, J. (2018): Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences of the United States of America*, 115(16): 3635–3644.

Hays, D. C. (1960): *Automatic Content Analysis*. Santa Monica: Rand Corp.

Hirschberg, J. – Manning, C. D. (2015): Advances in natural language processing. *Science*, 349(6245): 261–266. doi: 10.1126/science.aaa8685.

Ignatow, G. – Mihalcea, R. (2016): *Text Mining. A Guidebook for the Social Sciences*. Thousand Oaks, CA: Sage Publications.

Indig B. (2018): Közös crawlnak is egy korpusz a vége – Korpuszépítés a CommonCrawl .hu domainjából. In Vincze V. (szerk.) *XIV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2018)*. Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport, 125–134.

Kmetty Z. (2018): A szociológia helye a Big Data-paradigmában és a Big Data helye a szociológiában. *Magyar Tudomány*, 179(5): 683–692.

Kozlowski, A. C. – Taddy, M. – Evans, J. A. (2018): The Geometry of Culture: Analyzing Meaning through Word Embeddings. *American Sociological Review*, 84(5): 905–949. arXiv preprint arXiv:1803.09288 https://arxiv.org/abs/1803.09288 (letöltés: 2019. január 18.)

Liu, B. (2015): *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge: Cambridge University Press.

Magu, R. – Luo, J. (2018): Determining Code Words in Euphemistic Hate Speech Using Word Embedding Networks. In *Proceedings of the Second Workshop on Abusive Language Online*, Brussels: Association for Computational Linguistics, 93–100.

Martins, R. – Gomes, M. – Almeida, J. – Novais, P. – Henriques, P. (2018): Hate Speech Classification in Social Media Using Emotional Analysis. In *7th Brazilian Conference on Intelligent Systems (BRACIS)*. Sao Paulo: IEEE, 61–66. doi: 10.1109/BRACIS.2018.00019.

McFarland, D. A. – Ramage, D. – Chuang, J. – Heer, J. – Manning, C. D. – Jurafsky, D. (2013): Differentiating language usage through topic models. *Poetics*, 41(6): 607–625.

Mikolov, T. – Sutskever, I. – Chen, K. – Corrado, G. S. – Dean, J. (2013): Distributed Representations of Words and Phrases and their Compositionality. Burges, C. J. C. – Bottou, L. – Welling, M. – Ghahramani, Z. – Weinberger, K. Q. (eds.) *Proceedings of the Conference on Advances in Neural Information Processing Systems 26 (NIPS)*. La Jolla: Neural Information Processing Systems Foundation, 3136–3144.

Miner, G. – Elder, J. – Hill, T. – Nisbet, R. – Delen, D. – Fast, A. (2012): *Practical text mining and statistical analysis for non-structured text data applications*. Waltham: Academic Press of Elsevier.

Moreno, S. A. – Redondo, T. (2016): Text Analytics: the convergence of Big Data and Artificial Intelligence. *International Journal of Interactive Multimedia and Artificial Inteligence*, 3(6): 57–64. doi: 10.9781/ijimai.2016.369

Moretti, F. (2013): *Distant Reading*. London: Verso.

Németh R. (2015): A számok tényleg magukért beszélnek? *Replika*, 92–93(2015/3–4): 203–209.

Rehurek, R. (2011): *Scalability of Semanic Analysis in Natural Language Processing*. Dissertation. Brno: Faculty of Informatics, Masaryk University. https://radimrehurek.com/phd_rehurek.pdf (letöltés: 2019. január 18.)

Rice, D. R. – Zorn, C. (2019): Corpus-based dictionaries for sentiment analysis of specialized vocabularies. *Political Science Research and Methods*, 1–16.

Rosen-Zvi, M. – Chemudugunta, C. – Griffith, T. – Smyth, P. – Steyvers, M. (2008): Learning Author-Topic Models from Text Corpora. In *ACM Transactions on Information Systems*, 28(1): 1–38. http://psiexp.ss.uci.edu/research/papers/AT_tois.pdf (letöltés: 2019. január 18.)

Sebők M. (szerk.) (2016): *Kvantitatív szövegelemzés és szövegbányászat a politikatudományban*. Budapest: L'Harmattan Kiadó.

Szabó M. K. (2014): Egy magyar nyelvű szentimentlexikon létrehozásának tapasztalatai. In *„Nyelv, kultúra, társadalom" konferencia*, Budapest.

Szabó, M. K. – Berend, G. – Kiss, L. – Ring, O. – Vidács, L. – Kmetty, Z. (2019): Mapping the dynamic change of the concept "industry" and "agriculture" in the Hungarian socialist era using a word embedding model. In *2nd Annual POLTEXT Conference*, Tokyo.

Tikk, D. (szerk.) (2007): *Szövegbányászat*. Budapest: Typotex Kiadó.

Vijayarani, S. – Ilamathi, M. J. – Nithya, M. (2016): Preprocessing Techniques for Text Mining - An Overview. *International Journal of Computer Science and Communication Networks*, 5(1), 7-16.