Statisztikai Szemle

Közzététel: 2022. február 17.

A tanulmány címe:

Szóbeágyazási vektortérmodellek társadalomtudományi alkalmazása

Szerző:

KMETTY ZOLTÁN,
az Eötvös Loránd Tudományegyetem docense,
az Eötvös Loránd Kutatóhálózat Társadalomtudományi Kutatóközpont
CSS-Recens Kutatócsoportjának tudományos főmunkatársa
E-mail: kmetty.zoltan@tk.hu

DOI: https://doi.org/10.20311/stat2022.2.hu0105

Az alábbi feltételek érvényesek minden, a Központi Statisztikai Hivatal (a továbbiakban: KSH) *Statisztikai Szemle* c. folyóiratában (a továbbiakban: Folyóirat) megjelenő tanulmányra. Felhasználó a tanulmány vagy annak részei felhasználásával egyidejűleg tudomásul veszi a jelen dokumentumban foglalt felhasználási feltételeket, és azokat magára nézve kötelezőnek fogadja el. Tudomásul veszi, hogy a jelen feltételek megszegéséből eredő valamennyi kárért felelősséggel tartozik.

1. A jogszabályi tartalom kivételével a tanulmányok a szerzői jogról szóló 1999. évi LXXVI. törvény (Szjt.) szerint szerzői műnek minősülnek. A szerzői jog jogosultja a KSH.

2. A KSH földrajzi és időbeli korlátozás nélküli, nem kizárólagos, nem átadható, térítésmentes felhasználási jogot biztosít a Felhasználó részére a tanulmány vonatkozásában.

3. A felhasználási jog keretében a Felhasználó jogosult a tanulmány:

a) oktatási és kutatási célú felhasználására (nyilvánosságra hozatalára és továbbítására a 4. pontban foglalt kivétellel) a Folyóirat és a szerző(k) feltüntetésével;

b) tartalmáról összefoglaló készítésére az írott és az elektronikus médiában a Folyóirat és a szerző(k) feltüntetésével;

c) részletének idézésére – az átvevő mű jellege és célja által indokolt terjedelemben és az eredetihez híven – a forrás, valamint az ott megjelölt szerző(k) megnevezésével.

4. A Felhasználó nem jogosult a tanulmány továbbértékesítésére, haszonszerzési célú felhasználására. Ez a korlátozás nem érinti a tanulmány felhasználásával előállított, de az Szjt. szerint önálló szerzői műnek minősülő mű ilyen célú felhasználását.

5. A tanulmány átdolgozása, újra publikálása tilos.

6. A 3. a)–c.) pontban foglaltak alapján a Folyóiratot és a szerző(ke)t az alábbiak szerint kell feltüntetni:

„*Forrás*: *Statisztikai Szemle* c. folyóirat 100. évfolyam 2. számában megjelent, ***Kmetty Zoltán*** által írt, **'Szóbeágyazási vektortérmodellek társadalomtudományi alkalmazása'** című tanulmány (link csatolása)"

7. A Folyóiratban megjelenő tanulmányok kutatói véleményeket tükröznek, amelyek nem esnek szükségképpen egybe a KSH vagy a szerzők által képviselt intézmények hivatalos álláspontjával.

Kmetty Zoltán

Szóbeágyazási vektortérmodellek társadalomtudományi alkalmazása

How to use vector space models in social sciences

KMETTY ZOLTÁN,
az Eötvös Loránd Tudományegyetem docense,
az Eötvös Loránd Kutatóhálózat Társadalomtudományi Kutatóközpont
CSS-Recens Kutatócsoportjának tudományos főmunkatársa
E-mail: kmetty.zoltan@tk.hu

A tanulmány azt mutatja be, hogy miként használhatók a vektortérmodellek a társadalomtudományi kutatásokban. Az első két fejezet a módszer alapvetéseit és a történeti evolúcióját tárgyalja, illetve rövid áttekintést ad a matematikai hátteréről és az alkalmazását övező dilemmákról. Részletesen ismerteti a word2vec-, fastText- és GloVe-algoritmusokat, valamint a kontextuális szóbeágyazási modelleket. A harmadik fejezet példákkal illusztrálja a vektortérmodellekből kapott eredmények feldolgozásának lehetőségeit, az utolsó pedig arra fókuszál, hogy a társadalomtudományok milyen módon kapcsolódnak a bemutatott módszerhez.

TÁRGYSZÓ: szóbeágyazás, vektortérmodell, társadalomtudomány

This study demonstrates how vector space models can be used in social science research. Presenting the principles and historical development of this method, the first two parts give a brief overview of its mathematical background and the dilemmas concerning its application. The word2vec, fastText, and GloVe sub-algorithms and the contextual word embedding models are explained in detail. The third part illustrates the different ways of processing the results obtained from vector space models through examples. The final part focuses on how the social sciences relate to the method presented.

KEYWORD: word embedding, vector-space model, social science

A társadalomtudományok empirikus eszköztára talán sosem bővült annyira intenzíven, mint az elmúlt 20-30 évben. A számítógépes kapacitás (mind a tárhely, mind a sebesség) növekedése és a különböző programnyelvek használatának egyre szélesebb körű elterjedése lehetővé tette a korábban nem vagy csak nagyon nehezen megvalósítható módszerek ismertebbé válását. Ezzel párhuzamosan a kvantitatív elemzésekben egyre több új típusú adatforrás jelent meg. Az újfajta adatokat tekintve a kapcsolathálózati adatbázisok és a társadalmi hálózatelemzés jelentette az első hullámot, a szöveges adatok vizsgálata pedig a másodikat.[1] Bár e kettő közül egyik sem az elmúlt 30 évben kezdődött, az igazi társadalomtudományi felfutást mégis ebben az időszakban lehetett megfigyelni – a hálózati (network-) módszereknél az 1990-es, a szöveges adatok vizsgálata esetében pedig a 2010-es évektől. Népszerűbbé válásukat részben a már korábban említett számítógépes „forradalom" táplálta, részben pedig az azzal erősen összefüggő digitalizáció (*Kmetty* [2018]).

Az új típusú adathalmazok nagyban különböznek a tradicionális – elsősorban survey- vagy adminisztratív alapú adathalmazoktól[2] –, ami új elemzési módszereket indukált. A tanulmány fókuszában álló szöveges adatok feldolgozására és elemzésére speciális technikák alakultak ki. Ezt a módszertani irányt a természetes nyelvfeldolgozás (natural language processing, NLP), illetve a szövegbányászat (text mining) címkéjével látták el.

Se az NLP, se a szövegbányászat nem a társadalomkutatások oldaláról indult. Alapvetően számítógépes nyelvészek, statisztikusok, számítástudománnyal foglalkozók (computer scientists) és részben fizikusok, illetve más természettudósok álltak a módszertani fejlesztések élén az 1990-es és 2000-es években. A felhasználásban manapság is dominál az üzleti vonal, így nem véletlen, hogy a Google és a Facebook élen jár az algoritmusok fejlesztésében (lásd később). A tisztán nyelvészeti és üzleti felhasználás (például információkinyerés, spamklasszikfikáció stb.) mellett a már említett módszerek a társadalomtudományi kutatások előtt is új perspektívát nyitottak, ami rövid idő alatt egy olyan új interdiszciplináris tudományterület kialakulásához vezetett, amely számítógépes társadalomtudomány (computational social science, CSS) néven kezdte el intézményesülési útját. A folyamatot a tudományterülethez köthető kutatócsoportok, egyetemi szakok megjelenése és konferenciák megrendezése jelezte a 2010-es években (annak is elsősorban a második felében). Bár a szöveges

[1] A kettő természetesen össze is kapcsolódhat, szöveges adatokat is lehet elemezni network-módszerekkel.

[2] A területi és időbeli adatok elemzéséhez szintén kialakult egy speciális módszertan, de a megközelítésmódok közötti különbség itt nem olyan nagy, mint a network- vagy a szövegelemzés esetében.

adatokon keresztül jutottunk el a CSS fogalmáig, ez utóbbi jóval szélesebb adatkört takar – a szakterülethez köthetjük a digitális adatforrásokat függetlenül azok tartalmától, de a különböző szimulációs területeket (például ágensalapú szimulációt) is. Míg a numerikus adatok elemzéséhez nem kellett új eszköztárat építeni, „csak" a régit újrahangolni és kiegészíteni (*Kmetty* [2018]), a szöveges adatok elemzéséhez teljesen új eszközökre volt szükség (*Németh–Katona–Kmetty* [2020]). Jelen tanulmányban az utóbbiak egy elemét, a szóbeágyazási vektortérmodelleket (word-embedding model) fogom mélyebben bemutatni. Egy nagyon gyorsan fejlődő terület esetén persze kérdés, mit nevezhetünk újnak. Ahogy érzékelhető lesz majd a módszer bemutatásánál, az előzmények már az 1980-as évek végén megjelentek, de a robbanásszerű elterjedés a 2010-es évek elejére tehető, alapvetően egy jól implementálható, neurálisháló-alapú megközelítés megjelenéséhez kötődve (*Mikolov et al.* [2013]).

De mit is nevezünk szóbeágyazási vektortérmodellnek? Ez gyakorlatilag egy dimenziócsökkentő eljárás, amellyel a célunk annak meghatározása, hogy az adott szó milyen környezetben, milyen kontextusban szerepel. Dimenziócsökkentésre azért van szükség, mert a szavak egy szövegben rengeteg más szó mellett fordulhatnak elő, és a nyers szógyakoriság nem tud képet adni e kontextusról.

Tanulmányomban azt mutatom be, hogy a vektortérmodellek miként használhatók a társadalomtudományi kutatásokban. Először a módszer alapvetéseit és a történeti evolúcióját ismertetem, majd röviden áttekintem a matematikai hátterét és az alkalmazását övező dilemmákat. Kitérek arra is, hogy e modellek mit mérnek, és milyen jellemző társadalomtudományi alkalmazásaik találhatók a nemzetközi szakirodalomban. A dolgozat nemcsak a vektortérmodellek társadalomtudományi használatát tárgyalja, de lehetőséget ad az érdeklődőknek arra is, hogy maguk is kipróbálják a módszert. Ennek céljából a vektortérmodellek használatát tárgyaló fejezetben bemutatott példákhoz elérhetővé teszem a futtatásukhoz szükséges R kódokat.[3]

1. Vektortérmodellek – statisztikai alapok, módszertani megfontolások

1.1. A vektortérmodellek evolúciója

1.1.1. Kiindulás

A tanulmány középpontjában egy, a 2010-es évtizedben felfutó, több tudományterületen is használható statisztikai elemzési eljárás, a szóbeágyazás áll. Alaplogikájának megértéséhez érdemes visszalépni a kvantitatív szövegelemzés alapjaihoz.

[3] https://github.com/zkmetty/nlp

Ha meg akarjuk érteni, hogy egy nagy szöveghalmaz milyen tartalommal rendelkezik, akkor erre a legegyszerűbb megoldás a szövegben előforduló leggyakoribb szavak vizsgálata. Az általánosan megjelenő szavakon (például a névelőkön) „túljutva", rövid idő után meghatározhatók azok a szavak, amelyek már jelzést tudnak adni a szöveg tartalmáról. Ez a megközelítés viszonylag egyszerű vizsgálatokat eredményez, inkább kiindulópontként használható komplexebb munkák előtt. Ugyanakkor értékes kimenetet adhat önmagában is (*Fokasz et al.* [2015]; például beszédes lehet, hogy különböző típusú oldalak milyen szavakat használnak vagy nem használnak egyazon témában). Az adott szó használatának gyakorisága azonban csak nagyon korlátozott információval látja el az elemzőt, ennél sokkal érdekesebb az, hogy a szó milyen kontextusban fordul elő.

A kontextus megragadásának legegyszerűbb módja azoknak a szavaknak a megkeresése, amelyek gyakran szerepelnek az általunk vizsgálni kívánt szavak közelében. A nyelvek szókészlete ugyanakkor nagyon gazdag, tele szinonimákkal, ragokkal, igeidőkkel. A szövegek előfeldolgozása (lásd később) sokat tud ezen a problémán egyszerűsíteni, de nem ad megnyugtató választ arra, hogy egy szó teljes környezetét miként tudjuk feldolgozni. E nehézség a szóbeágyazási modellek segítségével küszöbölhető ki. A kérdést heurisztikusan megközelítve, a szóbeágyazási modellek alapvető célja annak meghatározása, hogy egy adott szó (vagy annak n-gramja – lásd később) milyen más szavakhoz van közel a felhasznált korpuszban.

A probléma megoldására definiált első megközelítések az ún. együttes szóelőfordulási mátrixokból (term co-occurrence matrix, TCM) indultak ki. A matematikai alapprobléma ebben az esetben az, hogy van egy nagyon ritka mátrixunk,[4] amelyben a szavak együttes előfordulását tároljuk. Az együttes előfordulást sokféleképpen lehet definiálni, általában a szóközelséggel szokták megadni. Tehát egy adott szó környezetét a +/–1 X távolságban levő szavak jelentik. Az X általában nem nagyobb mint 10, de persze speciális esetben lehet ennél nagyobb is. Egy nem különösen nagy, 100 000 szót tartalmazó szövegben a TCM mérete: 100 000 * 100 000, azaz 10 000 000 000 cellából áll. Viszonylag egyértelmű, hogy egy ilyen nagyságú, ám kifejezetten ritka mátrix esetében a direkt statisztikai elemzés fel sem merülhet. A cél tehát egy olyan redukált mátrix kialakítása, amely lehetőleg minél nagyobb arányban megőrzi az eredeti mátrix információtartalmát, és akkora méretű, ami már standard statisztikai eszközökkel kezelhető. Ez egy klasszikus dimenziócsökkentési eljárási probléma. A dimenziócsökkentés a társadalomtudósok számára sem ismeretlen, analógiaként hozhatjuk a főkomponens-elemzést, ahol hasonló logikájú dimenziócsökkentést végzünk. Az analógia azért is jó, mert a korai dimenziócsökkentési eljárások egy főkomponens-elemzéshez hasonló megoldást, a szinguláris érték-felbontást (singular value decomposition, SVD) használták, elsősorban dokumentációklasszifi-

[4] A ritka mátrixnak nincs definíciója, de olyan mátrixot érdemes magunk elé képzelni, amelyben a cellák kevesebb mint 0,1 százalékában van nullától eltérő érték.

kációs céllal – ezt a módszert nevezik látens szemantikus elemzésnek (latent semantic analysis, LSA; *Deerwester et al.* [1990]).

Az LSA-ben a kiindulópont egy dokumentum-szógyakorisági mátrix, de a számolás alapját jelentő SVD-módszer TCM-en is alkalmazható. Az SVD-re épülő mátrixfaktorizációs modellekre a disztribúciós szemantikus modellek (distributional semantic models, DSM) címkét is használják. Az SVD-modellek bemeneti adatait egy TCM képezi, kimenetként pedig egy alacsony dimenziós (100-1 000) vektorteret ad ki a modell. A szavak gyakorisága nagyon ferde eloszlást követ, akárcsak az együttes szóelőfordulás, és ez a szógyakoriság a vektorterekben is visszaköszön – a gyakoribb szavak közelebb vannak egymáshoz. Ennek kiküszöbölésére számos megoldás született, például a nagyon gyakori szópároknál egy bizonyos küszöbértékben maximalizálták az együttes előfordulások számát (vagy akár ki is hagyták a gyakori szópárokat), más szerzők pedig a TCM-táblában a logaritmusokat vagy a szavak Pearson-korrelációját használták a nyers számok helyett (*Rohde–Gonnerman–Plaut* [2006]).

Bengio et al. [2003] a mátrixfaktorizációs modellekhez képest egy másik megközelítést javasoltak; az elsők között használtak neurális hálókat a szókontextus becslésére. A DSM-ek alapvetően gyakoriságiak, míg a neurálisháló-alapú nyelvi modellek (neural network language model, NNLM) predikciósak, tehát szemben az előbbi heurisztikus megközelítéssel, a szavak kontextusát a szerzők a lokális környezetük alapján becsülték meg az NNLM-ekben.

Mind a DSM, mind az NNLM-ek (összefoglaló néven szóbeágyazási vagy vektortérmodellek) megmaradtak egy szűkebb tudományos közösségen belül a 2000-es években. Az akkori gépkapacitások mellett nagy szövegeken gyakorlatilag nem futottak le véges idő alatt, ezért csak egy nagyon szűk kör foglalkozott e nyelvfeldolgozási iránnyal.

1.1.2. Word2vec

Az igazi áttörést *Mikolov et al*. 2013-as cikke hozta. Az akkor a Google-nél dolgozó csapat egy olyan NNLM-et fejlesztett ki, amelynek használata szélesebb közönség számára vált elérhetővé. A word2vec névre keresztelt módszer nemcsak a feladatmegoldásban működött nagyon jól (lásd később), de a korábbi módszerekhez képest kevesebb erőforrást is igényelt. A munka alapjait *Mikolov* egyébként már 2007-ben ismertette MSc-szakdolgozatában.

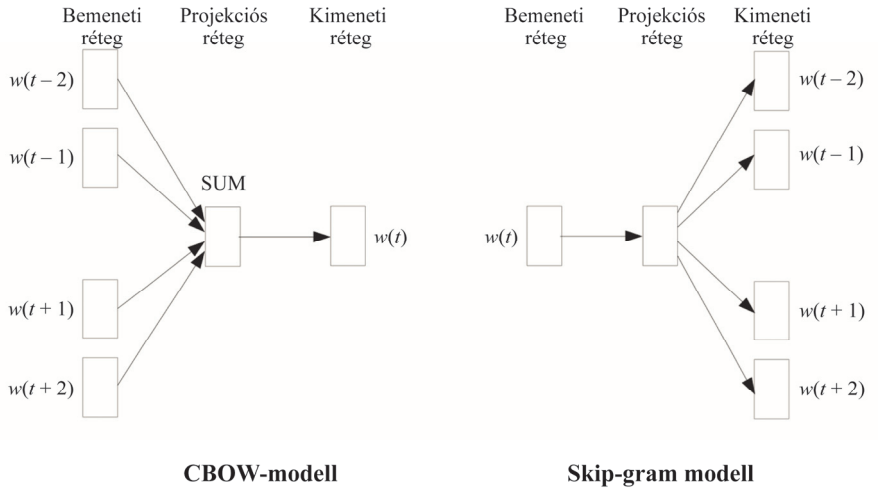
Bengio et al. [2003] korábban említett tanulmányukban a következő négy rétegből álló „mély” neurális háló alkalmazását javasolják a klasszifikációs modell kiszámítására: bemeneti, projekciós, rejtett és kimeneti réteg. Nagy szövegkorpusz esetén ez a megközelítés rendkívül hosszú számítási idővel jár. A probléma megoldására szintén használt rekurrens neurális hálókban (lásd *Mikolov et al*. [2013]) annyiban változtatnak az alapmodellen, hogy kimarad a projekciós réteg, viszont

a rejtett réteghez egy rekurrens mátrix kapcsolódik, amely gyakorlatilag folyamatosan új inputtal táplálja a rejtett réteget. Ebben a megközelítésben a számítási komplexitás csökken, de továbbra is nagyon magas.

Mikolov et al. [2013] ehhez képest egy egyszerűbb modellt javasolnak cikkükben. Mivel a korábbi modellek komplexitásáért elsősorban a rejtett réteg felel, ezt kihagyják a neurális hálóból, amely így esetükben három rétegből áll: bemeneti, projekciós és kimeneti réteg. A szerzők megközelítésének újszerűsége abban rejlik, hogy a neurális hálót egy klasszifikációs modellel vegyítik, gyakorlatilag klasszifikációs modellt építenek. A bemeneti adatot az adott szó kontextusa jelenti, a kimenetit pedig egy szó. A klasszifikációs modell valójában „egyszerű" logisztikus regresszió; a végeredmény szempontjából a szavak mellé rendelt súlyok fontosak, melyek gyakorlatilag a szavak vektorreprezentációját jelentik.

Mikolovék két megközelítést is bemutatnak tanulmányukban. A folytonos szózsák- (continuous bag of words, CBOW) modellben egy adott szó környezetéből indulnak ki, és abból próbálják megbecsülni a szót. A skip-gram modellben pedig fordított logikát követnek (*Mikolov et al*. [2013]): egy adott szóból becsülik meg annak kontextusát (hasonló megközelítés kapcsán lásd például *Mnih–Kavukcuoglu* [2013]). Az 1. ábra e két megközelítést szemlélteti.

1. ábra. A CBOW- és a skip-gram modell felépítése
(Logic of the CBOW and skip-gram models)



Forrás: *Mikolov et al.* [2013].

A CBOW-modell nagyon gyors, akár nagy adatbázisokon is jól használható, de csak a gyakrabban előforduló szavak esetén ad megbízható eredményt. Ezzel szemben a skip-gram modell jól működik ritkább szavakra is, bár inkább

kisebb korpuszok esetében javasolt a hosszabb számítási idő miatt. A word2vec végeredményét gyakorlatilag a projekciós (rejtett) rétegben kiszámított klasszifikációs modell súlyai adják, ezt használhatjuk a szövegünk további elemzésére.

Mikolovék modellje három ok miatt hozott igazi áttörést. Az első és legfontosabb, hogy nagyon jó eredményt tudott felmutatni azokban a feladatokban, amelyekkel rendszeresen tesztelik a modellek pontosságát (lásd később). Az is lényeges volt az elterjedésében, hogy a szerzők elérhetővé tettek olyan előkészített (angolban a „train" szót használják erre) vektortereket, amelyeket hatalmas szövegkorpuszokra építettek rá. Ezek a vektorterek önmagukban is alkalmasak arra, hogy különböző alkalmazásokban és tudományos kutatásokban használják őket. A harmadik fontos ok pedig az volt, hogy Mikolovék szabad hozzáférésűvé tették a word2vec kódjait[5], így gyakorlatilag jelentősen leszállították azt a „programozási küszöbtudást", amellyel bárki saját vektortereket hozhat létre.

1.1.3. FastText

A word2vec továbbfejlesztéseként tekinthetünk a 2016-ban publikált fastText-algoritmusra (*Joulin et al.* [2016], *Bojanowski et al.* [2017]), amelynek Facebookhoz köthető kutatói csapata részben a word2vec alkotóiból áll (például Mikolov személyében). Ez esetben az újítás lényege az volt, hogy szemben a word2vec-algoritmussal, amely a szavakat használta fel a vektortér kialakításakor, a fastText karakter n-gramokból indult ki. Például a piros szó a „pir", „iro", „ros", „piro", „iros" és „piros" n-gramokból építhető fel (ha 3 és 5 közötti n-gramokat használunk), és a szó pozícióját e 7 n-gram összegeként definiálhatjuk. Ennek több előnye is van. A modell képes jól megragadni akár ritka szavak pozícióját is, ha azok karakter n-gramjai megtalálhatók más szavakban, sőt akár olyan szavakét is, amelyek nem szerepelnek az eredeti korpuszban. A gyakorlatban ez nagyon előnyös tulajdonság, mert azokban az alkalmazásokban, ahol vektortérmodellek szerepelnek a háttérben, problémát okozhat, ha a felhasználók ritka vagy ismeretlen szavakat használnak. Erre egyszerű példaként a keresőmotorok említhetők. A felhasználók gyakran elütnek kifejezéseket, ami nehéz helyzet elé állítja a keresőalgoritmust. Ha azonban a keresést egy fastText-alapú nyelvi modell támogatja, sokkal nagyobb valószínűséggel juthat el a felhasználó az általa keresett tartalomhoz. *Joulin et al.* [2016] két példán is tesztelték az algoritmusukat. Mind érzelembesorolásban, mind címkepredikcióban jobban, ráadásul sokkal rövidebb futásidővel teljesített a fastText a riválisalgoritmusoknál. A word2vec-hez képest viszont lassabb, főleg akkor, ha széles karakter n-gram spektrumot állítunk be a tréningjéhez. A társadalomtudományi felhasználás kapcsán visszatérünk majd a két modell gyakorlati különbségeire.

[5] https://code.google.com/archive/p/word2vec/

1.1.4. GloVe

A word2vec- és fastText-algoritmusok mögötti megközelítés nemcsak statisztikai módszerében más (NNLM vs. mátrixfaktorizáció) a korábbi SVD-modellekhez képest, hanem abban is, hogy alapvetően egy szó lokális környezetéből indul ki, szemben az SVD-modellekkel, amelyeknél globális környezetet használnak a vektortérképzésre. A Stanford NLP-kutatócsoportja által kifejleszett GloVe[6] (*Pennington–Socher–Manning* [2014]) a két megközelítést ötvözi abból a szempontból, hogy egyszerre veszi figyelembe a vektortér kiszámításakor a lokális és a globális környezetet. *Penningtonék* 2014-es írásukban több hasonló modell eredményeivel is összevetik sajátjaikat. A GloVe az összehasonlítás alapján legalább olyan jól, vagy jobban teljesít, mint a riválisalgoritmusok.

Akárcsak a „klasszikus" SVD-algoritmust használó modellekben, a kiinduló adat a GloVe esetében is egy TCM. A GloVe viszont nem mátrixfaktorizációs megközelítést követ, hanem egy adott szópár együttes előfordulásának logaritmusát becsüli meg. Ez a gyakorlatban azt jelenti, hogy a GloVe algoritmusa egy olyan függvényt minimalizál, amelyben az adott szópárhoz tartozó szavak vektorainak szorzatából kivonjuk a szavak logaritmusát, majd ennek vesszük a négyzetét. A modell kapcsán két technikai érdekességet érdemes megjegyezni. Az első, hogy a GloVe a gyakori szavak szerepének csökkentése érdekében egy súlyfaktort alkalmaz, kiküszöbölve ezzel a TCM-alapú modellek egyik alapproblémáját. A második pedig, hogy a modell definiálásakor párhuzamosan két vektorteret is használnak. Ez nem következik szükségszerűen a modellfelírásból, de a gyakorlati eredmények alapján így robusztusabban működik az algoritmus. A két vektorteret általában összegzik, és a közös vektorteret használják a további számításokban. A GloVe mögött álló modellt a dupla vektoros megoldás miatt nevezik log-bilineáris regressziós modellnek (*Pennington–Socher–Manning* [2014]). A GloVe-módszer egyik nagy előnye, hogy nem csak egy lokális ablakot használ, mint a word2vec-alapú modellek, a másik pedig az, hogy jobban értelmezhető az általa kapott végeredmény, legalábbis abban az értelemben, hogy két adott szóhoz tartozó vektor szorzatának tartalmi jelentést tudunk adni.

1.1.5. Melyik modell jobb?

A bemutatott különféle szóbeágyazási modellek között nehéz rangsorolni, mivel „széles körű paraméterezhetőségük" miatt nem könnyű objektívan eldönteni, hogy melyikük teljesít jobban (akár a részfeladatokban). A vektortérmodelleket összehasonlító cikkek (lásd például *Spirling–Rodriguez* [2019]) legfőbb tanulsága, hogy megfelelő paraméterezés mellett nincs nagy különbség a „teljesítményük"

[6] https://nlp.stanford.edu/projects/glove

között. Kisebb korpuszok és ritkább szavak esetén a GloVe valamivel stabilabb, mint a lokális ablakot használó word2vec-modellek. Ez utóbbiak közül a skip-gram szinte minden helyzetben jobban teljesít, mint a CBOW. Futásidő szempontjából a GloVe gyorsabb, mint a word2vec (és főleg, mint a skip-gram), de több memóriát igényel, mivel a bemeneti oldalon el kell készíteni egy TCM-et, amely már közepes korpusznál is rendkívül nagy lehet. Az egyes megközelítések közötti választást befolyásolhatja a projekt célja, az elérhető korpusz nagysága, jellege, a rendelkezésre álló számítógépes erőforrás nagysága (gépek száma, CPU-[7]/GPU[8]-szám, memória) vagy olyan praktikusabb szempontok, mint, hogy a kutató milyen programnyelven tud jól kódolni, és az adott programnyelven a különböző módszerek miként érhetők el.

1.1.6. Kontextualizált vektortérmodellek

A bemutatott módszerek evolúciója alapján megállapítható, hogy a 2010-es évek elején gyorsult fel igazán a szóbeágyazási modellek fejlesztése. Mint már említettük, a fastText-et 2016-ban publikálta egy Facebook-hoz köthető kutatócsoport (*Joulin et al*. [2016]). Az azóta eltelt évek sem teltek el eseménytelenül a tudományterületen, több izgalmas modellt is bemutattak. Az eddig közzétett módszerek mind statikus vektortérmodellek voltak. Ez a gyakorlatban azt jelenti, hogy mindegyik szónak kontextustól függetlenül 1 vektortér-reprezentációja készül. A kontextualizált szóbeágyazási (contextualized word embeddings, CWE) modellek ezzel szemben adott kontextushoz kötik, hogy milyen vektortértéket kap egy szó. Bizonyos esetekben a statikus vektorterekre épülnek rá kétirányú rekurrens neurális modellek, ilyen a Flair (*Akbik–Blythe–Vollgraf* [2018]) vagy az ELMo[9] (*Peters et al*. [2018]).

A módszerek legújabb generációja viszont más logikát követ, és nem használ statikus beágyazást. Közös pontjuk az, hogy főleg fordítási feladatokban használt, ún. „transformerekre" (transzformátorokra) építenek. Ez az általános megnevezés olyan nyelvi modellekre utal, amelyekben van egy bekódoló (encoder) és egy kikódoló (decoder) elem. E kettő különböző neurális hálókból épül fel (rekurrens, konvolúciós stb.). *Vaswani et al*. [2017] egy olyan transformert vezettek be, amelyben a bekódoló elem egy összpontosító (attention) és egy nem visszacsatoló (feed forward) neurális hálóból áll, míg a kikódoló elemhez a megkülönböztető és a nem visszacsatoló réteg közé egy bekódoló-kikódoló összpontosító (encoder-decoder attention) réteget is hozzáadtak. A modell, mivel nem tartalmaz rekurrens elemet, rendkívül gyors futásra képes, az összpontosító réteg pedig nagy teljesítményt eredményez. Ez utóbbi réteg lényege az, hogy egy adott szövegrészen belül megadjuk, mely másik szóval tudjuk a célszó megértését segíteni, kvázi egyfajta súlyt rendelünk a kontextus-

[7] Központi feldolgozó egység (central processing unit, CPU).
[8] Grafikai processzor (graphics processing unit, GPU).
[9] Beágyazások nyelvi modellekből (embedding from language models, ELMo).

ban megjelenő szavakhoz. Ezt a logikát vitte tovább az OpenAI (*Radford et al.* [2018]) és a jelenlegi egyik legújabb módszer[10], a transzformátoralapú gépi tanulási technika (bidirectional encoder representations from transformers, BERT; *Devlin et al*. [2019]) is azzal a különbséggel, hogy csak a bekódoló elemet vették át. A BERT kiugróan magas teljesítményét részben annak köszönheti, hogy nemcsak balról jobbra tanul, de parallel jobbról balra is. Ezek a CWE-modellek főleg a többértelmű szavak esetén lényegesen jobb eredményt érnek el szinte minden nyelvi feladatban, mint a statikusak (*Wiedemann et al*. [2019]). Társadalomtudományi használatuk terjedőben van. Klasszifikációra már ma is lehet példát találni (*Samory et al*. [2020]), de az elkövetkező években derül ki, hogy a tartalmi fókuszú elemzésekben mennyire tudnak teret nyerni.

1.2. Technikai megfontolások

A tanulmány előző alfejezetében a szóbeágyazási módszerek rövid fejlődéstörténetét mutattam be. A technikai részletek tárgyalása után a téma kifejtését érdemes újra egy kissé heurisztikusabb nézőpontból folytatni. A kiindulópont az lehet, hogy a különböző tudományterületek és üzleti szereplők egymástól eltérő célból használhatják a szóbeágyazási módszereket. Társadalomtudományi perspektívából legtöbb esetben szövegeken keresztül szeretnének megérteni egy adott társadalmi jelenséget, annak beágyazottságát vagy temporális mintáit. Nyelvészként érdekes lehet például bizonyos szavak morfológiájának változása vagy a szavak értékváltása (*Szabó* [2019]), de akár az is, hogy miként lehet e nyelvi modelleket más nyelvészeti feladatokban használni (például emócióelemzés, lemmatizáció[11] stb. céljából). A szóbeágyazási módszerek azonban nem a tudományos felhasználhatóságuk miatt, hanem elsősorban azért terjedtek el, mert nagyon sok gyakorlati feladatban hasznosnak bizonyulnak. Fel lehet őket használni többek között fordítóprogramokban, szöveges keresések támogatására, szöveges botok programozására, dokumentumklasszifikációra stb. Könnyen belátható azonban, hogy más szempontoknak kell megfelelnie egy társadalomtudományi kérdésre választ kereső szóbeágyazási modellnek, mint egy keresési algoritmusnak. Az 1. táblázat a szóbeágyazási modellekkel szemben támasztott elvárásokat sorolja fel.

[10] Bár kevesebb mint 2 éve jelent meg a módszert bemutató cikk, csak a Google Scholar alapján már több mint 12 000 hivatkozása van.

[11] Szavak alapalakjának a megkeresése.

1. táblázat

Szóbeágyazási modellekkel kapcsolatos elvárások felhasználási területenként
(Aspects that word embedding models must meet, by application area)

Elvárás	Társadalomtudományi	Ipari
	alkalmazás	
Korpusz jellege	egyedi, az adott területhez köthető	általános
Korpusz nagysága	akár kis korpusz is	minél nagyobb
Előkészített vektorterek	inkább nem	inkább igen
Elgépelt szavak, ismeretlen szavak	kevésbé fontosak	fontosak
Ritka szavak	kevésbé fontosak	fontosak
Stemmelés/lemmatizáció	javasolt	nem javasolt
Módszer	word2vec, GloVe	fastText, CWE-k (ELMo, BERT)

Megjegyzés. Stemmelés jelentése: szavak megfosztása toldalékoktól, ragoktól, igeidőktől.

Bármilyen vektortérmodellről is beszélünk, minden esetben a korpusz a kiindulópont. Ez azoknak a szövegeknek az összességét jelenti, amelyekből elkészítjük a nyelvi modellünket. E szövegek lehetnek közösségimédia-tartalmak, online újságcikkek, kommentek, digitalizált tartalmak – gyakorlatilag bármilyen online elérhető szöveges tartalom. Az ipari alkalmazásokban jellemzően általános, nagy korpuszokból indulnak ki. Ilyen alapkorpusz lehet például a Common Crawl nonprofit szervezet korpusza, a CC, amelyet gyakorlatilag az internetes tartalmak egy mintájának is tekinthetünk (https://commoncrawl.org). A CC folyamatosan gyűjti erre írt speciális algoritmusokkal (crawlers) az online oldalakra kikerülő tartalmakat, és ezeket szabadon elérhetővé teszi. A szervezet oldalán több mint 40 nyelven érhető el szöveges adat (magyarul is), 7 évre visszamenőleg. Hatalmas, petabyte nagyságrendű adatmennyiségről van szó, amely azonban nem túl jó minőségű – a szövegek tisztítása már magában is nagy feladat (*Indig* [2018]). Szintén nyílt hozzáférésű, kiterjedt szöveges adatforrást jelent a Wikipedia (https://dumps.wikimedia.org/backup-index.html). Ennek tartalma is többnyelvű, tehát nem csak angol nyelven jelenthet megoldást a korpuszra. Mind a CC-t, mind a Wiki-korpuszt gyakran használják a szóbeágyazási modellekben,[12] mivel rengeteg témát lefednek, rengeteg unikális szót tartalmaznak, és már méretük miatt is nagyon robusztusak a belőlük kapott eredmények. Egy ipari alkalmazásban (például egy keresést támogató applikációban) az itt felsorolt tulajdonságok meglehetősen előnyösek. De vajon miért nem feltétlenül azok egy társadalomtudományi kutatásban? A válasz egyszerű: ugyanazon okból, amiért a legtöbb survey-módszertannal foglalkozó kutató (100-ból 99) előnyben részesít egy 1 000 fős reprezentatív mintát egy 1 000 000 fős kényelmi mintával szemben. Bár a CC és a Wiki-korpusz is nagy, a belőlük kapott eredmények nem (vagy csak

[12] A word2vec és fastText oldalán elérhető előkészített vektorterek is részben ezekre a korpuszokra épülnek.

korlátozottan) általánosíthatók, és a külső érvényességük alacsonyabb egy szelektált, célhoz szabott korpuszhoz képest. Társadalomtudományi kutatások is készülnek nagy, általános korpuszokon (*Kmetty–Koltai–Rudas* [2021]), de ezek esetében a kutatók gyakran saját, egyedi korpuszt használnak egy adott témához, mivel ezáltal lehetővé válik az általános korpuszokon nem lehetséges kérdések megfigyelése (*Szabó et al.* [2020]).

Az eddigiekkel szorosan összefügg az is, hogy mennyire jellemző az előkészített vektorterek használata egyes alkalmazási területeken. A vektorterek a szóbeágyazási algoritmusok végeredményei; soraikban szavak vannak, oszlopaikban pedig szavakhoz tartozó súlyok. Általában 100-500 dimenzióból állnak (lásd később). A word2vec és a fastText sikeréhez nagyban hozzájárult, hogy esetükben elérhetővé váltak előkészített vektorterek. Ez a gyakorlatban azt jelenti, hogy a fejlesztők hatalmas korpuszokon (lásd például CC, Wikipedia) elkészítették a szövegbeágyazást, a kapott vektortereket pedig szabadon letölthetővé tették. Informatikai/programozói szempontból az előkészített vektorterek kezelése nagyságrendekkel egyszerűbb feladat, mint egy nagy korpusz építése és beágyazása. Ez kifejezetten olyan nagyságú korpuszokra igaz, mint amilyenek hozzáférhetők a word2vec vagy a fastText oldalán. A vektorterek az ipari alkalmazások többségében igen, a társadalomtudományi projektekben viszont nem jól használhatók a korábban említett szempontok miatt. Szintén a használati céllal van összefüggésben, hogy az ipari alkalmazásokban előny, ha a modell robusztusan tudja kezelni a ritka, valamint a kiinduló korpuszban nem szereplő szavakat, illetve megoldást nyújt az elgépelésekre. Speciális elemzési célok kivételével ezeknek azonban nincs jelentősége egy társadalomtudományi projektben. A kutatásokban általában nem a kivételek az érdekesek, hanem a „masszív" trendek. Ebből következik, hogy az ipari projektben a fastText-hez hasonló karakter n-gram megközelítést használó algoritmusok kerülnek előtérbe, a társadalomtudományiakban ugyanakkor ezeknek nincsen hozzáadott értékük – sőt akár zavaró is lehet, ha a végeredményben két ellentétes jelentésű, de hasonló alakú szó közel kerül egymáshoz.

A bemeneti korpusz kapcsán a méret és a forrás mellett az is érdekes, hogy milyen lépéseken megyünk keresztül a beágyazásra való előkészítés során. A korpusztisztításnak nincs egy általános, minden projektre érvényes formulája (*Németh–Katona–Kmetty* [2020]), viszont léteznek olyan lépések, amelyeket érdemes elvégezni a tartalomelemzés előtt (elemzés alatt itt nemcsak a szóbeágyazásra, hanem minden más szövegbányászati megoldásra [például topikmodellezésre] is gondolok).

Az előfeldolgozási lépések jellemzően idő- és erőforrás-igényesek. Ha azonban társadalomtudományi célból akarunk felhasználni egy vektortérmodellt, akkor érdemes végrehajtani őket, mert jelentősen befolyásolhatják eredményeink érvényességét. Az ipari alkalmazásokban ez utóbbinak általában kisebb a fontossága, sőt a lemmatizáció/stemmelés általában kifejezetten kerülendő, hiszen a legtöbb esetben épp az a cél, hogy a nyelvi modell egy adott szó minden alakjára jól működjön.

Miután véglegesítettük a bemeneti korpuszt, és kiválasztottuk az algoritmust, a beágyazási modell paraméterezése a következő lépés. Ez értelemszerűen a választott

algoritmustól függ, de vannak olyan paraméterek, amelyek minden algoritmusban egyaránt választhatók. Ezek közül a két legfontosabb az ablak nagysága és a vektortér mérete. Először ez utóbbival foglalkozok.

A vektortér nagysága két elemből áll össze. Az egyik a szavak száma, amely alapvetően adott egy szótárban. Ha a nagyobb trendeket akarjuk vizsgálni egy témában, a ritka szavak akár ki is hagyhatók a korpuszból, csökkentve ezzel a futásidőt és növelve a modell stabilitását (lásd később). A másik dimenzió a szavakhoz rendelt súlyvektor hossza. Ez jellemzően 100 és 500 között mozog, leginkább 200-300 dimenziós vektortereket képeznek. Intuitív módon azt gondolhatnánk, hogy minél több dimenziónk van, annál jobb lesz a vektortér; ám ez a gyakorlatban erősen korpuszfüggő (*Spirling–Rodriguez* [2019]). Kis korpusz esetén a túl sok dimenzió csökkenti az eredmények stabilitását, tehát az adott szó pozíciójában nagyobb lesz a véletlen szerepe, de nagy korpusznál sem érdemes extra számú dimenziós vektorteret kialakítani, mert a futásidő növekszik, és nem kapunk jobb eredményeket.

Az ablak nagysága a másik olyan paraméter, amelyről minden modell kapcsán fontos döntenünk. Az ablak azt a szókörnyezetet jelenti, amelyet a modellek felhasználnak a vektortér kialakítására. A word2vec ezen az ablakon megy végig, a GloVe ezt az ablakot használja fel, hogy kiszámítja a TCM-ot. Az ablak általában kétoldali szimmetrikus, de vannak példák aszimmetrikus, illetve olyan ablakokra is, ahol súlyoznak a közelséggel. Ha túl pici ablakot használunk, nem jelenik meg, ha túl nagyot, akkor „szétfolyik" az adott szó kontextusa. A kis ablak jobban működik a lexikális egyezések vizsgálata során, míg a nagyobb az analógiák vizsgálatában (*Lison–Kutuzov* [2017]), és kisebb korpuszok esetében növeli a stabilitást (*Szabó et al.* [2020]). Közösségimédia-tartalmaknál (tweetek, kommentek) szintén érdemes nagyobb ablakot használni, hogy az adott szöveg teljes kontextusában beépüljön a vektortérbe (*Yang–Macdonald–Ounis* [2018]). Bár nincs egységes álláspont az ablak ideális mérete kapcsán, általában 5-10 közé tehető, de ahogy írtam, ez függ a korpusz jellegétől és nagyságától.

Bár még nem emeltem ki, az eddigiekből egyértelműen kiderül, hogy a szóbeágyazás – akárcsak más nyelvi modellek – erősen nyelvfüggő. Maga az algoritmus nyelvtől függetlenül ugyanazokat a lépéseket végzi, de a nyelv meghatározza az előfeldolgozást, a modell paraméterezését és értelemszerűen a kapott vektorteret is. A különböző módszerek összehasonlítását az is megnehezíti, hogy a legtöbb kiértékelés angol nyelvre készült, így az általánosan javasolt paraméterbeállítások nem biztos, hogy más nyelvek esetén is megfelelők. Az ablaknagyságnál is érdemes ezt figyelembe venni, és az adott nyelv egyedisége alapján meghatározni az ideális ablakot. Erősen nyelvfüggő például, hogy az egy mondaton belüli, egymásra ható szavak között mekkora az átlagos távolság (az ún. dependencialánc). Azokban a nyelvekben, ahol a dependencialánc hosszabb, érdemes nagyobb ablakot használni. Az átlagos távolság jóval nagyobb a magyar, a német és a kínai nyelvben, mint a románban vagy a japánban. Az angol nyelv ilyen szempontból átlagosnak számít (*Liu* [2008]).

A technikai bemutatás kapcsán az utolsó vizsgált jellemző a modellek stabilitása. Ez jelen kontextusban akként értelmezhető, hogy mennyiben térnek el egymástól az adott korpuszra lefutatott, egyazon módszert használó elemzések eredményei. Az eltérő vektorterek abból következnek, hogy a súlyvektorok illesztésekor az algoritmus egy véletlen inicializálásból indul ki, és a minimalizálni kívánt hibafüggvényt még véletlen hibatagokkal is kiegészíti. Az ilyen típusú instabilitás jellemző más NLP- (például a topik-) modellekre is, de a standard társadalomtudományi módszerek között is találunk rá példát (lásd *k*-közép klaszterezés). A stabilitás erősen összefügg a korpusznagysággal: minél kisebb a korpuszunk, annál instabilabb az eredmény. Bizonyos paraméterek beállításával (kisebb dimenziószám, nagyobb ablakméret, sok iteráció) csökkenthető, de nem tüntethető el az instabilitás, mert ez a módszer sajátja.

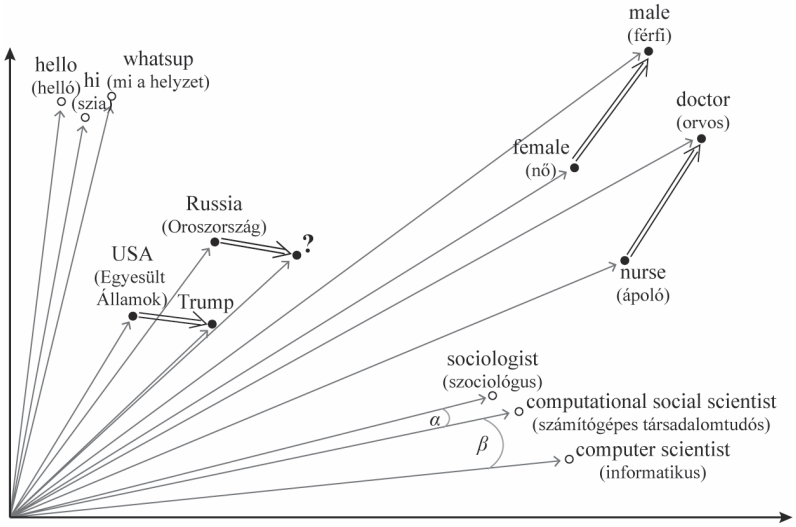
2. Mit kezdjünk a vektortérrel?

Az előző alfejezetekben bemutattam a szóbeágyazási modellek matematikai/statisztikai alapjait és azokat a technikai megfontolásokat, amelyeket érdemes átgondolni egy vektortérmodell készítése/használata előtt. Jelen alfejezetben egy használható vektortérből indulunk ki. De vajon mit lehet e vektortérrel kezdeni?

A kérdés megválaszolásához érdemes újra felidézni az alacsony dimenziós vektortér eredeti korpuszból való alkotásának célját: szeretnénk minél jobban megérteni egy adott szövegben szereplő szavak egymáshoz való viszonyát. Vajon melyek azok a szavak, amelyek jellemzően egy kontextusban szerepelnek/„taszítják” egymást? A szóbeágyazási modellek segítségével e kérdésre pontos válasz adható. Gyakorlatilag módszertől függetlenül az egyes szavak vektorait úgy optimalizáljuk, hogy az egymáshoz közel eső szavak e térben is közel legyenek egymáshoz. A közelség ebben az esetben akár a jelentést vagy a témát érintő, akár a szintaktikai közelséget is jelentheti (ha nem lemmatizáltuk a korpuszunkat). A 2. ábra egy egyszerűsített kétdimenziós térben mutat be erre néhány példát különböző fogalmak vonatkozásában. A közelség kiszámítására leggyakrabban a két vektor szögtávolságának koszinuszából[13] indulunk ki, és ezt fordítjuk át egy közelségmutatóra (lásd később).

[13] Két szó közelségének kiszámításakor triviális megoldásként egy egyszerű euklidészitávolság-metrika juthat elsőként eszünkbe. Ez jelen esetben azonban nem a legjobb megközelítés, mivel a vektorok hossza összefüggést mutat a szavak gyakoriságával és kontextusfüggőségével (*Schakel–Wilson* [2015]). Az euklidészi távolság helyett számos elemzés szögtávolságokat használ, pontosabban a szavak koszinuszközelségét vizsgálja. A koszinuszközelség 1, ha két szó között a bezárt szög 0, 90 fokos szögnél 0, 180 foknál pedig –1. Nincs egységes definíció arra, hogy mi számít magas vagy alacsony koszinuszközelségnek. Saját tapasztalataim alapján ennek meghatározására a következő általános hüvelykujjszabály használható: 0,2 alatt a közelség gyengének, 0,2 és 0,4 között közepesnek, 0,4 felett erősnek tekinthető. Ezek az értékek azonban egyaránt függnek a korpusztól és a beágyazási algoritmustól, ezért a legjobb, ha a koszinuszközelségek sorrendjére fókuszálunk és nem az abszolút értékére.

2. ábra. Példa a szavak/fogalmak közelségére
(Example of proximity of words/concepts)



Forrás: *Németh–Koltai* [2021].

A 2. ábrán közel helyezkednek el egymáshoz a köszönések (hello, hi, whatsup), az országok és a politikusok, valamint a tudományterületek. Például a számítógépes társadalomtudós közelebb van a szociológushoz, mint az informatikushoz.[14] A diagram egy érdekes lehetőségre, az analógiák vizsgálatára is rávilágít. Ha meghatározzuk a nő és a férfi vektor között bezárt szöget, és ezt kivetítjük az orvosra, akkor megkapjuk a szakma „női" megfelelőjét, az ápolót. Ugyanezzel a logikával azt is meg tudjuk nézni, hogy a USA → *Trump* párosításnak ki felel meg Oroszország viszonylatában – a kérdőjel helyére viszonylagos biztonsággal beírhatjuk *Putyin*t. E lehetőségek társadalomtudományi hasznára a későbbiekben térünk vissza.

Az eddigi példák elméleti síkon mutatták be, hogy miként működik a módszer. A gyakorlati demonstrációra egy előkészített vektorteret használok, amely a Wikinews korpuszán alapul. A Wikinews a 2017-es Wikipediából, a UMBC[15] korpuszából (50 000 oldal 100 000 000 weblapjának gyűjtése) és a stamt.org oldalról származó hírekből épül fel. A teljes korpusz 16 milliárd szót tartalmaz. Az előkészített, 300 dimenziós vektortér, amelynek képzéséhez a fastText-algoritmust használták, a korpusz 1 000 000 leggyakoribb szavát tartalmazza (ezek nincsenek stemmelve, és bennük kis- és nagybetűk egyaránt előfordulnak). A vektortér szabadon letölthető a fastText weboldaláról: https://fasttext.cc/docs/en/english-vectors.html.

[14] Az ábra demonstrációs céllal készült, mögötte nincsenek valós adatok.

[15] Marylandi Egyetem, Baltimore megye (University of Maryland, Baltimore County, UMBC) https://ebiquity.umbc.edu/blogger/2013/05/01/umbc-webbase-corpus-of-3b-english-words/

Az elemzést egy egyszerű lekérdezéssel kezdjük. Milyen szavak vannak közel a *sociology*-hoz (szociológiához)?

2. táblázat

A „sociology" (szociológia) szó 15 legközelebbi szomszédja
(The 15 closest neighbours of the word 'sociology')

Sor-rend	Szó	Koszinusz-közelség	Sor-rend	Szó	Koszinusz-közelség	Sor-rend	Szó	Koszinusz-közelség
1.	Sociology	0,79	6.	criminology	0,71	11.	linguistics	0,63
2.	psychology	0,77	7.	sociological	0,71	12.	ethnography	0,63
3.	anthropology	0,76	8.	economics	0,69	13.	theology	0,63
4.	sociologists	0,73	9.	philosophy	0,66	14.	ecology	0,63
5.	sociologist	0,71	10.	biology	0,64	15.	science	0,62

Megjegyzés. Itt és a továbbiakban: sociology: szociológia; psychology: pszichológia; anthropology: antropológia; sociologist(s): szociológus(ok); criminology: kriminológia; sociological: szociológiai; economics: közgazdaságtan; philosophy: filozófia; biology: biológia; linguistics: nyelvészet; ethnography: néprajz; theology: teológia; ecology: ökológia; science: tudomány.
Forrás: Wikinews alapján saját számítás.

Az angol nyelvű korpuszban a *sociology* (szociológia) szóhoz a *Sociology* van a legközelebb, tehát annak nagybetűvel kezdődő változata, de a sorrendben az első öt között van a *sociologists* (szociológusok) és a *sociologist* (szociológus) is. A tanulmány korábbi alfejezeteiben már többször említettem, hogy az általános korpuszokat használó, előképzett vektorterek esetében a készítők kevés előfeldolgozást végeznek. Ez nem hanyagság a részükről, inkább világos célt szolgál: számos alkalmazásban kifejezetten fontos, hogy a vektortér ne csak tisztított, lemmatizált szöveget tartalmazzon. Ez sok társadalomtudományi elemzésben zavaró tud lenni, ezért is érveltem korábban amellett, hogy érdemes az adott feladatunkhoz saját vektortereket képezni, jól előkészített korpuszt felhasználva. A 2. táblázatban szereplő top 15-ben természetesen feltűnnek a „rokon szakmák" is, elsőként a *psychology* (pszichológia) és az *antropology* (antropológia), utánuk pedig a *criminology* (kriminológia), az *economics* (közgazdaságtan) és a *philosophy* (filozófia). A legközelebbi természettudományi szakma, a *biology* (biológia) a 10. helyet foglalja el.

A kapott lista ránézésre logikusnak tűnik: a 2. táblázatban olyan tudományterületeket látunk, amelyek a szociológia rokon szakmáinak tekinthetők, de például a *political science* (politológia) hiányzik a listáról. Mivel ez utóbbi angol nyelven két szóból áll, vektorterünk e szóösszetételt nem tartalmazza. A top 15-ös listára még felférő *science* (tudomány) szó távolsága 0,62 volt a *sociology*-tól, a *political*-é (politikai) pedig 0,54. Érdemes megvizsgálni, hogy a *political* és a *science* szavak

közös vektora milyen távol van a *sociology* szótól. Ehhez elég egyszerűen összeadni a két szó vektorait és az összeadott vektor távolságát kiszámítani a *sociology* vektorával. A kapott értékünk 0,65 – ez már felfért volna a toplistára, a *philosophy* és a *biology* közé. E megoldás természetesen nem ekvivalens azzal a megközelítéssel, hogy már a korpusz szintjén összevonunk egybetartozó szavakat, de közelítő megoldásnak elfogadható.

Az elemzésben továbblépve megvizsgálhatjuk, hogy a kiválasztott 10 diszciplína milyen közel van egymáshoz. A 3. táblázat a tudományterületek közelségi mátrixát mutatja. Mivel a koszinuszközelség szimmetrikus, elég a mátrix egyik felét kitölteni.

3. táblázat

Tíz kiválasztott tudományterület koszinuszközelsége
(Cosine proximity of the ten selected disciplines)

Tudomány-terület neve	sociology	psychology	economics	philosophy	linguistics	biology	physics	mathematics	engineering	chemistry
sociology		**0,77**	**0,69**	**0,66**	**0,63**	**0,64**	0,57	0,55	0,52	0,52
psychology			**0,61**	**0,68**	**0,61**	**0,67**	0,58	0,56	0,53	0,56
economics				**0,65**	0,52	0,59	0,57	0,56	0,59	0,57
philosophy					0,58	**0,60**	**0,61**	**0,60**	0,57	0,55
linguistics						0,55	0,52	0,56	0,46	0,47
biology							0,69	0,59	0,56	**0,72**
physics								**0,70**	**0,60**	**0,71**
mathematics									**0,61**	**0,60**
engineering										0,57

Megjegyzés. A 0,6 feletti értékeket vastagítással jelöltük. Physics: fizika; mathematics: matematika; engineering: mérnöktudományok.

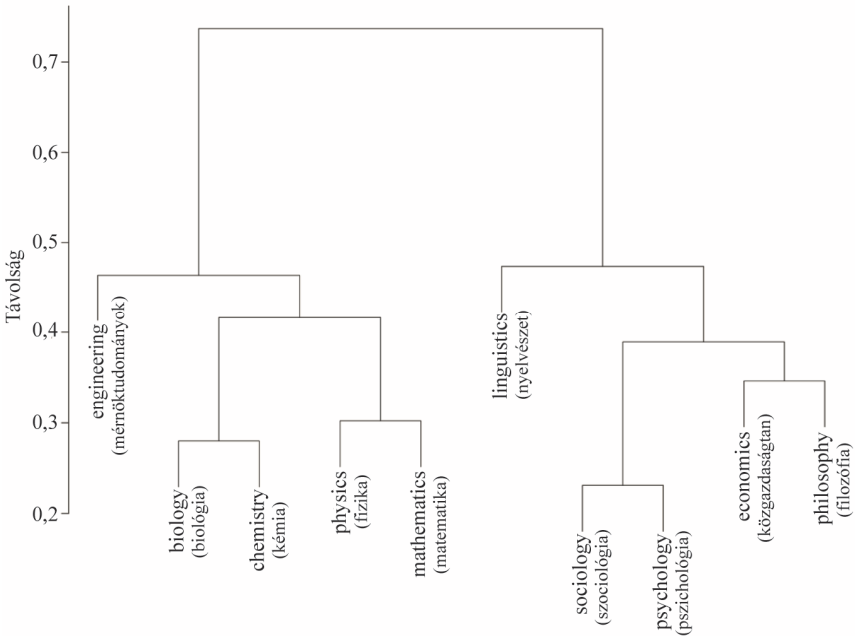
Forrás: Wikinews alapján saját számítás.

Kifejezetten alacsony koszinuszértékeket nem látunk a 3. táblázatban, de ez nem meglepő, hiszen alapvetően egy területre koncentrálunk. A mátrix alapján viszont jól kirajzolódik a tudományterületek belső összefonódása: a társadalom- és bölcsészettudományi diszciplínák nagyon közel kerülnek egymáshoz, de elválnak a természettudományi területektől. Ennek demonstrálására további elemzéseket végezhetünk a mátrixon.

A közelségmátrix jól használható mind a klaszterelemzésben (ilyenkor a közelséget át kell transzformálni távolságra), mind a kapcsolathálózat-elemzésben vagy

más dimenziócsökkentő eljárásban. A 3. ábra a 10 tudományterület hierarchikus klaszterdendrogramját szemlélteti. Az egyik „ágon” a nyelvészet, a szociológia, a pszichológia a közgazdaságtan és a filozófia helyezkedik el, míg a másikon a fizika, a matematika, a biológia, a kémia és a mérnöktudományok. Az egymással összekerülő párok triviálisnak tűnnek, ez alól talán egyedüli kivétel a közgazdaságtan és filozófia összekapcsolódása.

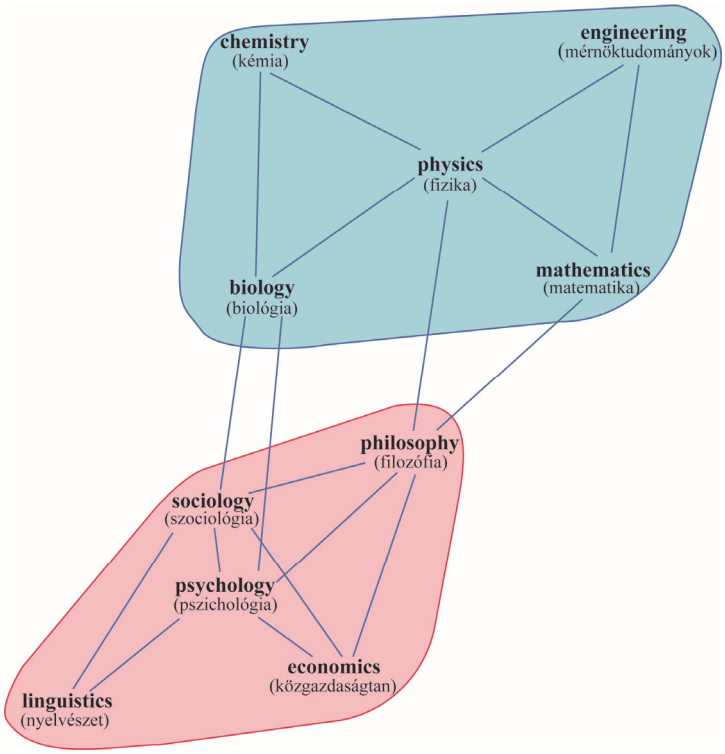
3. ábra. A tudományterületek távolságmátrixából képzett hierarchikus klaszterezés dendrogramja
(Dendrogram of hierarchical clustering formed from the distance matrix of disciplines)



Forrás: Wikinews alapján saját számítás.

A kapcsolathálózati megközelítésben érdemes egy küszöbérték alapján a közelségmátrixot 0/1 dichotóm értékeket felvevő bináris mátrixra transzformálni, ahol 1 jelenti két szó összekapcsolását a hálózatban. A klaszteres megoldással szemben a hálózati vizualizáció komplexebb összefüggéseket is képes megmutatni. A 4. ábrán szereplő hálózatnál a 0,6-es értéknél húztuk meg a határt. Ez esetben is jól kirajzolódik a két tudományterületi csoport elkülönülése, csakúgy, mint az összekapcsolódási pontok (például a biológia-szociológia-pszichológia hármasa vagy a filozófia-fizika-matematika közötti közös kapocs).

4. ábra. Tudományterületek közelségéből képzett hálózati vizualizáció
(Network visualization formed from the proximity of disciplines)

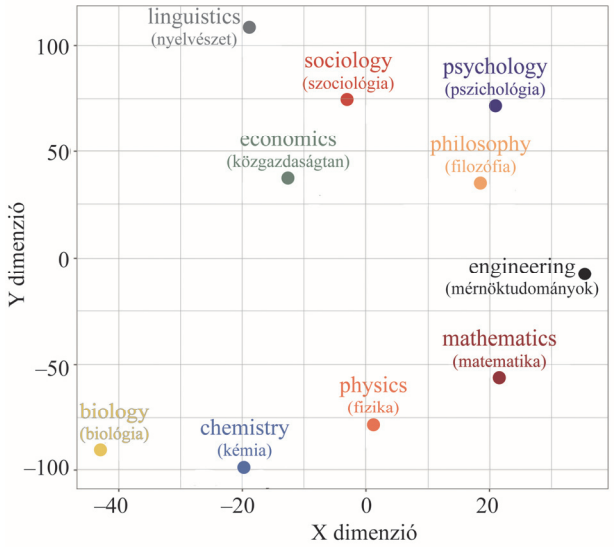


Forrás: Wikinews alapján saját számítás.

A klaszter- és a kapcsolathálózati elemzéshez képest másik megközelítést jelentenek a dimenziócsökkentő eljárások. Itt szóba jöhetnek egyszerűbb főkomponens-/faktorelemzés-alapú vagy olyan módszerek is, amelyek képesek megragadni a struktúra mögötti nemlineáris összefüggéseket. Az előbbieket inkább akkor használjuk, ha a kapott dimenziókat más elemzésekbe visszük tovább, az utóbbiakat pedig akkor, ha képet akarunk kapni egy szóhalmaz belső strukturálódásáról. A komplexebb megközelítések közül a tartalmi vizsgálat céljára leginkább a *t*-eloszlású sztochasztikus szomszéd beágyazás (*t*-distributed stochastic neighbour embedding, T-SNE) módszer terjedt el (*Maaten–Hinton* [2008]) a szöveges vektorterek esetében. Az 5. ábra a T-SNE segítségével kapott első két dimenzió szerint ábrázolja a tudományterületeket.

A dimenziók értelmezése egyáltalán nem triviális egy T-SNE modellben. Az Y a korábban már többször bemutatott természettudományok vs. társadalomtudományok dimenziója, az X ugyanakkor nehezen interpretálható.

5. ábra. Tudományterületek távolsága a T-SNE módszer alapján
(Distance of disciplines based on the T-SNE method)

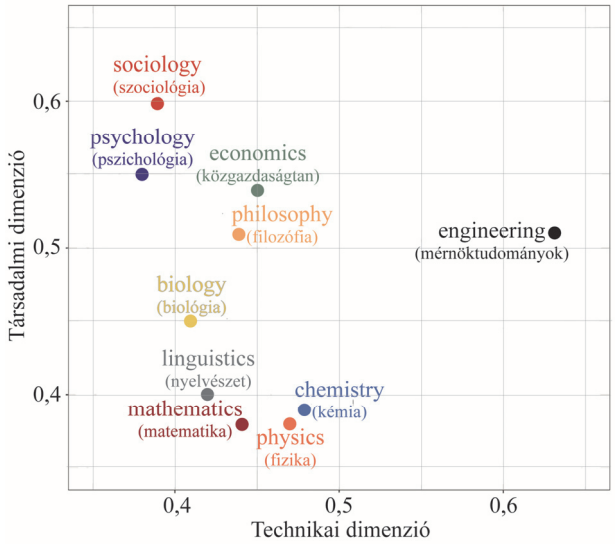


Forrás: Wikinews alapján saját számítás.

Az eddig ismertetett elemzésben a tudományterületek szavainak közelségéből indultam ki, és exploratív logikát követtem. A tudományterületek egymáshoz viszonyított pozícióját azonban külső dimenziók és konfirmatív logika alapján is lehet vizsgálni. A 6. ábrán a társadalmi és a technikai dimenziót határoztuk meg, melyeket e két szóhoz (*social* [társadalmi], *technical* [technikai]) vett közelséggel mértünk. Ezekben ugyancsak elválnak egymástól a tudományterületek. A várakozásoknak megfelelően a szociológia és a pszichológia a társadalmi dimenzióban vesz fel magas értéket, ezzel szemben a mérnöktudományok, a fizika és a kémia a technikaiban.

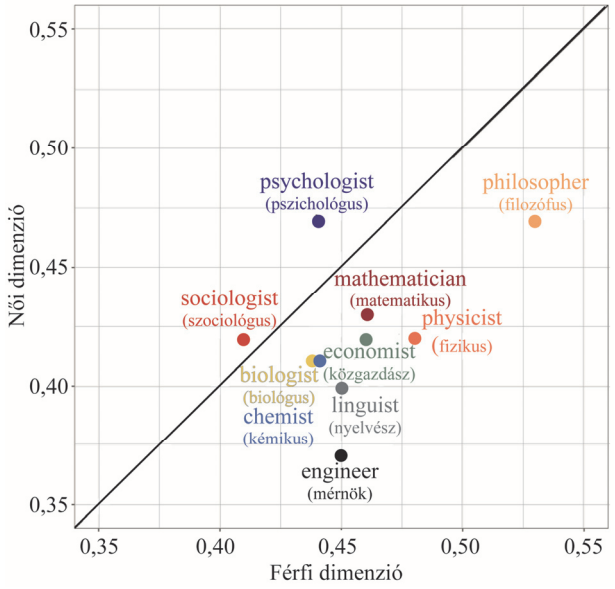
A vizsgált dimenziókban a tudományterületek elkülönülése logikusnak tűnik. Az eddig ismertetett eredmények társadalomtudományi szempontból kevésbé érdekesek, de rávilágítanak a módszerben rejlő lehetőségekre. A *social* és *technical* szavak helyére bármit behelyettesíthetünk – ezáltal komplex módon vizsgálhatjuk, hogy milyen a kommunikáció egy jelenséget érintően az online diskurzusban. Izgalmas kérdés például, hogy milyen gendervonzata van egyes szavaknak. A szociológusok lányok, a mérnökök pedig fiúk? Vagy fordítva? E kérdés „tétjére” a következő alfejezetben még visszatérek, itt csak az elemzési megközelítésre fókuszálok. Ez esetben a tudományterületek helyett a hozzájuk tartozó szakmáknak számolom ki a férfi-női távolságát. A férfi dimenzióhoz való közelséget a *he* (ő [férfi névmás]) szóval operacionalizálom, a nőit pedig a *she*-vel (ő [női névmás]). A komplexebb dimenzió meghatározására a későbbiekben térek ki.

6. ábra. Tudományterületek pozíciója a „social" (társadalmi) és a „technical" (technikai) szavakhoz viszonyítva
(Position of disciplines in relation to the words 'social' and 'technical')



Forrás: Wikinews alapján saját számítás.

7. ábra. Foglalkozások pozíciója a „he" és „she" szavakhoz viszonyítva
(Position of occupations in relation to the words 'he' and 'she')



Forrás: Wikinews alapján saját számítás.

A 7. ábra egyszerűbb interpretációját az átló segíti, amely elválasztja egymástól a női és a férfi oldalt. A legtöbb tudományterület, de leginkább a mérnöktudományok, a nyelvészet, a fizika és a filozófia, közelebb helyezkedik el a férfi dimenzióhoz. A női oldalon csak a szociológus és a pszichológus foglalkozásokat találjuk.

2.1. A szóbeágyazási vektortérmodellek társadalomtudományi szerepe és alkalmazásának lehetőségei

Ebben az alfejezetben a vektortérmodellek társadalomtudományi hasznosíthatóságára, valamint arra koncentrálok, hogy miért érdemes ezeket társadalomtudományi szempontból vizsgálni. Az első cél a társadalomtudományi kutatási keretben triviális, az utóbbi viszont nem feltétlenül az, ezért ennek tárgyalásával kezdek.

Ahogy a dolgozat korábbi részében már részletesen kifejtettem, számos mesterségesintelligencia- (MI-) alkalmazás használ nyelvi modelleket. E modellek egy részének kimenete maga is valamilyen szöveges tartalom (például egy fordítás), de sok esetben inkább valamilyen klasszifikáció (például spam-e az adott e-mail, vagy továbbjuthat-e az adott jelentkező az állásinterjú következő körébe). Egy MI-algoritmus különféle (nem szándékolt) torzításokat tartalmazhat (*Mehrabi et al.* [2019]), melyek negatívan befolyásolhatják a klasszifikációs modelljeink működését. A torzítást legegyszerűbben a fordítóprogramokon lehet vizsgálni, amit a Google Fordító működésével fogok demonstrálni (hasonló logikájú részletes elemzés kapcsán lásd *Prates–Avelar–Lamb* [2019]). A következő kis „játékot" bárki kipróbálhatja, tartalmát tetszőlegesen variálva. Vegyünk egy rövid magyar szöveget:

„Az iskolában mindenkiről készült egy jellemzés.
Szabóról a következőt mondták. Ő biztos, hogy **politikus** lesz."

A Google Fordítóval erre a következő angol szöveget kapjuk:

„A description was made of everyone in the school.
The following was said about Szabó. **He** is sure to be a **politician**."

Ebben az esetben nem érdekel minket, hogy mennyiben hibás tartalmilag a fordítás, az viszont igen, hogy a politikus a *he* szót hívja elő, tehát ha politikus lesz, akkor Szabó férfi. De mi történik akkor, ha a magyar szövegben kicseréljük a politikust tanárra?

„**She's** sure to be a **teacher**."

Ha tanár lesz, akkor nő az illető. Tehát a fordítóalgoritmus mögötti nyelvi modell azt az információt használja, hogy egyes szakmák inkább női vagy férfi környezetben fordulnak elő. A 4. táblázat néhány példát tartalmaz arra vonatkozóan, hogy különböző foglalkozásoknál férfi vagy női személyes névmást javasol-e a Google Fordító.

4. táblázat

Foglalkozásokhoz rendelt személyes névmás a Google Fordító alapján
(Personal pronoun assigned to occupations based on Google Translate)

Foglalkozás	Személyes névmás
doktor	férfi
sebész	férfi
bőrgyógyász	női
fogorvos	női
tudós	férfi
pszichológus	női
sofőr	férfi

A fordítóalgoritmus mögötti nyelvi modellt nem csak foglalkozások esetében lehet vizsgálni. Így, ha valaki okos, akkor férfi, ha érzelmes, akkor pedig nő a Google Fordító szerint. Azt gondolhatnánk, hogy az életben nincs jelentősége annak, hogy a program szerint a lányok sírósak, a fiúk pedig bátrak. A valóságban azonban nem így van, mivel ez épp annak a mechanizmusnak a manifesztálódása, amely megerősíti a szerephierarchiákat vagy a nemek közötti egyenlőtlenséget. Ráadásul, ha azt feltételezzük, hogy a Google más moduljai, például a reklámajánlati rendszer (Google Ads) is hasonló nyelvi modulokat használ, akkor a torzítás hatása már közvetlenül is megjelenik (*Datta–Tschantz–Datta* [2015]).

A klasszifikációs modelleknél nehezebb dolgunk van, ha rekonstruálni szeretnénk a torzítást. A legtöbb ipari alkalmazás ugyanis feketedobozként működik, nem nyilvánosak sem a bemeneti adatai, sem a feldolgozási algoritmusai. Még ha azonosítani is tudjuk bizonyos társadalmi csoportok hátrányos megkülönböztetését, akkor sem tudhatjuk, hogy pontosan miért jön létre a torzítás. Jó példa erre *Chen et al*. [2018] kutatása, amely azt vizsgálja, vajon állásközvetítői oldalon van-e különbség abban, hogy hányadik helyen jelennek meg a férfi és a női munkavállalók a keresőablakban. A szerzők elemzésükben kimutatják, hogy – bár az állásadók neme nincs külön regisztrálva (ők is név alapján következtettek a nemre) – minden lehetséges háttérhatás kiszűrése mellett is a női álláskeresők kissé hátrébb sorolódnak. Ennek egyik lehetséges oka, hogy az algoritmus „rátanul” olyan nyelvi elemekre az

önéletrajzokban, amelyek látens módon összefüggnek az álláskereső nemével (ennek kapcsán lásd még *De-Arteaga et al*. [2019]).

A társadalomtudósok szerepe ebben az esetben elsősorban az, hogy kritikusan vizsgálják azokat az algoritmusokat, amelyek akár napi szinten döntést hoznak az emberek életéről. Amennyiben az alkalmazások által tartalmazott torzításokat sikerül azonosítani, a készítők lépéseket tehetnek a kiküszöbölésük érdekében. A szóbeágyazási modellekkel kapcsolatban több módszertani javaslat is készült a torzítások felismerésére (*Caliskan–Bryson–Narayanan* [2017], *Garg et al.* [2018]), illetve kezelésére a vektorterekben (*Bolukbasi et al.* [2016], *Zhao et al.* [2018], *Manzini et al.* [2019], *Gonen–Goldberg* [2019]). E torzításokat természetesen nem a beágyazási algoritmus okozza, hanem a bemenetként használt korpuszok, amelyek legtöbb esetben rengeteg (nemi és etnikai) sztereotípiát hordoznak magukban (a Wikipedia kapcsán lásd például *Wagner et al.* [2015]).

A társadalomtudósok nem csak „kapuőrként" vizsgálhatják a vektortereket, azok felhasználói szempontból is értékesek lehetnek számukra. Ez utóbbi tekintetben technikai és tartalmi felhasználást különböztethetünk meg.

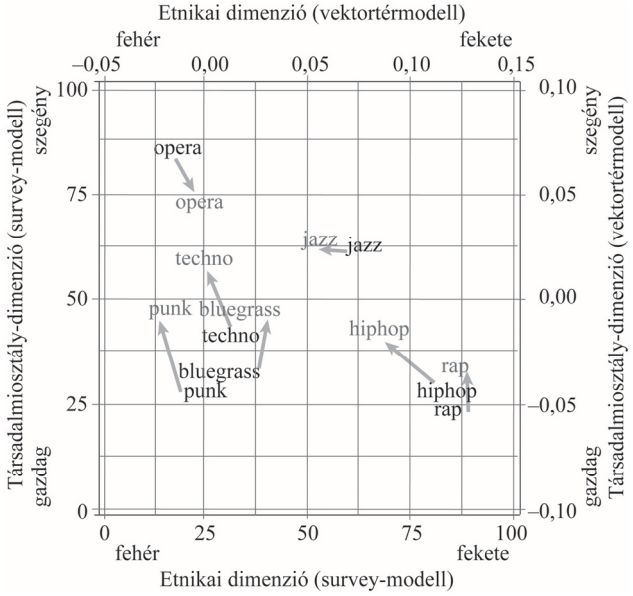
Az előbbi esetén nincs szükség hosszú magyarázatra; egyszerűen arról van szó, hogy valamilyen klasszifikációs modellben szóbeágyazási módszereket használnak. Az elemzés így nem a szóbeágyazásra fókuszál, hanem a klasszifikáció kimenetére (*Yang–Macdonald–Ounis* [2018], *Samory et al.* [2020]), de a klasszifikációs modell bizonyos paraméterei (például az, hogy egy adott kategóriába kerülést milyen szavak generálják) az elemzésben is érdekesek lehetnek (*Nakandala et al.* [2017]).

A technikai felhasználással szemben a tartalmiban már nem az algoritmus klasszifikációs ereje érdekel minket, hanem az, hogy a vektorterek milyen társadalmi összefüggések kimutatására alkalmasak. Ami az ipari alkalmazásban torzítás, az a szociológiai alkalmazásban a kibányászandó eredmény. Mely foglalkozások kötődnek erősen a nemhez vagy az etnikumhoz? Milyen szabadidő-eltöltési formák jellemzők a gazdagokra és a szegényekre? Hogyan alakult bizonyos fogalmak/csoportok társadalmi kontextusa az elmúlt 100 évben? Ez csak néhány azon társadalomtudományi kérdések közül, amelyek a vektortérmodellek segítségével megválaszolhatók.

A módszer tartami felhasználhatóságának elismerése kapcsán fontos mérföldkőnek tekinthetjük *Kozlowski*, *Taddy* és *Evans* „The geometry of culture: Analyzing the meanings of class through word embeddings" (A kultúra geometriája: az osztály jelentéseinek elemzése szóbeágyazásokon keresztül) című tanulmányának 2019-es megjelenését a szociológia zászlóshajó lapjában, az *American Sociological Review*-ban (Amerikai Szociológiai Szemle). Kozlowskiék munkája két szempontból is nagyon érdekes. A szerzők egyrészről azt vizsgálják, hogy különböző kulturális és szabadidős kérdések nemi/etnikai/társadalmiosztály-beágyazottsága mérhető-e vektortérmodellekkel. Ennek eldöntésére saját survey-kutatásukat használták fel, amely-

ben a felvételi mintába került személyektől azt kérték, hogy szemantikus differenciaskálákon osztályozzanak bizonyos kérdéseket. Az osztályozási szempont az volt, hogy a vizsgálati „objektum" mennyire férfias vagy nőies, fehér vagy afroamerikai, illetve alsó vagy felső társadalmi osztályhoz köthető. Kozlowskiék számos dimenziót bevontak az elemzésbe: ételeket, zenei stílusokat, foglalkozásokat, sportokat, járműveket és keresztneveket. A survey- és a szóbeágyazási eredmények erős összefüggést mutatnak. Leginkább a nemi bontást tekintve egyezett meg a két módszer, a survey- és a vektortérmodellekből kialakított genderskálák korrelációja 0,7 és 0,9 közötti volt. Az etnikai és a társadalmiosztály-dimenzió esetén a korrelációs értékek valamivel alacsonyabb bizonyultak, de ezeknél sem mértek 0,4 alatti összefüggést (*Kozlowski–Taddy–Evans* [2019]).

8. ábra. Zenei stílusok etnikai és társadalmiosztály-kötődése a survey és vektortérmodellekben
(Ethnic and social class 'attachment' of musical styles in survey and vector space models)



Forrás: *Kozlowski et al.* [2019].

Kozlowskiék tanulmányához hasonlóan *Joseph* és *Morgan* [2020] is survey- és vektortérmodell-eredményeket vetnek egybe, de szélesebb körű tételszettet használva. Eredményeik szerint azok a koncepciók mérhetők jól szóbeágyazással, amelyek esetében egy survey-ben is nagy az egyetértés a válaszadók között. Tehát minél extrémebb egy fogalom kulturális beágyazottsága, és azt minél kisebb szórással ítélik meg az emberek, annál erősebb az összefüggés a survey- és a vektortérmodellek

eredményei között. A szerzők elemzése arra is rávilágít, hogy fontosabb, hogy mit mérünk, mint az, hogy azt miként mérjük. A survey- és a vektortérmodellek közötti összefüggés erősségére nem igazán hatott *Joseph* és *Morgan* [2020] korpuszválasztása, de az sem, hogy melyik beágyazási algoritmust használták.

Kozlowski, *Taddy* és *Evans* [2019] tanulmányukban a külső validáció mellett a másik fontos szempont a történeti összevetés. A szerzők azt elemzik, hogy 1900-tól kezdve napjainkig miként változott a foglalkozások társadalmiosztály- és genderpozíciója, illetve összességében milyen módon alakult a társadalmiosztály- és genderfogalmak közös halmaza. Eredményeik szerint a társadalmi osztály kulturálisból „technikaibb" munkaerőpiaci kategóriává módosult, illetve e változás az Egyesült Államokban időben „elcsúszott" Angliához képest. A szerzők munkája jól illeszkedik azon tanulmányok sorába, amelyek vektortérmodellek segítségével, történeti perspektívában próbálják feltárni a fogalmi változásokat (a módszertan kapcsán lásd *Hamilton–Leskovec–Jurafsky* [2016a], [2016b], tartalmi típusú elemzés kapcsán pedig *Kulkarni et al*. [2015]).

Nagy történeti perspektívát fog át *Garg et al*. [2018] előítéletekre fókuszáló cikke is, amely 1900-tól elemzi egyes foglalkozások etnikai és genderkötődésének erősségét. A beágyazás révén kapott eredmények összhangban vannak a népszámlálási adatok alapján kirajzolódó foglalkozási mintákkal. A szerzők a különböző etnikumokhoz kötődő sztereotip kifejezéseket, valamint azt is elemzik, hogy azok a bevándorlási hullámokkal egyidejűleg milyen módon változtak. Garg-ék tanulmányával *Szabó et al*. [2020] munkáját állíthatjuk párhuzamba, amely a Kádár-korszak főbb fogalmainak változását tárja fel vektortérmodellekkel. Ez utóbbi tanulmány jó példa arra, hogy speciális korpuszokon miként lehet e módszert alkalmazni.

3. Nyelvi modellek és a társadalomtudományok – merre mutat a jövő?

Tanulmányom záró részében egy nehezen megválaszolható kérdést, a nyelvi modellek és a társadalomtudományok összekapcsolódásának lehetséges jövőbeli irányait próbálom elemezni. A kérdést nem lehet különválasztani a CSS fejlődésétől és szakmán belüli pozíciójától. Míg a 2000-es évek kulcsszava a hálózatkutatás volt, a 2010-es években a hangsúly áthelyeződött egy tágabb területre, ahol interdiszciplináris kutatócsapatok kvantitatív módszerekkel vizsgálják a digitális tartalmakat. A hálózatkutatás ilyen értelemben jó „előfutár" volt, hiszen már abban is együttműködtek egymással matematikusok, fizikusok és társadalomtudósok. A CSS terén e kör nyelvészekkel és számítógépes mérnökökkel bővült. Az egyes hálózatkutatási

területek (például a survey-alapú egonetwork-kutatások) meg tudtak maradni tisztán a szociológián belül; a CSS azonban olvasztótégelyként működik, nem lehet leválasztani belőle olyan tudományterületeket, ahol ne lenne relevanciája másoknak. Ebből következően a jövő útja az interdiszciplinaritás: azok a kutatások tudnak majd jelentőst hatást elérni, amelyek különböző tudományterületekről származó impulzusokat, ismereteket képesek összekapcsolni. Ez ugyanakkor nem jelenti azt, hogy a társadalomtudósoknak elég csak érteniük egy problémát, megoldást arra majd a mérnökök vagy a nyelvészek találnak. Ehelyett elvi szinten kell átlátniuk a modellek működését, és tisztában kell lenniük azzal, hogy milyen adatokra építve, mely módszerekkel, milyen kutatási kérdésekre lehet választ adni, és mely kutatási problémákat nem lehet azokkal megválaszolni. A módszerek „finomhangolását" természetesen rá lehet bízni a specialistákra, de ehhez érteni kell az alapokat.

A nyelvi modelleknek továbbra is három fő alkalmazási területük van a társadalomtudományokban. Elsőként a társadalomtudósoknak támogatniuk kell azokat a kutatásokat, amelyek arra irányulnak, hogy a nyelvi modelleken alapuló applikációk esetleges diszkriminációfelerősítő hatását megértsék és kiküszöböljék. Egyre több ipari alkalmazás mögött jelennek meg nyelvtechnológiai megoldások. Online ajánló rendszerek, chatbotok, fordítóprogramok – mindhárom olyan terület, ahol a rosszul tanított modellek könnyen vezethetnek diszkriminatív tartalmakhoz. E problémát a Google Fordító példáján demonstráltam a dolgozatom korábbi részében.

A második felhasználási terület különböző tartalmak klasszifikációjával kapcsolatos. Számos olyan társadalomtudományi kérdés fogalmazható meg, amelyekre elsősorban nagy szöveges adattartalmak csoportosításával lehet válaszolni. Legyen a téma például depresszió, szexizmus vagy káromkodás az online térben – nyelvi klasszifikációs modellekkel közelebb kerülhetünk a jelenségek megértéséhez. A napjainkban egyik fő megközelítésnek számító vektortéralapú módszer, a BERT 90 százalékos pontossággal meg tudja mondani, hogy egy tweet szexista-e, vagy sem (*Samory et al.* [2020]). A jelenleg alkalmazott módszerek hatásossága alig marad el a humán kódolók pontosságától. A tanító adatok megfelelő kiválasztása azonban elengedhetetlenül fontos ahhoz, hogy a nyelvi modellek jól működjenek. *Samory et al*. [2020] legújabb tanulmányukban azt mutatják be, hogy drámai mértékben javítja a szexista tweetek azonosítását, ha az annotátorok úgy írják át azokat, hogy ne legyenek szexisták. A „javított" tweetek tanuló adathalmazba keverése 10-15 százalékkal növeli a becslések pontosságát. E megközelítés jó példa arra, hogy ne tekintsünk úgy egy módszerre, mint ami mindent megold, hanem használjuk ki a társadalomtudományi területeken felhalmozott rengeteg tudást az algoritmusok minél eredményesebb tanításához.

Az utolsó, talán legkevésbé kiaknázott terület az NLP-módszerek és ezen belül a vektortérmodellek elemzési célú felhasználása, bár e tekintetben változást jelez, hogy az egyik szociológiai vezető folyóirat, az *American Sociological Review*

2019-ben közölt egy cikket a témában (*Kozlowski–Taddy–Evans* [2019]). Ugyanakkor továbbra is inkább a CSS-lapokban (lásd például *EPJ Data Science*) tudnak megjelenni a szociológiai fókuszú elemzések.

Több időnek kell eltelnie ahhoz, hogy meg tudjuk állapítani, vajon mennyire tud elterjedni a vektortérmodellek használata. E tekintetben sokat segíthet az, ha tisztában vagyunk a megfelelő alkalmazásukkal, a bennük rejlő lehetőségekkel, valamint a használatuk kapcsán felmerülő technikai kérdésekkel. Tanulmányomban ezekre igyekeztem részletesen kitérni. Dolgozatom internetes mellékleteként (https://github.com/zkmetty/nlp) elérhetővé teszem azokat a kódokat, amelyek segítségével elkezdhető a vektortérelemzés. Reményeim szerint munkám előmozdítja majd a módszer lehetséges alkalmazásaival kapcsolatos hazai diskurzust.

Irodalom

AKBIK, A. – BLYTHE, D. – VOLLGRAF, R. [2018]: Contextual string embeddings for sequence labelling. In: *Bender, E. M. – Derczynski, L. – Isabelle, P.* (eds.): *Proceedings of the 27th International Conference on Computational Linguistics.* Association for Computational Linguistics. Santa Fe. pp. 1638–1649. https://aclanthology.org/C18-1139.pdf

BENGIO, Y. – DUCHARME, R. – VINCENT, P. – JAUVIN, C. [2003]: A neural probabilistic language model. *Journal of Machine Learning Research*. Vol. 3. No. 3. pp. 1137–1155.

BOJANOWSKI, P. – GRAVE, E. – JOULIN, A. – MIKOLOV, T. [2017]: Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*. Vol. 5. June. pp. 135–146. https://doi.org/10.1162/tacl_a_00051

BOLUKBASI, T. – CHANG, K. W. – ZOU, J. Y. – SALIGRAMA, V. – KALAI, A. T. [2016]: Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: *Lee, D. D. – von Luxburg, U. – Garnett, R. – Sugiyama, M. – Guyon, I.* (eds.): *Proceedings of the 30th International Conference on Neural Information Processing Systems*. Curran Associates Inc. Red Hook. pp. 4356–4364. https://papers.nips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

CALISKAN, A. – BRYSON, J. J. – NARAYANAN, A. [2017]: Semantics derived automatically from language corpora contain human-like biases. *Science*. Vol. 356. Issue 6334. pp. 183–186. https://doi.org/10.1126/science.aal4230

CHEN, L. – MA, R. – HANNÁK, A. – WILSON, C. [2018]: Investigating the impact of gender on rank in resume search engines. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery. New York. pp. 1–14. https://doi.org/10.1145/3173574.3174225

DATTA, A. – TSCHANTZ, M. C. – DATTA, A. [2015]: Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *Proceedings on Privacy Enhancing Technologies*. No. 1. pp. 92–112. https://doi.org/10.1515/popets-2015-0007

DE-ARTEAGA, M. – ROMANOV, A. – WALLACH, H. – CHAYES, J. – BORGS, C. – CHOULDECHOVA, A. – KALAI, A. T. [2019]: Bias in bios: A case study of semantic representation bias in a

high-stakes setting. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery. New York. pp. 120–128. https://doi.org/10.1145/3287560.3287572

DEERWESTER, S. – DUMAIS, S. T. – FURNAS, G. W. – LANDAUER, T. K. – HARSHMAN, R. [1990]: Indexing by latent semantic analysis. *Journal of the American Society for Information Science*. Vol. 41. Issue. 6. pp. 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

DEVLIN, J. – CHANG, M. W. – LEE, K. – TOUTANOVA, K. [2019]: BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Burstein, J. – Doran, Ch. – Solorio, Th*. (eds.): *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics. Minneapolis. pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423

FOKASZ N. – TÓTH G. – MICSINAI I. – JELENFI G. – ELŐD Z. [2015]: Kampány és valóságkonstrukció. A 2010-es és a 2014-es választási kampányok összehasonlító elemzése a NOL és az MNO oldalakon megjelent kampánytémák dinamikája alapján. *Jel-Kép*. 36. évf. 3. sz. 25–63. old. https://doi.org/10.20520/Jel-Kep.2015.3.25

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*. Vol. 115. No. 16. pp. E3635–E3644. https://doi.org/10.1073/pnas.1720347115

GONEN, H. – GOLDBERG, Y. [2019]: Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In: *Burstein, J. – Doran, Ch. – Solorio, Th*. (eds.): *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Association for Computational Linguistics. Minneapolis. pp. 609–614. https://doi.org/10.18653/v1/N19-1061

HAMILTON, W. L. – LESKOVEC, J. – JURAFSKY, D. [2016a]: Diachronic word embeddings reveal statistical laws of semantic change. In: *Erk, K. – Smith, N*. (eds.): *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. Berlin. pp. 1489–1501. https://doi.org/10.18653/v1/P16-1141

HAMILTON, W. L. – LESKOVEC, J. – JURAFSKY, D. [2016b]: Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In: *Su, J. – Duh, K. – Carreras, X*. (eds.): *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics. Austin. pp. 2116–2121. https://doi.org/10.18653/v1/D16-1229

INDIG B. [2018]: Közös crawlnak is egy korpusz a vége – Korpuszépítés a CommonCrawl.hu domainjaiból. In: *Vincze V.* (szerk.): *XIV. Magyar Számítógépes Nyelvészeti Konferencia*. Szegedi Tudományegyetem. Szeged. 125–134. old.

JOSEPH, K. – MORGAN, J. H. [2020]: When do word embeddings accurately reflect surveys on our beliefs about people? In: *Jurafsky, D. – Chai, J. – Schluter, N. – Tetreau, J*. (eds.): *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics. pp. 4392–4415. https://aclanthology.org/2020.acl-main.405.pdf

JOULIN, A. – GRAVE, E. – BOJANOWSKI, P. – MIKOLOV, T. [2016]: Bag of tricks for efficient text classification. In: *Lapata, M. – Blunsom, Ph. – Koller, A*. (eds.): *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers.* Association for Computational Linguistics. Valencia. pp. 427–431. https://aclanthology.org/E17-2068.pdf

KMETTY Z. [2018]: A szociológia helye a Big Data-paradigmában és a Big Data helye a szociológiában. *Magyar Tudomány*. 179. évf. 5. sz. 683–692. old. https://doi.org/10.1556/2065.179.2018.5.11

KMETTY, Z. – KOLTAI, J. – RUDAS, T. [2021]: The presence of occupational structure in online texts based on word embedding NLP models. *EPJ Data Science.* Vol. 10. No. 55. pp. 1–20. https://doi.org/10.1140/epjds/s13688-021-00311-9

KOZLOWSKI, A. C. – TADDY, M. – EVANS, J. A. [2019]: The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*. Vol. 84. No. 5. pp. 905–949. https://doi.org/10.1177/0003122419877135

KULKARNI, V. – AL-RFOU, R. – PEROZZI, B. – SKIENA, S. [2015]: Statistically significant detection of linguistic change. In: *Proceedings of the 24th International Conference on World Wide Web*. Association for Computing Machinery. New York. pp. 625–635.

LISON, P. – KUTUZOV, A. [2017]: Redefining context windows for word embedding models: An experimental study. In: *Tiedemann, J. – Tahmasebi, N*. (eds.): *Proceedings of the 21st Nordic Conference on Computational Linguistics*. Association for Computational Linguistics. Gothenburg. pp. 284–288. https://aclanthology.org/W17-0239.pdf

LIU, H. [2008]: Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*. Vol. 9. No. 2. pp. 159–191. https://doi.org/10.17791/jcs.2008.9.2.159

MAATEN, L. V. D. – HINTON, G. [2008]: Visualizing data using t-SNE. *Journal of Machine Learning Research*. Vol. 9. November. pp. 2579–2605.

MANZINI, T. – LIM, Y. C. – TSVETKOV, Y. – BLACK, A. W. [2019]: Black is to criminal as Caucasian is to police: Detecting and removing multiclass bias in word embeddings. In: *Burstein, J. – Doran, Ch. – Solorio, Th*. (eds.): *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* Association for Computational Linguistics. Minneapolis. pp. 615–621. https://aclanthology.org/N19-1062.pdf

MIKOLOV, T. – CHEN, K. – CORRADO, G. – DEAN, J. [2013]: Efficient estimation of word representations in vector space. Poster presentation. *International Conference 'Learning Representations' 2013*. 2–4 May. Scottsdale.

MEHRABI, N. – MORSTATTER, F. – SAXENA, N. – LERMAN, K. – GALSTYAN, A. [2019]: A survey on bias and fairness in machine learning. *ACM Computing Surveys*. Vol. 54. No. 6. Article No. 115. pp. 1–35. https://doi.org/10.1145/3457607

MNIH, A. – KAVUKCUOGLU, K. [2013]: Learning word embeddings efficiently with noise-contrastive estimation. In: *Burges, C. J. C. – Bottou, L. – Welling, M. – Ghahramani, Z. – Weinberger, K. O.* (eds.): *Proceedings of the 26th International Conference on Neural Information Processing Systems.* Curran Associates Inc. Red Hook. pp. 2265–2273.

NAKANDALA, S. – CIAMPAGLIA, G. L. – SU, N. M. – AHN, Y. Y. [2017]: Gendered conversation in a social game-streaming platform. *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 11. No. 1. pp. 162–171.

NÉMETH R. – KATONA E. R. – KMETTY Z. [2020]: Az automatizált szöveganalitika perspektívája a társadalomtudományokban. *Szociológiai Szemle.* 30. évf. 1. sz. 44–62. old. https://doi.org/10.51624/SzocSzemle.2020.1.3

NÉMETH, R. – KOLTAI, J. [2021]: Discovering sociological knowledge through automated text analytics. In: *Rudas, T. – Péli, G*. (eds.): *Pathways Between Social Science and Computational Social Science – Theories, Methods and Interpretations*. Springer. New York.

PENNINGTON, J. – SOCHER, R. – MANNING, C. D. [2014]: GloVe: Global vectors for word representation. In: *Moschitti, A. – Pang, B. – Daelemans, W*. (eds.): *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).* Association for Computational Linguistics. Doha. pp. 1532–1543. https://doi.org/10.3115/v1/D14-1162

PETERS, M. E. – NEUMANN, M. – IYYER, M. – GARDNER, M. – CLARK, C. – LEE, K. – ZETTLEMOYER, L. [2018]: Deep contextualized word representations. In: *Walker, M. – Ji, H. – Stent, A.* (eds.): *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics. New Orleans. pp. 2227–2237. https://doi.org/10.18653/v1/N18-1202

PRATES, M. O. – AVELAR, P. H. – LAMB, L. C. [2019]: Assessing gender bias in machine translation: A case study with Google Translate. *Neural Computing and Applications*. Vol. 32. Issue 10. pp. 6363–6381. https://doi.org/10.1007/s00521-019-04144-6

RADFORD, A. – NARASIMHAN, K. – SALIMANS, T. – SUTSKEVER, I. [2018]: *Improving Language Understanding with Unsupervised Learning*. Technical Report. OpenAI. https://openai.com/blog/language-unsupervised/

ROHDE, D. L. – GONNERMAN, L. M. – PLAUT, D. C. [2006]: An improved model of semantic similarity based on lexical co-occurrence. *Communications of the ACM*. Vol. 8. pp. 627–633.

SAMORY, M. – SEN, I. – KOHNE, J. – FLOECK, F. – WAGNER, C. [2020]: 'Unsex me here': Revisiting sexism detection using psychological scales and adversarial samples. *Computer Science*. 27 April. Corpus ID: 216553394.

SCHAKEL, A. M. – WILSON, B. J. [2015]: *Measuring word significance using distributed representations of words.* arXiv:1508.02297.

SPIRLING, A. – RODRIGUEZ, P. L. [2019]: *Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research*. Working Paper. https://arthurspirling.org/documents/embed.pdf

SZABÓ M. K. [2019]: Az értékváltás jelensége a magyar nyelvben. A negatív emotív elemek egy sajátos használatáról. *Magyar Nyelv*. 115. évf. 3. sz. 309–323. old. https://doi.org/10.18349/MagyarNyelv.2019.3.309

SZABÓ, M. K. – RING, O. – NAGY, B. – KISS, L. – KOLTAI, J. – BEREND, G. – VIDÁCS, L. – GULYÁS, A. – KMETTY, Z. [2020]: Exploring the dynamic changes of key concepts of the Hungarian socialist era with natural language processing methods. *Historical Methods*. Online first. pp. 1–13. http://doi.org/10.1080/01615440.2020.1823289

YANG, X. – MACDONALD, C. – OUNIS, I. [2018]: Using word embeddings in Twitter election classification. *Information Retrieval.* Vol. 21. Nos. 2–3. pp. 183–207. https://doi.org/10.1007/s10791-017-9319-5

VASWANI, A. – SHAZEER, N. – PARMAR, N. – USZKOREIT, J. – JONES, L. – GOMEZ, A. N. – KAISER, L. – POLOSUKHIN, I. [2017]: Attention is all you need. In: *von Luxburg, U. – Guyon, I. – Bengio, S. – Wallach, H. – Fergus, R.* (eds.): *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Curran Associates Inc. Red Hook.

WAGNER, C. – GARCIA, D. – JADIDI, M. – STROHMAIER, M. [2015]: It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 9. No. 1. AAAI Press. Palo Alto. pp. 454–463. https://ojs.aaai.org/index.php/ICWSM/article/view/14628/14477

WIEDEMANN, G. – REMUS, S. – CHAWLA, A. – BIEMANN, C. [2019]: *Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings*. https://www.inf.uni-hamburg.de/en/inst/ab/lt/publications/2019-wiedemannetal-konvens-bert.pdf

ZHAO, J. – ZHOU, Y. – LI, Z. – WANG, W. – CHANG, K. W. [2018]: Learning gender-neutral word embeddings. In: *Blanci, E. – Lu, W*. (eds.): *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. Brussels. pp. 4847–4853.