Dr. Kmetty Zoltán

Szóbeágyazási vektortér modellek társadalomtudományi alkalmazása

Habilitációs dolgozat

Eötvös Loránd Tudományegyetem

Társadalomtudományi Kar

2020

Tartalomjegyzék

1 Bevezetés

A társadalomtudományok empirikus eszköztára talán sosem bővült annyira intenzíven, mint az elmúlt 20-30 évben. A számítógépes kapacitás (mind tárhely, mind sebesség) növekedése és a különböző programnyelvek használatának egyre szélesebb körű elterjedése lehetővé tette korábban nem, vagy csak nagyon nehezen implementálható módszerek általános elterjedését. Ezzel párhuzamosan a kvantitatív elemzésekben egyre több új típusú adatforrás jelent meg. A relációs network adatbázisok és a társadalmi hálózatelemzés jelentette az első hullámot az új típusú adatokban, a szöveges adatok elemzése pedig a második hullámot[1]. Sem a kapcsolathálózati, sem a szöveges adatok elemzése nem az elmúlt 30 évben kezdődött, de az igazi társadalomtudományi felfutás ebben az időszakban volt megfigyelhető – a network módszereknél a 90-es évektől, a szöveges adatok esetében pedig a 2010-es évektől. A felfutást ebben az esetben részben a már korábban említett számítógépes „forradalom" táplálta, részben pedig az ezzel erősen összefüggő digitalizáció (Kmetty 2018).

Az új típusú adathalmazok nagyban különböztek a tradicionális – elsősorban survey vagy adminisztratív alapú adathalmazoktól[2] – és ez a különbözőség új elemzési módszereket indukált. A dolgozat fókuszában álló szöveges adatok esetében mind az adatfeldolgozásra, mind az adatelemzésre speciális módszerek alakultak ki. Ezt a módszertani irányt a természetes nyelvfeldolgozás (Natural Language Processing – NLP) illetve a szövegbányászat (text mining) címkéjével látták el. A két fogalom tartalmi átfedéseinek és különbözőségeinek vizsgálata nem képezi a dolgozat lényegi

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

[2] A téri és temporális adatok elemzéséhez szintén kialakult egy speciális módszertan, de a távolság itt nem olyan nagy, mint a network vagy a szövegelemzés esetében.

részét, de ha mégis meg kell különböztetnünk a fogalmakat, akkor azt mondhatjuk, hogy a szövegbányászat (részben) NLP módszereket használ ahhoz, hogy feldolgozzon szabadszavas vagy strukturált szöveges tartalmakat (Kao – Poteet 2007).

Se az NLP, se a szövegbányászati módszertan nem a társadalomkutatások oldaláról indul. Alapvetően számítógépes nyelvészek, statisztikusok, számítástudománnyal foglalkozók („Computer scientist") és részben fizikusok, illetve más természettudósok vannak a módszertan fejlesztés élén a 90-es és 2000-es években. A felhasználásban dominál az üzleti vonal, nem véletlen, hogy a Google vagy a Facebook élen jár az algoritmusok fejlesztésében (lásd később). A tisztán nyelvészeti és üzleti felhasználás mellett (pl: információ kinyerés, spam klasszikfikáció, stb..) természetesen a módszerek új perspektívát nyitottak a társadalomtudományi kutatások előtt, ami rövid idő alatt egy új interdiszciplináris tudományterület kialakulásához vezetett, ami számítógépes társadalomtudomány (Computational Social Science - CSS) néven kezdte el intézményesülési útját. Az intézményesülést jól jelzi a területhez köthető kutatócsoportok, szakok és dedikált konferenciák megjelenése a 2010-es években (annak elsősorban második felében). Az egyes társadalomtudományi[3] diszciplínák eltérő mértékben kapcsolódtak be, kapcsolódtak hozzá az új területhez. A politikai kommunikáció, a tudománymetrika (Science of Science) vagy a kulturális fogyasztás mind olyan területek, amik hatékonyan kutathatók CSS módszerekkel (Kmetty 2018), és ezek a területek hangsúlyosan meg is jelennek a szakmai konferenciákon, szemben az olyan „hagyományos" szociológia témákkal, mint a szegénység, amely témában sokkal kevesebb CSS kutatás készült (eddig). Bár a szöveges adatokon keresztül

[3] Ebben a dolgozatban a társadalomtudományokra fókuszálunk és nem foglalkozunk a digitális bölcsészettel, ami módszertanában sok átfedést mutat a CSS-sel.

jutottunk el a CSS fogalmáig, utóbbi jóval szélesebb adatkört tartalmaz – a területhez köthetjük a digitális adatforrásokat függetlenül azok tartalmától, de a különböző szimulációs területeket is (pl: ágens alapú szimuláció). De a numerikus adatok elemzéséhez nem kellett új eszköztárat építeni, „csak” a régit kellett újrahangolni és kiegészíteni (Kmetty 2018). A szöveges adatok elemzéséhez viszont teljesen új eszköztárra volt szükség (Németh – Katona – Kmetty 2020). Habilitációs dolgozatomban ennek az új eszköztárnak egy újabb elemét fogom mélyebben bemutatni a szóbeágyazási vektortér modelleket (word-embedding model). Egy nagyon gyorsan fejlődő terület esetén persze kérdés, mit is nevezhetünk újnak. Ahogy látjuk majd a módszer bemutatásánál, az előzmények már a 80-as évek végén megjelennek de a robbanásszerű elterjedés a 2010-es évek elejére tehető, ami alapvetően egy jól implementálható neurális háló alapú megközelítés megjelenéséhez köthető (Mikolov et al 2013).

De mit is nevezünk szóbeágyazásnak vagy vektortér modellnek? Gyakorlatilag egy dimenziócsökkentő eljárásról beszélünk, ahol célunk az, hogy meghatározzuk egy adott szó milyen környezetben, milyen kontextusban szerepel. Dimenziócsökkentésre azért van szükség, mert egy szó egy szövegben rengeteg más szó mellett előfordulhat és egy nyers szógyakoriság nem tud képet adni arról, hogy mi ez a kontextus. Egy „kicsi” 100 000 egyedi szót tartalmazó szövegben is, egy 100 000 * 100 000 azaz 10 000 000 000 cellát tartalmazó mátrixot kellene kiszámolnunk ahhoz, hogy felrajzoljuk a szavak együttjárását. A szövegek pedig jellemzően ennél jóval több egyedi szót tartalmaznak. A dimenziócsökkentés során ezt a kontextust redukáljuk le egy szavak_száma*dimenzió mátrixra, ahol a dimenziószám jellemzően 200-400 között mozog. Ez a csökkentett dimenzió megtartja azt a „jó” tulajdonságát, hogy az

egymással megegyező kontextusban szereplő szavak közel vannak egymáshoz[4], és emellett a mátrix nagysága miatt egyszerűen elemezhető normál erőforrássokkal is.

A habilitációs dolgozatomban azt a célt tűzöm ki, hogy bemutassam, hogyan lehet használni a vektortér modelleket társadalomtudományi kutatásokban. A dolgozat első felében bemutatom a módszer alapvetéseit, történeti evolúcióját röviden áttekintem a matematikai hátteret, valamint a módszert övező dilemmákat. Kitérek arra, hogy mit mérnek ezek a modellek és milyen jellemző társadalomtudományi alkalmazások vannak a nemzetközi szakirodalomban. A dolgozat nem csak azt a célt szolgálja, hogy bemutassa a vektortér modellek társadalomtudományi használatát, hanem azt is, hogy lehetőséget adjon az érdeklődőknek arra, hogy maguk is kipróbálják a módszert. Ezért a vektortér modellek használatát tárgyaló fejezetben bemutatott példákhoz elérhetővé teszem a futtatásukhoz szükséges R kódokat (https://github.com/zkmetty/nlp).

A dolgozat második felében 3 olyan esettanulmányt mutatok be, amelyekben a szóbeágyazási módszert alkalmaztuk különböző társadalomtudományi kérdések megválaszolására.

1. A Koltai Júliával és Rudas Tamással közös munkánkban (Kmetty – Koltai – Rudas 2020) azt vizsgáltuk, hogy online szövegek alapján milyen foglalkozási hierarchiát lehet felállítani és a felállított foglalkozási hierarchia mennyiben vág egybe a klasszikus foglalkozási rangsorokkal. Tanulmányunkban olyan hierarchiát alakító aspektusok fontos szerepére hívjuk fel a figyelmet, ami eddig a szakirodalomban kevéssé volt hangsúlyos.

[4] A közelség definíciójára később visszatérünk

2. Az ELKH TK, CSS-Recens kutatócsoportjával Kádár-kori újságcikkeket elemeztünk vektortér modellekkel (Szabo et al 2020). Munkánk fókuszában a mezőgazdaság és ipar szerepének változása állt a Kádár-korszakban.
3. A Tóbiás Dániellel közös munkánkban pedig azt vizsgáltuk, hogy a Twitch videomegosztó portálon különböznek-e egymástól a férfi és női gamerek videói alatti reakciók – megjelenik-e objektifikáció a női gamerek videói alatt.

A három esettanulmány nem csak témájában különbözik, hanem a bemeneti adatok jellegében is. A foglalkozási pozíciót vizsgáló munkánkban nem nyers szövegekből indulunk ki, hanem előre elkészített vektorterekből. Ennek a megközelítésnek az előnyeit és hátrányait később részletesen körbejárjuk. A történeti fókuszú esettanulmány egy rendszerváltozás előtti magyar korpuszt dolgoz fel. Szemben a numerikus adatfeldolgozás univerzális nyelvével, a szövegek elemzése több esetben lokális megoldásokat követel meg. Az esettanulmány azt a célt is szolgálja, hogy bemutassuk magyar szövegek esetén milyen előfeldolgozási megoldások állnak rendelkezésre.

A Twitch fórumok elemzése ehhez képest teljesen eltérő megközelítést kíván, mivel itt jellemzően struktúráltalan, de nagyon rövid szövegeket kell vizsgálnunk, ahol ráadásul sok speciális szövegelem (pl. jellemző emojik, rövidítések) is megjelenik.

A dolgozat megértéséhez nincs szükség nagyon erős statisztikai háttérre, a módszereket elsősorban heurisztikusan mutatom be. A technikai részleteket külön szövegdobozban tárgyalom. Ezeket a szövegdobozokat elsősorban azoknak ajánlom elolvasni, akik szeretnék mélyebben megérteni a módszer statisztikai hátterét.

2 Vektortér modellek – statisztikai alapok, módszertani megfontolások

2.1 A vektortér modellek evolúciója

2.1.1 A kályhától...

A dolgozat középpontjában egy elmúlt évtizedben felfutó, több tudományterületen is használható statisztikai elemzési módszer áll – a szóbeágyazás. A módszer alaplogikájának megértéséhez érdemes visszalépni a kvantitatív szövegelemzés alapjaihoz.

Ha nagyon általánosan meg akarjuk érteni, hogy egy nagy szöveghalmaz milyen tartalommal rendelkezik akkor a legegyszerűbb megoldás az, ha megvizsgáljuk mik a szövegben előforduló leggyakoribb szavak. Ha az általánosan minden szövegben megjelenő szavakon túljutunk (pl: névelők) akkor rövid idő után el fogunk jutni olyan szavakig, amelyek már jelezhetik azt, hogy milyen tartalom van egy szövegben. Ez a megközelítés viszonylag egyszerű elemzéseket eredményez, inkább kiindulásnak használható komplexebb munkák előtt. Persze egy ilyen megközelítés is adhat önmagában értékes kimenetet (Fokasz et al. 2015), például beszédes lehet, hogy különböző típusú oldalak milyen szavakat használnak vagy nem használnak ugyanabban a témában (pl: migráns vs menekült).

Egy adott szó használatának gyakorisága azonban csak nagyon korlátozott információval látja el az elemzőt. Egy szó önmagában kevéssé érdekes, sokkal érdekesebb az, hogy milyen kontextusban fordul elő. Ugyanarról a politikusról írhatnak sokat kormánypárti és ellenzéki médiában is, de nagy valószínűség szerint a kontextus különböző lesz.

A kontextus megragadásának legegyszerűbb módja azoknak a szavaknak a megkeresése, amelyek gyakran szerepelnek az általunk vizsgálni kívánt szavak közelében. Az egyes nyelvek szókészlete viszont nagyon széles, teli szinonímákkal, ragokkal, igeidőkkel. A szövegek előfeldolgozása (lásd később) sokat tud ezen a problémán egyszerűsíteni, de nem ad megnyugtató választ arra, hogyan tudjuk egy szó teljes szókörnyezetét feldolgozni. Erre a problémára nyújtanak jó megoldást a szóbeágyazási modellek. Ha nagyon heurisztikusan akarjuk megközelíteni a kérdést, akkor azt mondhatjuk, hogy a szóbeágyazási modelleknek alapvetően az a célja, hogy megtalálják azt, hogy egy adott szó (vagy szó n-gram – lásd később) milyen más szavakhoz van közel a felhasznált korpuszban.

A probléma megoldására definiált első megközelítések együttes szóelőfordulási mátrixokból indultak ki. Az alapvető matematikai probléma ebben az esetben az, hogy van egy nagyon ritka mátrixunk[5], amiben a szavak együttes előfordulását tároljuk. Együttes előfordulást sokféleképpen lehet definiálni, de általában szóközelséggel szokták ezt megoldani. Tehát egy adott szó környezetét a +-1 „X” távolságban lévő szavak jelentik. Az X általában nem nagyobb mint 10, de persze speciális esetben lehet ennél nagyobbra is venni az ablakot. Egy nem különösen nagy, 100 000 szót tartalmazó szövegben az együttes előfordulási mátrix mérete: 100 000 * 100 000, tehát 10 000 000 000 cellából áll. Viszonylag egyértelmű, hogy egy ilyen nagyságú és kifejezetten ritka mátrix esetében a direkt statisztikai elemzés fel sem merülhet. A cél tehát egy olyan redukált mátrix kialakítása, ami lehetőleg minél nagyobb arányban megőrzi az eredeti mátrix információtartalmát és akkora méretű, amit már tudunk

[5] A ritka mátrixnak nincs definíciója, de olyan mátrixot érdemes magunk elé képzelni, amiben a cellák kevesebb mint 0.1 százalékában van nullától elétrő érték.

kezelni standard statisztikai eszközökkel. Ez egy klasszikus dimenziócsökkentési eljárási probléma. Ez a dimenziócsökkentés nem ismeretlen a társadalomtudósoknak sem, 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 (SVD) használták, elsősorban dokumentáció klasszifikációs céllal – ezt a módszert nevezik látens szemantikus elemzésnek (LSA – Deerwester et. al 1990).

LSA

Az LSA-ban a kiindulás egy dokumentum-szógyakoriság mátrix (document-term matrix – DTM), de a számolás alapját jelentő SVD módszer használható közös szó-előfordulási mátrixon is (term co-occurance matrix – TCM). Az SVD-n alapú mátrix faktorizációs modellekre a disztribúciós szemantikus modellek (DSM) címkét is használják. Az SVD modellek bemeneti adatai egy együttes szóelőfordulási mátrix (TCM) volt, kimenetként pedig egy alacsony dimenziós (100-1000) vektorteret adott ki a modell. A szavak gyakorisága nagyon ferde eloszlást követ, akárcsak az együttes szóelőfordulás, és ez a szógyakoriság a vektorterekben is visszaköszönt – a gyakoribb szavak közelebb voltak egymáshoz. Ennek kiküszöbölésére számos megoldás született, például a nagyon gyakori szópároknál egy küszöbértékben maximalizálták az együttes előfordulások számát (vagy akár ki is hagyták a gyakori szópárokat), más szerzők a TCM táblában a logaritmusokat használták a nyers számok. helyett, vagy a szavak Pearson korrelációját (Rohde et al 2006).

A mátrix faktorizációs modellekhez képest egy más megközelítést javasoltak Bengio és munkatársai (2003). Bengio-ék használtak elsők között neurális hálókat a szókontextus megbecslésére. A DSM modellek alapvetően gyakorisági modellek, míg a neurális háló alapúnyelvi modellek (NNLM) predikciósak, tehát szemben az előbbi

heurisztikus megközelítésével, a szavak kontextusát ezekben a neurális háló alapú modellekben a lokális környezetük alapján becsülték meg a szerzők.

Mind a DSM, mind a neurális háló alapú modellek (összefoglaló néven szóbeágyazási modellek, vagy vektortér modellek) megmaradtak egy szűkebb tudományos közösségen belül a 2000-es években. Az akkori gépkapacitások mellett ezek a modellek nagy szövegeken gyakorlatilag nem futottak le véges idő alatt, ezért csak egy nagyon szűk kör foglalkozott ezzel a nyelvfeldolgozási iránnyal.

2.1.2 Word2Vec

Az igazi áttörés Mikolov és munkatársai (2013)-as cikke hozta el. Az akkor Google-nél dolgozó csapat egy olyan neurális háló alapú módszert fejlesztett ki, ami szélesebb közönségnek elérhetővé tette a módszer használatát. A word2vec névre keresztelt módszer nem csak feladat megoldásban működött nagyon jól (lásd később), hanem a korábbi módszerekhez képest kevesebb erőforrást is igényelt. A munka alapjait Mikolov egyébként már 2007-ben letette MSC szakdolgozatában.

Word2vec

A korábban már említett Bengio et al (2003) cikkben egy négy rétegből álló „mély” neurális hálót javasoltak a szerzők: bemeneti, projekciós, rejtett, kimeneti a klasszifikációs modell kiszámolására. Nagy szövegkorpusz esetében ez a megközelítés nagyon magas számítási idővel jár. A problémára szintén használt rekurrens neurális hálókban (lásd Mikolov et al 2013) annyiban változtatnak az alap modellen, hogy kimarad a projekciós réteg, viszont a rejtett réteghez egy rekurrens mátrix kapcsolódik, ami gyakorlatilag folyamatosan új inputtal táplálja a rejtett réteget. Ebben a megközelítésben a számítási komplexitás csökken, de továbbra is nagyon magas.

Az úttörő, Mikolov et al (2013) cikkben a szerzők ehhez képest egy valamivel egyszerűbb modellt javasolnak. Mivel a korábbi modellek komplexitásáért elsősorban a rejtett réteg felel, ezt kihagyják a neurális hálóból és 3 réteget szerepeltetnek: bemeneti, projekciós[6], kimeneti. Az újítás az ő megközelítésükben az, hogy vegyítik egy klasszifikációs modellel a neurális hálót, gyakorlatilag egy klasszifikációs modellt építenek. A bemeneti adatot egy adott szó kontextus jelenti, a kimeneti adatot pedig egy szó. A klasszifikációs modell egy „egyszerű" logisztikus regresszió. A végeredmény szempontjából, ami fontos, az a logisztikus regresszióban az egyes szavak mellé rendelt súlyok. Ezek a súlyok jelentik gyakorlatilag az egyes szavak vektorreprezentációját.

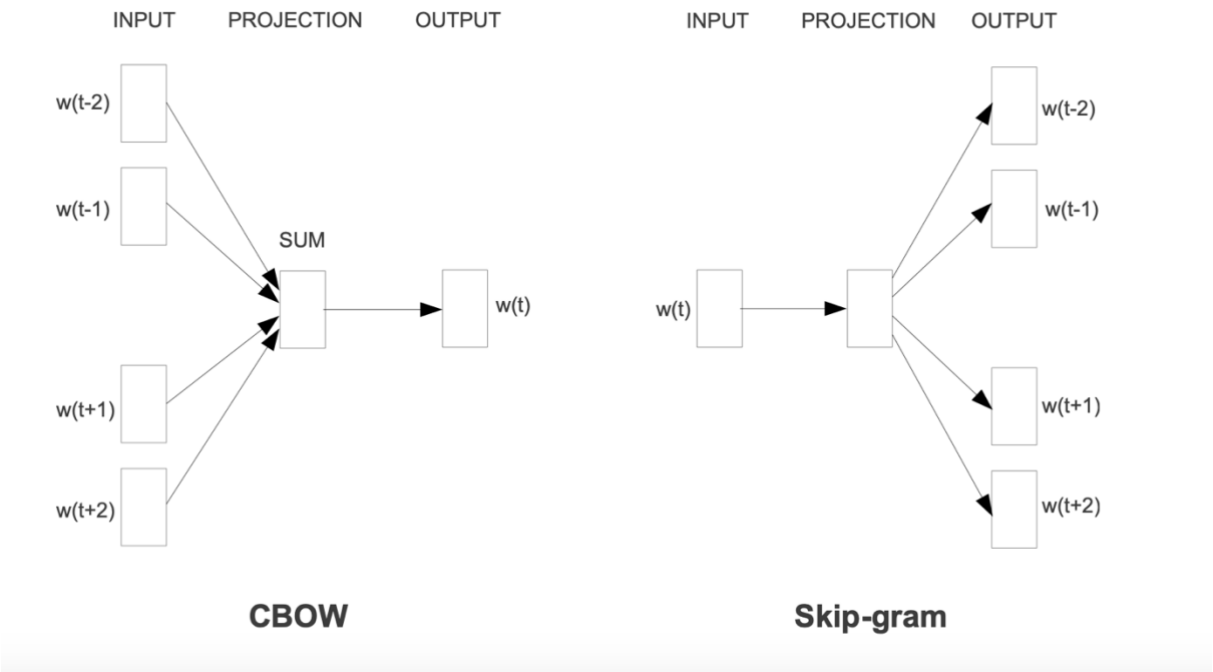
Mikolovék két megközelítést is bemutatnak cikkükben. A folytonos szózsák modellben (Continuous Bag of Words - CBOW) egy adott szó környezetéből indulnak ki, és abból próbálják megbecsülni a szót. A skip-gram modellben fordított logikát alkalmaznak (Mikolov et al 2013), egy adott szóból becsülik meg a szó kontextusát (hasonló megközelítés kapcsán lásd például: Mnih – Kavukcuoglu 2013). Az 1. ábra ezt a két megközelítést szemlélteti.

Negative Sampling

A word2vec modellek lokális ablakot használnak a becsléskor. A lokális ablakban azonban csak azt látjuk, hogy milyen szavak szerepelnek együtt egy kontextusba. A logisztikus regressziós modellhez kellenek olyan szavak is, amik nincsenek az adott kontextusban. Ha az összes szövegben szereplő szóval végigtesztelnénk, minden egyes ablakra, hogy az adott kontextus szó megjelenik-e a vizsgált szavunk mellett, akkor a futási idő extra magas lenne. Erre megoldásként egy mintavételes megközelítést javasolnak a szerzők. Véletlenszerűen kiválasztanak más szavakat is a korpuszból [k szó, k jellemzően kisebb. mint 10] és ezekkel a szavakkal illesztik a modellt. Ezt nevezik „negative sampling" megközelítésnek (Goldberg – Levy 2014).

[6] A Mikolov et al 2013 cikkben végig következetesen projekciós rétegnek nevezik, de más munkákban rejtett rétegként hivatkozzák.

A CBOW módszer nagyon gyors, akár nagy adatbázisokon is jól használható, de alapvetően csak a gyakrabban előforduló szavak esetén ad megbízható eredményt. Ezzel szemben a skip-gram modell jól működik ritkább szavakra is, de inkább kisebb korpuszok esetében javasolt, a magasabb számolási idő miatt. A word2vec modell végeredménye gyakorlatilag a projekciós (rejtett) rétegben kiszámolt klasszifikációs modell súlyai, ezt használhatjuk a szövegünk további elemzésére.



Ábra 1. A CBOW és a Skip-gram modell felépítése. Forrás: Mikolov et al 2013

Mikolovék modellje három ok miatt hozott igazi áttörést. Az első és legfontosabb az, hogy nagyon jó eredményt tudott felmutatni azokban a feladatokban, amikkel rendszeresen tesztelik a modellek pontosságát (lásd később). Nagyon fontos volt a modell elterjedésében az is, hogy elérhetővé tettek olyan előkészített (angolban a 'train' szót használják erre) vektortereket, amiket hatalmas szövegkorpuszokra építettek rá. Ezek a vektorterek önmagukban is alkalmasak voltak arra, hogy

különböző alkalmazásokban és tudományos kutatásokban felhasználják őket. És legalább annyira fontos volt az is, hogy a szerzők szabadon elérhetővé tették a word2vec kódjait (https://code.google.com/archive/p/word2vec/), így gyakorlatilag jelentősen leszállították azt a „programozási küszöbtudást", amivel bárki saját vektortereket hozhat létre.

A módszer validációja

A dolgozat korábbi részében említettem, hogy a word2vec azért is tudott elterjedni, mert az „elődeinél" jobbnak bizonyult feladatmegoldásban. Ahhoz, hogy ezt jobban megértsük érdemes bemutatni azt, milyen feladatokkal tesztelik ezeket a vektortereket. A beágyazási modell azt ígéri a felhasználónak, hogy nagy pontossággal meg tudja mondani egy szó kontextusát, környezetét. Ezt leggyakrabban analógiák keresésével tesztelik. Ezek az analógiák lehetnek szemantikaiak és szintaktikusak is. Klasszikus szemantikai analógia például a következő: Ha azt mondjuk, hogy „király" + „férfi" akkor mi lesz ennek a „női" analógiája. A tesztelés akkor zárul pozitív eredménnyel, ha a modellünk a fenti kérdésre azt válaszolja, hogy „királynő" (ennek a technikai részleteit később ismertetjük"). A szintaktikai tesztelésnél valamilyen nyelvtani szabályt tesztelünk, pl: „megy" -> „ment" analógiájára a „jön" szóra a „jött" válasz kell, hogy érkezzen. A tesztelés másik gyakorija példája az, hogy mondatokból kihagynak egy szót, és egy maghatározott szókészletből kell kiválasztani a megfelelő szót (lásd pl: Zweig-Burges 2011). Mikolov és munkatársai (2013) több korábbi algoritmussal is összevetették az eredményeiket. Mind a két word2vec algoritmus nagyon jól szerepel ezekben az összehasonlításokban, szinte az összes feladatban magasabb pontosságot értek el, ráadásul a futási idő töredékéért cserében. A két word2vec módszer közül a várakozásoknak megfelelően a skip-gram pontosabb bizonyult, mint a CBOW, utóbbi viszont gyorsabban futott le nagy adathalmazokon (Mikolov et al 2013).

2.1.3 Fasttext

A word2vec továbbfejlesztéseként tekinthetünk a 2016-ban publikált Fasttext algoritmusra (Joulin et al 2016, Bojanowski et al 2017), a Facebook-hoz köthető kutatói

csapat részben át is fedett a word2vec alkotóival (Mikolov személyében). Az újítás lényege az volt, hogy szemben a word2vec algoritmussal, ami a szavakat használta fel a vektor tér kialakításakor, a fasttext karakter N-grammokból indult ki. Például a piros szó, a „pir”, „iro”, „ros”, „piro”, „iros” és „piros” n-grammokból építhető fel (ha 3-5 közötti n-grammokat használunk), és a szó pozícióját ennek a hét n-grammnak az összegeként definiálhatjuk. Ennek több előnye is van. A modell képes jól megragadni akár ritka szavak pozícióját is, ha azok karakter n-grammjai megtalálhatóak más szavakban. Sőt akár olyan szavaknak is meg tudja találni a pozícióját, amelyek nem szerepelnek az eredeti korpuszban. Ez a gyakorlatban nagyon előnyös tulajdonság, mert azokban az alkalmazásokban, ahol vektortér modellek szerepelnek a háttérben, problémát okozhat, ha ritka vagy ismeretlen szavakat használnak a felhasználók. Egyszerű példaként gondolhatunk egy kereső motorra. A felhasználók gyakran elütnek kifejezéseket, ezzel nehéz helyzet elé állítva a kereső algoritmust. Ha azonban egy fasttext alapú nyelvi modell megtámogatja a keresést, akkor sokkal nagyobb valószínűséggel juthat el a felhasználó ahhoz a tartalomhoz, amit keresni szeretett volna. Joulin és munkatársai (2016) két példán is tesztelték az algoritmusukat. Mind érzelem besorolásban, mind címke predikcióban jobban teljesített a Fasttext mint a rivális algoritmusok, ráadásul sokkal rövidebb futásidővel. A word2vec-hez képest viszont lassabb a futási ideje főleg, ha széles karakter N-gramm spektrumot állítunk be a modell tréningjéhez. A társadalomtudományi felhasználás kapcsán visszatérünk majd a két modell gyakorlati különbségeire.

2.1.4 GloVe

A word2vec és Fasttext algoritmusok mögötti megközelítés a korábbi SVD modellekhez képest nem csak a statisztikai módszerben volt más (neurális háló alapú modell vs mátrix faktorizáció), hanem abban is, hogy alapvetően egy szó lokális környezetéből indult ki, szemben az SVD modellekkel, ahol globális környezetet használtak a vektortér képzésre. A Stanford NLP kutatócsoportja által kifejleszett GloVe[7] (Pennignton et al 2014) a két megközelítést ötvözi olyan szempontból, hogy egyszerre veszi figyelembe a vektortér kiszámolásakor a lokális és a globális környezetet. Penningtonék 2014-es cikkükben több hasonló modellel is összevetik az eredményeiket. Ezekben az összevetésekben a GloVe legalább olyan jól, vagy jobban teljesít, mint a rivális algoritmusok.

GloVe

Akárcsak a „klasszikus" SVD algoritmust használó modellekben, a kiinduló adat a GloVe esetében is egy TCM. A GloVe viszont nem mátrix faktorizációs megközelítést követ, hanem egy adott szópár együttes előfordulásának logaritmusát becsüli meg. Ez gyakorlatban azt jelenti, hogy a GloVe algoritmusa egy olyan függvényt minimalizál, amiben az adott szópárhoz tartozó szavak vektorainak szorzatából[8] kivonja a szavak logaritmusát, majd ennek veszi a négyzetét[9]. A modell kapcsán még két technikai dolgot érdemes megjegyezni. A gyakori szavak szerepének csökkentése érdekében egy súlyfaktort használnak a modellben, ezzel kiküszöbölve a TCM alapú modellek egyik alapproblémáját. A másik „érdekesség" a modell kapcsán, hogy párhuzamosan két vektorteret is használnak a modellek definiálásakor. Ez nem szükségszerű a modell felírásából, de a gyakorlati eredmények azt mutatták, hogy így robusztusabban működik az algoritmus. A két vektorteret általában összegzik és ezt a közös vektorteret használják a további számításokban. A dupla vektoros megoldás miatt nevezik a GloVe

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

[8] Ahhoz, hogy ez a szorzat skalár legyen értelemszerűen az egyik vektort transzponáljuk

[9] Az algoritmus mindkét szóhoz egy hibatagot is társít

mögött álló modellt, log-bilineáris regressziós modellnek (Pennington et al 2013). A GloVe módszer nagy előnye, hogy nem csak egy lokális ablakot használ, mint a word2vec alapú modellek és jobban értelmezhető a végeredmény is, legalábbis abban az értelemben, hogy két adott szóhoz tartozó vektor szorzatának tudunk tartalmi jelentést adni.

A fejezetben bemutatott különféle szóbeágyazási modellek között nehéz rangsorolni. A modellek széles paraméterezhetősége miatt nagyon nehéz objektívan összevetni, hogy melyik módszer jobb (akárcsak részfeladatokban). A vektortér modelleket összehasonlító cikkek (lásd pl: Spirling-Rodriguez 2019) legfőbb tanulsága az, hogy megfelelő paraméterezés mellett nincs nagy különbség az egyes modellek „teljesítménye" között. Kisebb korpuszok és ritkább szavak esetén a GloVe valamivel stabilabb, mint a lokális ablakot használó word2vec modellek. A word2vec modellek közül a Skip-gram szinte minden helyzetben jobban teljesít, mint a CBOW. Futási idő szempontjából a GloVe gyorsabb, mint a word2vec (főleg, mint a Skip-gram), de több memóriát is igényel, mivel a bemeneti oldalon el kell készíteni egy TCM-et, ami már közepes korpusznál is nagyon-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 szám, memória) vagy olyan praktikusabb szempontok is, hogy a kutató milyen programnyelven tud jól kódolni és az adott programnyelven melyik módszer hogyan van implementálva.

2.1.5 Kontextualizált vektortér modellek

Ha megfigyeljük a fejezetben bemutatott módszerek evolúcióját, akkor látható, hogy a 2010-es évek elején gyorsult fel igazán a szóbeágyazási modellek fejlesztése. A Fasttext-et 2016-ban publikálta egy Facebookhoz köthető kutatócsoport (Joulin et al

2016). Az azóta eltelt 4 év [a dolgozatot 2020 nyarán-őszén írom] természetesen nem telt el eseménytelenül ezen a tudományterületen, több izgalmas modellt is publikáltak. Az eddig bemutatott módszerek, mind statikus vektortér modellek voltak. Ez a gyakorlatban azt jelenti, hogy mindegyik szónak kontextustól függetlenül 1 vektortér reprezentációja készül. A kontextualizált szóbeágyazási modellek (CWE) ezzel szemben egy adott kontextushoz kötik azt, hogy milyen vektortér értéket kap egy adott szó. A modellek egy része a statikus vektorterekre épít rá két irányú rekurrens neurális modelleket, ilyen a Flair (Akbik et al 2018), vagy az ELMo (Peters et al 2018)). A modellek legújabb generációja viszont más logikát követ és nem használ statikus beágyazást. Az új modellekben a közös pont az, hogy egy „Transformer" modellre építenek. A Transformerek olyan nyelvi modelleknek az általános megnevezései, amiben van egy bekódoló (encoder) és egy kikódoló (decoder) elem. A két elem különböző neurális hálókból épül fel (rekurrens, konvolúciós, stb.) Ezeket a modelleket főleg fordítási feladatokban használják. Vaswani és munkatársai (2017) egy olyan Tranformert vezettek be, amiben a bekódoló elem egy összpontosító (Attention) és egy nem visszacsatoló (feed forwad) neurális hálóból áll, míg a kikódoló elemhez hozzáadtak a megkülönböztető és a nem visszacsatoló réteg közé egy kikódoló-bekódoló összpontosító (Encoder-Decode Attention) réteget is. A modell mivel nem tartalmaz rekurrens elemet nagyon gyors futásra képes, az összpontosító (Attention) réteg pedig nagyon magas teljesítményt eredményez. Az összpontosító réteg lényege az, hogy egy adott szövegrészen belül megadjuk azt, hogy a célszó megértését melyik másik szóval tudjuk segíteni. Heurisztikusan megközelítve egyfajta súlyt rendelünk a kontextusban megjelenő szavakhoz. Ezt a logikát vitte tovább az OpenAI (Radford et

al 2018) és a jelenlegi egyik state-of-the-art[10] módszer, a BERT is (Devlin et al 2018), azzal a különbséggel, hogy csak a bekódoló (encoder) elemet vették át. A BERT nagyon magas teljesítményét részben annak köszönheti, hogy a tanulásnál nem csak balról-jobbra tanul a modell, hanem parallel jobbról balra is. Ezek a kontextualizált vektortér modellek lényegesen jobb eredményt értek el szinte minden nyelvi feladatban, mint a statikus modellek főleg olyan szavak esetében, amelyek többértelműek (Wiedemann et al 2019). A társadalomtudományi használata ezeknek a módszereknek most bontakozik ki. Klasszifikációra már ma is lehet példát találni (Samory et al 2020), az elkövetkező években kiderül, hogy tartalmi fókuszú elemzéseben mennyire tudnak teret nyerni a CWE modellek.

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

2.2 Technikai megfontolások

A habilitációs dolgozat előző fejezetében körbejártam a szóbeágyazási módszerek rövid fejlődéstörténetét. A technikai részletek után érdemes újra egy picit heurisztikusabb nézőpontból megvizsgálni a módszereket. Az első premisszánk az, hogy a különböző tudományterületek és üzleti szereplők teljesen eltérő célokkal használhatják a szóbeágyazási módszereket. Társadalomtudományi perspektívából a kiindulás a legtöbb esetben az, hogy szövegeken keresztül szeretnénk megérteni egy adott társadalmi jelenséget, az adott társadalmi jelenség beágyazottságát vagy temporális mintáit. Nyelvészként érdekes lehet egyes szavak morfológiájának változása vagy szavak értékváltása (Szabó 2019), de akár az is, hogy hogyan lehet ezeket a nyelvi modelleket más nyelvészeti feladatokban használni (pl: emóció elemzés, lemmatizáció, stb..). A szóbeágyazási módszerek azonban nem tudományos felhasználhatóságuk miatt terjedtek el elsősorban, hanem azért, mert nagyon sok gyakorlati feladatban hasznosnak bizonyultak. Fel lehet őket használni fordító-programokban, szöveges keresés támogatására, szöveges bot-ok programozására, dokumentum klasszifikációra, stb. Azt gondolom, hogy egyszerűen belátható az, hogy más szempontoknak kell, hogy megfeleljen az a szóbeágyazási modell, ami egy társadalomtudományi kérdésre felel és más szempontoknak az, ami egy keresési algoritmust támogat. A következő táblázat (tábla 1.) ezeket a szempontokat sorolja fel.

	Társadalomtudományi	Ipari
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	kevéssé fontosak	fontosak
ritka szavak	kevéssé fontosak	fontosak
stemmelés/lemmatizáció	javasolt	nem javasolt
módszer	word2vec, GloVe	Fasttext, kontextualizált vektortér modellek (ELMO, BERT..)

Tábla 1. Szóbeágyazási modellekkel kapcsolatos elvárások bemutatása felhasználási területenként

Bármilyen vektortér modellről is beszélünk, a kiindulás minden esetben a korpusz. A korpusz azoknak a szövegeknek az összessége, amiből elkészítjük a nyelvi modellünket. Ezek lehetnek közösségi média tartalmak, online újságcikkek, kommentek, digitalizált tartalmak – gyakorlatilag bármilyen online elérhető szöveges tartalom. Az ipari alkalmazásokban jellemzően általános és nagy korpuszokból indulnak ki. Ilyen alap korpusz lehet például a „Common crawl” (CC) amit gyakorlatilag az internetes tartalmak egy mintájának is tekinthetünk (https://commoncrawl.org). A CC folyamatosan gyűjti erre írt speciális algoritmusokkal (Crawler) a különböző online oldalakra kikerülő tartalmakat és ezeket a tartalmakat szabadon elérhetővé teszi bárki számára. Hatalmas adatmennyiségről beszélünk – petabyte nagyságrendben. Az oldalon több mint 40 nyelven van elérhető szöveges adat (magyar is), 7 évre visszamenőleg. A hatalmas adatmennyiség azonban nem túl jó minőségű – a szövegek kitisztítása már magában is egy nagy feladat (Indig 2018). Szintén nagy, elérhető 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, mind a Wiki korpusz gyakran használt a szóbeágyazási modellekben,[11] mivel rengeteg témát lefednek, nagyon sok unikális szót tartalmaznak és már a méretük miatt is nagyon robusztusak a belőlük kapott eredmények. Egy ipari alkalmazásban (például egy keresést támogató applikációban) ezek a felsorolt tulajdonoságok nagyon előnyösek. De miért nem jó ez feltétlen egy társadalomtudományi kutatásban? Ugyanazért nem, amiért a legtöbb survey módszertannal foglalkozó kutató (100-ból 99) előnybe részesít egy 1000 fős reprezentatív mintát egy 1 000 000 fős kényelmi mintával szemben. Bár nagyok ezek a korpuszok, de nem (vagy csak korlátozottan) általánosíthatóak a belőlük kapott eredmények, alacsonyabb a külső érvényességük, egy szelektált, célhoz szabott korpuszhoz képest. Ahogy a későbbi nemzetközi és hazai kutatások kapcsán látjuk majd, társadalomtudományi kutatások is készülnek nagyobb általánosabb korpuszokon, de utóbbiak esetében legalább annyira jellemző, hogy egy adott témához saját egyedi korpuszt is használnak, mivel ezáltal lehetővé válik például olyan kérdéseknek is a megfigyelése, ami általános korpuszokon nem lehetséges (például hogyan kommunikálnak a migrációról/menekültkérdésről a kormánypárti/ellenzéki sajtóban).

Az eddigiekkel szorosan összefügg az is, hogy mennyire jellemző előkészített vektorterek használata az egyes alkalmazási területeken. A vektorterek a szóbeágyazási algoritmusok végeredményei. A vektorterek soraiban a szavak vannak, az oszlopokban pedig a szavakhoz tartozó súlyok. Általában 100-500 dimenziós

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

vektortereket képeznek (lásd később). A word2vec és a Fasttext sikeréhez nagyban hozzájárult, hogy elérhetővé tettek előkészített vektortereket. Ez gyakorlatban azt jelenti, hogy elkészítették a szövegbeágyazást hatalmas korpuszokon (lásd pl: CC, Wikipedia), a kapott vektortereket pedig szabadon letölthetővé tették. Informatikai/programozói szempontból az előkészített vektorterek kezelése nagyságrendekkel egyszerűbb feladat, mint egy nagy korpuszt építeni és azt beágyazni. Ez kifejezetten igaz azokra a nagyságú korpuszokra, amit a word2vec vagy fasttext oldalán elérhetővé tettek. Ezek a vektorterek jól használhatóak az ipari alkalmazások többségében, viszont pont a korábban említett szempontok miatt nem feltétlen használhatóak társadalomtudományi projektekben. Szintén a használati céllal van összefüggésben, hogy az ipari alkalmazásokban előny, ha a modell robusztusan tudja kezelni a ritka szavakat, megoldást nyújt az elgépelt szavakra, vagy akár olyan szavakat is tud kezelni, amelyek nincsenek benne a kiinduló korpuszban. Speciális elemzési célok kivételével ezeknek nincs azonban jelentősége egy társadalomtudományi projektben. A kutatásokban általában nem a kivételek az érdekesek, hanem a masszív trendek. Ebből következik, hogy az ipari projektben előtérbe kerülnek a Fasttext-hez hasonló karakter N-gramm megközelítést használó algoritmusok, míg a társadalomtudományi projektekben ennek nincs hozzáadott értéke – sőt akár zavaró is lehet, ha két ellentétes jelentésű, de hasonló alakú szó közel kerül egymáshoz a végeredményben.

A bemeneti korpusz kapcsán nem csak a mérete és a forrása érdekes, hanem az is, hogy milyen lépéseken megyünk keresztül, amikor ezt a korpuszt előkészítjük a beágyazásra. Nincs a korpusztisztításnak egy általános, minden projektre ráhúzható formulája (Németh – Katona – Kmetty 2020), de vannak olyan lépések, amiket

általában érdemes elvégezni mielőtt elemezni kezdjük a tartalmat (itt elemzés alatt nem csak szóbeágyazásra gondolunk, hanem bármilyen más szövegbányászati megoldásra (pl: topikmodellezés)).

Előfeldolgozás

Duplikáció szűrés: Az online leszedett tartalmakban gyakori, hogy ugyanaz a tartalom többször is szerepel a korpuszban[12], ezeket a duplikációkat érdemes eltávolítani.

Felesleges tartalmak kiszűrése: felesleges/koszos tartalmakon olyan szövegekben megtalálható szövegrészeket értünk, ami nem tartozik szorosan az adott tartalomhoz, de mégis bekerül a szövegbe. Ilyen lehet például egy oldalsáv szövege ajánlóval, egy cikk végén elhelyezett hirdetés (pl: támogasd az adott lapot), vagy egy http link egy másik oldalra

Speciális karakterek eltávolítása: a legtöbb szóbeágyazási modellben nem használjuk a központozást, tehát az összes pont/vessző/felkiáltójel/idézőkelt típusú karaktert eltávolítjuk a korpuszból

Kisbetűsítés: A nagybetűk átalakításra kerülnek kisbetűkre a szövegben, annak érdekében, hogy minél egységesebb legyen a korpusz.

Lemmatizáció/stemmelés: Egy szövegben ugyanaz a szó rengeteg alakban szerepelhet. Ez a sokszínűség nem mindig hasznos, ezért jellemző előfeldolgozási lépés a szavak egységesítése. Ennek két elterjedt megoldása van. A stemmelés egy adott szóból nyersen levágja a ragokat, nyelvi elemeket, ezzel szemben a lemmatizáció a szó alap alakját (lemmáját) keresi meg. Utóbbi megoldás általában komplikáltabb és időigényesebb, de jobb eredményt ad.

[12] Ennek sokszor technikai okai vannak (rosszul működött a crawler), de lehet akár az is, hogy egy adott oldalnak a mobilos és nem mobilos változata is bekerül a gyűjtésbe.

Név elem felismerés: A név elem felismerés során összevonjuk a különböző intézmény/tulajdonneveket, például a „pest” „megye” szópárból „pest_megye” lesz. Így ezekre az összevont entiásokra tudunk külön is fókuszálni az elemzésben.

Szignifikáns (bi)grammok: a szövegekben általában vannak olyan gyakori szópárok (bigrammok), amik nem tulajdonnevek, de mégis gyakran egymás mellett szerepelnek. Feladat függően érdemes lehet a gyakran előforduló szópárokat összevonni és közös entitásként kezelni őket (pl: az angol nyelvben nagyon sok foglalkozás egy szópárból áll, amit érdemes összevonni egy elemzés előtt: data scientist -> data_scientist).

Az előfeldolgozási lépések jellemzően idő és erőforrás igényesek. Ha azonban társadalomtudományi célból akarunk felhasználni egy vektortér modellt érdemes ezeket elvégezni, mert jelentősen befolyásolhatják az eredményeink érvényességét. Ezzel szemben az ipari alkalmazásokban általában ennek kisebb a jelentősége, sőt a lemmatizáció/stemmelés kifejezetten elkerülendő általában, hiszen a legtöbb esetben pont 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 a paraméterezés értelemszerűen függ a választott algoritmustól, de vannak olyan paraméterek, ami minden algoritmusban egyaránt választhatóak. Ebből a két legfontosabb az ablak mérete és a vektortér mérete. Utóbbival kezdjük.

A vektortér nagysága két elemből áll össze. Az egyik a szavak száma, ami alapvetően adott egy szótárban. Ha a nagyobb trendeket akarjuk vizsgálni egy témában, akkor a ritka szavak akár ki is hagyhatók a korpuszból, ezzel csökkentve a futásidőt és növelve a modell stabilitását (lásd később). A másik dimenzió a szavakhoz rendelt súlyvektor

hossza. Ez jellemzően 100-500 között mozog, leginkább 200-300 dimenziós vektortereket képeznek. Intuitívan azt gondolhatnánk, hogy minél több dimenziónk van, annál jobb lesz a vektortér. Ez a gyakorlatban azonban nem így van, erősen korpuszfüggő (Spirling-Rodriguez 2019). Kis korpusz esetén a túl sok dimenzió az eredmények stabilitását csökkenti, tehát egy adott szó pozíciójában nagyobb lesz a véletlen szerepe. De nagy korpusznál sem érdemes extra sok dimenziós vektorteret kialakítani, mert a futási idő növekszik, de cserébe nem kapunk jobb eredményeket.

Az ablak nagysága a másik olyan paraméter, amit minden modell kapcsán fontos eldöntenünk Az ablak azt a szókörnyezetet jelenti, amit a modellek felhasználnak a vektortér kialakítására. A word2vec ezen az ablakon megy végig, a GloVe ezt az ablakot használja fel, hogy kiszámolja a TCM-et. Az ablak általában kétoldali szimmetrikus ablak, de vannak példák asszimetrikusra is, illetve olyan ablakokra, ahol súlyoznak a közelséggel. Ha túl pici ablakot használunk nem jelenik meg az adott szó kontextusa, ha túl nagy ablakot akkor „szétfolyik” ugyanez a kontextus. Kis ablak jobban működik lexikális egyezések vizsgálatára míg a nagyobb ablakok jobbak analógiák vizsgálatára (Lison – Kutuzov 2017). A nagyobb ablak, kisebb korpuszok esetében növeli a stabilitást (Szabó et al 2020). Közösségi média tartalmaknál (tweetek, kommentek) szintén érdemes hosszabb ablakot használni, hogy az adott szöveg teljes kontextusában beépüljön a vektortérbe (Yang el at 2018). Bár nincs egységes álláspont az ablak ideális nagysága kapcsán, általában 5-10 közé teszik a méretet, de ahogy írtam, ez függ a korpusz jellegétől és nagyságától is.

Bár explicite eddig ezt nem emeltem ki külön, de az eddigi dolgozatból egyértelműen kiderül, hogy a szóbeágyazás – akárcsak más nyelvi modellek – erősen nyelvfüggő.

Maga az algoritmus nyelvtől függetlenül ugyanazokat a lépéseket végzi el, de a nyelv meghatározza ez előfeldolgozást, a modell paraméterezését és értelemszerűen a kapott vektorteret is. A különböző módszerek összehasonlítását megnehezíti az is, hogy a legtöbb kiértékelés angol nyelvre készült, így az általánosan javasolt paraméterbeállítások nem biztos, hogy ugyanúgy megfelelőek más nyelvekre (jó ellenpélda erre Bojanowski et al 2017 cikke). Az ablaknagyságnál is érdemes ezt figyelembe venni, és az adott nyelv egyedisége alapján meghatározni az ideális ablakot. Erősen nyelvfüggő, hogy egy mondaton belül lévő egymásra ható szavak között mekkora az átlagos távolság. Azokban a nyelvekben, ahol hosszabb ez a dependencia lánc érdemes hosszabb ablakot használni. Az átalagos nyelvi dependencia jóval hosszabb például a magyar a német vagy a kínai nyelvben, mint a románban vagy a japánban. Az angol nyelv ilyen szempontból átlagosnak számít (Liu 2008).

További paraméterek

A többi változtatható paraméter esetében általában hagyatkozhatunk az algoritmusok alapbeállításaira. A word2vec esetében a mintavételes szavak (negative sampling) számára az 5 javasolt (Mikolov et al 2013), a GloVe esetében a simító paraméterként szereplő Alpha[13]-ra 0.75 (Pennington et al 2014). Bármelyik modellt választjuk, érdemes több iterációt is engedni, hogy növeljük a modell konvergenciáját. Az iterációk számának növelése kapcsán csak a gépidő jelenthet korlátot.

A modellek technikai bemutatása kapcsán az utolsó vizsgált terület a modellek stabilitása. A stabilitást úgy lehet értelmezni ebben a kontextusban, hogy ugyanarra a korpuszra lefutatott ugyanazt a módszert használó elemzés eredményei mennyiben

[13] Ez a kitevőben szereplő paraméter csökkenti az adott szópár fontosságának szerepét a hiba minimalizációban

térnek el egymástól. 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 modellekre is (pl: topikmodell), 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 az instabilitás, de nem tüntethető el, mert ez a módszer sajátja.

Instabilitás kezelése

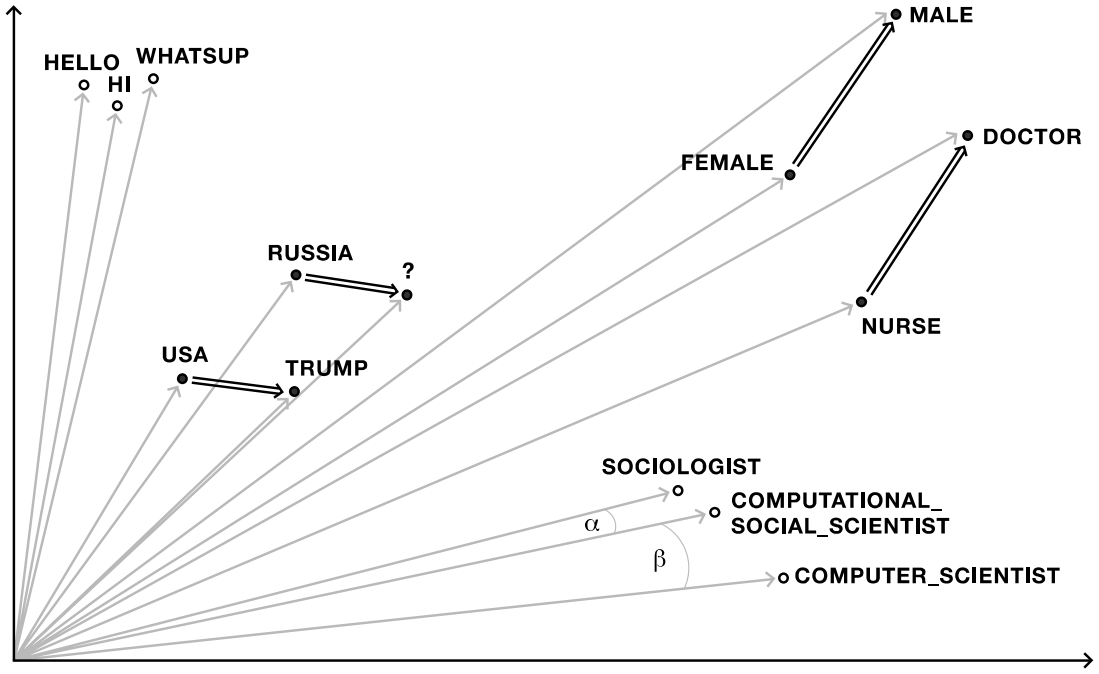
Az instabilitást kétféleképpen lehet kezelni a tudományos elemzésekben. Ez a két kezelés nem zárja ki egymást, együtt is alkalmazható. Az egyik megoldás az, hogy sokszor újrafuttatjuk a vektortér modelleket és amikor eredményt közlünk akkor a több futás eredményét átlagoljuk. Ez a bootstrap megközelítés lehetőséget ad arra is, hogy konfidencia intervallummal egészítsük ki az eredményeinket. Erre mutatok példát a dolgozat későbbi részében a második esettanulmányban. A másik lehetőség egy kis korpusz esetében az lehet, hogy egy meglévő általános korpuszon képzett vektorteret felhasználunk az egyedi korpuszuknál. A machine learning (ML) algoritmusokban ezt nevezik transfer learning-nek. Abból indulhatunk ki, hogy bizonyos szókapcsolatok nagyon általánosak (pl: a pizza és a tészta közel van egymáshoz mert mindkettő étel) és ezeket a szókapcsolatokat felesleges újra megtanítani egy modellnek. Így logikus lépés, hogy a kiinduló vektorsúlyoknak vegyünk egy általános és nagy korpuszra már elkészített vektorteret és erre a vektortérre tanítsuk rá a saját korpuszunkat. Technikailag ez többféleképpen is megoldható. Lehet akár építeni olyan mély neurális hálót, ahol külön réteget kap az általános és a specifikus adathalmaz[14]. Ennél azonban jóval egyszerűbb, ha inputként használjuk az általános vektortér súlyait a specifikus modellben.

[14] Képtanító algoritmusoknál ez bevett megoldásnak számít

2.3 Mit kezdjünk a vektortérrel?

Az előző fejezetekben bemutattam a szóbeágyazási modellek matematikai/statisztikai alapjait és azokat a technikai megfontolásokat, amelyeket érdemes átgondolni egy vektortér modell készítése/használata előtt. Ebben a fejezetben onnan indulunk, hogy van egy használható vektorterünk. De vajon mit lehet ezzel a vektortérrel kezdeni?

Érdemes újra felidézni, hogy mi az a jellemző tulajdonság, amiért hasznos az eredeti korpuszból egy alacsony dimenziós vektorteret alkotni. Abból indulunk ki, hogy az a célunk, hogy minél jobban megértsük egy adott szövegben szereplő szavaknak az egymáshoz való viszonyát. Mik azok a szavak, amik jellemzően egy kontextusban szerepelnek, mik azok a szavak, amik „taszítják” egymást? A szóbeágyazási modellek erre a kérdésre tudnak nagyon pontos választ adni. Gyakorlatilag módszertől függetlenül az egyes szavak vektorait úgy optimalizáljuk, hogy az egymáshoz közel eső szavak egymáshoz közel legyenek ebben a térben. A közelség ebben az esetben vagy jelentésben tapasztalható közelséget jelent, vagy témaközelséget vagy akár szintaktikai közelséget (ha nem lemmatizáltuk a korpuszunkat). A 2. ábra néhány példát mutat különböző fogalmak esetében a közelségre egy egyszerűsített két-dimenziós térben. A közelség kiszámolására leggyakrabban a két vektor szögtávolságának koszinuszából indulnak ki, és ezt fordítják át egy közelség mutatóra (lásd később).



Ábra 2. Példa, szavak/fogalmak közelségére (forrás: Németh – Koltai 2021)

A példán közel helyezkednek el egymáshoz a köszönések (hello, hi, whatsup), az országok és politikusok, valamint a tudományterületek. A számítógépes társadalomtudós közelebb van a szociológushoz, mint az informatikushoz (computer scientist)[15]. De az ábra rávilágít egy még érdekesebb lehetőségre, mégpedig az analógiák vizsgálatára. Ha megnézzük a nő és a férfi vektor között bezárt szöget és ezt kivetítjük a doktorra akkor megkapjuk a szakma „női” megfelelőjét – az ápolót. Ugyanezzel a logikával meg tudjuk nézni, hogy az USA Trump párosításnak ki felel meg Oroszország viszonylatában – a kérdőjel helyére viszonylagos biztonsággal beírhatjuk Putin-t. Ezeknek a lehetőségeknek a társadalomtudományi hasznára a későbbiekben térünk vissza.

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

Az eddigi példák elméleti síkon mutatták be, hogyan működik a módszer. A működés gyakorlati demonstrálására egy előkészített vektorteret használok, ami a Wikinews korpuszon alapul. A Wikinews a 2017-es Wikipediából, a UMBC[16] korpuszból (50 000 oldal, 100 000 000 weblapjának gyűjtése) és a stamt.org oldalról származó hírekből épül fel. A teljes korpusz 16 milliárd tokent tartalmaz. A 300 dimenziós vektortér képzéséhez a fasttext algoritmust használták. A vektortér a korpusz 1 000 000 leggyakoribb szavát tartalmazza (nincs stemmelve, kis és nagybetűk egyaránt előfordulnak a vektorteret alkotó szavakban). Az előkészített vektortér szabadon letölthető a Fasttext weboldaláról: https://fasttext.cc/docs/en/english-vectors.html.

Programnyelv

Jelenleg a két legtöbbet használt programnyelv NLP-vel kapcsolatos feladatokra a Python és az R. Mind a két programnyelv ingyenes és nagyon széles használói bázissal rendelkezik a programozók, adattudósok között. Mind a két programnyelvben elérhetőek olyan csomagok, amikkel magas minőségű NLP elemzéseket lehet végezni. Python esetében az előfeldolgozásra elsősorban az nltk és a Spacy csomagot használják, vektortér képzésre pedig elsősorban a Gensim-et[17]. R-ben előkészítésnél a tidytext és tokenizers csomagok nagyon hasznosak, a vektorképzésnél pedig a text2vec a FastTextR illetve a word2vec. Nincs gyakorlatilag olyan szóbeágyazáshoz köthető feladat, amit egyik programnyelvben meg lehet csinálni a másikban pedig nem, ezért egyéni preferencia (programnyelv ismeret) alapján érdemes választani. Kollaboratív projektekben akár a két programnyelv keverése is elképzelhető, mivel R környezetbe is lehet python kódokat beágyazni és ez fordítva is működik. A dolgozatban bemutatott elemzések R-ben készültek. A fejezet összes kódja szabadon elérhető a következő Github oldalon: https://github.com/zkmetty/nlp

[16] https://ebiquity.umbc.edu/blogger/2013/05/01/umbc-webbase-corpus-of-3b-english-words/

[17] A Keras csomaggal saját neurális hálókat is készíthetünk, komplexebb probléma eseténben (pl: tranfer learning) ez kifejezetten hasznos tud lenni

Az elemzést egy egyszerű lekérdezéssel kezdjük. Milyen szavak vannak közel a szociológiához?

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

Tábla 2. A „sociology” szó 15 legközelebbi szomszédja, Wikinews, saját számolás

Az angol nyelvű korpuszban a „sociology” szóhoz a „Sociology” van a legközelebb, tehát ugyanannak a szónak a nagybetűvel kezdődő változata. De az első ötben van a „sociologists” és a „sociologist” szó is. A dolgozat korábbi fejezeteiben többször említettem már, hogy az általános korpuszokat használó előképzett vektorterek esetében kevés előfeldolgozást végeznek. Ez nem hanyagság a készítők részéről, világos célja van, nagyon sok alkalmazásban kifejezetten fontos, hogy a vektortér ne csak kitisztított, lemmatizált szöveget tartalmazzon. Ez sok társadalomtudományi elemzésben zavaró tud lenni, ezért is érveltem korábbam amellett, hogy érdemes az adott feladatunkhoz saját vektortereket képezni jól előkészített korpuszt felhasználva. A top 15-ben természetesen feltűnnek a „rokonszakmák”, elsőként a pszichológia és antropológia, utána a kriminológia a közgazdaságtan és a filozófia. A legközelebbi természettudományi szakma a biológia – a 10 helyen van ezen a toplistán.

Koszinusz közelség

Két szó közelségének kiszámolásakor triviális megoldásként egy egyszerű euklidészi távolság metrika juthat elsőként eszünkbe. Ez ebben az esetben azonban nem a legjobb megközelítés, mivel a vektorok hossza összefüggést mutat a szavak gyakoriságával és a szavak kontextus függőségével (Schakel – Wilson 2015). Az euklidészi távolság helyett szögtávolságokat használnak az elemzésekben, pontosabban a szavak koszinusz közelségét vizsgálják. A koszinusz közelség 1, ha két szó között bezárt szög 0, 90 fokos szögnél 0 a közelség, 180 foknál pedig -1. Nincs arra egységes definíció, hogy mi számít magas vagy alacsony koszinusz közelségnek. A saját tapasztalatok alapján azt az általános hüvelykujjszabályt javaslom, hogy a 0.2 alatti közelséget gyengének, a 0.2-0.4 közötti értéket közepesnek a 0.4 feletti értéket pedig erősnek fogadjuk el. De az értékek függenek a korpusztól és a beágyazási algoritmustól, ezért a legjobb eljárás, ha a koszinusz közelségek sorrendjére fókuszálunk, nem az abszolút értékére.

A kapott lista külső ránézésre logikusnak tűnik, olyan tudományterületeket látunk, amelyek valóban a szociológia rokonszakmáinak tekinthetők. De hiányérzetünk is lehet, például a politológia hiányzik a listáról. A „political science” az angolban két szóból áll, ezért a vektorterünk ezt a szóösszetételt nem tartalmazza. A top15-ös listára még felérő „science” szó távolsága 0.62 volt a „sociology”-tól, a „political” távolsága 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 kiszámolni a „sociology” szó vektorával. A kapott értékünk 0.65 – ez már felfért volna a top listára, a filozófia és a biológia közé. Ez a megoldás természetesen nem ekvivalens azzal a megközelítéssel, hogy már a korpusz szintjén összevonunk olyan szavakat, amik egybe tartoznak (lásd korábban: névelem felismerés, szignifikáns bi-grammok), de egy közelítő megoldásnak elfogadható.

Az elemzésben továbblépve megvizsgálhatjuk, hogy a kiválasztott 10 diszciplína milyen közel van egymáshoz. A következő táblázat a tudományterületek közelségi mátrixát mutatja. Mivel a koszinusz közelség szimmetrikus, ezért elég a mátrix egyik felét kitölteni. A 0.6 feletti értékeket kiemeltük a cellákban.

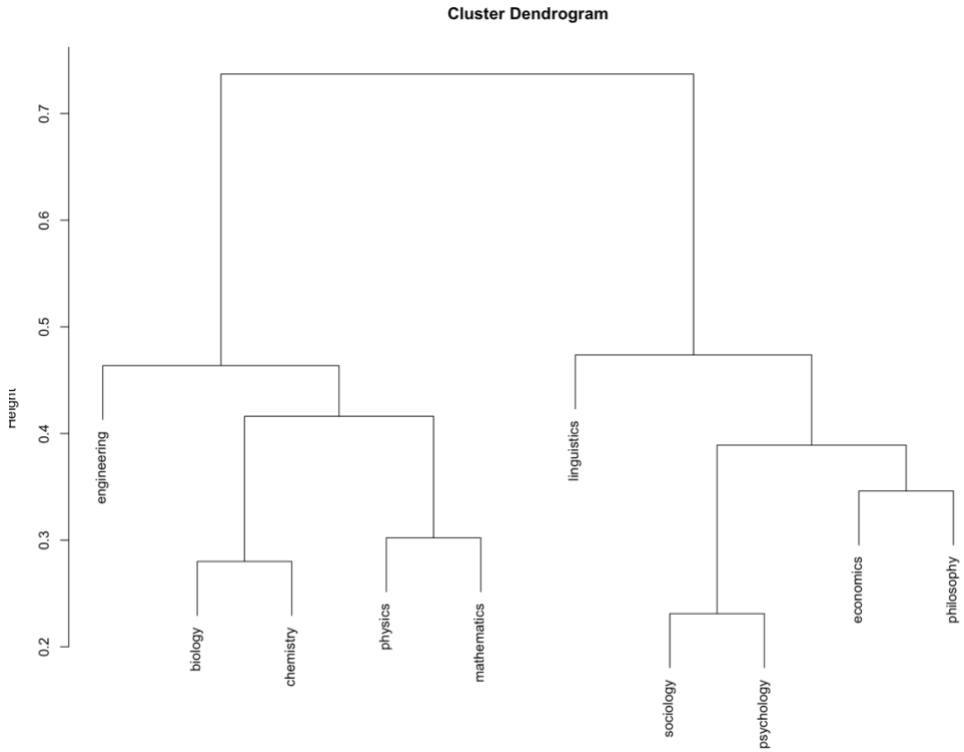
	sociology	psychology	economics	philosophy	linguistics	biology	physics	mathematics	engineering	chemistry
sociology		**0,77**	**0,69**	**0,66**	**0,63**	**0,64**	0,57	0,55	0,52	0,52
psychology			**0,61**	**0,68**	**0,61**	**0,67**	0,58	0,56	0,53	0,56
economics				**0,65**	0,52	0,59	0,57	0,56	0,59	0,57
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

Tábla 3. 10 kiválasztott tudományterület koszinusz közelsége, Wikinews, saját számítás

Kifejezetten alacsony koszinusz értékeket nem látunk a táblában, de ez nem is meglepő, hiszen alapvetően egy területre koncentrálunk. A mátrix alapján viszont jól kirajzolódik egy belső összefonódása a tudományterületeknek, ahol láthatóan nagyon közel kerülnek egymáshoz társadalomtudományi és bölcsészettudományi diszciplínák

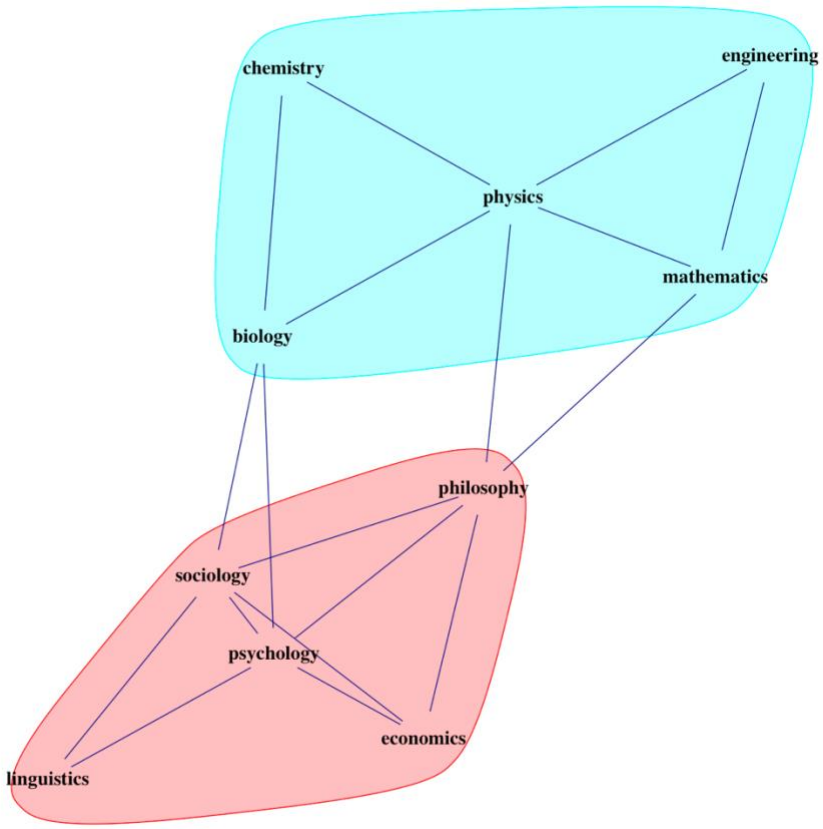
és ezektől elvállnak a természettudományos tudományterületek. Ennek a demonstrálására további elemzéseket végezhetünk a fenti mátrixon.

A közelség mátrix jól használható mind klaszterelemzésben (a közelséget át kell transzformálni távolságra), mind kapcsolathálózat elemzésben vagy más dimenziócsökkentő eljárásokban. A következő ábra a 10 tudományterület hierarchikus klaszter dendogrammját mutatja. Az egyik „ágon” helyezkedik el a nyelvészet, a szociológia a filozófia a pszichológia és a közgazdaságtan, míg a másik ágon a fizika, matematika, biológia, kémia és a mérnöktudományok. Az egymással összekerülő párok triviálisnak tűnnek, ez alól talán egyedüli kivétel a közgazdaságtan és filozófia összekapcsolódása.



Ábra 3. Tudományterületek közelségéből képzett hierarchikus klaszterezés dendogrammja, Wikinews, saját számítás

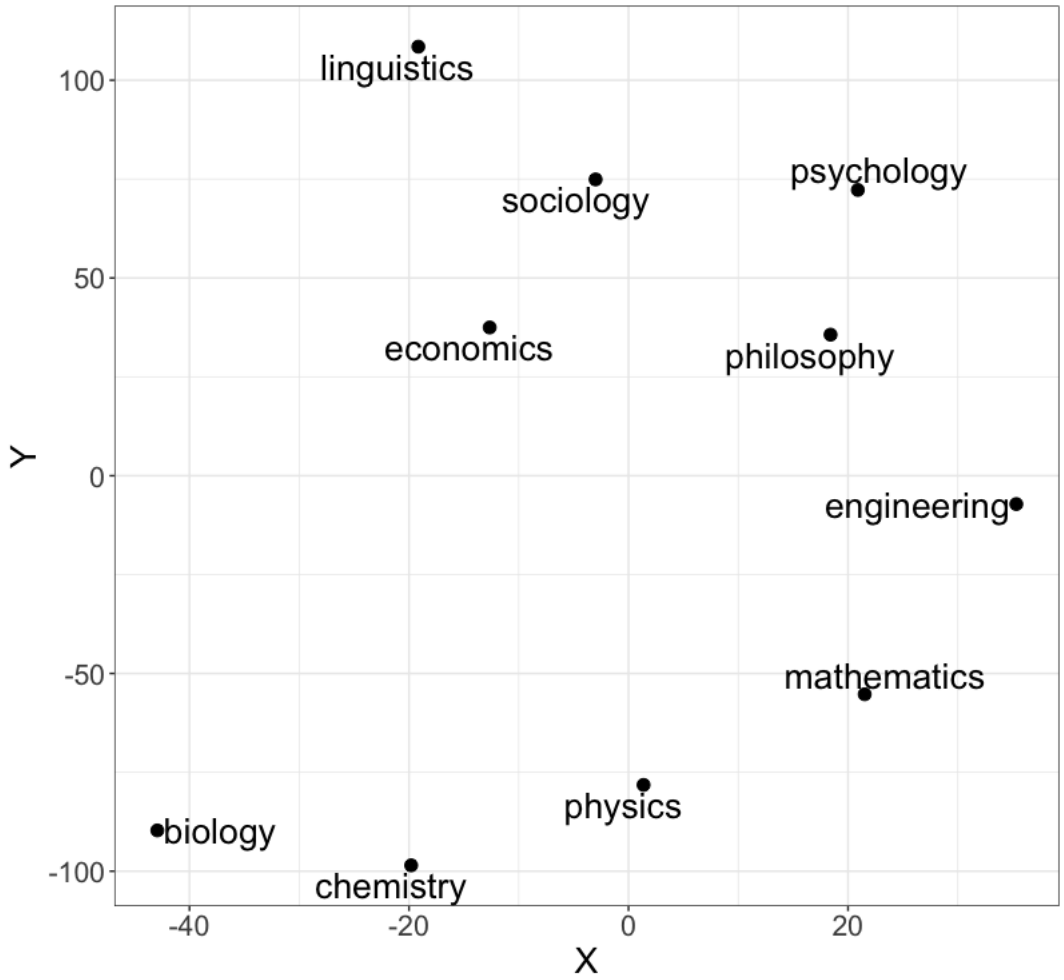
A kapcsolathálózati megközelítésben érdemes egy küszöbérték mentén áttranszformálni a közelség mátrixot egy 0/1 dichotóm értékeket felvevő bináris mátrixra, ahol 1 jelenti azt, hogy két szót összekapcsolunk a hálózatban. A klaszteres megoldással szemben a hálózati vizualizáció komplexebb összefüggéseket is képes megmutatni. A következő ábrán szereplő hálózatnál 0.6 értéknél húztuk meg a határt. Ezen az ábrán is jól kirajzolódik a két tudományterületi csoport elkülönülése, de az is, hogy hol vannak az összekapcsolódási pontok – például a biológia-szociológia-pszichológia hármasa vagy a filozófia-fizika-matematika közötti közös kapocs.



Ábra 4. Tudományterületek közelségéből képzett hálózati vizualizáció, Wikinews, saját számítás

A klaszter és kapcsolathálózati elemzéshez képest egy harmadik megközelítést jelentenek a különböző dimenziócsökkentő eljárások. Itt szóba jöhetnek egyszerűbb főkomponens/faktor elemzés alapú megközelítések, vagy akár olyan módszerek is,

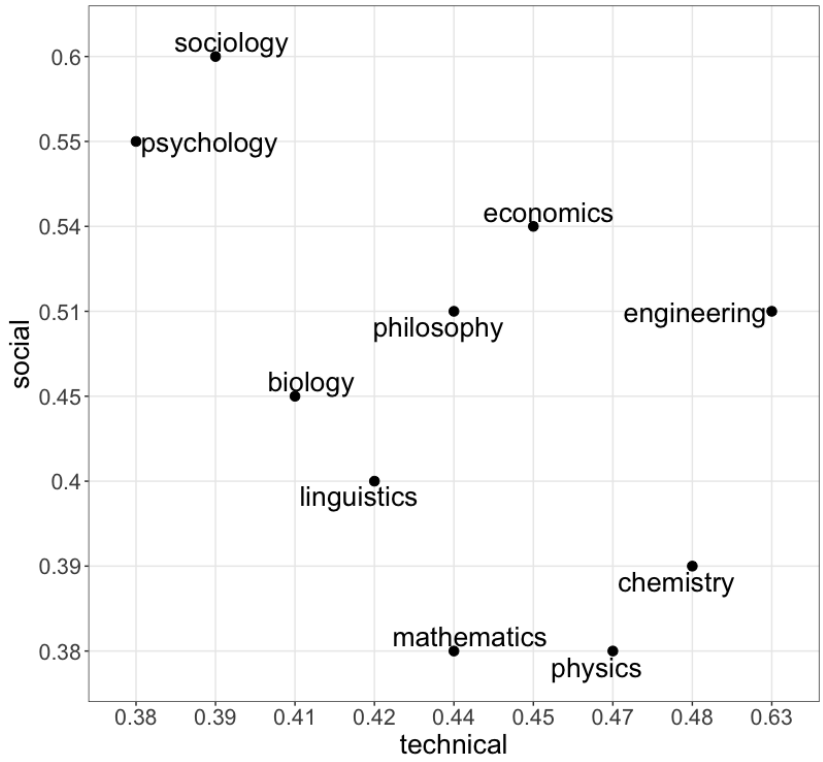
amelyek képesek megragadni a struktúra mögötti nem-lineáris összefüggéseket is. Előbbit inkább akkor használják, ha a kapott dimenziókat tovább viszik más elemzésekbe (erre az esettanulmányok között mutatunk majd példát), utóbbiakat pedig inkább akkor, ha képet akarunk kapni egy szóhalmaz belső strukturálódásáról. A komplexebb megközelítések közül leginkább a T-SNE (T-Distributed Stohastic Neighbour Embedding: Maaten – Hinton 2008) használják szöveges vektorterek esetében tartalmi vizsgálatra. A következő ábrán az ezzel a módszerrel kapott első két dimenzió mentén ábrázolt tudományterületek láthatóak.



Ábra 5. Tudományterületek távolsága T-SNE módszer alapján, Wikinews, saját számítás

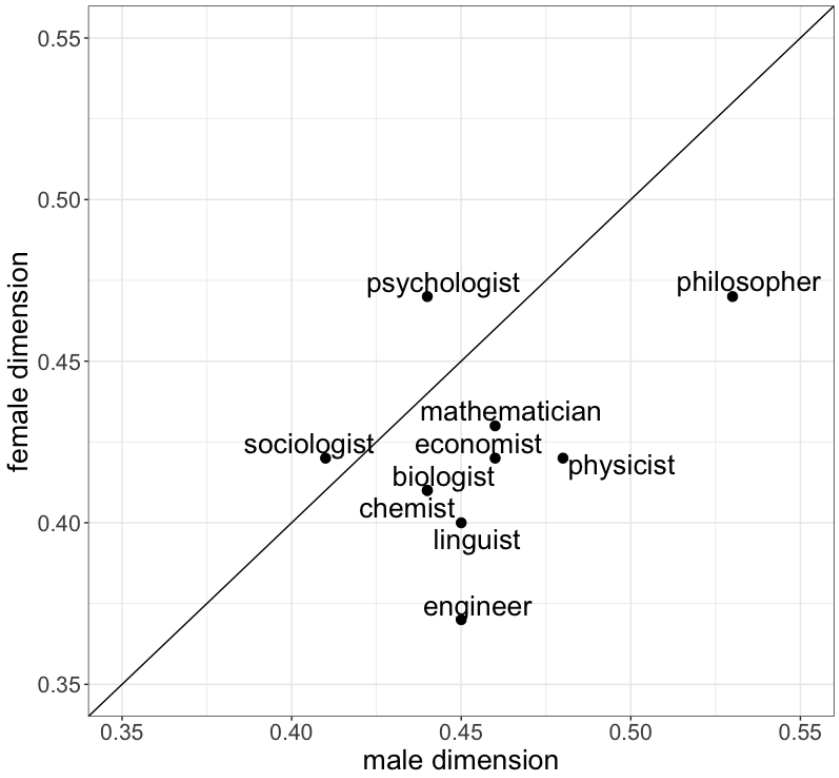
A dimenziók értelmezése egyáltalán nem triviális egy T-SNE modellben. Az Y dimenzió a korábban már többször bemutatott természettudományok vs humán-társdalomtudományok dimenziója, az X dimenzió viszont nehezen interpretálható.

Az eddig bemutatott elemzésben a tudományterületek szavainak közelségéből indultam ki és exploratív logikát követtem. De a tudományterületek egymáshoz viszonyított pozícióját lehet külső dimenziók mentén is vizsgálni konfirmatív logika mentén. A következő ábrán két dimenziót határoztunk meg, az egyik a társadalmi („social”) a másik pedig a technikai („technical”). Ezeket a dimenziókat a két szóhoz vett közelséggel mértük. Ebben a két dimenzióban is elvállnak 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 technikai dimenzióban.



Ábra 6. Tudományterületek pozíciója a „social” és „technical” szavakhoz viszonyítva, Wikinews, saját számítás

A vizsgált dimenziókban a tudományterületek elkülönülése logikusnak tűnik. Az eddig bemutatott eredmények társadalomtudományi szempontból kevésbé érdekesek, de a módszerben rejlő lehetőségeket jól mutatják. A „social” és „technical” szavak helyére bármit behelyettesíthetünk – a módszer lehetővé teszi, hogy komplex módon megvizsgáljuk, hogy egy jelenség körül milyen kommunikáció van az online diskurzusban. Az egyik izgalmas kérdés például, hogy milyen gender vonzata van egyes szavaknak. A szociológusok lányok, a mérnökök pedig fiúk? Vagy fordítva? Ennek a kérdésnek a „tétjére” a következő fejezetben még visszatérek, a mostani részben csak az elemzési megközelítésre fókuszálok. A tudományterületek helyett az adott területekhez tartozó szakmáknak számolom ki a férfi-női távolságát. A férfi („male”) dimenzióhoz való közelséget a „he” szóval operacionalizálom, a női („female”) dimenziót a „she” szóval. A komplexebb dimenzió meghatározásra visszatérek majd a későbbiekben.



Ábra 7. Foglalkozásos pozíciója a „he” és „she” szavakhoz viszonyítva, Wikinews, saját számítás

Az ábra egyszerűbb interpretációját segíti a behúzott egyenes, ami elválasztja a női és férfi oldalt. A legtöbb tudományterület közelebb van a férfi dimenzióhoz, leginkább a mérnöktudományok, a nyelvészet, a fizika és a filozófia. A női oldalon egyedül a szociológust és a pszichológust találjuk.

Az eddig bemutatott példákban tudományterületeket, illetve a területekhez köthető foglalkozásokat vizsgáltam exploratív (klaszter, network) és konfirmatív (távolság általunk meghatározott szavaktól) megközelítéssel. A fejezet utolsó részében azt mutatom meg, hogyan lehet ezeket a módszereket kombinálni egy elemzésben. Ugyanazt a beágyazási modellt használom, mint a fejezet eddigi részében, de új fókusszal – a korrupció szó szemantikai környezetét elemzem részletesebben.

A korrupció szöveganalitikai vizsgálata azért kifejezetten érdekes kérdés, mert a szó jelentésének nincs jól körülhatárolt alakja. Bár vannak korrupciós definíciók, de az hogy ki mit tekint korrupciónak nagyon eltérő lehet azonos kultúrkörön belül is, de országok/kultúrák között még inkább.

A Wikinews korpuszra épülő szóbeágyazásban a korrupcióhoz legközelebbi szavak között találjuk a megvesztegetést („bribery”, „venality”) a barátoknak („cronysm”) és rokonoknak („nepotims”) nyújtott (nem megérdemelt) előnyöket leíró szavakat, valamint a korrupció különböző nyelvi alakjait és szinonimáit („corruptions”, „graft”).

Érdekes azt megvizsgálni, hogy egyes országokban milyen speciális korrupcióval kapcsolatos kifejezések bukkannak fel. A [„corruption” + „China” – „country”] vektorművelet előhív olyan szavakat, mint például a „guanxi”, ami Kínában azt a

speciális kapcsolthálózati rendszert jelenti, ami az üzleti (és más) viszonyokat működtet. De a topszavak között előjön például a (környezet)szennyezés („pollution”) is. Ha ugyanezt megvizsgáljuk Brazília esetében akkor az erdőírtást találjuk („deforestation”) Magyarország esetében pedig a baráti és rokoni „segítségnyújtást” („nepotism”, „cronysm”).

De vajon mutat-e valamit az, ha a korrupciót jelző szavak közel szerepelnek egy országhoz? Megfordítva a kérdést: azok az országok, amik korruptabbak közelebb szerepelnek a korrupciót jelző szavakhoz? Erre a kérdésre nem tudunk egzakt választ adni, de arra lehetőségünk van, hogy egy szóbeágyazásból készített korrupciós indexet összevessünk egy külső korrupciós index-szel. Utóbbihoz a Transparency International 2018-as korrupciós indexét (TCI) használom fel[18]. Az elemzésbe a következő 15 ország került be: USA, Németország, Magyarország, Franciaország, Olaszország, Görögország, Spanyolország, Svédország, Dánia, Anglia, Brazília, Kína, Oroszország, Románia, Szlovákia.

A korrupciós index legegyszerűbb mérése a vektortérben a vizsgált országoknak és a korrupció [corruption] szó koszinusz közelségének a kiszámolása. A TCI-vel ez a mutató -0.48-as korrelációt mutat. Mivel a TCI esetében az alacsonyabb érték jelent magasabb korrupciót a két mutató összefüggése a várt irányban alakult. Az elemzés korábbi részében azonban már rámutattunk, hogy a korrupció eltérő tartalmú lehet egyes országokban, ezért érdemesebb lehet komplexebb módon mérni a vektortérben. A komplexebb mérésre több lehetőségünk van. A kiindulópont minden esetben az, hogy meghatározunk olyan szavakat, amik az adott fogalmat mérik (ezt

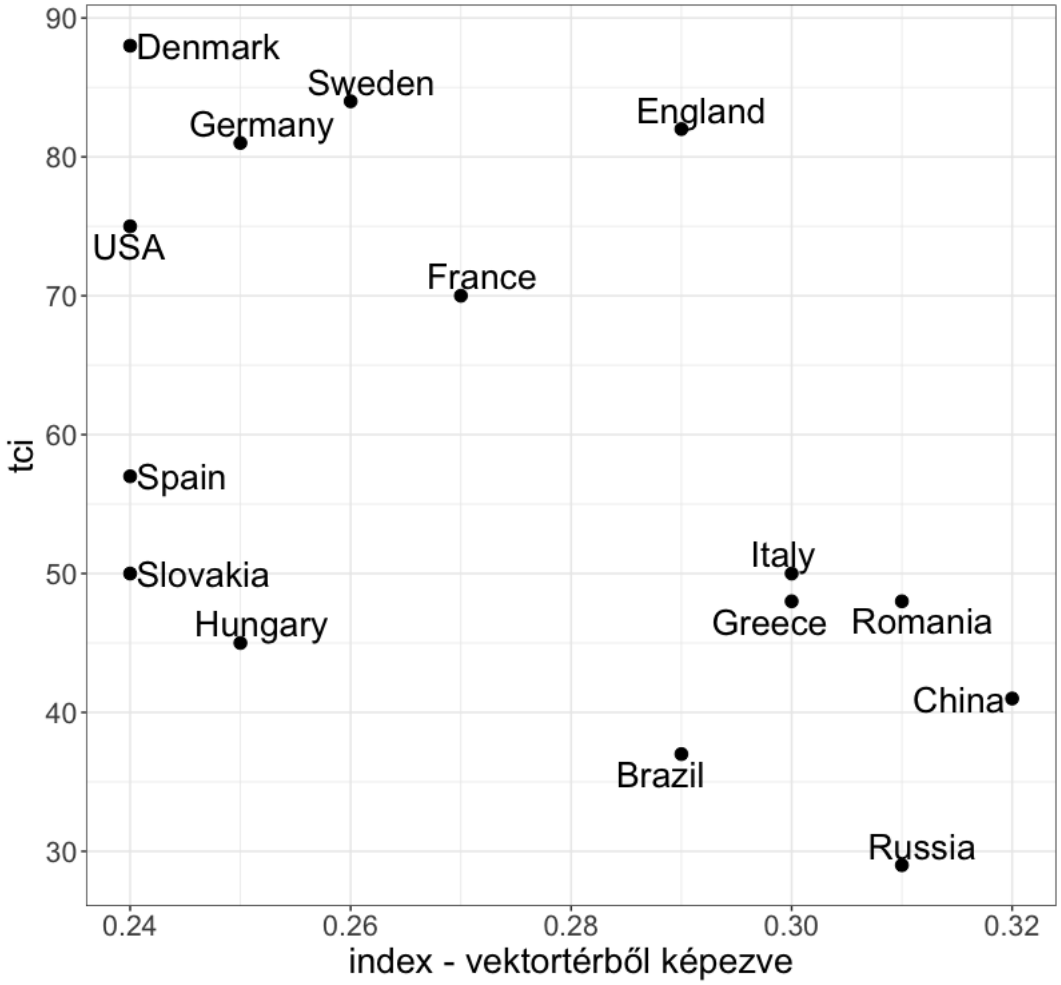
[18] https://www.transparency.org/en/cpi#

nem lehet teljesen adatvezérelten megoldani, a végső lépésben kell egy kutatói döntés). A korrupció esetében 7 szót választottunk ki a fogalom mérésére: corruption, bribery, cronyism, nepotism, "malfeasance", "venality","graft".

Követhetünk olyan megközelítést, hogy kiszámoljuk mind a hét korrupciót mérő szóval az egyes országok távolságát, és a 7 távolság átlagát használjuk az indexnek. Ezzel a megközelítéssel az összefüggés valamennyit erősödik (-0.53). Ha kivesszük a legkevésbé jól működő szót [bribery], akkor az összefüggés még erősebb lesz a TCI-vel: -0.6.

Szintén érvényes megközelítés az, ha a korrupciót mérő szavakat már a vektortérben összeadjuk és az összegvektorral számoljuk ki az egyes országok koszinusz közelségét. Az így kapott korreláció -0.49 lett. Ha az összeadásból kivesszük itt is a megvesztegetés [bribery] szót, akkor -0.56-os értéket kapunk. Bármelyik megoldást is válasszuk, érdemes a robusztusabb megoldást használni, tehát egy adott dimenziót több szóval mérni. Ez a logika nagyon hasonló azzal, ahogy survey-ek esetében képzünk indexet (lásd: indikátorok felcserélhetőségének elve).

A vektortér összeadásra épülő módszerből kapott index („bribery” nélkül) és a TCI által kifeszített kétdimenziós térben érdemes vizuálisan és megnézni az egyes országok helyzetét.



Ábra 8. Országok pozíciója a TCI indexhez és a vektortérből kiszámolt korrupciós indexhez képest , Wikinews, saját számítás

Az ábra bal-felső és jobb-alsó sarkában elhelyezkedő országokat konzisztensen méri a TCI és a vektortér index. Spanyolország, Szlovákia és Magyarország esetében a TCI „szigorúbb” mint a szóbeágyazási modellből kapott mutató, Anglia esetében viszont utóbbi becsül rosszabb értéket. De egy általános korpuszon képzett beágyazási modell esetében nem is várhatunk ennél erősebb korrelációt. A korrupció területi (országok közötti) különbségeit azt vélelmezhetjük, hogy tudja mérni a beágyazás. Viszont a korrupciónak van egy erős temporális volatilitása. Érdemes csak Magyarországra gondolni, ahol a 2000-es évek közepi 40. hely környékéről a 70. helyre esett vissza az ország a korrupciós rangsorban. A wikinews egyik alapja a Wikipedia, ami számos történelmi bejegyzést tartalmaz (lásd Anglia rossz értéke a

beágyazási modellben) a másik pedig a UMBC korpusz, ami egy 2007-es gyűjtés. Utóbbi magyarázhatja Magyarország, Szlovákia és Spanyolország esetében is a különbséget. Érdekes irányvonal lehetne egy frissebb korpuszon megvizsgálni az előbb bemutatott összefüggést, ez azonban már túlmutatna jelen fejezet alapvetően demonstrációs célú keretein[19].

A fejezetben számos olyan technikát megmutattam, amivel a vektortereinket elemezhetjük. A következő fejezetben a módszer társadalomtudományi hasznosíthatóságára koncentrálok, valamint arra, hogy miért is érdemes ezeket a modelleket társadalomtudományi szempontból is vizsgálni.

[19] A téma iránt érdeklődőknek javaslom Axelsson és Dahlberg (2018) érdekes cikkét. A svéd kutatók több országban az adott ország saját nyelvén megjelenő online tartalmak alapján vizsgálták azt, hogy hol milyen korrupciós formák jellemzőek.

2.4 A társadalomtudományok szerepe és lehetőségei

Az előző fejezetet azzal zártam, hogy a módszer társadalomtudományi hasznosíthatóságra fogok fókuszálni a következő részben, valamint arra, hogy miért érdemes a vektortér modelleket társadalomtudományi szempontból vizsgálni. Az első cél triviális egy társadalomtudományi kutatási keretben, utóbbi viszont nem feltétlenül az, ezért ezzel kezdem ezt a fejezetet.

Ahogy a dolgozat korábbi részében részletesen kifejtettem, számos mesterséges intelligencia (MI) alkalmazás használ nyelvi modelleket. A modellek egy részének kimenete maga is valamilyen szöveges tartalom lesz – például egy fordítás. De számos esetben a kimenet valamilyen klasszifikáció: Pl spam-e az adott email vagy tovább jusson-e az adott CV a következő körbe. Egy MI algoritmus számos (nem szándékolt) torzítást tartalmazhat (Mehrabi et al 2019), ezek a torzítások pedig negatívan befolyásolhatják a klasszifikációs modelljeink működését. A torzítást legegyszerűbben a fordítóprogramokon lehet vizsgálni. Ezt a google fordító működésével fogom demonstrálni (hasonló logikájú részletes elemzés kapcsán lásd: Prates et al 2019). A következő kis „játékot” bárki megcsinálhatja, tetszőlegesen variálva a tartalmat. Vegyünk egy rövid magyar szöveget, amit le szeretnénk angolra fordítani:

’Az iskolában mindenkiről készült egy jellemzés. Szabó-ról a következőt mondták. Ő biztos, hogy **politikus** lesz.’

A következő fordítást kapjuk vissza:

„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 érdekes, hogy a politikus a „he” szót hívja elő, tehát ha politikus lesz, akkor ő férfi. De mi történik, ha a magyar szövegben kicseréljük a politikust, tanárra?

„**She's** sure to be a **teacher**.”

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

Foglalkozás	személyes névmás
doktor	férfi
sebész	férfi
bőrgyógyász	nő
fogorvos	nő
tudós	férfi
pszichológus	nő
sofőr	férfi

Tábla 4. Foglalkozásokhoz rendelt személyes névmás a google fordító alapján, saját számítás

A modellt nem csak foglalkozások esetében lehet kiértékelni. Ha valaki okos akkor férfi, ha érzelmes akkor pedig nő a google fordító szerint. Vállrándítással elintézhetjük, hogy nincs igazán jelentősége az életben annak, hogy a google fordító szerint a lányok sírósak, a fiúk pedig bátrak. A valóságban azonban ez nem igaz, mert ez pont annak a mechanizmusnak a jól megfogható manifesztálódása, ami megerősíti a szerep hierarchiákat ezáltal további erősítve a gender egyenlőtlenséget. Ráadásul, ha azt feltételezzük, hogy a Google más moduljai, például a reklám ajánlati rendszer (google ads) használ hasonló nyelvi modulokat, akkor a torzítás hatása már direktben is megjelenik (Datta et al 2015). De a klasszifikációs modelleknél nehezebb dolgunk van, ha rekonstruálni szeretnénk a torzítást. A legtöbb ipari alkalmazás teljes feketedobozként működik, nem nyilvánosak sem a „bemeneti" adatok, sem a feldolgozási algoritmusok. Ha esetleg azt tudjuk azonosítani, hogy az algoritmus hátrányosan megkülönböztet bizonyos társadalmi csoportokat, még akkor sem tudhatjuk, hogy pontosan miért jön létre a torzítás. Erre jó példa Chen és munkatársai (2018) kutatása, amiben azt vizsgálták, hogy állás közvetítő oldalon, van-e abban különbség, hogy hányadik helyen jelennek meg a férfi és női munkavállalók a kereső ablakban. Elemzésükben kimutatták, hogy annak ellenére, hogy az állásadók neme nincs külön regisztrálva (a kutatók is név alapján következtettek a nemre) a női álláskeresők kis mértékben hátrébb rangsorolódnak a keresésekben, akkor is, ha minden lehetséges háttérhatást kiszűrnek. Ennek egyik lehetséges oka az, hogy az algoritmus „rátanul" olyan nyelvi elemekre az önéletrajzokban, amik látens módon összefüggenek az álláskereső nemével (ennek kapcsán lásd még: De-Arteaga et al 2019).

A társadalomtudósok szerepe ebben az esetben elsősorban az, hogy kritikai szemmel vizsgálják azokat az algoritmusokat, amelyek akár napi szinten döntést hoznak az

emberek életéről. Ha ezeket a torzításokat sikerül azonosítani, akkor lehetőség lesz lépéseket tenni annak érdekében, hogy ezeket megszüntessék az alkalmazás készítői. A szóbeágyazási modellek esetében több olyan módszertani javaslat készült, ami felismeri a vektorterekben a torzításokat (Caliskan et al 2017, Garg et al 2018) és ezeket kezeli is (Bolukbasi et al 2016, Zhao et al 2018, Manzini et al 2019, Gonen – Goldberg 2019). A torzításokat természetesen nem a beágyazási algoritmus okozza, hanem a bemenetként használt korpuszok, amik legtöbb esetben magukban hordanak rengeteg nemi és etnikai sztereotípiát (a Wikipedia kapcsán lásd pl: Wagner et al 2015).

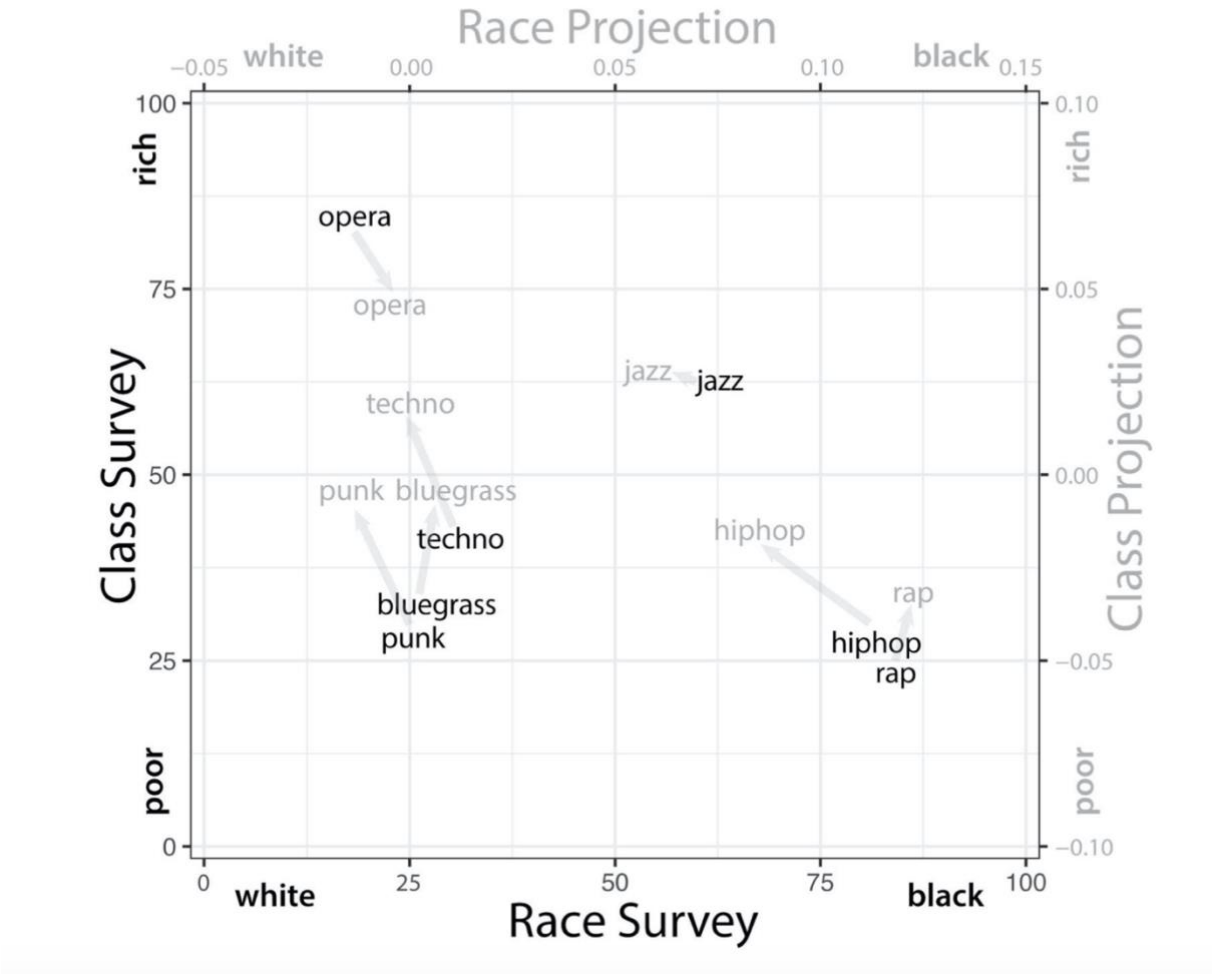
A társadalomtudósok nemcsak „kapuőrként" vizsgálhatják a vektortereket, hanem felhasználói szempontból is értékesek ezek a modellek. Felhasználó szempontból megkülönböztethetjük a technikai és a tartalmi felhasználást.

A technikai felhasználás nem érdemel hosszú magyarázatot. Itt egyszerűen arról van szó, hogy valamilyen klasszifikációs modellben szóbeágyazási módszereket használnak fel. Ezekben az esetekben az elemzés nem a szóbeágyazásra fókuszál, hanem a klasszifikáció kimenetére (Yang et al 2018, Samory et al 2020), de a klasszifikációs modell bizonyos paraméterei (pl. egy adott kategóriába kerülést milyen szavak generálták) érdekesek lehetnek az elemzésben is (Nakandala et al 2016).

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 szociológia alkalmazásban maga a kibányászandó eredmény. Mik azok a foglalkozások, amik erősen kötődnek nemhez vagy etnikumhoz? Milyen szabadidő

eltöltési formák jellemzőek a gazdagokra és szegényekre? Hogyan alakult bizonyos fogalmak/csoportok társadalmi kontextusa az elmúlt 100 évben? Számos olyan társadalomtudományi kérdés, amire képesek választ adni vektortér modellek.

A módszer tartami felhasználhatóságának elismerése kapcsán fontos mérföldkőnek tekinthetjük Kozlowski, Taddy és Evans ”The Geometry of culture” című tanulmányának 2019-es megjelenését a szociológia zászlóshajó lapjában az American Sociological Review-ban. Kozlowski-ék munkája két szempontból is nagyon érdekes. A szerzők egyrészről azt vizsgálták, hogy különböző kulturális és szabadidős kérdések nemi/etnikai/osztály beágyazottsága mérhető-e vektortér modellekkel. A módszer kiértékeléséhez felhasználtak egy saját survey kutatást, amiben a minta tagoktól azt kérték, hogy szemantikus differencia skálákon osztályozzanak kérdéseket. Az osztályozási szempont az volt, hogy a vizsgálati „objektum” mennyire férfias vagy nőies, fehér vagy afro-amerikai illetve alsó- vagy felső osztályhoz köthető. Sok dimenziót bevontak az elemzésbe: ételeket, zenei stílusokat, foglalkozásokat, sportokat, járműveket és még keresztneveket is. A survey és a szóbeágyazási eredmények erős összefüggést mutattak. Leginkább a nemi bontásban egyezett meg a két módszer, itt 0.7 és 0.9 közötti korrelációt kaptak a szerzők a survey-ből és a vektortérből kialakított gender skála között. Ezek az értékek valamivel alacsonyabb voltak etnikai bontásban és osztály dimenzióban, de még utóbbi esetben sem mértek 0.4 alatti összefüggést (Kozlowski e at 2019).

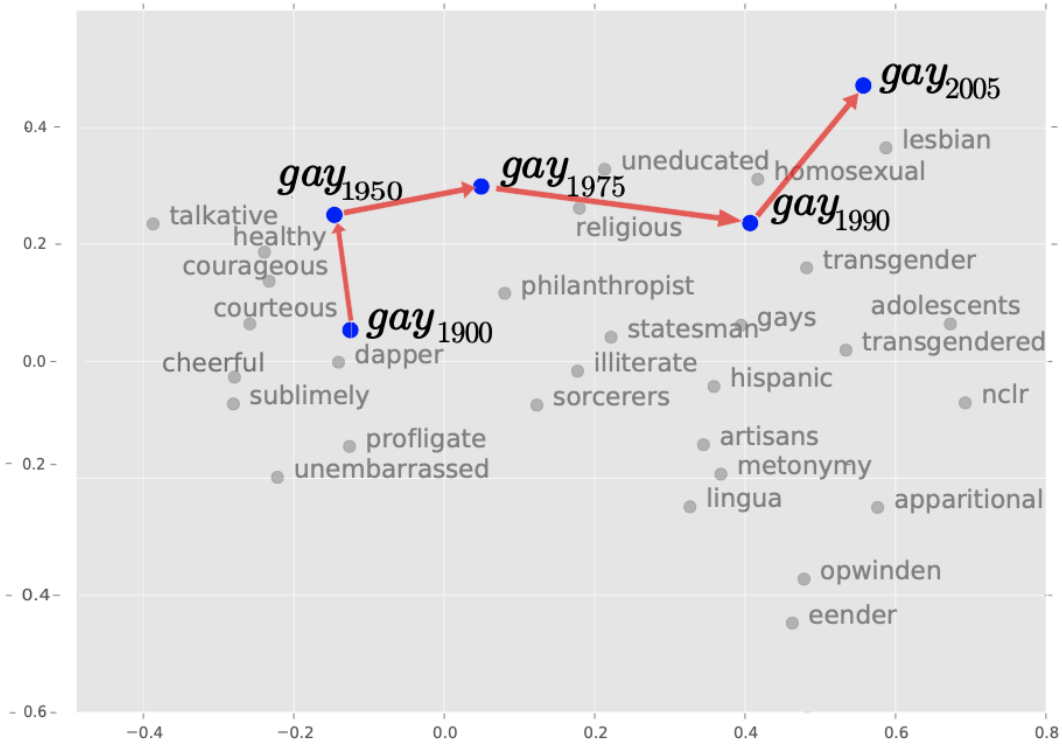


Ábra 9. Zenei stílusok nemi és etnikai kötődése a survey és vektortér modellekben, Kozlowski et al 2019

Kozlowskiék tanulmányához hasonlóan Joseph és Morgan (2020) is survey eredményekkel vetette egybe a vektortér modellek eredményeit, de egy jóval szélesebb item szettet használva. Eredményeik azt mutatták, hogy azokat a koncepciókat lehet jól mérni szóbeágyazással, amik esetében egy survey-ben is nagy az egyetértés a válaszolók között. Tehát minél extrémebb egy fogalom kulturális beágyazottsága és minél kisebb szórással ítélik meg az emberek ezt a beágyazottságot, annál erősebb az összefüggés a survey és a vektortér modellek eredményei között. Joseph és Morgan (2020) elemzései arra is rámutattak, hogy fontosabb az, hogy mit mérünk, mint az, hogy hogyan mérjük. A survey és a vektortér

modellek közötti összefüggés erősségére nem igazán hatott a korpusz választás, vagy az, hogy melyik beágyazási algoritmust használták.

A külső validáció mellett a másik izgalmas irány Kozlowskiék elemzésében egy történeti összevetés volt. Elemzésükben azt vizsgálták, hogyan változott egyes foglalkozások osztály és gender pozíciója 1900-tól kezdve napjainkig, illetve összességében hogyan alakult a gender és osztály fogalmak interszekciója. Az elemzésükben bemutatták, hogyan ment át az osztály egy kulturális kategóriából egy technikaibb munkaerő piaci kategóriává és ez az átalakulás időben hogyan csúszott el az USA-ban Angliához képest (Kozlowski et al 2019). Kozlowskiék munkája jól illeszkedik abba a sorba, amelyben vektortér modellek segítségével vizsgálnak, történeti perspektívában fogalmi változásokat (módszertan kapcsán lásd Hamilton et al 2016a, Hamilton et al 2016b.) Az egyik első hasonló tematikájú elemzésben Kulkarni és szerzőtársai (2015) a Google n-gram korpuszon vizsgálták szavaknak a jelentésváltozását. A leginkább hivatkozott példájukban a meleg („gay”) szó kapcsán mutatták be, hogyan alakult át a vidám/jól öltözött jelentése a szónak az 1970-es években egy szexuális csoportot leíró szóvá.



Ábra 10. A meleg (gay) szó kontextusváltozása 1900 és 2005 között, Kulkarni et al 2015

Hasonlóan nagy történeti perspektívát fog át Garg és munkatársainak (2017) előítéletekre fókuszáló cikke. A szerzők azt vizsgálták, hogy milyen erős etnikai és gender kötődése volt egyes foglalkozásoknak 1900-tól kezdve. A beágyazásból kapott eredmények összecsengtek a népszámlálási adatokban is látható foglalkozási mintákkal. Gargék azt is elemzték milyen sztereotip kifejezések kötődtek a különböző etnikumokhoz és ez hogyan változott a különböző bevándorlási hullámokkal párhuzamban.

Gargék tanulmánya több szálon is kapcsolható a következő fejezetekben bemutatott saját esettanulmányokhoz. Az elsőként bemutatott Rudas Tamással és Koltai Júliával közös tanulmányunk (Kmetty – Koltai – Rudas 2020) a foglalkozások egymáshoz való pozícióját vizsgálja előre kialakított vektorterekben. Tanulmányunkban olyan

hierarchiát alakító aspektusok fontos szerepére hívjuk fel a figyelmet, ami eddig a szakirodalomban kevéssé volt hangsúlyos.

A történeti elemzések irányához kapcsolódnak az ELKH TK, CSS-RECENS kutatócsoportjában végzett elemzéseink, amelyben a Kádár-rendszerben vizsgáltuk egyes szavak és fogalmak jelentésváltozását (Szabó et al 2020). A jelen dolgozatban bemutatott második esettanulmányban az elemzési keretünk módszertani aspektusára koncentrálok – és azt mutatom be, hogyan lehet a fogalmak jelentésváltozását bemutatni egy viszonylag kis magyar korpuszon.

A harmadik esettanulmány egy speciális adathalmazt vizsgál nemi sztereotípiák után kutatva – a Twitch videomegosztó oldalt. Tóbiás Dániellel közös tanulmányunkban női és férfi gamerek videói alatt megjelenő kommentekben vizsgáljuk a nemi objektifikáció megjelenését. Elemzésünkben azt demonstráljuk hogyan lehet konfirmatív kutatásokban felhasználni a vektortér reprezentációt.

3 Esettanulmány 1 - Foglalkozások pozíciója a szemantikus térben

Eredeti tanulmány[20]

Kmetty, Z., Koltai, J., & Rudas, T. (2020). The presence of occupational structure in online texts based on word embedding NLP models. arXiv preprint arXiv:2005.08612. https://arxiv.org/pdf/2005.08612

3.1 Problémafelvetés

A társadalmi rétegződés és osztályszerkezet vizsgálat, a szociológia alapkutatási kérdései közé tartozik több mint 100 éve. Többfajta megközelítést használ a szakma a társadalmi pozíció vizsgálatára; egyes modellekben az elmélet dominál, más modellek inkább adat vezéreltek; megkülönböztethetünk folytonos és kategoriális besorolásokat. De megközelítéstől függetlenül igaz, hogy a modellek jelentős részében a foglalkozási helyzetből indulnak ki a társadalmi pozíció meghatározásakor. Az ipari társadalmakban a foglalkozás nagyon jól méri a társadalmi helyzetet – sokkal stabilabb indikátor, mint a vagyon, vagy az aktuális jövedelem (Connelly et al 2016). A megközelítések sokféleségéből adódik, hogy nincs arra egységes definíciós keret, hogy használják a társadalomtudósok a foglalkozásokat a társadalmi pozíció mérésekor. A modellek egy részében horizontális kategóriákba sorolják a társadalom tagjait a foglalkozások alapján (Goldthorpe et al 1982, Rose-Harrison 2007), más kutatók graduális skálákat szerkesztenek (Ganzeboom - Treiman 1996). Bukodi

[20] A habilitációs dolgozatban a tanulmányunk rövidített változatát mutatom be. A szóbeágyazási módszer bemutatását teljes mértékben elhagyom, mivel ezt részletesen kifejtem ebben a dolgozatban.

szerzőtársaival (2011) két nagy csoportra osztotta a foglalkozási rétegződés/osztály modelleket aszerint, hogy objektív vagy szubjektív indikátorokat használnak a társadalmi pozíció mérésekor. Goldthorpe és Gold (1972) szubjektív skálákon mérték, hogy egyes foglalkozások mekkora presztízzsel rendelkeznek az emberek szerint. Treiman (1977) szintén szubjektív kérdésekkel vizsgálta a foglalkozások egymáshoz képesti pozícióját. A kutatási alapján kialakított SIOPS skálát mai napig széles körben használják az elemzésekben. Ezzel szemben az ISEI (International Socio-Economic Index - Ganzeboom-Treiman 1996) és a Cambridge skála (CAMSIS - Prandy-Lambert 2013, Meraviglia et 2016) objektív indikátorokat használ. Az ISEI az egyes foglalkozásokhoz rendelt iskolai végzettség és átlagos jövedelem alapján képez hierarchiát a foglalkozások között. A CAMSIS házastársak foglalkozási mátrixából képzi a hierarchiát. Utóbbihoz nagyon hasonló Chan és Goldthorpe (2004) megközelítése, annyi különbséggel, hogy barátok foglalkozási mátrixából indulnak ki a szerzők. Meraviglia és szerzőtársai (2016) elemzései alapján, az elméleti különbségek ellenére a folytonos foglalkozási skálák között nagyon magas empirikus korreláció figyelhető meg.

Az eltérő kiindulás miatt abban sincs egyetértés a modellek kapcsán, hogy a foglalkozásoknak melyek azok az elemei, amik a hierarchiáért felelnek. Az egyik legszélesebb körben használt EGP (Erikson et al 1979) modellben a munkavállalók munkaerőpiaci pozíciójából indulnak ki. Az adott munkakör stabilitása, jövedelmi pozíciója, illetve a munkakörhöz kötődő autoritási szint az elsődleges tényezők a kategória besorolásban (Conelly et al 2016). A korábban említett ISEI skálákában a jövedelem mellett, az iskolai végzettség játszik döntő szerepet a hierarchia alakításában. A SIOPS esetében pedig egy elvontabb presztízs fogalmat használnak, aminek elemei a jövedelem, a tudás, de a munkához kapcsolódó hatalom is.

A tanulmányunkban bemutatott újszerű megközelítés az adatvezérelt modellek közé tartozik, leginkább a CAMSIS modellel mutat hasonlóságot. A társadalmi közelség helyett, azonban a foglalkozások szemantikai közelségét használjuk fel a foglalkozási hierarchia kialakítására. A szóbeágyazási módszeren alapuló exploratív modellünk arra is lehetőséget teremt, hogy feltárjuk milyen tényezők mentén különülnek el egymástól a foglalkozások az online szövegek szemantikai terében.

3.2 Adatok és módszerek

A foglalkozási struktúra szemantikai vizsgálatához előkészített vektortereket használtunk fel. Mivel nem egy speciális, hanem egy általánosabb kutatási kérdésünk volt, ezért megfelelőnek gondoltuk ezeket a vektortereket az elemzéshez. Mivel ezek a vektorterek szabadon hozzáférhetőek és letölthetőek, ezért az eredményeink bárki számára viszonylag egyszerűen reprodukálhatóak.

Három előkészített vektorteret használtunk. Az első beágyazási modell a Common Crawl (CC) webarchívumra épült, 2011 és 2017 között letöltött oldalakat felhasználva. Angol nyelvű oldalakat tartalmaz a korpusz, de mivel nincs geolokáció az oldalak a világ bármelyik országából származhatnak. 600 milliárd szóból állt a korpusz, ebből a vektortérbe 2 000 000 egyedi szó került be, 300 dimenziós térbe tömörítve. A második korpusz a Wikinews volt. Mivel a habilitációs dolgozat korábbi részében már használtam ezt a korpuszt, nem mutatom be hosszabban. 16 milliárd szóból állt a korpusz, ebből a vektortérbe 1 000 000 egyedi szó került be, 300 dimenziós térbe tömörítve. Mind a két vektorteret a Fasttext algoritmussal (Joulin et al 2016) ágyazták

be. A CC és az alap Wikinews vektortérben nem használták fel a beágyazás során azt, hogy a Fasttext karakter n-grammokat is be tud építeni a futáskor, ezért ezek a modellek a klasszikus word2vec modellhez vannak közelebb. A harmadik, szintén wikinews korpuszra épülő vektortér képzésekor viszont a karakter n-grammra bontást is felhasználták. A későbbiekben ezt „wikinews subwords” vektortérként hivatkozzuk.

Összesen 234 foglalkozást választottunk ki (lásd melléklet M1 tábla). A 234 foglalkozás kiválasztásánál arra törekedtünk, hogy vertikálisan és horizontálisan is jól lefedjék a teljes foglalkoztatási mezőt. Az angol nyelvben sok foglalkozás két szóból áll (pl. Data Scientist). Mivel előkészített vektortereket használtunk az elemzésben, ezért a két szóból álló foglalkozásokat ki kellett hagynunk. A vektortér modellek esetében az alacsonyabb említésű szavak pozíciója kevésbé stabil, ezért a robusztusság növelésének érdekében úgy döntöttünk, hogy a vektorterekből a leggyakoribb 200 000 szót használjuk fel. A szűkítés miatt valamelyest csökkent az elemzésben felhasználható foglakozások száma is. A CC korpuszban 204, a Wikinews korpuszban 207 foglalkozás maradt bent – 202 foglalkozás volt benne mind a két szűkített korpuszban.

Mind a három vektortér esetében ugyanazt a módszertat követtük. Első lépésben kiszámoltuk az összes foglalkozás egymással vett koszinusz közelségét. Néhány példa foglalkozás esetében ezt a mátrixot mutatja a következő tába (Tábla 5).

	doctor	cardiologist	sociologist	historian	shopkeeper	barmaid
doctor	1.00	0.61	0.29	0.25	0.34	0.25
cardiologist		1.00	0.32	0.27	0.20	0.11
sociologist			1.00	0.62	0.31	0.25
historian				1.00	0.26	0.20
shopkeeper					1.00	0.48
barmaid						1.00

Tábla 5. Hat kiválasztott foglalkozás szemantikai közelsége (koszinusz közelség, CC vektortér)

Mivel hasonló szemantikai tartalmakat vizsgálunk (foglalkozások), nem meglepő módon pozitív koszinusz közelséget kapunk. Az értékeken belül viszont viszonylag nagy a szórás. Nem meglepő módon a doktor meg a kardiológus közel van egymáshoz, akárcsak a szociológus és a történész, illetve a bolti eladó és a pultoslány. Ez a kis minta már előrevetíti, hogy lehetséges lesz különböző strukturáló tényezőket azonosítani a foglalkozási közelség mátrixban. A legfontosabb dimenziók megtalálásához egy klasszikus módszert, a faktoranalízist használtuk. A bemenet ebben az esetben azonban nem egy korrelációs mátrix volt, hanem a korábban bemutatott közelség mátrix. Többfajta faktoranalízist is teszteltünk – de nagyon kis különbség volt csak az eredmények között. A tanulmányunkban a minres (minimum residual) módszert használtuk, varimax rotációval (Revelle 2018). A későbbi elemzésben az egyes foglalkozásokhoz rendelt faktorsúlyokat használjuk fel.

Mivel vállaltan exploratív jellegű a tanulmányunk, ezért nem volt arra hipotézisünk, hogy hány értelmes faktorba rendeződhetnek a foglalkozások. A végső fakorszám meghatározásánál egyrészről empirikus tesztekre támaszkodtunk, és praktikus

szempontokat is figyelembe vettünk. Mivel a korábbi témában született tanulmányok azt hangsúlyozták, hogy egynél több fontos dimenzió húzódik meg a foglalkozások szerveződésében, ezért a minimális faktorszámot kettőben határoztuk meg. Az interpretálhatóság érdekében viszont nem akartunk 5-nél több faktort kinyerni a modellekből. A modelleket végigteszteltük 2-5 közötti összes faktorszámra. Mind a khi-négyzet próbák mind az RMSR értékek alapján a 2 és 5 közötti faktorszám jól illeszkedett a kiinduló modellre. A részletes elemzések azt mutatták, hogy a három faktoros megoldás adja a legígéretesebb eredményeket, ezért a dolgozatban ennek az eredményét mutatjuk be részletesebben, de a többi esetre is kitérünk majd.

Egy újszerű megközelítés esetében nagyon fontos kérdés, hogy mennyire robusztusak a kapott eredmények. A robusztusságot két irányból is teszteltük. Egyrészről megvizsgáltuk, hogy a különböző vektorterekben kapott foglalkozási pozíciók között mennyire erős az átfedés. Ez praktikusan azt jelentette, hogy a foglalkozások faktormodellekből kapott súlyait korreláltattuk össze a vektorterek között.

Másrészről azt is megvizsgáltuk, hogy a foglalkozások kontextusa mennyire egyezik meg a vektorterekben. Ha veszünk két egymástól függetlenül képzett vektorteret akkor abban a megegyező szavak koszinusz közelsége nulla lesz. Ahhoz, hogy megvizsgáljuk a kontextus hasonlóságot a két vekorteret egymáshoz kell igazítani. Ehhez első lépésben leszűkítettük a CC és Wikinews vektortereket azokra a szavakra, amik mind a két korpuszban benne voltak. Ezután Procrustes rotációval egymáshoz igazítottuk a mátrixokat. A kiigazított vektorterekben a foglalkozások egymáshoz viszonyított közelsége megmaradt, és már értelmezhetővé vált ugyanazon foglalkozás koszinusz közelségének az értéke a két vektortér között. Ez a koszinusz közelség

sosem lesz 1, mivel a modellek között eltér a kontextus és még valamekkora véletlen zaj is keletkezik a modellek illesztésekor (lásd korábban). De a magasabb koszinusz közelség egyértelmű indikátora a kontextus hasonlóságnak (Hamilton – Leskovec – Jurafsky 2016).

3.3 Eredmények

3.3.1 Commow Crawl

A CC korpuszban a doktor volt a kiválasztott foglalkozások közül a leggyakoribb, de szintén gyakran használták a szövegekben a sofőr az író, a szakács a bíró az ügyvéd szavakat is. A kiválasztott foglalkozások közül 204 olyan volt, ami benne volt a korpusz 200 000 leggyakoribb szavában. Ennek a 204 foglalkozásnak számítottuk ki a koszinusz közelség mátrixát és ezt a mátrixot használtuk fel a faktoranalízisben. Első lépésben a két faktoros megoldást vizsgáltuk. Az RMSR érték 0.07 volt[21], ami megfelelő illeszkedést mutat. A két faktor hasonló megmagyarázott varianciával rendelkezett (ez a rotálás miatt nem meglepő). Mind a két faktor esetében a tudás látszik az egyik legfontosabb szempontnak. Az első faktor közelebb van valamennyire a média és művészet világához, a második faktor pedig a tudományhoz (lásd tábla 6).

[21] 0.10 alatt elfogadható, 0.05 alatt jó az illeszkedés

Factor 1	Factor 2
curator	historian
editor	biologist
geographer	zoologist
professor	sociologist
sociologist	geographer
biologist	physicist
chairperson	journalist
historian	ornithologist
environmentalist	lecturer
commentator	writer

Tábla 6. A 10 legmagasabb faktorsúlyú foglalkozás listája (CC)

Az eredményeink külső érvényességének teszteléséhez kiszámoltuk a faktorsúlyok korrelációját az ISEI skálával. Az első faktor korrelációja az ISEI-vel 0.64 volt, a második faktoré 0.79. Ezek az eredmények azt mutatják, hogy mindkét kapott dimenziónak nagyon erős a vertikális rendező ereje.

A három faktoros megoldásban az RMSR érték valamivel jobb volt (0.06), mint két faktor esetében. A megmagyarázott varianciák között azonban már nagyobb volt a különbség. Az első faktor korrelált legerősebben az ISEI-vel (0.71), de a második és harmadik faktor korrelációja is erős volt az ISEI skálával: 0.59/0.45. A két faktoros és három faktoros megoldások első faktorai közötti korreláció 0.9 volt, és hasonló értéket kaptunk a második faktorok esetében is. Ezek az eredmények azt mutatják, hogy a főbb dimenziók megegyeznek faktorszámtól függetlenül.

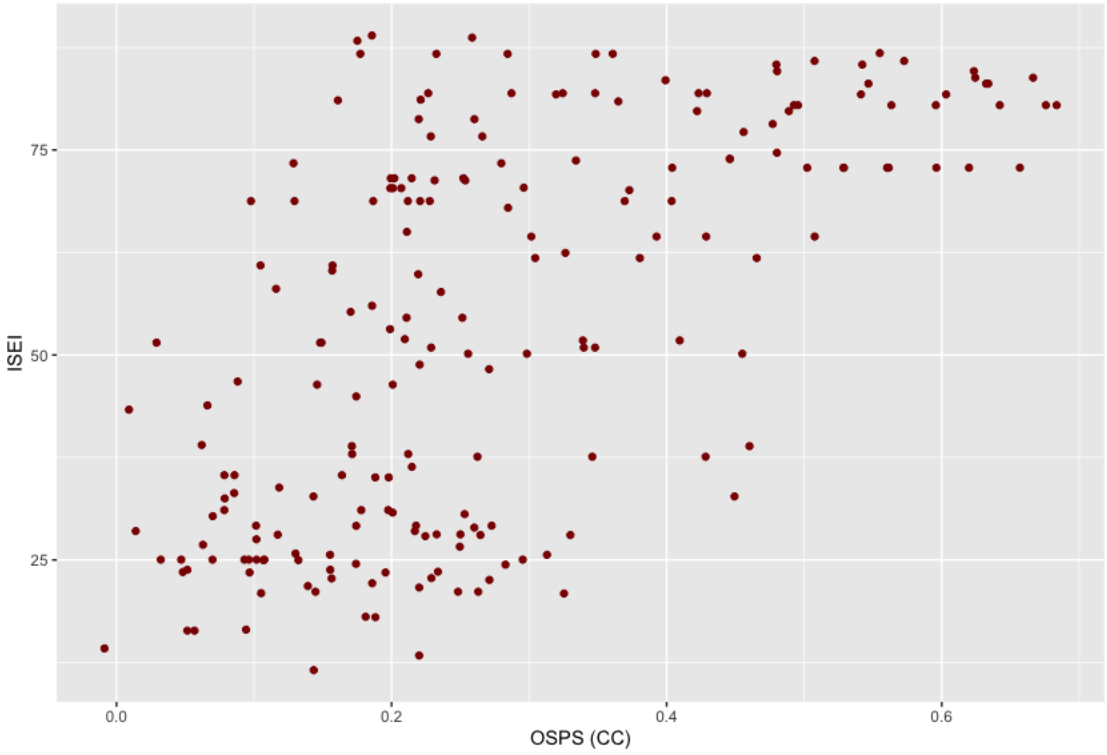
A következő tábla (Tábla 7.) a három faktorhoz tartozó legalacsonyabb és legmagasabb faktorsúlyú foglalkozásokat mutatja be.

First factor				Second factor				Third factor			
Highest loadings		Lowest loadings		Highest loadings		Lowest loadings		Highest loadings		Lowest loadings	
rank order	ISEI	rank order	ISEI	rank order	ISEI	rank order	ISEI	rank order	ISEI	rank order	ISEI
chairperson	71.29	dressmaker	23.47	ecologist	80.46	courier	30.34	secretary	44.94	brazier	28.52
ecologist	80.46	electrician	36.35	historian	83.81	stewardess	46.76	commissioner	78.76	animator	79.74
professor	85.41	waiter	25.04	biologist	80.46	waiter	25.04	treasurer	73.38	tattooist	50.15
chancellor	70.34	shopkeeper	35.34	writer	72.83	vendor	23.53	mayor	68.77	plasterer	18.02
advocate	86.72	roofer	22.16	philosopher	83.81	driver	26.85	chancellor	70.34	cleaner	16.38
dean	65.01	maid	14.21	geographer	83.09	babysitter	24.98	prosecutor	86.72	acrobat	37.59
director.general	71.29	barman	25.04	zoologist	80.46	cleaner	16.38	dean	65.01	potter	24.43
commentator	72.83	barmaid	25.04	novelist	72.83	housemaid	16.38	senator	68.77	dancer	61.82
neurologist	81.92	housemaid	16.38	sociologist	83.09	barmaid	25.04	rector	70.34	painter	61.82
historian	83.81	barber	31.08	physicist	84.61	brazier	28.52	governor	68.77	weaver	28.95
commissioner	78.76	plumber	29.16	mathematician	81.78	constable	51.5	chairperson	71.29	bender	25.78
environmentalist	80.46	blacksmith	25.63	ornithologist	80.46	receptionist	39.02	clerk	43.33	cook	24.53
curator	77.19	plasterer	18.02	poet	72.83	waitress	25.04	attorney	86.72	assembler	27.91
biologist	80.46	carpenter	26.62	journalist	72.83	clerk	43.33	congressman	68.77	dishwasher	16.5
sociologist	83.09	bricklayer	22.57	botanist	80.46	maid	14.21	constable	51.5	welder	28.52

Tábla 7. A 15 legmagasabb és legalacsonyabb faktorsúlyú foglalkozás listája (CC)

Akárcsak a két faktoros megoldásban az első két faktor itt is a tudás köré épül. Az első faktoron az intézményi hatalom jobban megjelenik a második faktoron pedig már inkább a végzettség, iskolázottság. A harmadik faktornak van egy közéleti dimenziója, de leginkább a szervezeti erő, szervezési hatalom az, ami leginkább strukturálja ezt a dimenziót.

Az első faktor ISEI skálával való összefüggését részletesebben is megvizsgáltuk. Erre a faktorra a későbbiekben OSPS (foglalkozási pozíció szemantikus skálájának angol mozaikszava) névvel hivatkozunk. A következő ábra mutatja az ISEI és az OSPS értékeinek a grafikus összefüggését.



Ábra. 11. Foglalkozások pozíciója az ISEI és OSPS skálákon (CC)

Az ábrán is jól kirajzolódik a két mutató közötti erős összefüggés. A foglalkozás párok 75%-ra igaz, hogyha a SIOPS-ban az egyik foglalkozás magasabb státuszú, mint a

másik, akkor az OSPS-ben is ezt a sorrendet látjuk. Ez az eredmény is alátámasztja, hogy a digitális térben a foglalkozások strukturálódásában erős szerepet játszanak a SIOPS-t alkotó tényezők, mint az iskolázottság és a jövedelem. Ez az eredmény azért figyelemre méltó, mert a foglalkozások digitális szegmentálásához nem használtunk semmilyen külső információt, a modellnek nem volt min megtanulni azt, hogy miként rendezze a foglalkozásokat. De az is látható az adatokból, hogy vannak azért olyan foglalkozások, amiket eltérően sorol be a SIOPS és az OSPS. Olyan foglalkozások, mint a doktor, fogász, gyógyszerész vagy ügyvéd alacsony pontszámot kapott az OSPS skálán pedig a SIOPS szerint magas a státuszuk. Ez alapvetően abból következik, hogy egy foglalkozás online térbeli pozícióját erősen befolyásolja az, hogy az adott mezőhöz milyen diskurzus tartozik. A különböző orvos szakmák nyilvánvalóan magas státuszúak, de az orvosokkal a társadalom minden tagja érintkezik, ezért nagyon változatos diskurzusokban fordulhatnak elő ezek a foglalkozások. Ráadásul betegnek lenne nem jó dolog, ami szintén visszatükröződhet a foglalkozást körülvevő szavakban. Ez a fenti logika pedig ugyanúgy alkalmazható az ügyvédekre is. Tehát a foglalkozások szemantikai rangsorát az adott mező társadalmi kötései is meghatározzák.

A három faktoros modellben, nem várt eredmény volt számunkra, hogy a jövedelem nem jelent meg jól kézzelfoghatóan egyik dimenzióban sem. Ezért azt vártuk, hogy 4-ik vagy 5-ik dimenzióként megjelenik majd. Az újabb faktor modellek azonban cáfolták ezt a várakozásunkat.

A 4 faktoros modellben az első három faktor tartalmilag megegyezett a korábban már bemutatott eredménnyel. A negyedik dimenziót pedig egyértelműen a gender alapú szétválasztás dominálta. A legmagasabb pontszámot a babysitter, a manikűrös és a

pincérnő kapták, valamint a fodrász (hairdresser és nem barber!). Az 5 dimenziós faktor modellben az ötödik dimenzió pedig egy mező specifikus elkülönítést hozott be, alapvetően az egészségüggyel kapcsolatos foglalkozások kaptak magas faktor értéket ezen a dimenzión. Fontos megjegyezni, hogy bár a dimenzió szám növelésével a korábbi faktorok tartalmilag hasonlók maradnak, de az első faktorok ISEI-vel való korrelációja folyamatosan csökken a faktorszám növelésével.

3.3.2 Wikinews

A Wikinews vektortéren megismételtük ugyanazokat a számolásokat, amiket elvégeztünk a CC vektortér esetében. Egy picit más volt a leggyakoribb foglalkozások listája ebben a korpuszban. A leggyakoribb foglalkozás a szerkesztő volt (editor) ezt követte a bíró, a politikus és az ügyvéd. Szintén gyakori foglalkozás volt az újságíró az író és az énekes. Tehát ebben a korpuszban a leggyakoribb foglalkozások a médiához, közélethez, kultúrához kötődnek.

A három faktoros modell eredménye erősen egybevágott a CC vektortéren kapott értékekkel. A két vektortérben elkészített faktoranalízisekben az első faktorok közötti korreláció 0.97 volt, a második faktorok között 0.93 a harmadikok között 0.82. Ezek az eredmények azt mutatják, hogy nagyon hasonlóan strukturálódnak a foglalkozások a két vektortérben.

A részletes elemzés rámutat azonban apróbb eltérésekre. Néhány fizikai foglalkozás, mint a lakatos vagy a mosogató magasabb besorolást kapott a Wikinews modellben, ezzel szemben irodalomhoz/kultúrához köthető foglalkozást (költő, novella író, zeneszerző, festő) magasabbra soroltunk be a CC vektortér alapján. A következő táblázat (Tábla 8.) a három faktor esetében mutatja a legmagasabb és legalacsonyabb faktor súlyt kapó foglalkozásokat.

First factor				Second factor				Third factor			
Highest loadings		Lowest loadings		Highest loadings		Lowest loadings		Highest loadings		Lowest loadings	
rank order	ISEI	rank order	ISEI	rank order	ISEI	rank order	ISEI	rank order	ISEI	rank order	ISEI
chairperson	71.29	bartender	25.04	biologist	80.46	bender	25.78	secretary	44.94	acrobat	37.59
chancellor	70.34	dressmaker	23.47	mathematician	81.78	dishwasher	16.5	prosecutor	86.72	assembler	27.91
dean	65.01	shopkeeper	35.34	zoologist	80.46	maid	14.21	mayor	68.77	bricklayer	22.57
advocate	86.72	blacksmith	25.63	philosopher	83.81	barman	25.04	governor	68.77	jeweller	28.12
commentator	72.83	hairdresser	31.08	physicist	84.61	barista	25.04	chairperson	71.29	goldsmith	28.12
ecologist	80.46	roofer	22.16	botanist	80.46	courier	30.34	commissioner	78.76	shoemaker	18.07
director - general	71.29	barmaid	25.04	historian	83.81	waiter	25.04	senator	68.77	potter	24.43
professor	85.41	locksmith	33.16	ornithologist	80.46	janitor	21.82	attorney	86.72	beekeeper	28.04
historian	83.81	carpenter	26.62	geographer	83.09	clerk	43.33	treasurer	73.38	optician	59.85
sociologist	83.09	barman	25.04	ecologist	80.46	stewardess	46.76	lawyer	86.72	roofer	22.16
biographer	72.83	bricklayer	22.57	sociologist	83.09	babysitter	24.98	chancellor	70.34	tanner	28.08
editor	72.83	waiter	25.04	writer	72.83	barmaid	25.04	dean	65.01	tattooist	50.15
governor	68.77	waitress	25.04	geologist	86.81	waitress	25.04	ambassador	78.76	weaver	28.95
geographer	83.09	plumber	29.16	poet	72.83	cleaner	16.38	councillor	68.77	welder	28.52
marshal	60.92	plasterer	18.02	novelist	72.83	receptionist	39.02	professor	85.41	plasterer	18.02

Tábla 7. A 15 legmagasabb és legalacsonyabb faktorsúlyú foglalkozás listája (CC)

Ahogy a magas keresztkorrelációk előre vetítették az első három faktor interpretációja nagyon hasonló, mint amit a CC vektortérben már bemutattunk. Az első faktor a hatalom és tudás orientált foglalkozásokat emeli ki, a második faktoron pedig a tudományos foglalkozások kapnak magas értéket. A harmadik dimenzió mögött pedig itt is az intézményi hatalom és a szervezési képességek állnak. Az első faktor korrelációja az ISEI-vel 0.71 volt és a foglalkozási párok 74%-ban ugyanúgy rendeződtek sorrendbe az első faktoron, mint az ISEI skálán. Ebben a vektortérben is igaz volt, hogy elsősorban a különböző egészségügyi foglalkozások szerepeltek rosszabbul az első dimenzión az ISEI rangsorhoz képest.

A Wikinews vektortérben is teszteltük a 4- és 5 faktoros modelleket. A negyedik dimenzió ebben a vektortérben is a gender köré szerveződött (pincérnő, gondozónő, recepciós) az ötödik dimenzió pedig ismét mezőspecifikus lett, de nem az egészségügyi foglalkozások szervezték a dimenziót, hanem a médiához és kultúrához köthető foglalkozások kaptak itt magas faktor súlyt.

3.3.3 Wikinews – subwords

A harmadik vizsgált vektortér modell szintén a Wikinews korpuszra épült, de azzal a lényeges különbséggel, hogy a Fasttext beágyazó algoritmus a karakter n-grammokat is felhasználta a vektortér képzésekor. Ezzel a megközelítéssel a hasonló karakterfelépítésű szavak egymáshoz valamivel közelebb kerülnek a vektortérben. A korábbiakban már bemutatott három faktoros megoldást ebben a vektortérben is kiszámoltuk. A kapott eredmények egybevágtak a korábbiakkal. Az első faktor esetében a tudás és az intézményi hatalom rajzolódott ki, a második faktoron a végzettség és a tudás, a harmadik faktoron pedig a szervezési erő, szervezési hatalom dimenzióban összekötve a politikai hatalommal.

Az első faktor korrelációja az ISEI skálával 0.78 volt; a foglalkozási párok 77%-a volt ugyanolyan sorrendben a két skálán. A szemantikus skálán ebben a modellben is hasonló foglalkozások lettek alulbecsülve, mint a korábbi két modellben. Főleg az orvosi foglalkozásoknál kaptunk alacsonyabb besorolást (doktor, sebész, gyógyszerész, fogorvos), de gazdasági foglalkozások is lejjebb sorolódtak, mint a bankár vagy a könyvelő.

Ebben a vektortérben is teszteltük a 4 és 5 faktoros megoldást. A negyedik dimenzió erős gender karakterisztikával rendelkezett ebben a modellben is (fodrász, babysitter, pincérnő). Az ötödik dimenzión, pedig hasonlóan az alap Wikinews vektortérhez, médiához és kultúrához köthető foglalkozások kaptak magas faktorsúlyt (novella író, zenész, költő, zeneszerző, író).

3.4 Robusztusság

Az eredmények robusztusságát jól mutatja, hogy a különböző vektorterekben kialakított foglalkozási faktorsúlyok között nagyon magas a korreláció. A CC és a Wikinews modell első faktorai között 0.97, a második faktorok között 0.93, a harmadik faktorok között pedig 0.82 volt a keresztkorreláció.

A foglalkozási pozíciók stabilitását úgy is megvizsgáltuk, hogy összevetettük a különböző vektorterekben az egyes foglalkozások pozícióját. Ehhez a korábban már leírt módon egymásba forgattuk a CC és Wikinews vektortereket (Procrustes rotálással). Ezzel a megoldással lehetőségünk nyílt arra, hogy megvizsgáljuk mennyiben hasonló ugyanannak a foglalkozásnak a pozíciója a két vektortérben. Ehhez foglalkozásonként kiszámoltuk a koszinusz közelségeket. Az átlagos koszinusz közelség 0.79 volt. Nincs arra standard érték, hogy mi számít nagyon magas koszinusz

közelségnek, de a tapasztalatok azt mutatják, hogy csak egymáshoz nagyon közeli szavaknál tudunk 0.7 feletti koszinusz közelséget mérni. A CC korpuszban például két kutyafajta a Labrador és a Beagle közötti koszinusz közelség 0.7. Tehát a 0.79-es érték erős stabilitást jelez a foglalkozási pozíciókban a korpuszok között. De természetesen vannak olyan foglalkozások, ahol ennél valamivel alacsonyabb értéket mértünk, de 0.65 alatti értéket nem kaptunk. Ilyen foglalkozások voltak például a masszőr, a mosogató, az állatgondozó, a szerkesztő, a fogorvos vagy a lakatos.

Kiszámoltuk a stabilitási mérőszám korrelációját az egyes foglalkozások korpusz gyakoriságával (Wikinews). A kapott 0.59-es együttjárás megerősíti a szakirodalomban más máshol is olvasható összefüggést (lásd elsősorban: Hamilton – Leskovec – Jurafsky 2016), hogy a gyakoribb szavak stabilabb pozíciót vesznek fel a vektorterekben. A stabilitás szintén korrelált az ISEI pontszámmal (r=0.36, p=0.00). Ez az összefüggés bár gyengül, de akkor is fennáll, ha kontrolállunk a foglalkozások gyakoriságával (r=0.19, p=000). Tehát a magasabb presztízsű foglalkozások pozíciója stabilabb a korpuszok között.

3.5 Összefoglalás

A tanulmányunk elején két kérdést fogalmaztunk meg. Egyrészről azt vizsgáltuk meg, hogy a szemantikus térből kialakított foglalkozási skála mennyiben fed át a klasszikus hierarchia skálákkal. Másrészről arra fókuszáltunk, hogy találunk-e olyan szervező dimenziókat a foglalkozási térben, amik korábban nem jelentek meg hangsúlyosan a szociológia elméletekben. Az eredményeink alapján mind a két kérdésre pozitív; megerősítő választ kaptunk.

A kapott szemantikus OSPS skála erősen egyebvágott azokkal a klasszikus foglalkozási rangsorokkal, amiket jelenleg használnak a társadalmi struktúra vizsgálatára. Tanulmányunkban csak az ISEI skálával kapcsolatos eredményeket emeltük ki hely hiányában, de a SIOPS esetében is nagyon hasonló eredményeket kaptunk.

Az eredményeink sok szempontból egybevágnak és megerősítik a társadalmi rétegződés kutatás legfontosabb alapvetéseit, de felmerült egy nem várt dimenzió, ami a szervezési erővel, szervezési hatalommal van összefüggésben. A hatalom szerepe a foglalkozási hierarchiában nem új elem (lásd pl: Johnson 2016), de a mi eredményeink ennél többet mondanak, mert a tudás és a szervezeti erő összekapcsolódának a fontosságát hangsúlyozzák. A két elem közül legalább egyik, mindegyik dimenzióban megjelent az összes, három faktoros modellben. A tudás és végzettség valamilyen formában szinte az összes (neo-) weberiánus rétegződés és osztálymodellben szerepet kap. A szervezeti pozíció és a szervezési erő viszont nem általános eleme ezeknek a modelleknek, bár van társadalomtudományi előzménye. Freidson (1984) két csoportra bontotta az elitet klasszikus munkájában: tudás és adminisztratív elitre. Waring (2014) tovább vitte Friedson modelljét és kiegészítette két további elit csoporttal: a kormányzatival és a vállalatival. A harmadik faktorunk leginkább a tágan vett kormányzati elitnek felel meg, de adminisztratív jegyeket is mutat.

A megközelítés szempontjából fontos eredmény, hogy a kapott eredmények nagyon stabilnak bizonyultak. A különböző vektorterekben kibontott faktorok között nagyon magas volt a keresztkorreláció. A vektorterek egymásba forgatása szintén azt mutatta,

hogy a foglalkozok pozíciója nagyon hasonló a különböző vektorterekben. Ez egybevág azokkal a kutatási eredményekkel, amit azt mutatják, hogy általánosabb kutatási kérdések esetén mind a korpusznak, mind a választott beágyazási módszernek csak kis hatása van az eredményekre (Joseph és Morgan 2020). Az viszont érdekes kérdés, hogy a korpuszok milyen időszakot fednek le, mert a foglalkozások időbeli pozíciója változik (Garg et al 2018).

Tanulmányunk vállaltan exploratív munka. Az eredményeink jól alátámasztják azt, hogy egy ilyen jellegű megközelítés és egy társadalomtudományi szempontból új módszer hogyan vezethet minket új minták felfedezéséhez, ami később új hipotézisek felállítását teszi lehetővé (Nelson 2020). Véleményünk szerint a foglalkozások strukturálódásának megértéséhez közelebb juthatunk, ha a digitális térben keletkező adatokat is bevesszük a vizsgálati eszköztárba.

4 Esettanulmány 2 – Kulcsfogalmak jelentésváltozása a Kádár-korszakban

Eredeti tanulmány[22]

Szabó, M. K., Ring, O., Nagy, B., Kiss, L., Koltai, J., Berend, G., … Kmetty, Z. (2020). Exploring the dynamic changes of key concepts of the Hungarian socialist era with natural language processing methods. Historical Methods, Online first, 1–13. http://doi.org/10.1080/01615440.2020.1823289

4.1 Bevezetés

A Kádár-korszak egy sokat vizsgált időszaka a magyar társadalomtörténetnek. A legtöbb elemzés kvalitatív megközelítést használ a korszak elemzéséhez; a korszakban keletkezett szövegek alapján próbálják azonosítani a legfontosabb narratívákat (lásd pl: Gyáni 2016, Jakab 2012, Pap 2015 Pap 2016, Kovai 2016). A kvantitatív szövegelemzés módszere viszont egyáltalán nem jelenik meg az időszak kapcsán. A nemzetközi tapasztalatok azonban azt mutatják, hogy a szövegbányászati megoldásokkal megtámogatott elemzések teljesen új összefüggések feltárását is segíthetik (Miller 2013, Christianini et al 2018). NLP eszközök használatával felrajzolható akár egy hosszabb periódus diskurzus-dinamikája és egyes kulcsfogalmak szemantikai pozíciójának változása.

A legnagyobb kihívás egy ilyen jellegű elemzés esetében az, hogy nem áll rendelkezésre nagy mennyiségű digitális szöveges adat, amiből kiinduló korpuszt lehet készíteni. Az adott korszak dokumentumai (pl: újságcikkek, párthatározatok) sok esetben nincsenek digitalizálva és ha még digitalizálva is vannak, akkor is kép és nem

[22] Az eredeti tanulmány történeti kontextusba ágyazza a vizsgált jelenségek összekapcsolódásának változását. A habilitációs dolgozatban a tanulmánynak csak azon (általam írt) módszertani részét emelem ki, ami a speciális magyar történeti korpusz elemzésének gyakorlati aspektusaira fókuszál.

szöveg formátumban. Munkánk első lépésében ezért egy megfelelő korpuszt kellett készíteni, ami lehetővé teszi a kvantitatív alapú történeti elemzéseket. Tanulmányunkhoz az 1956 és 1989 között megjelenő Pártélet című újságból készítettünk egy korpuszt és ezt elemeztük tovább NLP eszközökkel. A fő célja a munkánknak az volt, hogy megvizsgáljuk, hogyan változott a Kádár-rendszer két kulcsfogalmának a mezőgazdaságnak és az iparnak a szemantikai pozíciója az 50-es évek végétől a 80-as évek végéig. Habilitációs dolgozatomban azt mutatom be részletesebben, hogyan alkalmaztuk a szóbeágyazási módszert a kutatási fókuszban lévő jelenségek vizsgálatára.

4.2 A korpusz

A Pártélet című újság a Magyar Szocialista Munkáspárt (MSZMP) központi pártlapja volt, így jó indikátora annak, hogy mi volt egy időszakban a párt főbb narratívája egyes kérdésekben. 1956 és 1989 között jelent meg havonta, az utolsó szám 1989 áprilisában jött ki a nyomdából. A lap célja leginkább az aktuális politikai ideológia propagálása volt. A lapban pártszervezéssel, gazdasági normatívákkal és ideológia alapvetésekkel kapcsolatos tartalmak jelentek meg elsősorban. A lapot nem a „köznépnek” szánták, hanem a pártfunkcionáriusok kapták meg. A lapot több mint 50 000 példányban nyomtatták ki havonta, ami azt jelzi, hogy nem csak a párt csúcsvezetői kaptak belőle, hanem a párthierarchia alsóbb szintjei is hozzájutottak.

A munka első lépése a korpusz előállítása volt. A scannelt pdf formátumú dokumentumokat[23] egy több lépéses folyamat során át kellett alakítani szöveges

[23] A Pártélet az Arcanum Digitheca (https://adtplus.arcanum.hu/en/collection/Partelet/) oldaláról érhető el. Az Arcanum sok ezer lap digitalizált változatához biztosít felhasználóinak hozzáférést. A korpusz készítés során az oldalról letöltött pdf dokumentumokból indultunk ki.

tartalmakká. Első lépésben a szkennelt képek minőségét kellett feljavítani, ezt követte az optikai karakter felismerés (Optical Caracter Recognization – OCR). A kapott szöveges dokumentumokból kivettük az oldalszámokat, a felesleges space-eket, javítottuk az elválasztásokat. Ezt követte a szöveges tartalom nyelvi előfeldolgozása. Ehhez a JAVA alapú magyarlanc-ot[24] használtuk (Zsibrita 2013). A szövegeket első lépésben mondatokra, majd szavakra bontottuk és a szavakat pedig lemmatizáltuk. Az elemzés előtt még kivettük a szövegből a pontokat és alkalmaztunk egy stopszó listát (pl: névelők kiszűrése). A létrejött korpuszt megvizsgáltuk egy magyar helyesírás ellenőrzővel a Hunspell-el. A korpuszban lévő egyedi szavak 14 százalékát nem ismerte fel a Hunspell. Az összes olyan fel nem ismert szót, ami legalább 10-szer előfordult manuálisan megvizsgáltunk. A problémás szavaknál három lehetőség volt. Vagy olyan szóval vagy névvel találkozott a hunspell ami létezik, de nem ismerte, mert csak az adott korra jellemző. Ez a problémás szavak 10%-ra volt igaz. Ezekben az esetekben nem kellett javítani. Ha rosszul volt írva a szó (OCR hiba miatt), akkor kijavítottuk a helyes szóra (pl: szocilista - szocialista). De voltak olyan esetek is, ahol nem volt egyértelműen megállapítható, hogy mi az eredeti szó, ezeket az eseteket nem tudtuk javítani. A javítások után az ismeretlen szavak száma a teljes korpuszban 10 százalék alá esett. Bár ez továbbra is viszonylag magas szám, a kutatásban alkalmazott szóbeágyazási módszer robusztusságának köszönhetően ez nem okozhatott jelentős torzítást a későbbi elemzésekben.

A Pártéletet 1956-ban adták ki először és 1988-ban jelent meg az utolsó szám. Az elemzéseinket éves alapon végeztük, amihez az volt az előfeltétel, hogy minden évben legyen egy kellően nagy korpuszunk. Ahogy a habilitációs dolgozat első részében részletesen is kifejtettem, a szóbeágyazási módszerek stabilitása erősen összefügg a

[24] http://www.inf.u-szeged.hu/rgai/magyarlanc

korpusz mérettel. Az előzetes elemzések után az 1956-58 közötti számok, valamint az 1989-es csonka év nem került be az elemzési korpuszba. Az 1959 és 1988 közötti időszakban a szavak száma nem mutatott nagy fluktuációt éves bontásban. A végső korpuszban 9 432 200 szó szerepelt, amiből 609 905 volt egyedi szó.

4.3 Módszertan

Tanulmányunkban fő célunk az volt, hogy megvizsgáljuk hogyan változott két kulcskoncepció – az ipar és a mezőgazdaság – szemantikai környezete a Kádár-rendszerben. Az elemzéshez szóbeágyazási modelleket használtunk. A módszer ilyen jellegű társadalomtörténeti használata bár nem egyedülálló (Hamilton Leskovec and Jurafsky 2016a, 2016b; Garg et al. 2018), de nem is túl gyakori. A modell illesztéséhez a GloVe algoritmust használtuk (Pennington et al 2014). Két ok is szólt a GloVe mellett. A GloVe a többi módszerhez képest stabilabb és robusztusabb eredményeket ad kicsi korpusz esetében (Spirling – Rodriguez 2019). A mi esetünkben ez különösen igaz, főleg az éves bontások esetében. A másik okunk praktikusabb volt. A GloVe R implementációja jól van dokumentálva és stabilan fut. Mivel a későbbi elemzéseink R-ben készültek nem akartunk a lépések között ugrálni az R és a Python között.

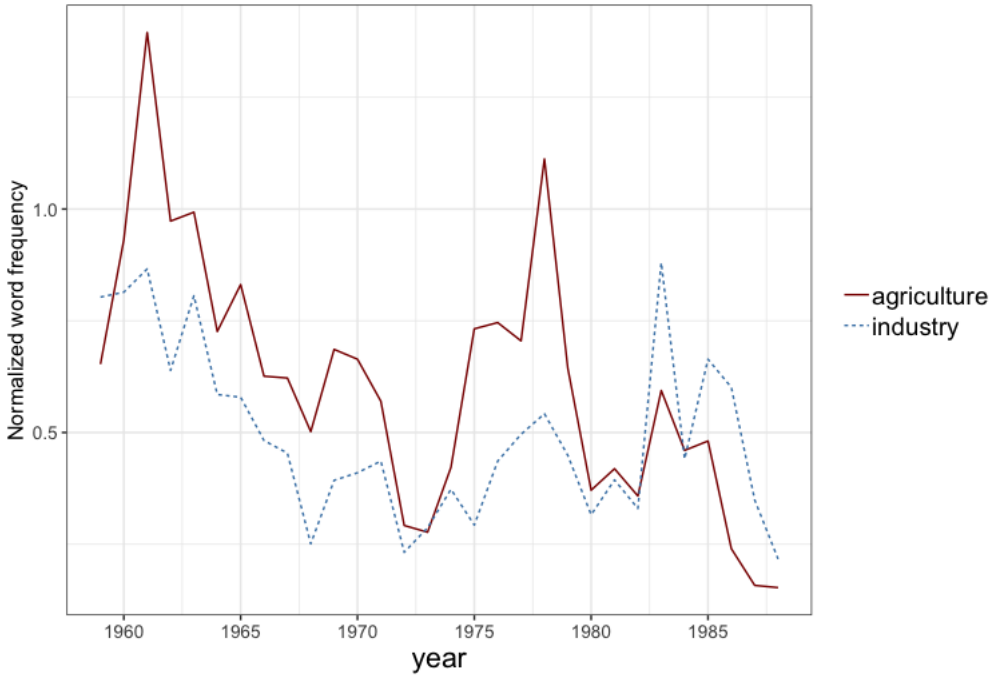
A GloVe-hoz szükséges TCM mátrix kiszámolása előtt kivettük azokat a szavakat, amik 6-nál kevesebbszer fordultak elő a teljes időszakban. Ez a lépés nem csak a futási időt csökkentette jelentősen, hanem a hibás szavak nagy részétől is meg tudtunk válni. A konkrét beágyazási modell paramétereit érdemes az adott korpuszra finom hangolni. Társadalomtudományi alkalmazásnál, főleg egy speciális időszakot lefedő kutatásban nincs olyan külső validitási szempont (pl: analógia keresés) ami automatikusan alkalmazható a modellekre. Ezért a paraméterek teszteléséhez egy

belső stabilitási mutatót vettünk figyelembe. Hatféle paraméter beállítást teszteltünk. Minden évben az összes paraméter beállításra 10-szer lefuttattuk a beágyazást. Az egyik paraméter a dimenzió méret volt (200, 250, 300) a másik pedig az ablak méret, amin belül két szót egy kontextusba helyeztünk (4, 7, 10). Azt vizsgáltuk, hogy ugyanabban az évben ugyanazzal a paraméterbeállítással mennyire stabil a 10-10 futás eredménye a kulcsszavak (mezőgazdaság, ipar) esetében. A stabilitás teszthez az adott évhez és paraméterbeállításhoz tartozó 10-10 vektorteret egymásra forgattuk Procusteres rotálással és kiszámoltuk a kulcsfogalmak koszinusz közelségét a vektorterek között. Ez a megközelítés megegyezik azzal, amit a foglalkozások vektorteres vizsgálata kapcsán már bemutattam. A legstabilabb megoldást mind az ipar mind a mezőgazdaság szaval esetében akkor kaptuk, amikor 200 dimenziós volt a vektortér és 10 szavas az ablak. A mezőgazdaság esetében a fenti paraméter beállításokat használva az átlagos koszinusz közelség 0.67 volt, az iparnál pedig 0.64. A stabilitás a korpuszunkban elsősorban az ablak méretétől függött és kevésbé a dimenziószámtól.

A paraméterek beállítása után minden évre lefuttattuk a beágyazást 100-szor. A 100 futásnak az átlagát használjuk a tanulmányban az egyes szavak szemantikai közelségének megállapítására. A 100 futás azt is lehetővé teszi, hogy az eredményekhez konfidencia intervallumot rendeljünk (Antoniak – Mimno 2018). Ahhoz, hogy a szavak időbeli pozíciójának a változását is megvizsgálhassuk minden évnek a vektorterét hozzáigazítottuk az előző év vektorteréhez a korábban már bemutatott procrusteres rotálással. A mezőgazdaság és ipar szemantikai környezetének vizsgálatához exploratív (Hengechen, Ros and Marjanen 2019) és konfirmatív megközelítést is használtunk.

4.4 Eredmények[25]

Tanulmányunkban a korszak két kulcsfogalmának, a mezőgazdaságnak és az iparnak a szemantikai pozícióját vizsgáltuk a Kádár-korszakban a Pártélet újságon keresztül. A mezőgazdaság szó valamivel gyakrabban jelent meg, mint az ipar (0.6 vs 0.5 100 szóból) a rezsim utolsó éveit leszámítva.



Ábra 12. Mezőgazdaság és ipar szavak gyakorisága évente 100 szóra vetítve

Érdekes látni a nagy kiugrást a 70-es évek végén a mezőgazdaság szó használatában. Az olajválságok miatt az ország exportlehetőségei jelentősen csökkentek, ami magával hozta az életszínvonal esését. Ezért kénytelen volt a Párt az 1968-as reform néhány mezőgazdasági elemét visszahozni, hogy növeljék a fogyasztást.

[25] A habilitációs dolgozatomban nem mutatom be az összes tanulmányban leírt eredményünket, alapvetően a módszer lehetőségeire fókuszálok.

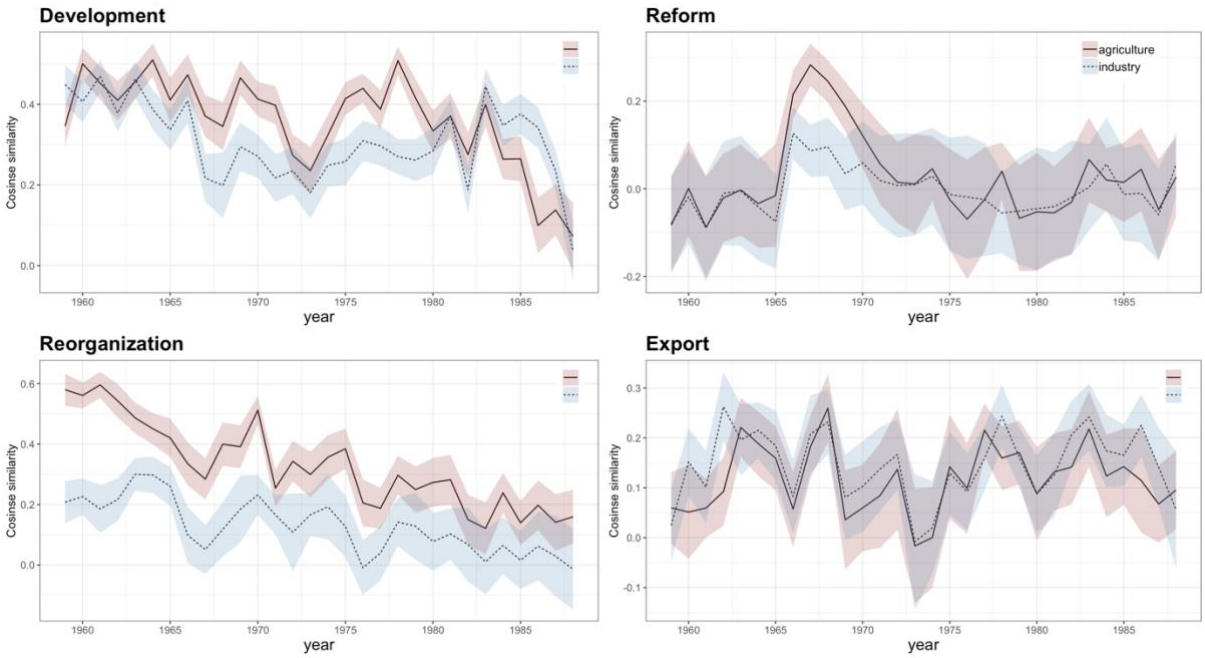
Az évek között egymásra forgatott vektorterek lehetőséget adnak arra, hogy mind a mezőgazdaságnak, mind az iparnak a szemantikai stabilitását vizsgálhassuk. A mezőgazdaság esetén az átlagos koszinusz közelség az évek között 0.44 volt, az ipar esetében 0.42. A stabilitási mutató csökkent az idő előre haladtával, de ez nem jelenti azt, hogy valóban csökkent a szemantikai stabilitása. A stabilitás összefügg a szógyakorisággal is. A szógyakorisággal kontrolált stabilitási idősor stabilnak bizonyult az időben.

Az exploratív elemzés keretében megvizsgáltuk, hogy mi a 20-20 legközelebbi szó az ipar és a mezőgazdaság esetében évente és a teljes időszakra. A mezőgazdasághoz az ipar a termelés, a fejlődés a fejlesztés és a szocialista szavak voltak legközelebb a teljes időszakban. Érdekes volt azoknak a szavaknak a listája, amelyek csak időnként bukkantak fel. A terv és a cél szavak inkább a 60-as években voltak jellemzőek a mezőgazdaságra, a 70-es években az eredmény a 80-as években a termék szavak kerültek közel a központi fogalmunkhoz.
Az ipar esetében a mezőgazdaság, a termelés, a fejlődés és fejlesztés szavak mellett megjelent a népgazdaság is mint közeli fogalom. A 60-as években a szocialista és a tervezés szó volt közel az iparhoz a 70-es években a munkavállaló a 80-as években pedig a hazai. Ezek a példák jól mutatják, hogy bár voltak állandó elemei a mezőgazdaság és ipar szavak kontextusának, minden időszakban voltak jellemző új fogalmak, amik jól mutatják a rendszer folyamatos átalakulását.

A létrejövő korpusz és az azt feldolgozó szóbeágyazási modell arra is lehetőséget ad, hogy általunk kiválasztott fogalmak viszonyát vizsgáljuk az elemzés központi fogalmaihoz képest (Ábra 13). Az exploratív elemzésből láttuk, hogy a **fejlődés** mind

az iparhoz, mind a mezőgazdasághoz közel helyezkedtek el a teljes időszakban. A részletes dinamikai elemzés jól mutatja, hogy elősorban a 60-as évek elején volt közel az iparhoz és a mezőgazdasághoz a fejlődés aztán főleg az ipar esetében nagyon lecsökkent. Az olajválságok után nagyon fontos volt, hogy a mezőgazdaság a belföldi keresletet el tudja látni, ezért is látunk egy második felfutást a fejlődés és mezőgazdaság kapcsolatába a 70-es évek második felétől. Főleg a rendszer elején nagy lendülettel kezdték az egyes szektorok átalakítását. Az **átszervezés** szó fontos korai szerepe (elsősorban a mezőgazdaság esetében) jól mutatja a rendszer kezdeti lendültét ezen a téren és kifulladását a 80-as évekre. A **reform** szó időbeli felfutása mutatta talán a legkevesebb meglepetést. A 68-as gazdasági reform mind a mezőgazdaság mind az ipar esetében nyomott hagy a szavak közelségén, de mindkét szó esetében a 70-es évek elejére visszatér a közelség a korábbi szintre és onnan nem is nagyon mozdul el. Az **export** esetében pedig kevésbé beszélhetünk trendekről, inkább az rajzolódik ki az adatokból, hogy a hazai és nemzetközi helyzet gyorsan változott és időlegesen hol nőtt, hol csökkent az export szerepe a mezőgazdaság és ipar kapcsán.

A fejezet alapját jelentő cikkünkben az eredményeket bővebb történelmi kontextusba helyeztük és további szavakat is vizsgáltunk. Jelen dolgozatban csak a módszer adta lehetőségeket szerettem volna kiemelni.



Ábra 13. Mezőgazdaság és ipar szavak koszinusz közelsége kiválasztott kulcsfogalmakkal éves bontásban. A 95%-os konfidencia intervallumot az évente lefuttatott 100 beágyazásból számoltuk ki.

4.5 Összefoglalás

A röviden bemutatott tanulmányunkban azt vizsgáltuk, hogyan változott a Kádár-korszak két kulcsfogalmának a mezőgazdaságnak és az iparnak a szemantikai kontextusa 1959 és 1988 között. Az elemzéshez a Pártélet című lapot dolgoztuk fel. A munka jól mutatja milyen összetett feladat lehet szövegbányászat szempontjából egy történeti projekt. A vektortér modellek már csak az elemzés végső részét jelentették, nagyobb munka volt eljutni a bemeneti korpusz előállításáig. A digitalizált, de szövegként nem tárolt adatok szöveggé alakítása volt az első lépés, amit követett a korpusz tisztítása, lemmatizációja és minőségellenőrzése. Két olyan eszközt is használtunk a munkában a Magyarláncot és a Hunspellt, amelyek nagy segítséget nyújtanak magyar nyelvű szövegek NLP feldolgozásakor. A szövegek előkészítése nem univerzális, érdemes átgondolni, hogy milyen elemzéseket akarunk egy szövegen elvégezni és ehhez igazítani az előfeldolgozást. A szóbeágyazás egy kellően robusztus módszer ahhoz, hogy „koszosabb” korpuszon is jól működjön. Az a tény,

hogy a korpuszunk 10%-a hibás szavakból állt nem okozott komoly problémát az elemzésünkben. A saját vektortér modell illesztés előnyét is jól mutatja az elemzésünk. Lehetőségünk volt a korpuszhoz illeszkedő legjobb paraméterekkel futtatni a modelljeinket és arra is volt lehetőségünk, hogy bootstrap megközelítést alkalmazva konfidencia intervallumot „rajzoljunk” az átlagaink köré.

A számítógépes nyelvészet a szociológia és a történelemtudomány metszetében készült elemzés arra is jó példa milyen gyümölcsöző lehet egyes tudományterületek együttműködése és milyen izgalmas kutatási kérdésekre adhat választ egy ilyen multidiszciplináris projekt.

5 Esettanulmány 3 – Szia: leszel a feleségem? Hogyan kommunikál a gamer közösség a női játékosokkal a Twitch-en?

Eredeti tanulmány

Tóbiás, D., Kmetty, Z. (2020). Gendered discourse on Twitch (kézirat)

5.1 Bevezetés

A nőkkel szembeni diszkrimináció nyelvi megjelenésével már foglalkoztam a dolgozatom korábbi részében. A Google fordító példája jól mutatja, hogy nem szándékoltan is, hogyan tudnak nyelvi modellek diszkriminatív módon működni. Ezeknek a nyelvi modelleknek a „hibája" alapvetően abból következik, hogy a kiinduló korpuszokban a nők rosszabb pozícióban vannak ábrázolva, mint a férfiak és a nyelvi modellek erre a klasszifikációra tanulnak rá. De a nőkkel szembeni negatív megkülönböztetésnek van egy másik eleme is, amit direktben nem mutatnak ki a vektortér modellek – ez pedig a női szereplésekhez köthető szöveges reakciók. Mások a kommentek egy hír alatt, ha férfi vagy nő írja a cikket? Mások a megjegyzések egy online videóhoz, ha férfi vagy nő szerepel rajta? Megjelennek-e például ezekben a tartalmakban olyan kommentek, amik nem a témával foglalkoznak, hanem a nőket tárgyiasítják? Utolsó esettanulmányunk ezt a kérdést járja körül online gamer-ek videóihoz érkező kommentek vizsgálatával. Pontosabban specifikálva azt vizsgáltuk Tóbiás Dániellel közös munkánkban, hogy eltérő módon kommentelnek-e a Twitch csatornán férfi és női játékosok videói alatt. A habilitációs dolgozatban azt emelem ki, hogy egy ilyen kérdés vizsgálatára hogyan használhatók a vektortér modellek. A

vektortér modellek előtt röviden ismertetem a kutatás menetét és a korpusz egyedi jegyeit, mert ezek nélkül nem értelmezhető a későbbi elemzés sem.

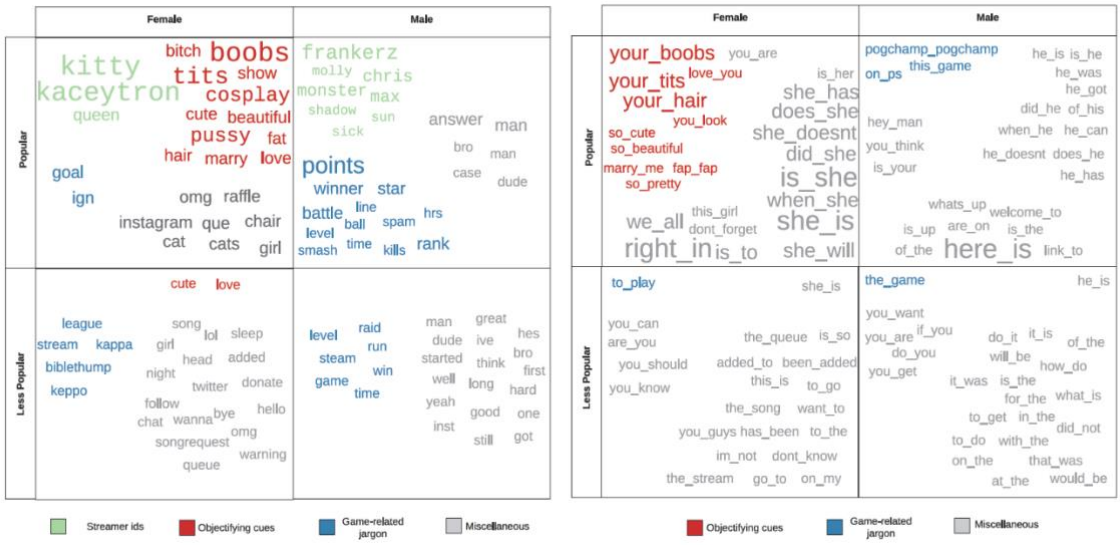
5.2 Twitch és online játékok

A Twitch elsősorban egy online közösségi videómegosztó platform. Ha párhuzamosságot keresünk akkor a Youtube juthat eszünkbe és valóban alapfunkcióit tekintve nagy a hasonlóság. A Twitch-et már kezdetektől fogva arra használták, hogy online stream-et közvetítsenek rajta, még a Youtube-nak az utóbbi funkciója csak később vált elterjedtté és a legtöbben továbbra is nem élő tartalmakat néznek a Youtube-on. A Twitch már kezdetektől fogva a gamer világra fókuszált. Bár elérhetők rajta általános csatornák is, továbbra is a játékokkal kapcsolatos tartalmak a leggyakoribbak. A Twitchnek átlagosan több mint 2 millió nézője van[26] és ez a szám dinamikusan növekszik (részben a koronavírus hatására is). Bár nem az összes játékos a Twitch-en közvetít, de a közvetített játéktartalmak 2/3-a Twitch-en zajlik. Tehát a Twitch elsősorban egy olyan platform, ahol online játékokat közvetítenek és ezeket a közvetítéseket pedig széles közönség nézi. A kutatásunk szempontjából azonban az a legfontosabb jellemzője a platformnak, hogy a közönség nemcsak passzívan tudja nézni az adást, hanem kommunikálhat is a játékosokkal chat formában. Ezzel el is érkeztünk a szűkebb témánkhoz és a kutatási kérdésünkhöz. Vajon ha női játékosokat látnak a kommentelők, akkor máshogy kommentelnek összevetve azzal, mintha férfi játékosok videóit nézik?

Munkánk nem az első ebben a sorban. Nakandala et al. (2016) 200 férfi és 200 női twitch csatornán keletkezett tartalmakat vetett össze egymással. A csatornák kiválasztásánál figyeltek arra, hogy hasonlóan népszerű férfi és női csatornák

[26] https://twitchtracker.com/statistics

legyenek kiválasztva. Utóbbira azért van szükség, mert a csatorna népszerűsége befolyásolja a szöveges tartalmakat, a népszerűbb csatornákon rövidebbek az üzenetek és kevésbé alakul ki valódi diskurzus (Nematzadeh et al 2016). Mivel a férfi csatornáknak átlagban több a követője, ezért a csatornák kiválasztásánál fontos szempont a népszerűségük szerinti rétegzés. A kiválasztott férfi-női csatornák különbségét több módszerrel is vizsgálták, többek között vektortér modellel is. Egy klasszifikációs modellt építettek arra, hogy férfi vagy női a játékos és ebben a modellben a szavak vektortér reprezentációit használták háttérváltozóként. A modell 87%-os pontossággal jósolta meg a streamelő nemét, ami jól mutatja a férfi és női csatornákhoz kötődő kommentek különbséget. Az eltérő nyelvezet azonban még nem probléma. A tartalmi elemzések azonban jól mutatták, hogy a női játékosok videói alatt megjelennek objektifikáló kifejezések.



Ábra 14. Férfi és női csatornákhoz kötődő jellemző szavak (unigram (bal) és bigram (jobb)) normaizált szógyakoriságok alapján. Forrás: Nakandala et al. (2016)

Mind a normalizált szógyakoriságokra épülő elemzés, mind a vektorteres klasszifikáció azt mutatta, hogy a női csatornákon sokkal inkább megjelentek objektifikáló kifejezések. Szexualizáló tartalmak, kéretlen közeledések, kinézetre utaló szavak

jellemzőbbek voltak a női játékosok esetében, a játékokhoz jobban köthető szavak a férfi játékosok esetében. Az is lényeges eredménye volt a kutatásunknak, hogy a népszerű csatornákon volt a különbség megfogható, a kevésbé népszerű csatornákon ez nem volt látható (Nakandala et al. 2016)

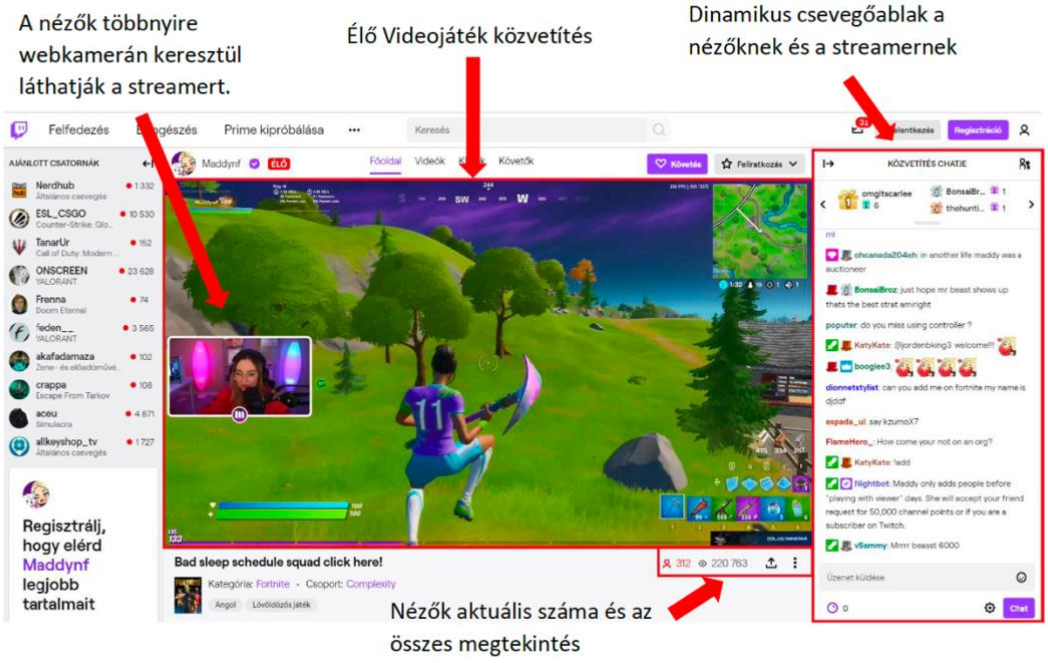
A fenti kutatásnak azonban volt legalább két gyenge pontja is. Ezek azzal voltak kapcsolatosak, hogy nem volt semmilyen tartalom szűrés, tehát bármelyik csatorna bekerülhetett. Ennek egyik következménye az volt, hogy nem vették figyelembe, hogy ki milyen játékot közvetít. Egy lövöldözős, egy stratégiai és egy sport játék alatt eltérőek lesznek a kommentek, mindegyik játéknak megvan a sajátos nyelvezete. Ha feltesszük azt, hogy a férfi és női játékosok nem ugyanolyan arányban játszanak az egyes játékokkal, akkor ez már önmagában eredményezhet egy jól klasszifikáló modellt. A másik talán sokkal fontosabb gond, hogy nem szűrték ki azokat a női játékosokat, akik videóikban szexualizálják magukat. Mivel a csatorna nézőnként fizet, ezért a női játékosok között megjelent egy olyan kisebbség, akik részben hiányos öltözékben játszottak. Ezt a viselkedést egyébként a Twitch jelenleg már szigorúan tiltja[27], de 2014-ben, amikor a jelzett kutatás adatai keletkeztek, még nem volt ennyire szigorú a platform.

5.3 Adatok és előfeldolgozás

A rövid elméleti bevezetőben is kitértünk arra, hogy a csatornák nyelve erősen függ a játéktól és a népszerűségtől. Ezért a kutatásba bekerülő csatornák esetében ezt a két szempontot vettük figyelembe. Két játékot választottunk ki, az egyik jelenleg legnépszerűbb lövöldözős játékot, a Fortnite-ot és egy klasszikusnak számító stratégia

[27] https://web.archive.org/web/20190330070237/https://www.twitch.tv/p/legal/community-guidelines/sexualcontent/

játékot a Starcraftot. A Fortnite esetében kevésbé kell ismerni a játékmenetet, alkalmi játéknézőknek is izgalmas lehet, míg a Starcraft esetében azt várjuk, hogy több lesz a „rendszeres” követő és kevesebb az alkalmi belenéző. Bár egy játékos több játékkal is játszhat, általában egy rövidebb időszakban ritka a játék váltás, professzionalizálódnak a játékosok. A játékosok kiválasztásánál a másik szempont a csatorna nézettsége volt. Mind a női mind a férfi játékosoknál egyaránt beválogattunk népszerű és kevésbé népszerű csatornákat. Az adatfelvétel 2020 január 8 és február 14 között zajlott. A Twitch API-jának segítségével ebben az időszakban folyamatosan követtük a kiválasztott profilokat és ha streameltek a játékosok az algoritmusunk élőben gyűjtötte a kommenteket. A megoldásunk olyan szempontból is előnyös, hogy ezzel megelőztük az esetleges tartalom moderációt.



Ábra 15. Példa egy élő Twitch adásra (Maddynf a Fortnite-tal játszik) – forrás: Tóbiás (2020)

Az 1 hónapos időtartam alatt sajnos három női játékos is felfüggesztette a fiókját. Így a végső mintánk 17 profil 1 hónapos követéséből állt össze. Az adatbázis végül

3 739 557 üzenetből állt az adattisztítás előtt. Az adattisztítás több lépcsőből állt. A Twitch csatornákon működnek chatbotok, amik automatikusan válaszolnak a nézők egyes kérdéseire. Ezek az üzenetek mindig felkiáltójellel kezdődnek. Mivel a chatbot válaszok semmilyen releváns információt nem hordoztak számunka, ezért ezeket eltávolítottuk az adatokból. A következő lépésben egységesítettük az emojikat. Erre azért volt szükség mert minden csatornánk lehetőség van saját emojik létrehozására. Tehát az egyszerű „kedvelés” emoji máshogy nézhet ki a különböző csatornákon, ami elemzési szempontból nem szerencsés. Az egységesítés során nem tudtunk teljeskörű megoldást használni, de a leggyakoribb emojikat minden csatorna esetében egységes formátumra hoztuk. A speciális tisztítási lépések után eltávolítottuk a speciális karaktereket a szövegekből, kivettük a stopszavakat, majd stemmeltük a tartalmat. A végső korpuszban 199 088 aktív néző 3 438 160 üzenete maradt meg. Az adatgyűjtés részletesebb leírása Tóbiás Dániel szakdolgozatában olvasható (Tóbiás 2020).

5.4 A vektortér modell

A vektortér modellek kiválóan alkalmasak arra, hogy megmutassák egy szöveges korpuszon milyen közel helyezkedik el két szó, vagy tágabb értelemben két koncepció. A kutatásunk elsősorban objektifikációra fókuszált, de az adatok lehetővé tették azt is, hogy több szempontból vizsgáljuk a női és férfi Twitch csatornák különbségeit. A korábbi szakirodalmi eredmények és az adatok exploratív elemzése után a következő elemzési kategóriákat határoztuk meg.

- Objektifikáció
- Vonzódás
- Női becézések (benevolent cat calling)

- Köszönés
- Játékkal kapcsolatos kifejezések
- Inzultálás

A Twitch-nek nagyon sajátos nyelve van, ami akár játékonként is változik. A közösség rengetek rövidítést használ, vagy átértelmez bizonyos szavakat. A GG például a „good game” azaz a jó játék rövidítése, de például a „cheese” se sajtot jelent, hanem egy jellemző játék technikát. Az egyedi szókészlet miatt nem volt reális opció, hogy előre definiált szótárat használjunk, ezért saját annotáció mellett döntöttünk. Legyűjtöttük az összes olyan szót, ami legalább 50-szer megjelent a korpuszban és egyenként bekódoltuk ezeket a szavakat a fenti kategória rendszer szerint. A kódolást külön végeztük és csak azokat a szavakat hagytuk bent végül az egyes kategóriánál, ahol mindketten egyetértettünk a besorolással. Az egyes kategóriához tartozó szavakat a 8-as táblázat tartalmazza.

Objektifikáció	anal, babe, baby, bikini, bitch, boob, booty, cum, cummy, cute, cuty nake, nippl, porn, pornstar, sex, sexy, sexyoffice, strip, virgin, whore
Vonzódás	boyfriend, amor, husband, red_heart
Női becézések (benevolent cat calling)	bonita, chickó, female, female_sign, girl, girly, girlygirl, gurl, honey, lady, lesbian, maid
Köszönés	adio, bye, byee, byeee, byeeee, byeeeee, cya, goodby, goodmorn, goodnight, greet, hello, helloo, hellooo, helloooo, hey, heyyy, heyyyi, heyyyyy, hi, hihi, hii, hiii, hola, howd, howdy, wassup, wave, wave_sg, waving_hand
Játékkal kapcsolatos kifejezések (FN)	action, arena, box, f, follow, gg, gooo, goooo, gooooo, grind, hide, hit, hp, item, jump, kill, loot, pew, pog, run, shoot, shot, shotgun, skin, snipe, sniper, stay, tryhard

Játékkal kapcsolatos kifejezések (SC)	archon, army, attack, campaign, chees, herro, ladder, rush
Inzultálás	babyrag, dead, ded, dumb, dummy, dump, git, l, lame, lll, llll, ose, loser, rip

Tábla 8. Vizsgált dimenziókat meghatározó szavak

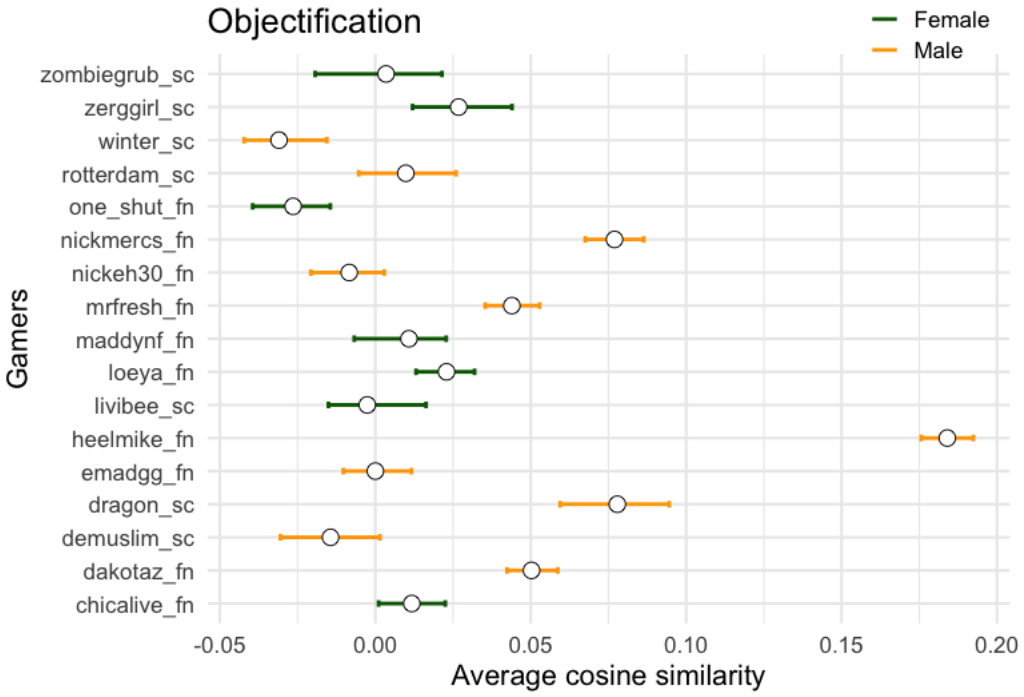
A szótár szavai megadják, hogy milyen szavakkal szeretnénk összekötni a játékosokat. De a beágyazási modellben alapban az az információ nem szerepel, hogy mely csatorna alatt volt az adott szöveg, tehát a csatorna és kommentek összekötése nem triviális feladat. A probléma megoldására egy trükkhöz folyamodtunk. Minden kommenthez hozzáfűztük a csatorna nevét. Ezzel gyakorlatilag hozzátapasztottuk a jellemző kommentekhez a csatornát. Mivel a kommentek általában nagyon rövidek - 1-2 szavasak - ezért a kiválasztott ablakunkban (10-es ablak) benne volt a teljes komment és a csatorna neve is.

A vektortér képzésre a GloVe algoritmust használtuk (Pennignton et al 2014). Csak azokat a szavakat vettük be a modellbe, amelyek legalább 50-szer szerepeltek a korpuszban. A vektortér nagyságát 200-ra állítottuk és 100 iterációt futtattunk.

Az egyes dimenziók és játékosok közelségét a következő módon számoltuk ki. Az adott dimenzió minden szavára kiszámoltuk a koszinusz közelségét az összes játékossal. Csak azokat a szavakat hagytuk bent, amelyek legalább 1 játékos esetében elérték a 0.1-es értéket. A bent maradt szavaknak pedig vettük az átlagát játékosonként. Mivel 100 iterációt futattunk, 100 átlag értékünk volt. Ennek a 100 iterációnak vettük az átlagát, illetve megbecsültük a felső és alsó konfidencia intervallumát. Ez az intervallum-becslés mutatja meg, hogy az adott dimenzió milyen közel van az adott játékoshoz. Az ábrákon a csatornák népszerűség alapján vannak sorba rendezve. Felül vannak a kevésbé népszerű csatornák, alul pedig a népszerűbb csatornák.

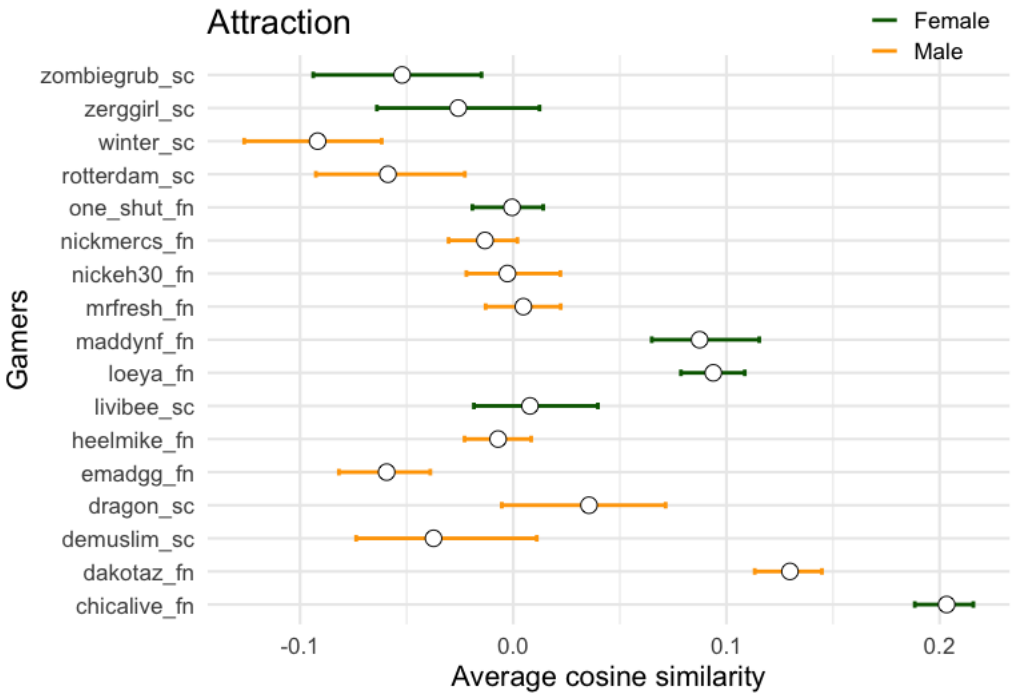
5.5 Eredmények

A tanulmányunk kiinduló kérdése ara vonatkozott, hogy lehet-e objektifkációt megfigyelni a Twitch-en a női játékosok esetében. A vektortér modellek alapján nem beszélhetünk objektifikációról, legalábbis szavak szintjén ez nem mérhető.



Ábra 16. Objektifikáció megjelenése Twitch csatornánként

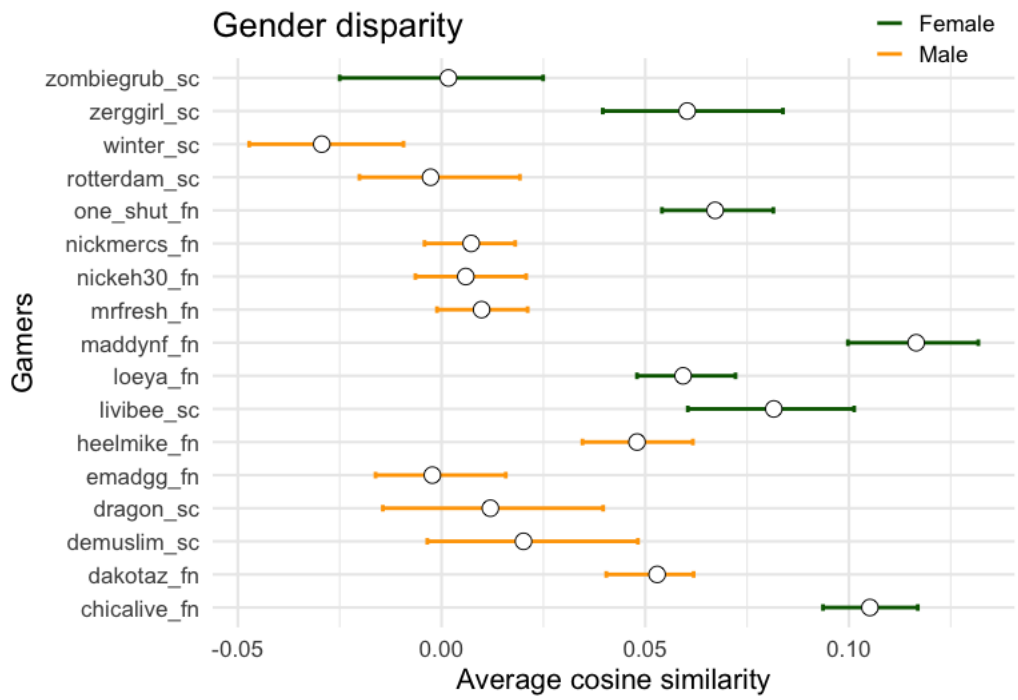
A legtöbb objektifikáló szó egy férfi gamer csatornáján (Heelmike) jelent meg. A csatornára nagyon jellemző a trágárság, az inzultálás is ezen a csatornán volt a leginkább jellemző (lásd melléklet ábra M1). A fenti eredmény azonben nem jelenti azt, hogy nem volt semmilyen különbség a férfi és női csatornák között. A vonzódás dimenziójában már látunk nemi eltéréseket.



Ábra 16. Vonzódás megjelenése Twitch csatornánként

Négy olyan csatorna volt, ahol a vonzódást mérő szavak átlagos közelsége 0.1 körüli vagy magasabb értéket vett fel. Ebből a négy csatornából három női streamer-hez volt köthető, főleg „chicalive” esetében mértünk magas értéket. Külön érdekesség, hogy a második helyre kerülő dakotaz bár férfi játékos, de szokott női karakterekkel is játszani, tehát a vonzódás itt akár a karakterére is vonatkozhat. Fontos kiemelni azt is, hogy mind a négy magasabb értéket produkáló csatorna játékosa Fortenite-tal játszott, a Starcraft játékosoknál nem figyeltünk meg különbséget a vonzódás értékei között.

A vonzódás mellett jól látható különbség volt a női becéző szavakban.



Ábra 16. Női becézések megjelenése Twitch csatornánként

A három legmagasabb koszinusz közelséget női csatornák produkálták. A két legmagasabb értékű női játékos Fortenite-tal játszott, a harmadik viszont a Starcrafttal, tehát itt nem volt olyan jelentős a játék hatása, mint a vonzódás esetében.

A korábban már idézett Nakandala et al (2016) tanulmány azt állapította meg, hogy a női csatornákon több a köszönés, a férfi csatornákon pedig több a játékkal kapcsolatos szó. A mi elemzésünk az előbbit csak részben erősítette meg, az utóbbit viszont inkább cáfolta. Bár a köszönés két női csatornán volt a legmagasabb, az eredmények azt mutatják, hogy a játéknak nagyobb a szerepe, egyértelműen a Fortnite esetén erősebb ez a minta. Itt feltételezhetjük, hogy kevesebb a tartalmi interakció a játék közben és ezért gyakoribbak az általános kommunikációs elemek (ábra M2). A Fortnite és a Starcraft esetében különböző játékhoz kapcsolódó szavak a jellemzőek, ezért itt külön szótárat építettünk játékonként. Az eredmények itt nagyon vegyesek voltak, nem volt

jellemző, hogy a férfi játékosok közelebb lettek volna a játékhoz kapcsolódó szavakhoz.

Összefoglaló

A Twitch mind tartalmi, mind nyelvi szempontból egy nagyon izgalmas kutatási terep. Az e-sport ágazat most bontakozik ki, főleg a fiatalabb generációk esetében óriási népszerűségnek örvendenek egyes játékok. A piac kitermeli a saját sztárjait, akiket akár milliók is követnek. Nagyon fontos folyamatosan vizsgálni azt, hogy egy ilyen növekvő népszerűségi piacon a nők milyen pozícióban vannak. A kutatásunk exploratív szakaszából egyértelműen látszott, hogy a nők jelentősen alul vannak reprezentálva a top játékosok között, nehéz volt megfelelő nézettségű női játékosokat találni. Ez már magában is mutatja, hogy jelenleg egy inkább férfiak által dominált terepről van szó. Ez még inkább fontossá teszi azt, hogy a női játékosok ne kapjanak a kommentelőktől becsmérlő, vagy objektifikáló megjegyzéseket – ne a nemük alapján ítéljék meg őket, hanem a játékuk alapján.

Az eredményeink azt mutatják, hogy nagyon erős objektifikáció nincsen a tartalmakban, de a szexizmus megjelenik, a női játékosok gyakrabban kapnak vonzódást jelző tartalmakat és különböző gyakran pejoratív női jelzők is megjelennek a komment folyamukban. Kutatásunk annyiban megerősíti a korábbi eredményeket, hogy elsősorban mi is a több nézőt vonzó csatornák esetében láttunk erősebb hatásokat. A kisebb csatornáknak valószínűleg állandóbb a közönsége, jobban moderálják magukat a hozzászólók. Az eredményeink azt is mutatják, hogy nagyon fontos figyelembe venni azt is, hogy milyen játékot közvetítenek a gamerek. A Fortnite-hoz kevesebb játék ismeret kell, ezzel szemben a Starcraft közvetítéseket akkor lehet igazán élvezni, ha ismerjük jól a játékmenetet. Ez már magában is nagyon különböző

nézőket tud bevonni. Vonzódást jelző szavakat a női Starcraft játékosoknál nem mértünk, ezzel szemben a női Fortnite játékosoknál jellemző elem volt. A továbbiakban érdemes lehet kiterjeszteni a kutatást más játékstílisokra is, hogy még komplexebben megérthessük a játékok a nyelvezet és a nemi diszkrimináció összefüggését.

A habilitációm középpontjában a vektortér modellek vannak, ezért az összefoglalás végén ezekre térek ki. A vektortér modellek jó módszernek bizonyultak arra, hogy egy nagyon speciális nyelvezetű és nagyon töredezett tartalmú korpuszon megvizsgálhassuk a Twitch és a nemi diszkrimináció összefüggését. Ebben a tanulmányban egy konfirmatív megközelítést használtunk, tehát előre definiáltuk az egyes fogalmakat alkotó szavakat és ezeknek a helyzetét vizsgáltuk meg csatornák mentén. A modellhez azonban egy kicsit módosítani kellett a bemenő adatokat, hozzá kellett „tapasztanunk” a szövegekhez a csatorna nevét. Ez a megközelítés más módon is használható. Hozzáadhatunk akár időcímkét vagy kulcsszavakat is egy szövegrészhez. Kifejezetten hasznos lehet ez az irány, ha kommenteket ágyazunk be és a kommenteket össze akarjuk kötni azzal a tartalommal, amire a kommentek érkeztek. Ilyen esetben például a cikk főbb kulcs szavaival egészíthetjük ki a kommentet. A legfontosabb üzenetet azonban elsősorban az, hogy ismerni és érteni kell a korpuszunkat ahhoz, hogy a megfelelő modellt ki tudjuk választani. Bár maga a módszer fekete dobozként működik, de a bemenő adatokra befolyással lehetünk, ezért fontos nagyon pontosan megérteni, hogy mi az a szöveg, amiből az elemzésünk kiindul.

A vektortér modellek nem nyújtanak univerzális megoldást minden NLP problémára, de a nemi diszkrimináció kimutatására kifejezetten alkalmasak. A gamer közösséggel kapcsolatos elemzésünk erre a felhasználási területre mutatott egy mintát. A módszer

maga azonban tetszőlegesen kiterjeszthető minden olyan esetre, ahol nők és férfiak jelennek meg online térben és a megjelenéseket szöveges reakciók kísérik. Ez egy fontos területe lehet a módszer alkalmazásának.

6 Nyelvi modellek és a társadalomtudományok – merre mutat a jövő?

A habilitációs dolgozatom záró részében egy nehezen megválaszolható kérdést próbálok körbejárni – a nyelvi modellek és a társadalomtudományok jövőbeli összekapcsolódásának lehetséges irányait. A kérdést nem lehet külön választani a számítógépes társadalomtudomány (CSS) fejlődésétől és szakmán belüli pozíciójától. Míg a 2000-es évek kulcsszava a hálózatkutatás volt a 2010-es években azonban áthelyeződött a hangsúly egy részben tágabb területre, amely az elsősorban digitális tartalmakat kvantitatív módszerekkel vizsgálja interdiszciplináris kutatócsapatokkal. A hálózatkutatás ilyen értelemben jó előfutár volt, hiszen azon a területen is találkoztak már a matematikusok a fizikusok és a társadalomtudósok. A CSS területen beléptek ebbe a körbe a nyelvészek, illetve a számítógépes mérnökök. A hálózatkutatás egyes területei meg tudtak maradni tisztán szociológián belül (ilyenek például a survey alapú ego network kutatások). A CSS terület azonban még inkább olvasztótégelyként működik nem lehet leválasztani belőle olyan területeket, ahol ne lenne relevanciája más tudományterületeknek. Ebből következően a jövő útja az interdiszciplinaritás ezen a területen. Azok a kutatások tudnak majd nagy hatást elérni, amelyek képesek lesznek keverni az egyes területekről jövő impulzusokat és tudásokat. Ez nem jelenti azt, hogy a társadalomtudósok hátradőlhetnek és azt mondhatják, hogy elég a problémát értenem, majd megoldják a mérnökök vagy a nyelvészek a számolást. Ezzel szemben a témával foglalkozó társadalomtudósoknak legalább elvi szinten érteni kell a modellek működését, és tisztába kell lenniük azzal, hogy milyen adatokból, milyen módszerekkel, milyen kutatási kérdésekre lehet választ adni és mik azok a kutatási problémákat, amiket nem lehet ezekkel a módszerekkel megválaszolni. A módszerek finomhangolását rá lehet bízni a specialistákra, de ehhez érteni kell az alapokat.

A nyelvi modellek társadalomtudományi felhasználása továbbra is három területen fog haladni. Egyrészről a társadalomtudósoknak támogatni kell azokat a kutatásokat, amik arra irányulnak, hogy a nyelvi modelleken alapuló applikációk esetleges diszkrimináció felerősítő hatását megérték és kiküszöböljék. Egyre több és több ipari alkalmazás mögött fognak megjelenni nyelvtechnológia megoldások. Online ajánló rendszerek, chatbotok, fordító programok – mind olyan terület ahol rosszul tanított modellek könnyen vezethetnek diszkriminatív tartalmakhoz. A Google fordító példáján demonstráltam a problémát a dolgozat korábbi részében.

A második felhasználási terület különböző tartalmak klasszifikációjával kapcsolatos. Számtalan olyan társadalomtudományi kérdést meg tudunk fogalmazni, amire nagy szöveges adattartalmak csoportosításával lehet elsősorban válaszolni. Legyen a téma depresszió, szexizmus, vagy káromkodások az online térben – nyelvi klasszifikációs modellekkel közelebb érhetünk a jelenségek megértéséhez. A klasszifikációs modelleken lehet leginkább nyomon követni az egész terület fejlődésének elképesztő sebességét. A jelenlegi egyik vezető megközelítésnek számító vektortér alapú módszer a BERT, 90 százalékos pontossággal meg tudja mondani, hogy egy tweet szexista vagy sem (Samory et al 2020). A jelenlegi 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 és szerzőtársai (2020) legújabb tanulmányunkban azt mutatják be, hogy milyen drámai mértékben javítja a szexista tweetek azonosítását az a módszer, ha az annotátorokkal úgy átíratnak szexista tweeteket, hogy azok ne legyenek már szexisták. A „kijavított” tweetek tanuló adathalmazba keverése 10-15%-kal növelte a becslések pontosságát.

Ez a megközelítés arra jó példa, hogy ne tekintsünk úgy a módszerre, mint ami minden megold, hanem használjuk ki a társadalomtudományi területeken felhalmozott rengeteg tudást ahhoz, hogy minél okosabban tudjuk tanítani az algoritmusokat.
Az utolsó terület az NLP módszerek és ezen belül a vektortér modellek elemzési felhasználásával kapcsolatos. A habilitációs dolgozatomban mind a három példa ehhez a területhez kapcsolódott. Talán ez a leginkább kihasználatlan oldala a módszernek, bár az, hogy a szakma zászlóshajó lapja az AJS lehoz egy ilyen módszert bemutató cikket jelezhet egyfajta változást ebben (Kozlowski et al 2019).

Több időnek kell ahhoz eltelnie, hogy lássuk pontosan mennyire tud elterjedni a vektortér módszerek használata. Utóbbiban sokat segíthet az, ha tisztán látjuk hogyan lehet jól használni a modelleket, milyen lehetőségek rejlenek benne és milyen technikai kérdések merülnek fel a használatával kapcsolatban. Habilitációs dolgozatomban ezekre a felmerülő kérdésekre, dilemmákra és a módszer lehetőségeire igyekeztem részletesen kitérni. A dolgozat mellékleteként elérhetővé tettem azokat a kódokat, amikkel el lehet indulni egy vektortér elemzésében. Reményeim szerint a munkám facilitálja majd a módszer lehetséges felhasználásával kapcsolatos hazai diskurzust.

7 Irodalomjegyzék

Akbik, A., Blythe, D., & Vollgraf, R. (2018, August). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1638-1649).

Antoniak, M. and Mimno, D. 2018. "Evaluating the stability of embedding-based word similarities." *Transactions of the Association for Computational Linguistics* 6: 107-119.

Axelsson, S. – Dahlberg, S. (2018): Corruption Talk: Mapping the Word Corruption in Online Text Data Across the World. General Conference of the European Consortium for Political (kézirat)

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, *3*(Feb), 1137-1155.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135-146.

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 *Advances in neural information processing systems* (pp. 4349-4357).

Bukodi, E., Dex, S., & Goldthorpe, J. H. (2011). The conceptualisation and measurement of occupational hierarchies: a review, a proposal and some illustrative analyses. Quality & Quantity, 45(3), 623-639.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183-186.

Chan, T. W., & Goldthorpe, J. H. (2004). Is there a status order in contemporary British society? Evidence from the occupational structure of friendship. European Sociological Review, 20(5), 383-401.

Chen, L., Ma, R., Hannák, A., & Wilson, C. (2018, April). Investigating the impact of gender on rank in resume search engines. In *Proceedings of the 2018 chi conference on human factors in computing systems* (pp. 1-14).

Cristianini, N., Lansdall-Welfare, T., and Dato, G. 2018. Large-scale content analysis of historical newspapers in the town of Gorizia 1873–1914, *Historical Methods: A Journal of Quantitative and Interdisciplinary History,* DOI: 10.1080/01615440.2018.1443862

Connelly, R., Gayle, V., & Lambert, P. S. (2016). A Review of occupation-based social classifications for social survey research. Methodological Innovations, 9, 2059799116638003.

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*, *2015*(1), 92-112.

De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., ... & Kalai, A. T. (2019, January). 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* (pp. 120-128).

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*, *41*(6), 391-407.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Erikson, R., Goldthorpe, J. H., & Portocarero, L. (1979). Intergenerational class mobility in three Western European societies: England, France and Sweden. The British Journal of Sociology, 30(4), 415-441.

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 megjelentkampány-témák dinamikája alapján. Jelkép, Vol. 36 (3), pp. 25-63.

Freidson, E. (1984). The changing nature of professional control. *Annual review of sociology*, *10*(1), 1-20.

Ganzeboom, H. B., & Treiman, D. J. (1996). Internationally comparable measures of occupational status for the 1988 International Standard Classification of Occupations. Social science research, 25(3), 201-239.

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*, *115*(16), E3635-E3644.

Goldberg, Y., & Levy, O. (2014). word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Goldthorpe, J. H., & Hope, K. (1972). Occupational grading and occupational prestige. Social Science Information, 11(5), 17-73.

Goldthorpe, J. H., Halsey, A. H., Heath, A. F., Ridge, J. M., Bloom, L., & Jones, F. L. (1982). Social mobility and class structure in modern Britain.

Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*.

Gyáni, G. (2016). *A történelem mint emlék(mű)*. Budapest: Kalligram.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016a). Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016b). Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing* (Vol. 2016, p. 2116). NIH Public Access.

Hengchen, S., Ros, R., & Marjanen, J. (2019). A data-driven approach to the changing vocabulary of the nation in English, Dutch, Swedish and Finnish newspapers, 1750-1950. In In Proceedings of the Digital Humanities (DH) conference.

Indig, B. (2018): Közös crawlnak is egy korpusz a vége – Korpuszépítés a CommonCrawl.hu domainjaiból. In: Vincze, V. (szerk): XIV Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Magyarország : Szegedi Tudományegyetem, Informatikai Intézet, (2018) pp. 125-134.

Jakab, A. Zs. (2012) *Emlékállítás és emlékezési gyakorlat. A kulturális emlékezet reprezentációi* Cluj-Napoca, 2012. Kriza János Néprajzi Társaság–Nemzeti Kisebbségkutató Intézet.

Johnson, T. J. (2016). *Professions and Power (Routledge Revivals)*. Routledge.

Joseph, K., & Morgan, J. H. (2020). When do Word Embeddings Accurately Reflect Surveys on our Beliefs About People?. *arXiv preprint arXiv:2004.12043*.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Kao, A., & Poteet, S. R. (Eds.). (2007). *Natural language processing and text mining*. Springer Science & Business Media.

Kmetty, Z. (2018). A szociológia helye a big data-paradigmában és a big data helye a szociológiában *Magyar Tudomány*, *179*(5), 683-692.

Kmetty, Z., Koltai, J., & Rudas, T. (2020). The presence of occupational structure in online texts based on word embedding NLP models. arXiv preprint arXiv:2005.08612.

Kovai M. 2016. *Lélektan és politika. Pszichotudományok a magyarországi államszocializmusban 1945-1970* Károli Gáspár Református Egyetem - L'Harmattan.

Kulkarni, V., Al-Rfou, R., Perozzi, B., & Skiena, S. (2015, May). Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 625-635).

Lison, P., & Kutuzov, A. (2017). Redefining context windows for word embedding models: An experimental study. *arXiv preprint arXiv:1704.05781*.

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, *9*(2), 159-191.

Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, *9*(Nov), 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. *arXiv preprint arXiv:1904.04047*.

Meraviglia, C., Ganzeboom, H. B., & De Luca, D. (2016). A new international measure of social stratification. *Contemporary Social Science*, *11*(2-3), 125-153.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

Miller, I. M. 2013. Rebellion, crime and violence, in *Qing China: a topic modeling approach. Poetics,* 41:6, 626–649. DOI:10.1016/j.poetic.2013.06.005

Mnih, A., & Kavukcuoglu, K. (2013). Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in neural information processing systems* (pp. 2265-2273).

Nakandala, S., Ciampaglia, G. L., Su, N. M., & Ahn, Y. Y. (2016). Gendered conversation in a social game-streaming platform. *arXiv preprint arXiv:1611.06459*.

Nematzadeh, A., Ciampaglia, G. L., Ahn, Y. Y., & Flammini, A. (2016). Information overload in group communication: From conversation to cacophony in the twitch chat. *Royal Society open science*, *6*(10), 191412.

Német, R.; Katona, E.; Kmetty, Z. (2020): Az automatizált szöveganalitika perspektívája a társadalomtudományokban. Szociológia Szemle 2020/2 (megjelenés alatt)

Nelson, L. K. (2020). Computational grounded theory: A methodological framework. Sociological Methods & Research, 49(1), 3-42.

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. New York, NY, Springer. (forthcoming)

Pap, M. (2015). Kádár demokráciája. *Politikai ideológia és társadalmi utópia a Kádár-korszakban* Budapest, Nemzeti Közszolgálati Egyetem.

Pap, M. (2017). A népitől a szocialista demokráciáig A korai Kádár-korszak demokráciafogalma a pártfolyóiratok tükrében. *Múltunk* 2017/1.

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Prandy, K., & Lambert, P. (2003). Marriage, social distance and the social space: an alternative derivation and validation of the Cambridge Scale. Sociology, 37(3), 397-411.

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*, 1-19.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding with unsupervised learning. *Technical report, OpenAI*.

Revelle, W. (2018) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 1.8.12.

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*, *8*(627-633), 116.

Rose, D., & Harrison, E. (2007). The European socio-economic classification: a new social class schema for comparative European research. European Societies, 9(3), 459-490.

Samory, M., Sen, I., Kohne, J., Floeck, F., & Wagner, C. (2020). " Unsex me here": Revisiting Sexism Detection Using Psychological Scales and Adversarial Samples. *arXiv preprint arXiv:2004.12764*.

Schakel, A. M., & Wilson, B. J. (2015). Measuring word significance using distributed representations of words. arXiv preprint arXiv:1508.02297.

Spirling, A. and Rodriguez, P. L. 2019. Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research. Working paper.

https://www.nyu.edu/projects/spirling/documents/embed.pdf

Szabó. M., K. (2019): Az étékváltás jelensége a magyar nyelvben. A negatív emotív elemek egy sajátos használatáról. Magyar Nyelv 115(3) 309-323

Szabó, M. K., Ring, O., Nagy, B., Kiss, L., Koltai, J., Berend, G., … Kmetty, Z. (2020). Exploring the dynamic changes of key concepts of the Hungarian socialist era with natural language processing methods. Historical Methods, Online first, 1–13. http://doi.org/10.1080/01615440.2020.1823289

Tóbiás, D. (2020). A nemi diszkrimináció megjelenésének elemzése Twitch.tv csatornákon szövegbányászati módszerek segítségével (ELTE Szociológia MA, szakdolgozat) https://rc2s2.eu/wp-content/uploads/2020/06/Tóbiás-Dániel_oikcfl_szakdolgozat_2019-20-2.pdf

Treiman, D.J. (1977). Occupational Prestige in Comparative Perspective. Academic Press, New York

Yang, X., Macdonald, C., & Ounis, I. (2018). Using word embeddings in twitter election classification. *Information Retrieval Journal*, *21*(2-3), 183-207.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008).

Wagner, C., Garcia, D., Jadidi, M., & Strohmaier, M. (2015). It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. *arXiv preprint arXiv:1501.06307*.

Waring, J. (2014). Restratification, hybridity and professional elites: questions of power, identity and relational contingency at the points of 'professional–organisational intersection'. *Sociology Compass*, *8*(6), 688-704.

Wiedemann, G., Remus, S., Chawla, A., & Biemann, C. (2019). Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. *arXiv preprint arXiv:1909.10430*.

Zhao, J., Zhou, Y., Li, Z., Wang, W., & Chang, K. W. (2018). Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496*.

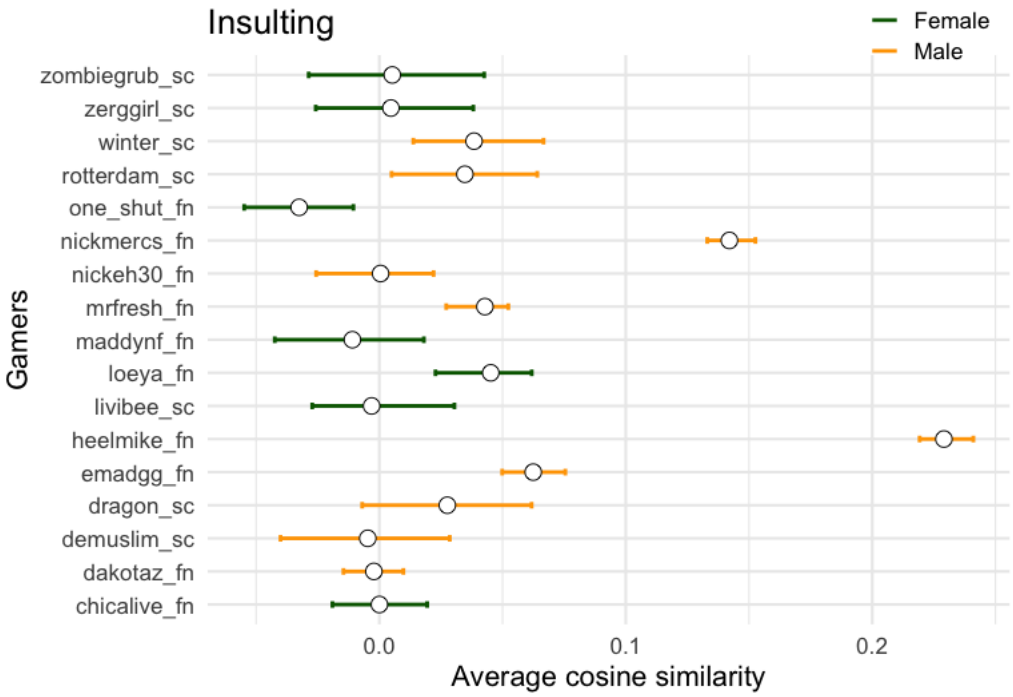
Zweig, G., & Burges, C. J. (2011). The microsoft research sentence completion challenge. *Microsoft Research, Redmond, WA, USA, Tech. Rep. MSR-TR-2011-129*.

Zsibrita, J., Vincze, V., and Farkas, R. 2013. magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In *Proceedings of RANLP 2013*, 763–771.

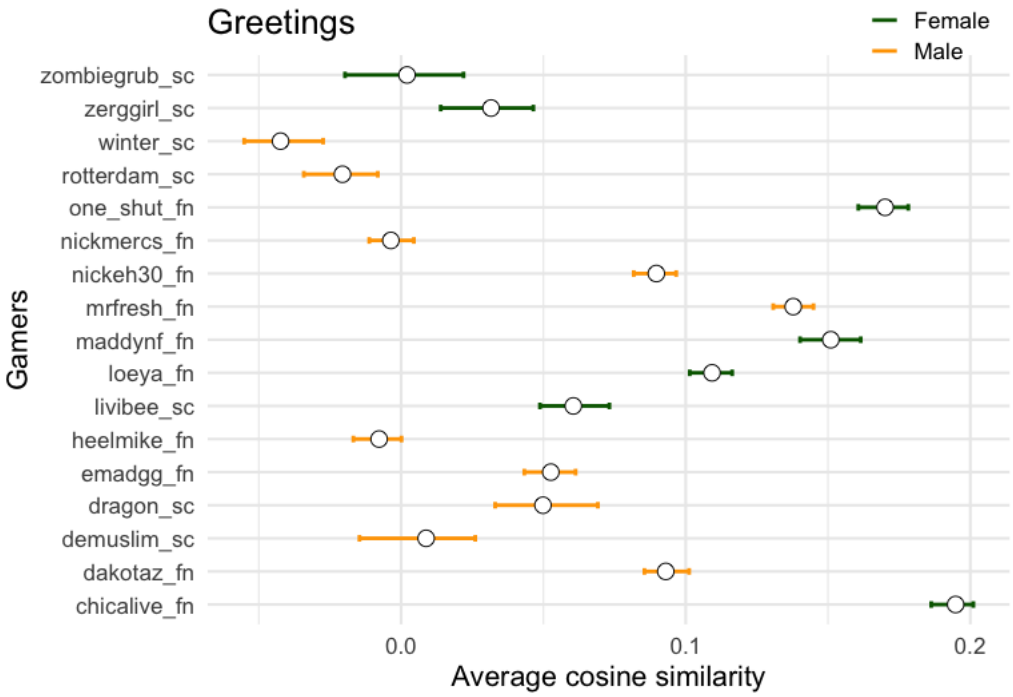
8 Melléklet

accompanist, accountant, acrobat, actor, actuary, admiral, advocate, agriculturist, agrologist, agronomist, allergist, ambassador, anaesthesiologist, anatomist, animator, appraiser, archaeologist, architect, assembler, astrologer, astronaut, athlete, attorney, auditor, babysitter, baker, ballerina, banker, barber, barista, barkeeper, barmaid, barman, bartender, beekeeper, bender, biographer, biologist, bishop, blacksmith, blocklayer, boatman, bodyguard, bookkeeper, bookmaker, botanist, boxer, brazier, breeder, brewer, bricklayer, broker, butcher, cardiologist, carer, carpenter, cellist, ceo, chairperson, chancellor, chaplain, chef, chemist, cleaner, clerk, coalman, coastguard, coder, comedian, commentator, commissioner, composer, congressman, congresswoman, constable, cook, copywriter, coroner, corporal, councillor, courier, curator, dancer, dean, dentist, director-general, dishwasher, dockmaster, doctor, doorkeeper, dramatist, dressmaker, driller, driver, dustman, ecologist, editor, electrician, environmentalist, etcher, farmer, firefighter, fireman, fisher, flamecutter, footballer, forger, friar, furrier, gaoler, gardener, geodesist, geographer, geologist, goatherd, goldsmith, governor, grazier, grocer, hairdresser, head-teacher, historian, hooker, providing sexual services, housemaid, innkeeper, janitor, jeweller, journalist, judge, juggler, lawyer, lecturer, librarian, locksmith, lyricist, macroeconomist, maid, managing-director, manicurist, marketer, marshal, masseur, mathematician, mayor, mechanic, meteorologist, midwife, miner, money-lender, monk, nanny, neurologist, nightwatchman, novelist, nurse, optician, ornithologist, painter, paratrooper, parliamentarian, pastry-cook, pharmacist, philosopher, photographer, physicist, physiotherapist, planter, plasterer, plumber, poet, policeman, policewoman, politician, postman, postmaster, potter, priest, professor, programmer, proofreader, prosecutor, prostitute, psychiatrist, psychologist, psychotherapist, publicist, rabbi, radiographer, rancher, receptionist, rector, retailer, rheumatologist, roofer, sailor, secretary, senator, setter-operator, shepherd, shoe-polisher, shoemaker, shopkeeper, signwriter, singer, sociologist, soldier, solicitor, sommelier, sous-chef, stationmaster, statistician, steward, stewardess, stonecutter, storekeeper, surgeon, tailor, tanner, tattooist, telemarketer, telephonist, tiler, translator, treasurer, typist, vendor, waiter, waitress, weaver, webmaster, welder, writer, zookeeper, zoologist

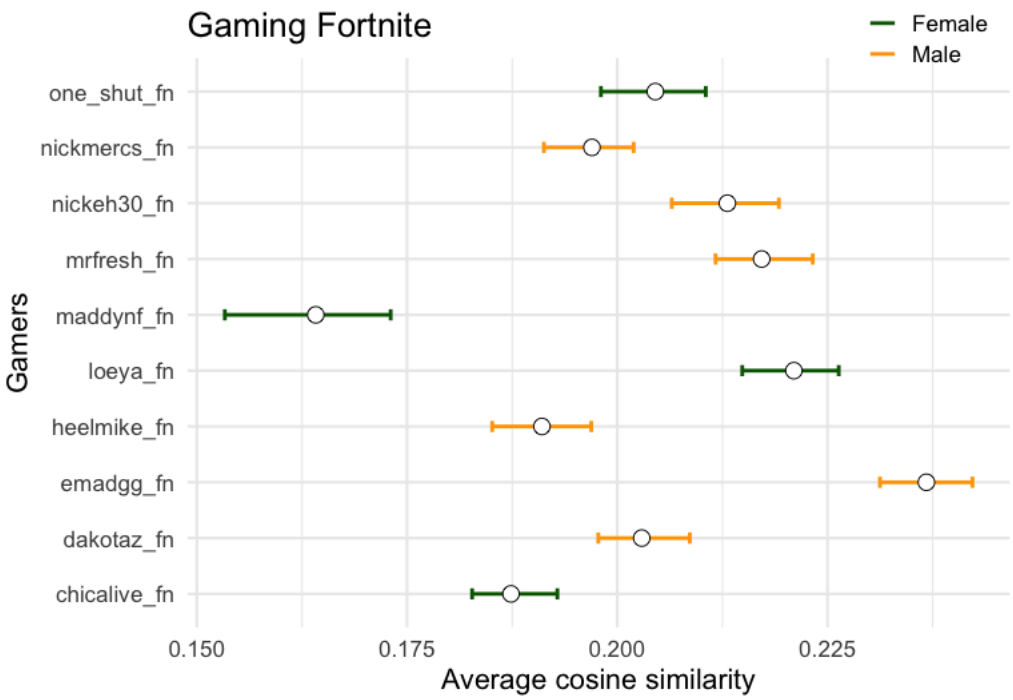
Tábla M1. Foglalkozási lista



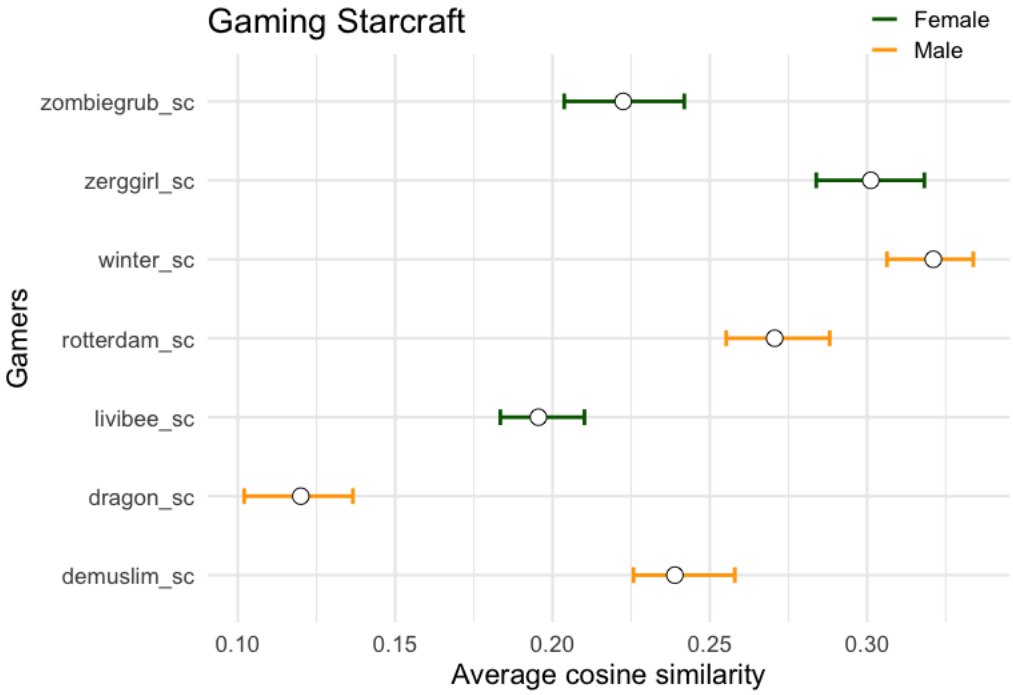
Ábra M1. Inzultálás megjelenése Twitch csatornánként



Ábra M2. Köszönés megjelenése Twitch csatornánként



Ábra M3. Játékkal (Fortnite) kapcsolatos szavak megjelenése Twitch csatornánként



Ábra M4. Játékkal (Starcraft) kapcsolatos szavak megjelenése Twitch csatornánként