Eötvös Loránd Tudományegyetem

Társadalomtudományi kar

ALAPKÉPZÉS

„Word embedding használhatósága a társadalomkutatásban”

Konzulens:

Dr. Németh Renáta

Készítette:

Érsek Boglárka

MQDLQE

szociológia szak

2024. április

Absztrakt

Dolgozatomban a word embedding (szóbeágyazási) modellek társadalomtudományos felhasználhatóságát járom körül. Célom annak a bemutatása, hogy a kutatók eddig milyen jellegű vizsgálatokhoz és miként alkalmazták ezt a módszert. Ezen felül azt is vizsgálom, hogy egy programozni nem tudó kutató számára milyen lehetőség van a módszer használatára. Írásomban először elhelyezem a témát a társadalomtudományos kutatási módszerek között, majd leírom a módszer lényegét és a lehetséges felhasználási módokat. A korábbi kutatások ismertetésével bemutatom, hogy a módszer egyaránt alkalmas technikai és tartalmi felhasználásra, illetve a nyelvi modelleken alapuló algoritmusok kritikus vizsgálatára is. Ezen felül bemutatom a magyar nyelvű szövegek felhasználhatóságát is. Pilot kutatásomban pedig példázom, hogy a WebVectors nevű online elérhető word embedding modell segítségével miként lehet programozói tudás nélkül is használni a módszert.

Tartalomjegyzék

Bevezetés

A digitális kor rengeteg új eszközt és technológiát hozott magával, melyek használhatók a társadalomkutatás során. Például: online felmérések, kérdőívek, interjúk vagy akár kísérletek is folyhatnak digitális környezetben a közösségi média platformokon, online közösségekben. Amellett, hogy ezeknek a már korábban is használt módszereknek az alkalmazása átterjedt az online térbe, nyílt egy teljesen új terület is a kutatók előtt az által, hogy a digitális platformok és technológiák drasztikusan növelték az adatok elérhetőségét. A kutatók már hozzáférhetnek az online felhasználók által generált hatalmas információhoz, a közösségi média interakcióktól az online viselkedésig (Salganik 2019).

Ennek a rengeteg összegyűlt adatnak (Big Data) az elemzésére új módszerek jöttek létre, az ilyen típusú új módszerek közé tartozik a számítógépes szövegelemzés is, ami a szakdolgozatom központi témája. A digitális átalakulás paradigmaváltást eredményezett a kutatási területeken. A hagyományos módszereket digitális megközelítésekkel egészítik ki, vagy váltják fel, ami azt jelenti, hogy a kutatóknak alkalmazkodniuk kell az új tervezési, adatgyűjtési és eredmény elemzési módokhoz (Salganik 2019).

A számítógépes szövegelemzési technikák ismeretének több előnye is van napjaink társadalomkutatásában. Valószínűsíthető, hogy a közeljövőben egyre olcsóbb lesz a digitális adatokhoz való hozzáférés és egyre drágább lesz minőségi kérdőíves adatokat létrehozni, ennek következtében hangsúlyosabb szerep fog jutni a Big Data-ra alapuló kutatásoknak, és kevésbé hangsúlyos a kérdőívekre épülőknek. Ezen kívül az is előnye a digitális adatok használatának, hogy olyan információkhoz is hozzájuthatunk, melyekhez kérdőívekkel csak nagyon nehezen lehetne, például a témák kényessége miatt. Ilyen témák a depresszió vagy az előítéletesség kutatása (Németh – Katona – Kmetty 2020). Ezeknek az előnyöknek és jövőbeli perspektíváknak az ismeretében fontosnak tartom, hogy a számítógépes szövegelemzési módszerek minél több hangot kapjanak a társadalomtudományos diskurzusban. Ezt elősegítendő dolgozatomban bemutatom a számítógépes szövegelemzési módszerek egy típusát, és annak használhatóságát a társadalomkutatásban.

A számítógépes szövegelemzésbe sokféle különálló módszer tartozik, vannak köztük egyszerűbbek és összetettebbek egyaránt. Kmetty, Németh és Katona (2020) írnak például szentiment- és emócióelemzésről (a módszerek lényege a szerző egy tárggyal kapcsolatos

véleményének vizsgálata, kategóriákba való besorolás segítségével), klaszterelemzésről (a szövegek kategorizálása a hasonlóságuk vagy különbözőségük alapján), topikmodellekről (olyan eljárások, melyeknek célja a korpusz témáinak azonosítása) és word embedding modellekről (a korpusz szavainak egy digitális vektortérben való elhelyezése a jelentésük alapján), melyek mindegyike alkalmas lehet társadalomkutatásra. Több módszer társadalomtudományos használhatóságának bemutatása nem férne bele szakdolgozatom terjedelmébe, ezért kiválasztottam egyet, a word embedding modelleket (magyarul szóbeágyazási modelleknek nevezik őket, de az angol megnevezés terjedt el inkább, ezért használtam a címben is a word embeddinget) és ezek társadalomtudományos használhatóságát járom körül szakdolgozatomban. A választásom azért esett a word embedding modellekre, mert ezt a módszert találtam a legsokrétűbbnek abból a szempontból, hogy lehet alkalmazni a megfelelő programozói tudás meglétével összetettebb elemzésekre, ugyanakkor az interneten bárki számára elérhető demo-kat használva egy programozni nem tudó kutató is tudja alkalmazni a módszert. Dolgozatomban nem célom a word embedding modellek módszertanának teljes körű bemutatása, a modellek működésének részletes leírása (erről például Kmetty (2022) ír részletesen), dolgozatomban inkább a kutatásokban való használhatóságukra koncentrálok.

A dolgozat célja

A dolgozatom célja kettős: először szeretném összefoglalni korábbi kutatások ismertetésén keresztül, hogy hogyan használhatóak a word embedding modellek a társadalomkutatásban, ezt követően egy saját pilot kutatás segítségével szeretném bemutatni, hogy programozói tudás hiányában miként lehet alkalmazni a modelleket.

Tehát a két kutatási kérdésem a következő:

- Milyen módokon használják a szóbeágyazási (word embedding) modelleket a kutatók a társadalom kutatásban?
- Miként tudja felhasználni a társadalomkutatásban a word embedding modellek egyszerűbb, bárki számára hozzáférhető változatait egy programozni nem, vagy csak alig tudó kutató?

A számítógépes szövegelemzési módszerek helye a társadalomtudományos kutatási módszerek között

Ebben a részben szeretném elhelyezni a számítógépes szövegelemzési módszereket a társadalomkutatás egyéb módszerei közt, ismertetem az egyes módszerekhez való hasonlóságát, és bemutatom, hogy mennyiben jelent újdonságot ez a fajta megközelítés. Igyekszem példákkal szemléltetni, hogy a különböző módszer típusok miként jelennek meg a szövegelemzés gyakorlatában.

Anthony Giddens (2008) szerint a tudomány „az empirikus vizsgálat szisztematikus módszereinek használata, az adatok elemzése, elméleti gondolkodás és az érvek logikai értékelése, amelyek révén egy bizonyos kérdésről ismeretanyagot alakíthatunk ki" (Giddens 2008: 76). Tehát ahhoz, hogy a szociológiára tudományként tekinthessünk, szükségszerű különböző kutatásokat végezni.

Több féle csoportosítás szerint tudjuk elkülöníteni az egyes kutatási módszereket. Leggyakrabban a kvantitatív és a kvalitatív megközelítési módokról szoktunk beszélni, mind a két fajta megközelítést szokták használni a szövegelemzésben is.

A kvantitatív kutatások célja a numerikus adatok gyűjtése, illetve azok elemzése. Az adatgyűjtést általában kérdőívekkel, számszerűsített megfigyelésekkel és egyéb mennyiségi adatgyűjtési módszerekkel végzik. Az adatok értelmezése és a következtetések levonása legtöbbször valamilyen statisztikai módszer segítségével történik. Az ilyen típusú módszerek legnagyobb előnye a reprezentativitás, illetve, hogy könnyű a kapott adatokat csoportokba rendezni és összehasonlításokat végezni rajtuk. Ezen kívül elmondható, hogy ezek világosabb, részletesebb és érthetőbb megfigyelések. (Babbie 1999). Néhány példa a kvantitatív kutatásokra: Felvételi kutatás (kérdőívek vagy interjúk segítségével gyűjtött adatok elemzése), adatok másodelemzése (egyéb kutatásokból származó adatok számszerű elemzése, például különböző statisztikai módszerek alkalmazásával), társadalmi hálózatelemzés (a kapcsolathálókat vizsgáló kvantitatív módszer, amely általában grafikus modellek segítségével ábrázolja a kapcsolatokat).

Napjainkban egyre népszerűbb kvantitatív kutatásra használni a számítógépes szövegelemzési módszereket. Ezek a módszerek arra épülnek, hogy a szövegekben lévő

szöveges adatokból valamilyen módon (általban különböző szoftverek segítségével), számszerűsített, numerikus adatokat készítenek, és ezeket a numerikus adatokat elemzik tovább, sok esetben statisztikai módszerekkel. A kvantitatív szövegelemzést a két világháború között kezdtek el először alkalmazni a tömegmédia elemzésére. Ekkor a módszer alapvetően úgy működött, hogy a szövegekben kvalitatív módon, a kutatók által azonosított kódokat számszerűsítették. Emellett a kódokon kívül feldolgozatlan szövegrészleteket, metaadatokat is használtak a kutatásokban. A kvantitatív szövegelemzés legelterjedtebb módszerei közé tartoznak a szózsákmodellekre alapuló módszerek. A szózsákmodellek arra épülnek, hogy megszámolják azt, hogy az egyes szavak hányszor szerepelnek a szövegben, tehát a szavak halmazaként tekint a szövegre. Ez a modell nyelvészeti szempontból leegyszerűsített, ugyanis nem veszi figyelembe a szavak sorrendjét a szövegen belül. Erre alapulnak például a szóelőfordulási gyakoriságokkal dolgozó modellek, vagy a topik-modellek is (Németh – Katona – Kmetty 2020).

Barna és Knap (2019) is kvantitatív szövegelemzéssel, egészen pontosan topik-modell segítségével vizsgálta azt, hogy a kuruc.info „zsidó" szót tartalmazó cikkei tartalmilag milyen témákra bonthatók. A topik-modellek úgy működnek, hogy az adott dokumentum gyűjteményben, jelen esetben a kuruc.info „zsidó" szót tartalmazó cikkeiben, megtalálható témákat azonosítják automatizált eljárásokkal. A kutatás során 15 témát/topic-ot azonosítottak. Például új antiszemitizmusra és összeesküvés-elméletekre (holokausztagadás, holokauszt-biznisz) épülő, faji és vallási alapú témákat különítettek el. Egy másik példa a kvantitatív tartalomelemzésre Szabó és társai (2020) munkája, akik a Kádár-korszak legfontosabb fogalmainak változását kutatták. Kutatásukhoz szóbeágyazási módszereket használtak, egészen pontosan azt vizsgálták, hogy a Pártélet című, 1956 és 1989 között kiadott havilapban, hogyan változott a politikai diskurzus bizonyos mezőgazdasággal és iparral kapcsolatos témákban. A kutatás technikai részleteiről, a word embedding magyar nyelven című részben írok részletesebben.

A kvalitatív kutatások részletesen vizsgálják a jelenségeket, és a résztvevők észleléseire, véleményeire fókuszálnak. A kvantitatív kutatásokhoz képest, kevés vizsgálati alannyal dolgoznak a kutatók, de jóval mélyebben, aminek az a következménye, hogy a módszer által kapott eredmények nem, vagy csak nehezen számszerűsíthetőek. Leginkább az interjúk, fókuszcsoportok és más, leíró megközelítések alkalmazása tartozik ehhez a típusú

megközelítéshez. Árnyaltabb, gazdagabb jelentéssel bíró, ugyanakkor több értelmű eredményhez juthatunk. Kezdetben a megfigyelések jelentős része ilyen típusú volt. (Babbie 1999). Néhány példa a kvalitatív kutatásokra: résztvevő megfigyelés (a kutató aktívan részt vesz a vizsgált csoport vagy közösség életében, és megfigyeli a jelenségeket), interjúk (az emberekkel való egyéni, vagy csoportos beszélgetések, amelyek célja a mélyebb megértés és a személyes tapasztalatok feltárása), életútinterjú (az egyének élettörténetének részletes elemzése, ami segít megérteni az életút alakulását és a társadalmi kontextust), tartalomelemzés (a szövegek, beszédek vagy más tartalmak részletes, minőségi elemzése, hogy feltárják a mögöttük álló jelentéseket).

A kvalitatív szövegelemzést gyakran használják a társadalomtudományokban, humán tudományokban és más területeken is, ahol fontos a részletes és mélyebb megértés a szövegek tartalmáról és jelentéséről. Két fontos szempont van amelyek alapján kvalitatívnak tekinthetünk egy szövegelemzést. Az egyik lényeges pont az, hogy az elemzés során a kutatók ténylegesen elolvassák a vizsgált szövegeket. A másik fontos kitétel a kutatás közbeni kategóriaképzés. Az egyik legismertebb és legkidolgozottabb kvalitatív szövegelemzési paradigma a Grounded Theory (GT), ennek lényege az, hogy az adatok elemzése során jutunk el az elmélet megfogalmazásához. A módszer által létrejött elméleteknek jól érthetőnek kell lenniük, illeszkedniük kell a megfigyelésekhez és magyarázatot kell adniuk valamilyen társadalmi folyamatra. A GT kutatásielveket alkalmazó szövegelemzésben nyílt kódolást és állandó összehasonlítást alkalmaznak. Ez azt jelenti, hogy nincsenek előre definiált kategóriák, hanem a kutató a szöveg kódolása során folyamatosan összehasonlítja egymással a nyers adatokat és az alakuló fogalmakat, így alakítva ki a kódokat. A módszer használatának nagy előnye az, hogy így sokkal kevésbe lesznek befolyásoltak az eredmények azáltal, hogy a kutatónak előismeretei, előfeltevései voltak (Sallay 2015: 13).

Fontos kiemelni, hogy a kvalitatív szövegelemzésnek nem kell kizárólag „kézzel”, a kutatók által történnie, a számítógépes szoftverek használata nem jelenti azt, hogy a kutatás kvantitatív lesz. A kvalitatív adatelemző szoftverek (QDA), vagy más néven számítógép által támogatott kvalitatív adatelemző programok (CAQDAS), valamilyen struktúrába szervezik az adatokat, ezzel könnyítik az elemzést. Így több szöveget, gyorsabban tudnak elemezni a kutatók. A módszer által javul a szöveg feldolgozásának minőségége is, hiszen kevésbé lankad a kutatók figyelme úgy, ha a program előre rendezi számukra valamilyen szempont szerint a szöveget. Megfigyelhető, hogy a programok elterjedésével nőtt a kvalitatív szövegelemzést

használó kutatások száma is. A legnépszerűbb ilyen szoftverek közé tartozik például a NVivo, az ATLAS.ti és a NUD*IST (Sulyok – Juhász – Erdei 2019).

Kvalitatív szövegelemzésre kiváló példa Tóth Olga (2001) kutatása. Ő a Nők Lapja 1989 és 1999 között kiadott lapjait vizsgálta kettős céllal. A lapban megjelenő nőképet, illetve annak változását vizsgálta, ezen kívül kíváncsi volt a lap szerkezetében végbement változásokra is. Véletlenszerűen kiválasztott havonta egy lapszámot, és ezeket a következő szempontokra koncentrálva elemezte: Figyelte az adott lap szerkezetét az alapján, hogy milyen arányban jelennek meg benne aktualitások, szolgáltatások, okosítások és riportok. Ezen kívül vizsgálta a híres emberek, szakértők és hétköznapi emberek megjelenését az újságban. A kutató különösen figyelt a sztereotip nemi szerepek megjelenésére is. Kutatása során arra az eredményre jutott, hogy a lapban az idő előrehaladtával egyre inkább megjelenik a fogyasztó centrikus szemlélet, mivel egyre több reklámmal és új ezoterika, fitness rovatokkal találkozott. Egyre több hír volt a sztárokról és a csillogó életükről és ezzel arányosan egyre kevesebb a hétköznapi emberekről. A lapban egyre hangsúlyosabban jelent meg a hagyományos család és az individuum fontossága, a munkáról és politikáról szóló témák viszont egyre inkább a háttérbe szorultak.

A számítógépes szövegelemzési módszerek alapvetően kvantitatívak, hiszen úgy működnek, hogy a kutatók nem olvassák el ténylegesen a szövegeket, hanem a szöveges forrásokban rejlő adatokat különböző módszerekkel numerikus adatokká alakítják, és ezeket a numerikus adatokat elemzik tovább, sok esetben statisztikai módszerekkel (Németh – Katona – Kmetty 2020).

Dolgozatom témáját egy másik kategorizációban elhelyezve szeretném megemlíteni a Giddens (2008) által elkülönített négy fő szociológiai kutatási módszert, melyek között vannak alapvetően kvantitatív és kvalitatív módszerek is. A négy módszer a terepmunka, a felmérés, a kísérlet és a dokumentumkutatás. A számítógépes szövegelemzés ezek közül a módszerek közül a dokumentumkutatásra hasonlít legjobban, hiszen mindkettő egy valamilyen módon létrejött, írott szöveges forrás-halmazból indul ki. A hagyományos dokumentumkutatás gyakran történeti jellegű, hiszen „sokszor csak valamilyen időperspektívából értelmezhetjük az adott problémáról gyűjtött anyagot” (Giddens 2008: 86). Ezzel szemben a számítógépes szövegelemzés során friss szövegeket tartalmazó korpuszokat is használhatunk (például közösségi médiából származó szövegek).

Giddens (2008) a dokumentumkutatás korlátjaként említi azt, hogy nehéz megállapítani, mennyire reprezentálnak a valóságban is megfigyelhető tendenciákat a források. Ezek a reprezentativitással kapcsolatos kételyek a számítógépes szövegelemzés esetében is fennállnak, annak ellenére is, hogy a források szerzőinek köre a digitális korban jelentősen kitágult, hiszen korábban csak könyvek, tanulmányok, újságcikkek lehettek a források, ma pedig elegendő egy telefon és internetelérés ahhoz, hogy valaki írjon egy kommentet valamelyik közösségi médiás platformon és ezáltal egyből „szerzővé" is váljon. Hiába van meg a lehetősége több embernek a forrássá váláshoz, sok esetben még így sem általánosíthatunk, hiszen sokan nincsenek fent a közösségi médiában, vagy nem kommentelnek, vagy például bizonyos szavak felül reprezentáltabbak egyes internetes hírportálokon, mint a való életben.

Erről ír Németh Renáta (2015) is, a hedonométer konkrét példáján szemléltetve a Big Data-ra alapuló kutatásokkal kapcsolatban felmerülő aggályokat. A reprezentativitás kérdéskörén túl az objektivitásról is ír, az online térben hagyott lábnyomoknak ugyanis önmagukban nincs jelentésük, minden esetben a kutató alkotja meg hozzájuk a jelentést. Mindezek ellenére nem kell használhatatlannak tekinteni a Big Data-ra alapuló módszereket a társadalomkutatásban, hisz sok esetben igen jól alkalmazhatóak, sőt vannak olyan témák, amelyeket más módszerekkel nem is lehetne vizsgálni. Arra viszont nagyon oda kell figyelni, hogy óvatosan kezeljük az eredményeket, ne vonjunk le belőlük általános következtetéseket.

Más szempontból viszont a számítógépes szövegelemzés sokkalta közelebb áll a felmérések, survey-k által gyűjtött adatok kiértékeléséhez, elemzéséhez, hiszen itt is gyakran statisztikai módszerek segítségével jutunk el a kutatásunk eredményéhez.

Napjainkban már létezik átfedés a társadalomkutatásban kvantitatív és kvalitatív kutatások, vagy a Giddens által említett módszerek között. A kutatók célja az, hogy a világot egyre komplexebben tudják leírni, ezért arra törekednek, hogy a töredezettség és a módszertani korlátok ne akadályozzák meg őket ebben (Király et al. 2014: 95). Elsősorban azért szükséges a különböző módszerek használata, mert külön-külön mindegyiknek megvannak a maga korlátai, így viszont lehetőség nyílik arra, hogy ellenőrizzék, kiegészítsék az egyes módszerekkel kapott eredményeket (Giddens 2008). Fontos, hogy ne keverjük össze a több módszertant használó kutatásokat a kevert módszertant alkalmazó kutatásokkal, az előbbiben lehetséges, hogy az összes általunk használt módszer kvalitatív/kvantitatív, ezzel

szemben az utóbbinak pont az eltérő módszertanok összehangolása a célja (Király et al. 2014: 95-96).

A 20. század második felében egyre inkább elkezdett egyesülni a korábban említett két alapvető megközelítési mód, a kutatók egyre gyakrabban kombinálták a kvantitatív és kvalitatív elemzéseket, hogy teljesebb képet kaphassanak a vizsgált társadalmi jelenségekről. A kvalitatív kutatásból származó eredmények segíthetik a kvantitatív adatok értelmezését, a kvantitatív eredmények pedig validálhatják a kvalitatív kutatásokból származó eredményeket. A módszerek keverése történhet az adatgyűjtés, az adatelemzés vagy az interpretáció fázisában, vagy akár az összes fázisban is. A keveredés módja lehet összekapcsolás (például egy kérdőív összeállításához kvalitatív adatgyűjtést veszünk alapul), integrálás (az adatok merge-elése meghatározott szempont alapján) vagy beágyazás (pl.: előzetes vagy utólagos kvalitatív mélyinterjú egy hatáselemzésnél) (Király et al. 2014:98).

A kevert kutatási módszerek alkalmazására a számítógépes szövegelemzésben jó példa Virágh és Szepesi (2022) kutatása, akik a vállalkozók reprezentációját vizsgálták 2015 és 2019 között az online média meghatározó orgánumaiban. Kutatásuk egy kvantitatív és egy kvalitatív szövegelemzésből állt. A kvantitatív elemzéshez cimkéskereséssel felkutatták a cikkeket, összesen 485-öt, majd ezeket előre meghatározott szempontok szerint kódolták egy két szintű adatbázisba, ezután SPSS segítségével elemezték az adatokat. A kutatás kvalitatív részében 45, egy-egy vállalkozót/vállalkozást bemutató portrét elemeztek az ATLAS.ti szövegelemző szoftver segítségével, öt dimenzió szerint (vállalkozói tulajdonságok, motivációk, mi a siker, sikertényezők, vállalkozó háttere) csoportosították az adatokat. A kutatásuk során a következő eredményekre jutottak: A kvantitatív elemzés során elemzett cikkek legnagyobb része, 41,2 százaléka, semlegesen ítéli meg a vállalkozókat, 32,4 százalékuk inkább negatívan, 26,4 százalékuk pedig inkább pozitívan. Ha általánosságban írnak a cikkek a hazai vállalkozókról, vállalkozásokról, akkor hajlamosabbak negatív módon megjeleníteni őket, mint akkor ha konkrét személyekről, cégekről van szó. A kvalitatív elemzésből is látható, hogy a bemutatott, meginterjúvolt alanyokat alapvetően pozitív megítéléssel mutatják be. A kutatás során sikerült elkülöníteniük hat vállalkozótípust: a született vállalkozót, a népmesehőst, az ötletvezéreltet, a lassú építkezőt, a hobbivállalkozót és a társadalmi vállalkozót.

A kutatási módszerek folyamatosan alakulnak és adaptálódnak a társadalomtudományok fejlődésével, és a kutatási kérdések változásával. A szociológia maga

is változik az új módszerek megjelenésével. Ez megfigyelhető volt például a survey-k megjelenésekor is, ekkor ugyanis felértékelődött a statisztika szerepe a társadalomkutatásban. Ezért az sem meglepő, hogy a Big Data megjelenése is változást von maga után a szociológiában, társadalomkutatásban, és megjelennek a különböző, erre épülő katatási módszerek, többek között a számítógépes szövegelemzés, amelyek idővel beépülnek a kutatók eszköztárába (Kmetty 2018).

A következőkben egy újabb szempontrendszer alapján helyezem el dolgozatom témáját. A társadalomtudományos kutatásokhoz alapvetően két logikai megközelítésből állhatunk hozzá. Kiindulhatunk valamilyen sejtésből, hipotézisből és a kutatás, megfigyelés által begyűjtött információk segítségével következtethetünk arra, hogy a kezdeti hipotézisünk megfelelően írja-e le a fennálló megfigyelt összefüggést. Ezt a logikai megközelítést deduktív módszernek nevezzük, tehát ebben az esetben az általánostól a speciális felé haladunk, az elméletet egy konkrét esetre alkalmazzuk. A másik, induktívnak nevezett megközelítésmód során, egy már meglévő megfigyelésből indulunk ki, és az összefüggések megtalálásával jutunk el a következtetéshez. Tehát az egyes részletektől az általános elvek, a tényektől az elmélet felé haladunk (Babbie 1999).

A survey módszertannal készülő kutatások hagyományosan deduktívak, ezzel szemben a Big Data-ra alapuló kutatások többségében inkább induktívnak nevezhetők. Így a természetes nyelv feldolgozás (NLP) módszerei leginkább induktív kutatásra használhatók, egyelőre még ritkábban alkalmazzák őket deduktív vizsgálatokban. Az induktív logika használata az adat alapú társadalomtudományos kutatásokban újszerű gondolkodásmódot igényel a szociológusoktól, mivel megfordítja a megszokott logikát (Tóbiás 2020). Sik Domonkos és társai depressziós fórumok vizsgálatával foglalkoznak, több kutatást is készítettek a témában. Az egyik kutatásukban topikmodellezéssel próbáltak tematikus csomópontokat találni a fórumokon (Sik et al. 2021). Ebben az esetben nem volt egy előre meghatározott hipotézisük/elméletük a kutatóknak, amit tesztelni szerettek volna, hanem az volt a cél, hogy beazonosítsák azokat a témákat, amelyekkel a bejegyzések foglalkoznak. Itt tehát induktív logikát követtek. Egy másik, a témában született kutatásukban azt vizsgálták, hogy a COVID-19 járványnak milyen hatása van az online depressziós fórumokon zajló diskurzusra. Ebben az esetben már volt egy előre meghatározott hipotézisük a témában, azt várták, hogy a COVID-nak fontos szerepe lesz a diskurzusban, és a témák e köré fognak

csoportosulni (Németh et al. 2023). Tehát ez a kutatás deduktív logikai szempontból közelítette meg a témát.

Mi a word embedding?

A digitalizáció és a számítógépes forradalom miatt egyre több új típusú adatforrás áll rendelkezésünkre. Ezek az új típusú adatok nagyon sokban különböznek a tradicionális adatoktól, ezért feldolgozásukhoz is új elemzési módszerek, speciális technikák szükségesek. Például a kapcsolathálózati adatbázisok elemzésére létrehozták a társadalmi hálózatelemzés módszerét (Kmetty 2022: 106). A számítógépes szövegelemzésben használt adatokról is elmondható, hogy nagyban eltérnek a korábban használt adatoktól, ugyanis korábban a kvantitatív társadalomkutatásban jól strukturált numerikus adatokkal dolgoztak. Az adatok keletkezésének módja miatt ez számítógépes szövegelemzésre már nem jellemző. A használt adatok nem kifejezetten elemzésre készültek, hiszen a professzionális szereplők által készített szövegek (könyvek, újságcikkek) mellett laikus felhasználók által írt tartalmak (blogok, kommentek) is elemezhetők. Az ilyen típusú szövegek strukturálatlanok, viszont nagyon sok információval szolgálhatnak az emberek véleményéről, attitűdjéről, motivációiról (Németh – Katona – Kmetty 2020).

A Big Data kutatás egy speciális ága a számítógépes szövegelemzés. A módszer háttere alapvetően nem a társadalomkutatásból indult, az 1990-es, 2000-es években számítógépes nyelvészek, statisztikusok, számítástudománnyal foglalkozók tették le a módszer alapjait. Napjainkban a módszer használata leginkább az üzleti szférában jelentős, a Google és a Facebook is rengeteg energiát fektet az algoritmusok fejlesztésébe. Létrejött egy új, interdiszciplináris tudományterület a számítógépes társadalomtudomány, a 2010-es évek második felében pedig megjelentek a tudományterülettel foglalkozó kutatócsoportok, konferenciák és egyetemi szakok (Kmetty 2022: 106).

A természetesnyelv feldolgozási (NLP) módszerek már meghaladják a szógyakoriság-elemzés alapú kvantitatív szövegelemzéseket, ezek a módszerek a gépi tanulás (machine learning) paradigmáján alapuló modellezési logikát követik (Németh – Katona – Kmetty 2020: 44). A természetesnyelv azt jelenti, hogy spontán alakult ki, az emberek közti nyelvi kommunikáció eredményeként. Az NLP módszerek nem alkalmasak a szövegek teljes tartalmi

megértésére, lényegük inkább az, hogy létrehozzanak módszereket, melyek képesek nagy mennyiségű, természetes nyelven íródott szöveg elemzésére. Ez a tudományterület az informatika, a mesterségesintelligencia-kutatás és a nyelvészet határterülete. Ahogy már korábban is említettem, ezek a módszerek strukturálatlan adatokkal, tehát olyan adatokkal melyeknek nincs adatbázis jellege, dolgoznak. Ezeket a strukturálatlan adatokat először előfeldolgozni kell, hogy elemzésre alkalmas numerikus adatbázist kapjunk. Az előfeldolgozás alapvetően több lépésből áll, de a magyar nyelv esetében ez még összetettebb, mert a magyar agglutináló nyelv, erről word embedding magyar nyelven című részben írok részletesebben (Németh – Katona – Kmetty 2020).

Az NLP módszerek többféle nyelvi modellt alkalmaznak. A legtöbb módszer a már korábban is említett szózsákmodellt veszi alapul, az ilyen típusú modellek nem veszik figyelembe a szavak sorrendjét a szövegben, a szavak halmazaként tekintenek a szövegekre. Szózsákmodellre épülhetnek például topikmodellek, klaszterelemzések és szentimentelemzések is. Vannak olyan módszerek, melyek már meghaladják a szózsákmodelleket abból a szempontból, hogy a szavak mondaton belüli környezetét is figyelembe veszik. Ilyen módszerrel dolgoznak a word embedding (szóbeágyazási) modellek is, melyekről részletesebben írok dolgozatomban (Németh – Katona – Kmetty 2020).

„A szóbeágyazási modellek (word embedding models) a vizsgált korpusz látens szemantikai struktúrájának reprezentálására szolgáló, a gépi tanulásban elterjedt neurális hálókat használó módszerek.” (Németh – Katona – Kmetty 2020: 57). Tehát a word embedding modellek a bevitt korpuszt feldolgozva, neurális hálók segítségével egy többdimenziós vektortérbe helyezik a szavakat úgy, hogy a jelentésük határozza meg a szavak elhelyezkedését. A vektortérben egy vektor egy szónak felel meg, a jelentések pedig úgy határozzák meg a szavak elhelyezkedését, hogy a hasonló jelentésű szavak közel, míg a különböző jelentésű szavak távol kerülnek egymástól (Németh – Katona – Kmetty 2020: 57).

Nagyon fontos tisztázni, hogy mit is értünk jelentés alatt, ugyanis ebben az esetben nem a hétköznapi értelemben vett jelentésről van szó. A disztribúciós szemantikai értelemben vett jelentésről beszélünk itt, hasonló jelentésű szavak alatt a hasonló környezetben elhelyezkedő szavakat értjük. A vektortér a szavak jelentésének kapcsolatán alapszik, a disztribúciós szemantika a szavak használatával azonosítja a szavak jelentését. A modellek általában a szót megelőző és követő 3-10 szót vizsgálják. Egy példán szemléltetve ezt: a férfi

és a nő szavak viszonylag közel találhatóak egymáshoz, mivel gyakran szerepelnek hasonló környezetben a mondatokban. A vektortérben úgy határozzuk meg a szavak közelségét, hogy megvizsgáljuk a szavaknak megfelelő vektorok által bezárt szöget. Ez a szög, illetve általában e szög koszinusza (a koszinusz nagyság) mutatja meg nekünk a közelséget (Németh – Katona – Kmetty 2020: 57).

A modell alkalmazásakor egyfajta dimenziócsökkentő eljárás történik, annak érdekében, hogy ki tudjuk szűrni az adott szó mellett szereplő rengeteg szó közül a releváns együttes szó előfordulásokat, így tehát a végén nem fog minden szó önálló dimenziót megjeleníteni. Ehhez hasonló dimenziócsökkentő műveletet végzünk a főkomponens elemzéskor is (Kmetty 2022).

A word embedding modellek előzményei már az 1980-as években megjelentek, de a 2010-es évekig várni kellett az áttörésükre. A robbanás Mikolov és társai cikkéhez volt köthető akik egy word2wec nevű, neurálisháló-alapú (NNLM) nyelvi modellt fejlesztettek ki, ami képes volt meggyorsítani a számítási időt, és ez által több ember által elérhetővé tenni a módszert (Kmetty 2022). Mikolovék (2013) célja az volt, hogy létrehozzanak egy, a kifejezések szövegben való megtalálására alkalmas egyszerű módszert, ami képes szavak millióinak a vektortérben való elhelyezésére. Kmetty (2022: 111) szerint a modell sikeressége három dolognak köszönhető: jó eredményeket ért el a tesztelt feladatokban, a szerzők nyilvánosságra hozták az előkészített vektortereket, amiket hatalmas szövegkorpuszokra építettek és szabadon hozzáférhetővé tették a word2vec kódjait, ami által kevesebb programozói tudással is képesek lehetnek a kutatók saját vektorterek létrehozására. A word embedding modelleket azóta is folyamatosan fejlesztik, a word2vec továbbfejlesztéseként megjelent például a fastText, illetve a részben Word2vec-re épülő GloVe. Megfelelő paraméterezés mellett nincs nagy különbség ezen módszerek hatékonyságában (Kmetty 2022).

A módszer alkalmas lehet szemantikai kapcsolatok megfigyelésére, például a foglalkozások nevei egymáshoz közel helyezkednek el. Ezen kívül megfigyelhetőek olyan csoportok, amelyek jól láthatóan elkülönülnek egymástól, például a női jellegű foglalkozások jól láthatóan elkülönülnek a férfi jellegűektől (Németh – Katona – Kmetty 2020: 57). A közelség-távolság megvizsgálásán túl analógiás tesztek elvégzéséhez is használhatóak a word embedding modellek. Többek között választ kaphatunk arra a kérdésre, hogy „Mi Magyarország fővárosa?”, ha például a Németország-Berlin relációt már azonosítottuk a

vektortérben. Nem csak ilyen konkrét technikai esetben képes a word embedding a válaszadásra. Azt is megtudhatjuk például, hogy „van-e az általunk vizsgált korpuszban a foglalkozásneveknek a nemi különbségeket reprezentáló nyelvhasználati különbsége” (Németh – Katona – Kmetty 2020: 58). Ezt például úgy tudjuk megvizsgálni, hogy megnézzük, melyik foglalkozást kapjuk akkor, ha az orvost eltoljuk ugyanolyan távolságra és irányba, mint ahogy a férfiből a nőt kapjuk. A kérdés az, hogy így például ápolót kaptunk-e? (Németh – Katona – Kmetty 2020)

Az egyes tudományterületek és gazdasági szereplők eltérő módon használhatják a word embedding módszerét. A nyelvészek számára például a szavak morfológiájának változása érdekes, a társadalomtudósoknak az, hogy megértsék a társadalmi jelenségeket a szövegeken keresztül. Ezeken kívül használhatók még a word embedding módszerei egyebek mellett szöveges keresések támogatására vagy fordító programokban is. A különböző típusú használatokhoz, különböző modellek kellhetnek, ezért nagyon fontos a megfelelő szempontok szem előtt tartása a modell kiválasztásakor (Kmetty 2022: 114).

A megfelelő modellválasztáson túl nagyon fontos a korpusznak, azaz a szövegek összességének a kiválasztása is. A korpusz bármilyen digitális formában fellelhető írott szövegből állhat, például közösségi médiás posztokból és kommentekből, folyóiratokból, újságokból, blogposztokból. A használt szövegek lehetnek már eleve digitális módon megjelentek vagy nyomtatásban megjelent művek digitalizált változatai is (Kmetty 2022).

Léteznek nagy, általános, többnyelvű korpuszok, például a CC (Common Crawl) és a Wikikorpusz. Ezek használata elterjedt az ipari alkalmazásban, hiszen ott előnyös, hogy nagyon sok unikális szót tartalmaznak, rengeteg témát lefednek és rendkívül robusztusok a belőlük kapott eredmények. Azonban ezek a társadalomkutatásban nem feltétlenül számítanak előnynek, hiszen a korpuszok nagysága ellenére sem lehet belőlük általánosítani, és a külső érvényességük is kisebb, mint a célhoz szabott, szelektált korpuszoknak (Kmetty 2022: 115-116). Nem csak előre elkészített korpuszok, hanem előkészített vektorok is léteznek, ezeknek nagy előnye az, hogy programozói tudás nélkül is jól használhatók. Ezek alkalmazása is inkább az ipari projektekben elterjedt, a társadalomkutatásokban kevésbé célszerű a használatuk (Kmetty 2022). Mindezek ellenére én a saját pilot kutatásomban egy nagy általános korpuszt, és előkészített vektorokat fogok használni, mivel arra vagyok kíváncsi, hogy egy programozni

nem, vagy csak kevéssé tudó, nagy erőforrásokkal nem rendelkező kutató, hogyan tudja használni a word embedding módszerét társadalomkutatásra.

Korábbi társadalomtudományos kutatások a módszer használatával

Mára egyre több társadalomtudományos kutatásban alkalmazzák a word embedding modelleket. Ebben a részben korábbi kutatások ismertetésének segítségével szeretném bemutatni, hogy milyen módok lehet alkalmazni ezt a módszert. Először is bemutatom, hogy mi az három terület ahol a társadalomtudósok alkalmazhatják a szóbeágyazási modelleket, majd a tartalmi felhasználáson belül mutatok különböző lehetséges módokat szóbeágyazási modellek használatára.

Kmetty (2022) három területet különít el, ahol a társadalomtudósok valamilyen módon használhatnak word embedding modelleket a kutatásaikban. Az első ilyen felhasználási mód a különböző nyelvi modelleken alapuló algoritmusok kritikus vizsgálata. Sok mesterséges intelligenciára épülő alkalmazás használ nyelvi modelleket, ezek a mesterséges intelligencia algoritmusok torzításokat tartalmazhatnak. A torzítást legegyszerűbben a Google Fordító segítségével lehet bemutatni, ezt a jelenséget Kmetty (2022: 127) példáján demonstrálom. Ha beírjuk a fordítóba magyarul, hogy „Az iskolában mindenkiről készült egy jellemzés. Szabóról a következőt mondták. Ő biztos, hogy politikus lesz.” és angolra fordítjuk akkor a következőt kapjuk „A description was made of everyone in the school. The following was said about Szabó. He is sure to be a politician.”. Tehát a program automatikusan férfi személyes névmást rendelt a politikus szóhoz. Ha az előbbi példamondatban kicseréljük a politikus szót tanárra, akkor a következő fordítást kapjuk: „She’s sure to be a teacher.”, tehát a program női személyes névmást rendelt a tanár foglalkozáshoz. Ezekben az esetekben a program mögött álló nyelvi modell abból indul ki, hogy egyes szakmák jellemzően női vagy férfi környezetben szerepelnek-e. Ezt a helyzetet azóta igyekeztek módosítani a Google Fordító készítői, ezért ma már csak ritkán lehet megfigyelni a jelenséget, az esetek nagy részében egymás alatt külön jeleníti meg az egyes nemekre vonatkozó speciális fordításokat a program. Ilyen és ehhez hasonló torzítások sok esetben állhatnak fent, a társadalomtudósok beazonosíthatják az ilyen torzításokat és ezzel elősegíthetik azt, hogy a készítők lépéseket tegyenek a kiküszöbölésükért.

Fontos kiemelni, hogy a torzításokat elsősorban nem a beágyazási algoritmus okozza, hanem az, hogy a bemeneti korpuszok sok sztereotípiát tartalmaznak. Az ilyen jellegű munkák elsősorban módszertani jellegűek (Kmetty 2022). Erről szól például Bolukbasiék (2016) írása is, akik munkájukban bemutatják, hogy a Google News cikkein tanított szóbeágyazások is felfedezhetők férfi és női nemi sztereotípiákat. Ennek javítására egy olyan módszert dolgoztak ki, amivel módosítani tudják a beágyazásokat úgy, hogy eltávolítják a nem kívánt nemi sztereotípiákat, például azt hogy asszociáció van a nő és a recepciós szavak között, és megtartják az olyan kívánt sztereotípiákat, mint a királynő és a nő szavak közötti asszociáció.

A társadalomtudósok számára a word embedding modellek alkalmazásának egy másik módja a technikai felhasználás. Ez azt jelenti, hogy a kutatók valamilyen klasszifikációs modellben használják a módszert. Az ilyen típusú kutatásokban a fókusz nem a szóbeágyazásokon, hanem a klasszifikáció kimenetén van (Kmetty 2022: 128). Yang, Macdonald és Ounis (2018) például egy ilyen jellegű kutatásban használják a szóbeágyazási modelleket, egészen pontosan azzal foglalkoznak, hogy a Twitteren lévő választással kapcsolatos tweetek felderítéséhez használnak word embedding modelleket és összehasonlítják a különböző paramétereken, például különböző háttér korpuszokon (Wikipédia cikkek, Twitter posztok), tanított modellek teljesítményét. A kutatásuk során arra jutottak, hogy a háttéradatoknak minél inkább igazodniuk kell a Twitter klasszifikációs adatbázishoz (adattípusban és időbeliségben is), a lényegesen jobb teljesítmény elérésének érdekében. Emellett azt is megfigyelték, hogy a nagyobb kontextus ablakkal és dimenzióval tanított modellek általában jobban teljesítenek.

A harmadik, Kmetty (2022: 128) által említett alkalmazási mód a tartalmi felhasználás. Az ilyen esetekben konkrétan a vektorok által kimutatott társadalmi összefüggésekre kíváncsiak a kutatók. Az algoritmusok kritikus vizsgálatával kapcsolatban említett torzítások ebben az esetben nem a kijavítandó hibát jelentik, hanem a kutatás központi elemét, valamilyen összefüggés jelét mutatják. Ilyen módon lehet vizsgálni például azt, hogy mely sportok/játékok/zenei stílusok kötődnek az egyes nemekhez/etnikumokhoz. Lehet történeti összevetésekben is használni word embedding modelleket, például miként változott az egyes foglalkozások kötődése a nemekhez az elmúlt száz évben. Rengeteg témában lehet készíteni kutatást a szóbeágyazási modellek tartalmi felhasználásával, mégis az ilyen típusú felhasználás eleinte kevésbé volt népszerű, mint a korábban említettek. Ebben Kozlowski, Taddy és Evans

„The geometry of culture: Analyzing the meanings of class through word embeddings” (2018) című munkája hozott változást. Ők saját kérdőíves kutatást is készítettek azzal a céllal, hogy megvizsgálják egyes témák vektormodellek általi kutathatóságát. Demonstrálni szerették volna a word embedding erejét a társadalomkutatásban, ezért létrehoztak egy survey-t, ami segítségével mérni tudták az asszociációkat a hétköznapi „objektumok” és a rassz, a gender és az osztály kulturális kategóriák között. A kérdőívben rasszra, genderre és osztályra vonatkozó nullától százig terjedő skálákon kellett beazonosítani a kérdezetteknek, hogy egy adott „objektum”, például a steak, hol helyezkedik el. Hét témában (foglalkozás, étel, ruházat, közlekedési eszköz, zenei stílus, sport, keresztnév) kellett ötvenkilenc különböző szót elhelyezniük a résztvevőknek a skálán. A survey és a word embedding eredmények között erős összefüggés fedezhető fel, a genderrel kapcsolatban egyeztek meg leginkább a két módszer által kapott eredmények (Kozlowski–Taddy–Evans 2018). A survey-vel való összevetésen túl azért tudta e kutatás szélesebb körben ismertté tenni a módszer alkalmasságát a társadalomtudományos témák tartalmi kutatására, mert 2019-ben a szociológia egyik legjelentősebb folyóiratában, az American Sociological Review-ban (Amerikai Szociológiai Szemle) jelent meg (Kmetty 2022). Kozlowskiék munkája nem csak ebből a szempontból számít úttörőnek, technikai szempontból is újdonságokat tartalmaz, amikre még a későbbiek során kitérek.

Következőkben három, word embeddinget különböző módon használó tanulmányt szeretnék bemutatni, melyek mind tartalmi felhasználásra alkalmazzák a modelleket. A legkevésbé összetett szóbeágyazási modellekre épülő elemzések a szavak távolságából indulnak ki. Ezt a módszert alkalmazza Kmetty, Koltai és Rudas (2021) is, kutatásukban a foglalkozási struktúrát vizsgálták online szövegeken alapuló word embedding modellekben. Eredményeiket összehasonlították a társadalmi rétegződésről szóló klasszikus eredményekkel, hogy kiderítsék, hasonló eredményeket kaptak-e a word embedding modellek használatával, és hogy megtudják, van-e olyan aspektusa a témának ami korábban a hagyományos vizsgálatok során nem merült fel. Három előre tanított vektorteret és nagy általános korpuszokat alkalmaztak. Az első vektortér a Common Crawl (CC) angol nyelvű szövegein alapult, a második nagyrészt a Wikinews korpuszon, melyben az angol Wikipédia oldalai találhatók, a harmadik szintén a Wikinews korpuszon alapult, de ebben figyelembe vettek szó alatti információkat is, ami azt jelenti, hogy a részben egyforma vagy azonos

gyökereken alapuló szavak, például „sociology” és „society” közelebb kerülhettek egymáshoz ebben a vektortérben.

A továbbiakban a kutatók példáján keresztül szeretném bemutatni, hogy mire alapulnak ezek a modellek. A kutatók feltételezik, hogy bizonyos kulturális tevékenységek közelebb állnak bizonyos foglalkozásokhoz. Ezt a foglalkozás pénzzel és státusszal való összefüggésére alapozzák. Például egy szenátor nagyobb valószínűséggel megy színházba, mint egy gépíró, míg bowlingozni inkább egy gépíró jár, mint egy szenátor. Ahhoz, hogy tesztelni tudjuk a foglalkozások közelségét bizonyos elfoglaltságokhoz, a szavakhoz tartozó vektorok által bezárt szög koszinuszát kell vizsgálnunk. Ez konkrétan a Common Crawl vektorterében úgy nézett ki, hogy a szenátor és színház szavak koszinusz hasonlósága 0,21, a gépíró és színház szavaké 0,12, a szenátor és bowling szavaké 0,05, a gépíró és bowling szavaké pedig 0,16. Ez megerősíti azt a feltételezést, hogy a különböző foglalkozásokhoz eltérő valószínűségek tartoznak az egyéb elfoglaltságokkal kapcsolatban (Kmetty – Koltai – Rudas 2021: 15). A kutatók az írásukban alapvetően inkább a módszertani részletekre helyezték a hangsúlyt, nem pedig az elemzésre, ennek ellenére találtak izgalmas eredményeket, melyek további vizsgálatára érdemes lenne részletesebb kutatásokat végezni. A következőkre jutottak a kutatók az elemzésük során: Alapvető hasonlóság fedezhető fel a szövegelemzés által megfigyelt foglalkozási struktúra és a presztízsskálák és társadalmi távolság skálák által leírt struktúrák között. Felfedezték a foglalkozási struktúra egy olyan dimenzióját, mely tudomásuk szerint eddig nem volt tárgya a rétegződésről szóló diskurzusnak a szociológiában. Ez a hatalmi és a szervezeti szempontok szerepe a foglalkozási struktúra alakulásában. Azzal korábban is foglalkoztak már, hogy a hatalom mértéke kulcsfontosságú a foglalkozás presztízsének szempontjából, újdonságértékkel leginkább az bír, hogy a kutatók felfedezték azt, hogy a szervezeti kapacitásnak önmagában lényeges hatása van a foglalkozás presztízsére (Kmetty – Koltai – Rudas 2021: 26).

A második felhasználási módra térve: klaszterelemzéshez is használhatók a word embedding modellek. A klaszterek kialakítása során „bizonyos jellemzők szerinti hasonlóságuk-különbözőségük alapján sorolják be csoportokba a vizsgált egyedeket” (Németh – Katona – Kmetty 2020: 54). Az ilyen típusú elemzések célja a korpuszunkban valamilyen struktúra felfedezése és az egyes szövegek, esetleg szavak csoportokba rendezése. A szövegek hasonlóságának meghatározásához a legegyszerűbb és legelterjedtebb mód a

szózsákmodellekre épülő klaszterezés alkalmazása (Németh – Katona – Kmetty 2020). Comito, Forestiero és Pizzuti (2019) cikke a szózsákmodellek használatára kínál alternatívát. Egy új megközelítést javasolnak a közösségi médiában megtalálható témák felismeréséhez azzal, hogy bemutatják a szóbeágyazások klaszterezésre való használatának lehetőségét. A Word embedding Clustering (WeC) egy olyan online klaszterezési módszer, ami azonos témákat tárgyaló közösségi média bejegyzéseket csoportosít a bejegyzések tartalmának szemantikai és lexikai jellemzőit egyaránt figyelembe véve. Ez a megközelítés előre tanított szóbeágyazási modelleket használ a rövid szövegek szemantikus reprezentációinak előteremtéséhez. A posztok csoportosítására egy hasonlósági mérőszámot alkalmaz, ami magába foglalja a bejegyzések tartalmának lexikai és szemantikai jellemzőit, tehát az arra vonatkozó információkat, hogy az egyes szavak hányszor szerepelnek a szövegben (lexikai), illetve azokat az információkat melyek a szavak word embeddingre épülő jelentését is tartalmazzák (szemantikai), illetve a poszt kiírásának időpontjára vonatkozó adatokat is. A kutatásban egy nagyjából 50’000, 2015 és 2016 szeptembere között az USA-ban keletkezett tweetet tartalmazó adatbázisból indultak ki, melyben kiszűrték az egészséggel kapcsolatos kulcsszavakat tartalmazó bejegyzéseket és ezek alkották a végső korpuszt. A szóbeágyazási modellek közül egy Word2Vec skip-gram modellt használtak, amit két korpuszon tanítottak be. A posztokban lévő szavak vektorreprezentációit úgy kapták, hogy a skip-gram modellt nagy korpuszon tanították be, és ezzel ellenőrizték, javították a kisebb méretű korpuszokon tanított tématérképeket. A kutatók arra jutottak, hogy a szóbeágyazási vektortérmodelleket használva jelentős pontossági javulás válik lehetővé a klaszterelemzés terén, hiszen olyan nyelvi mintákat és törvényszerűségeket is találtak így, amiket a kizárólag szintaktikai megközelítést alkalmazó elemzésekkel nem (Comito – Forestiero – Pizzuti 2019).

Végül a szóbeágyazás használatának harmadik módját mutatnám be. Korábban is említettem már Kozlowski, Taddy és Evans „The geometry of culture” (2018) című tanulmányát, ez a munka, a survey-vel való validáció mellett, több más, elsősorban technikai szempontból is jelentős. Kutatásuk során történeti összevetést alkalmaznak a teljes 20. századból származó szövegek elemzésével, valamint behoznak egy új módszert a word embedding modellek elemzésébe: egyfajta tengelyeket (dimenziókat) alkalmaznak, melyeket az ellentétes szópárokkal (például: férfi-nő, gazdag-szegény, fekete-fehér, liberális-konzervatív) hoznak létre. Ezek a tengelyek a vektorterekben szorosan megfelelnek a kultúra

különböző dimenzióinak. A kutatók kérdőíves és történelmi adatokkal is megerősítették, hogy a szavak elhelyezkedése ezeken a tengelyeken tükrözi a széles körben elterjedt kulturális dimenziókat. A szavak vektorainak elhelyezkedése a kulturális dimenziók mentén megmutatja, hogy az egyes fogalmak hogyan kapcsolódnak egymáshoz a kulturális kategóriákon belül. A kutatók példáján szemléltetve ez a következőképpen néz ki: a foglalkozások gender tengelyre való vetítésével látjuk, hogy a hagyományosan női foglalkozások (például: dadus, ápoló) a tengely egyik végén helyezkednek el, míg a tradicionálisan férfiasok (például: mérnök, ügyvéd) a másikon. Ez azért van, mert az „ápoló" („nurse") szó környezetében több feminin szó (például: „she", „her", „woman") található a korpusz szövegeiben, így ez a szó a nemi dimenziók női pólusa felé tolódik, míg a „mérnök" („engineer") szó kontextusában olyan szavak vannak inkább amik maszkulinok (például: „his", „him", „man") (Kozlowski–Taddy–Evans 2018: 4). Ez a „dimenziós" megközelítés nem csak a szavak közötti távolságból indul ki, hanem ennek a távolságnak az irányából is, ez pedig rengeteg plusz információhoz juttathat minket. A kulturális tengelyek vektortérbeli felfedezésével lehetőséget nyitottak arra a kutatóknak, hogy megfigyeljék, hogy az egyes „objektum"-ok hol helyezkednek el a különféle kulturális dimenziókban, és hogy ezek a dimenziók hogyan helyezkednek el egymáshoz képest az adott vektortérben. Például azzal, ha az „opera" és „jazz" szavak vektorait rávetítjük a gender dimenzióra megtudhatjuk, hogy a „jazz" vagy az „opera" nőiesebb-e (Kozlowski–Taddy–Evans 2018: 13).

Kozlowskiék (2018) a kutatásukhoz több szóbeágyazási modellt tanítottak be különböző korpuszokon. Az időbeli összevetéshez a Google Ngram korpuszt évtizedekre bontották, így tíz egymástól függetlenül felépített szóbeágyazási modellt kaptak. Ezeknek a modelleknek az összehasonlításával képesek voltak lekövetni a makrokulturális változások mintáit a huszadik században. A nemzet- és kultúraközi összevetéshez két külön modellt tanítottak, az egyiket az Egyesült Államokból, a másikat Nagy Britanniából származó korpuszokon. Azért ezt a két országot választották, mert a közös nyelv és a meglévő kulturális különbségek ideálissá teszik őket az összehasonlításra.

A kutatók számos érdekes eredményre jutottak. Általánosan elmondható, hogy a Google Ngram korpuszon leginkább a gender és a társadalmi osztály tengelyeket sikerült jól megragadni, a rasszt kevésbé. A kulturális dimenziókban zajló történelmi változások során a legtöbb „asszociáció" megmaradt az egyik évtizedről a másikra. A gender tengelyen az „ápoló" következetes, fokozatos mozgást mutat az erős női asszociációtól a gyengébb felé. A „mérnök"

lassú változást mutat, az idő előrehaladtával csökken a férfiassága. Az „újságíró” szakma drámaibb változáson megy át, teljesen megváltozik a nemi konnotációja, a huszadik század elején férfias foglalkozásnak számít, de a század végére inkább nőiessé válik. A társadalmi osztály változások alapvetően nem mozognak együtt a gender asszociációk változásaival. A „nővér” minél kevésbé feminin, annál inkább emelkedik a társadalmi osztályt tekintve. A „mérnök” változatlanul a spektrum közepén található, a felső osztály és a munkásosztály között. Az „újságíró” a középosztályból ellépett a felső osztály felé, ahogy egyre nőiesebb lett (Kozlowski–Taddy–Evans 2018: 39-41). Nem csak a dimenziókon belüli változásokról tettek megállapításokat a kutatók, hanem a dimenziók egymáshoz való viszonyának az idők során bekövetkezett változásairól is. Általában a hasonló jelentésű tengelyek (például: erős-gyenge és nagy-kicsi) koszinusz hasonlósága is nagy, ez azt jelenti, hogy párhuzamosak és, hogy nagy mértékű korrelációt mutatnak az egyes szavak különböző dimenziókban megjelenő vetületei. Erre a következő példát hozták a kutatók: A huszadik század elején a társadalmi osztály és a gender dimenziók szoros összefüggést mutattak úgy, hogy a nőiesség tartozott a felsőbb osztályokhoz, a maszkulinitás pedig az alsóbbakhoz. Ez stabilan így volt a huszadik század első felében, ezt követően a század utolsó évtizedeire ez a kapcsolat gyorsan egy össze nem függő ortogonalitássá csökkent (Kozlowski–Taddy–Evans 2018: 42). A nemzetközi összehasonlítás során, többek között azt vizsgálták, hogyan hat a társadalmiosztályra és a genderre a többi kulturális dimenzió az Amerikai Egyesült Államokban és Nagy Britanniában. Mind a két ország esetében felfedezték azokat a tengelyeket, melyek a legjobban hasonlítanak a gender és az osztály dimenziókra. A hasonlóságok például a következőképpen alakultak a gender esetében. A következő öt kulturális dimenzió áll a legközelebb a genderhez a tizenkilencedik és a huszadik század fordulóján: az Egyesült Államokban a masszív-finom, halk-hangos, gyengéd-kemény, félénk-merész, lágy-kemény dimenziók, míg Nagy Britanniában a gyengéd-kemény, a félénk-merész, a masszív-finom, a halk-hangos és az édes-száraz dimenziók. Az USA-ban a legközelebbi öt tengelyből négy szintén szerepel a Nagy Britanniában legnépszerűbb ötben, tehát mind a két ország esetében alapvetően hasonló szavak tartoznak a férfias és a nőies pólusokhoz (Kozlowski–Taddy–Evans 2018: 44).

Word embedding magyar nyelven

A szóbeágyazási modellek erősen nyelvfüggők. Természetesen ez alatt nem a modell algoritmusának változását kell érteni, hanem az előfeldolgozásnak, a modell paraméterezésének és a kapott vektortérnek a nyelvek szerinti változását. A modell paraméterezésénél például az ablakok nagyságának az adott nyelv alapján való megfelelő kiválasztására kell odafigyelni. Ez az adott nyelvre jellemző dependencialánc (azt méri, hogy az egymásra ható szavak között mekkora az átlagos távolság a nyelvben) hosszától függ, a hosszabb dependencialánccal rendelkező nyelvek, a magyar ilyen, nagyobb méretű ablakot igényelnek (Kmetty 2022:117).

Ahhoz, hogy magyarul lehessen szóbeágyazási modelleket alkalmazni, kellenek magyar nyelvű korpuszok. Szerencsére több nagy méretű, részben vagy egészben az internetről származó magyar nyelvű korpusz létezik, ilyen például a Webkorpusz, a Magyar Nemzeti Szövegtár (MNSZ1 és MNSZ2), a Pázmány korpusz, illetve a Common Crawl-nak is vannak magyar nyelvű szöveges adatai, bár ez utóbbi minősége nem a legjobb, tisztítása lassú és bonyolult folyamat (Indig 2018).

A szöveg előfeldolgozása egyértelműen változik a nyelvvel, hiszen egy magyar nyelvű szövegben a magyar nyelvi szabályok alapján kell azonosítani például a szótöveket. A magyar nyelv agglutináló, azaz ragozó, ezért elemzése nem egyszerű feladat, a megfelelő előfeldolgozás elengedhetetlen. Ezen kívül azért is nehéz a magyar nyelvvel való munka, mert sokkal kisebb a nyelvtechnológiai fejlesztőközösség magyar nyelven (szemben például az angollal), és ez azt eredményezi, hogy a programok sem dolgoznak az angolhoz hasonló pontossággal. Hazánkban a legjelentősebb műhely a MTA-SZTE Mesterséges Intelligencia Kutatócsoport és ennek a Nyelvtechnológiai csoportja (Németh – Katona – Kmetty 2020: 49-50).

Ők készítették a magyarláncot, ami a magyar szövegek nyelvi előfeldolgozására kifejlesztett eszköztár (Zsibrita – Vincze – Farkas 2013). Ezen kívül ehhez a kutatócsoporthoz kötődnek a Szántó, Vincze és Farkas (2017) által fejlesztett magyar nyelvű publikusan elérhető szóbeágyazási vektortérmodellek is. A modellekhez kapcsolódó vizsgálatukban arra jutottak, hogy az olyan morfológiailag gazdag nyelvek esetében, mint amilyen a magyar, a karakterszintű szóbeágyazási modellek használata sokkal előnyösebb, mint a szószintűeké. A kutatók a modelljeiket minél nagyobb és változatosabb korpuszra szerették volna építeni, ezért az MNSZ2 korpuszt kiegészítették a Hunglish korpusz magyar anyagaival, origo.hu és index.hu cikkekkel, gyakorikerdesek.hu-ról származó szövegekkel és az OpenSubtitles oldal

magyar rajongó feliratával. A szövegek szavakra bontását a magyarlánc beépített tokenizálójával végezték, így egy 4,29 milliárd szóból álló korpuszt kaptak. Ezen a korpuszon tanították be a szó- és karakterszintű szóbeágyazási modelljeiket, melyek nyilvánosan elérhetőek.

A következőkben szeretném bemutatni Szabó és társai (2020) már korábban is említett munkáját, főleg az előfeldolgozás lépéseinek hosszú és bonyolult sorára koncentrálva. A tanulmányban a Pártélet folyóiratból (1956-1989) készített nagy korpuszban vizsgálták az ipar és a mezőgazdaság témaköréhez kapcsolódó egyes fogalmak szemantikai kapcsolatát. Ennek kutatása azért különösen érdekes, mert a Pártélet a kormánypárt hivatalos lapja volt, így meg lehet vizsgálni a kormány hivatalos álláspontjának változásait is. A folyóirat szkennelt pdf formájában elérhető az Arcanum oldalán, ebből indultak ki a kutatók. Ezt először is NLP eszközökkel elemezhető digitális szöveggé kellett alakítani, ehhez egy szövegfelismerő (Optical Character Recognition, OCR) programot alkalmaztak. A kapott szöveget magyarlánc segítségével feldolgozták. Ezt követően egy hunspell nevű helyesírás-ellenőrző segítségével átnézték a szavakat, a rendszer nem ismerte az egyedi szavak tizennégy százalékát. Ha az adott ismeretlen szó minimum tízszer megjelent a korpuszban és minimum három karakterből állt, akkor a kutatók kézzel is ellenőrizték, így csökkenteni tudták az ismeretlen szavak arányát. Végül egy 9’432’200 tokenből álló korpuszt kaptak. Ezen a korpuszon tanították a GloVe algoritmusra épülő word embedding modeljüket. A vektortérben megvizsgálták mind a mezőgazdasághoz, mind az iparhoz legközelebb álló szavakat a korszak során. A mezőgazdasághoz a legtöbb évben az „ipar”, a „termelés”, a „fejlesztés”, a „haladás” és a „szocialista” szavak voltak a legközelebb. Ezeken kívül voltak olyan szavak, melyek bizonyos időszakokban nagyon közel kerültek a mezőgazdasághoz, aztán eltűntek, ilyen volt a „terv” és a „célkitűzés” a 60-as években, az „eredmény” a 70-esekben és a „termék” a 80-asokban. Az iparhoz a legtöbb évben a „mezőgazdaság”, a „termelés”, a „haladás”, a „fejlődés” és a „népgazdaság” szavak voltak legközelebb. Ezeken túl a korszak első éveiben feltűntek a „terv” és a „szocialista” szavak is. A 70-es években a „munkavállaló”, a 80-asokban a „haza” szavak is nagyon közel voltak az iparhoz. Összességében az derült ki a kutatásból, hogy az ipar és a mezőgazdaság szemantikailag közel voltak egymáshoz a teljes időperiódus alatt. Az egész időszak viszonylag stabil közelségeket mutat, amik csak a korszak végén kezdtek meredeken csökkenni. Ez valószínűleg az egyre közelebb kerülő szabadpiaci kapitalizmusnak volt köszönhető (Szabó et al. 2020: 7).

Saját pilot kutatás

A pilot kutatásom elkészítése előtt több, angol nyelvre működő online demó programot vizsgáltam meg, például Liu (2024), a Turku Egyetem (2024), a Carnegie Mellon Egyetem (2024) és a WebVectors (2024) modelljét. A WebVectors-al kapcsolatban találtam a legtöbb megbízható irodalmat, részletes leírást a modellek készítéséről, a korpuszok tartalmáról és jól követhető, felhasználóbarát használati utasítást az online elérhető demók használatához. Ezen kívül nagyon érdekesnek találtam az oldalnak azt a funkcióját, hogy online egyszerre akár négy különböző modell eredményeit is meg lehet tekinteni egy adott kérdéssel kapcsolatban, így akár össze is lehet vetni a különböző korpuszokon tanított modellek eredményeit. A következőkben bemutatom a WebVectors-t, az oldalon elérhető modelleket és azok funkcióit, majd elkészítem saját pilot kutatásom ezen a felületen.

Fares, Kutuzov, Oepen és Velldal (2017) létrehoztak egy online tárházat, melyen meg lehet osztani különböző adatokat a modellek betanításához, a szövegek előfeldolgozásának részleteit, vagy akár kész, előre tanított szóbeágyazási modelleket. A szerzők több okból is nagyon indokoltnak tartják egy ilyen jellegű weboldal létrehozását. A word embedding modellek betanítása időigényes lehet, illetve programozói ismereteket igényel, tehát a már előfeldolgozott korpuszok és kész vektorterek felhasználásának lehetősége egy saját modellben, vagy a kész modellek használatának lehetősége nagy segítség lehet a kutatók számára, ezen kívül az eredmények összehasonlíthatóságát és reprodukálhatóságát is javítja.

A cikk szerzői ezért kezdeményezik egy olyan mindenki számára hozzáférhető tárház létrehozását, ami nagy méretű szöveges forrásokat tartalmaz szóbeágyazási modellek számára, beleértve az előfeldolgozott korpuszokat és az előre tanított vektortérmodelleket, melyek sok féle keretrendszer és megközelítési mód számára jól használhatóak. Egy interaktív webalkalmazás segítségével a felhasználók online is felfedezhetik és összehasonlíthatják a különböző előre betanított modelleket. Azért fontos ezeknek az előre tanított modelleknek és előkészített vektortereknek az elérhetősége, mert ez biztosíthatja, hogy ugyanazokat a beágyazásokat használhassák több kutatásban is, és így jobban össze lehessen hasonlítani az eredményeket, valamint külön lehessen tesztelni az egyéb faktorok hatását az eredményre.

Az előkészített vektorterek jelenleg elérhetőek jónéhány algoritmus (például: GloVe, fastText, skip-gram, CBOW) számára. Még abban az esetben is nehéz megállapítani, hogy minek köszönhető az eltérés/hasonlóság, ha azonos adatbázison tanított szóbeágyazásokat szeretnénk összehasonlítani, mert a kapott eredményt nagyon befolyásolja a korpusz előfeldolgozásának módja és a használt word embedding algoritmus. A szerzők szerint, a tanulmány írásakor az elérhető előkészített vektorterek elég limitáltak voltak a választható korpuszok, illetve azok előfeldolgozottsága szempontjából. Ennek ellenére sokan használták őket az általuk nyújtott kényelem miatt. Ez sok esetben azt eredményezte, hogy a kutatásokban nem az adott feladatra kifejezetten alkalmas beágyazásokat alkalmazták. A lehető legpontosabb eredmény eléréséhez ugyanis nem elegendő a megfelelő szóbeágyazási algoritmus és korpusz kiválasztása, az előfeldolgozás lépéseinek gondos megválogatása gyakran fontosabb tényező ezeknél a teljesítmény szempontjából (Fares – Kutuzov – Oepen – Velldal 2017).

A kutatók által létrehozott weboldal (http://vectors.nlpl.eu) segítséget nyújt a szóbeágyazási modell létrehozásához szükséges lépések, például a megfelelő vektortér, a korpusz és a előfeldolgozási mód kiválasztásában. A tárhelyre fel lehet tölteni többek között tanító adatokat, előfeldolgozási folyamatokat és előre tanított szóbeágyazási modelleket is. Az így létrejövő információhalmaz segíti a kutatókat a modellek építésében, a megfelelő döntések meghozatalában, javítja a kutatások reprodukálhatóságát és lehetővé teszi a már létrehozott korpuszok és modellek újrahasználhatóságát egy másik kutatásban (Fares – Kutuzov – Oepen – Velldal 2017).

Az oldal biztosítja a WebVectors webszolgáltatást is, amely előkészített modelleket tartalmaz angol, norvég valamint egy kiegészítő együttműködés révén orosz nyelvre. Ez az eszköztár megkönnyíti a word embedding modellek használatát a mindennapi kutatásban, hiszen egy könnyen kezelhető és gyors rendszerről van szó, amit az egyéni igények alapján lehet alakítani (Fares – Kutuzov – Oepen – Velldal 2017).

A WebVector használata ingyenes és nem kötött semmiféle regisztrációhoz, így bárki számára könnyen hozzáférhető. Ezen kívül az is előnye a rendszernek, hogy nyílt forráskóddal rendelkezik. Használata lehetővé teszi a számítógépekhez kevésbé értő, programozni nem tudó kutatók számára is a szóbeágyazási modellek használatát, hozzáférést biztosít a megfelelő eszközökhöz a nyelvvel különböző módokon foglalkozó kutatók számára, ezzel

elősegítve a kutatásokat a különböző tudományterületeken (Fares – Kutuzov – Oepen – Velldal 2017).

A következőkben bemutatom a WebVectors eszköztár fő funkcióit. Lehetőség van az oldalon szemantikai hasonlóságok kiszámítására szópárok között, ezt a vektorok koszinusz hasonlóságának kiszámításával teszi a rendszer. Tehát a hasonlóság értéke egy 1 és -1 közötti szám lesz. A 0 érték azt jelenti, hogy a szavaknak nincs hasonló kontextusa, jelentésük nem kapcsolódik egymáshoz. Az 1 érték azt jelenti, hogy a szavak kontextusa teljesen azonos, így jelentésük nagyon hasonló, míg -1 esetén jelentésük ellentétes. Az oldalon lehetséges a szemantikai asszociációk keresése is. Ez azt jelenti, hogy a kérdezett szóhoz legközelebb álló tíz szót és a hozzájuk tartozó hasonlósági értéket felsorolja számunkra a program. A WebVectors képes algebrai műveletek, például összeadás, kivonás elvégzésére, az eredményt a művelet eredményéhez legközelebb eső szavak listájaként és a hozzájuk tartozó közelség értékeként kapjuk meg. Ezt a funkciót lehet alkalmazni analógiák elkészítéséhez is, ami nagyon gyakran használt módja a szóbeágyazások használatának. Az oldal alkalmas a szavak közötti szemantikai kapcsolat vizualizációjára is. Ez úgy történik, hogy ha beírunk egy szókészletet, az alkalmazás elkészíti a választott modellben az adott szavak között lévő belső kapcsolat térképét, majd visszaadja számunkra ennek a kétdimenziós változatát. A WebVectors ötödik funkciója pedig a kérdezett szóhoz tartozó nyers vektor bemutatása (Kutuzov – Kuzmenko 2017).

Mindegyik korábban említett funkció esetében lehetőség van part of speech (PoS) szűrők alkalmazására (ez a szófajok és egyéb nyelvészeti kategóriák jelölését jelenti (Németh – Katona – Kmetty 2020)), persze csak azokban az esetekben, ha abban a korpuszban amire a modell épül címkézték ezeket az elemeket. Ezen kívül mindig lehetőségünk van arra, hogy bejelöljük, hogy a korpuszban mennyire gyakori szavak jelenjenek meg nekünk, itt három opció közül választhatunk, megjelenhetnek a gyakori, a közepesen gyakori és a ritka szavak is, illetve ezek közül egyszerre többet is tudunk választani. Lehetséges az oldalon egyszerre több modell eredményeit is behívni, ezek egymás mellett fognak megjelenni a képernyőn. E funkció hasznos lehet a különböző, például eltérő korpuszon tanított vagy különböző előfeldolgozottságú, modelleket összehasonlító kutatások során (Kutuzov – Kuzmenko 2017).

Az oldalon összesen négy angol nyelvű modell érhető el, ezek mindegyike skip-gram algoritmus által lett tanítva. A következőkben ezt a négy modellt, illetve a hozzájuk használt

korpuszokat fogom bemutatni. Az English Wikipedia korpuszon alapuló modell a 2021 novemberéig angol nyelven megjelent Wikipédia cikkeket tartalmazza. Nagyjából három milliárd tokent tartalmaz és 199’430 különböző lemmatizált, a szavak alapalakjának megkeresésével szótövesített (Kmetty 2022) angol szót ismer (WebVectors Models 2024). Előfeldolgozottságát tekintve a korpusz mondatokra, illetve kifejezésekre, szavakra van bontva, tehát tokenizálva (Németh – Katona – Kmetty 2020) van. Ezen kívül lemmatizálva van és el van látva PoS-címkékkel, illetve a stop word removal is el van végezve rajta, ami a tartalommal nem rendelkező szavak, például névelők eltávolítását jelenti a szövegből (Németh – Katona – Kmetty 2020). Végül a szorosan kapcsolódó több tagból álló kollokációkat, például United Kingdom (Egyesült Királyság), kijelölték a korpuszban, hogy a részek egyben is megkaphassák a helyüket a modellben (Kutuzov – Kuzmenko 2017).

A második modell az English Gigaword korpusz ötödik kiadásán alapul, ami 2010 decemberéig tartalmazza különböző nemzetközi hírügynökségek angol nyelvű szövegeit. A következő forrásokat tartalmazza a korpusz: Agence France-Presse, Associated Press Worldstream, Central News Agency of Taiwan, Los Angeles Times/Washington Post Newswire, Washington Post/Bloomberg Newswire, York Times Newswire, Xinhua News Agency. Tehát szerepel benne Franciaországból, az Egyesült Államokból, Tajvanból és Kínából származó szöveg is (English Gigaword Fifth Edition 2024). Összesen 4.8 milliárd tokent tartalmaz és 297’790 lemmatizált angol szót (WebVectors Models 2024). Előfeldolgozottsága megegyezik az English Wikipedia korpuszéval, tehát tokenizálva, lemmatizálva és PoS-címkézve van, ezen kívül ki vannak benne jelölve a szorosan összefüggő kollokációk és el vannak távolítva a jelentés nélküli szavak (Kutuzov – Kuzmenko 2017).

A harmadik és egyben utolsó olyan modell melyet Kutuzovék maguk tanítottak a British National korpuszra alapul, amit arra hoztak létre, hogy írott és szóbeli széleskörű források által képviselje a brit angol széles keresztmetszetét a késő huszadik században (British National Corpus 2024). Ez a korpusz 98 millió tokenből épül fel, és összesen 163’476 angol szót tartalmaz. Előfeldolgozása megegyezik az előző kettőével (WebVectors Models 2024).

Az utolsó angolnyelvű modell a Google News adatbázison alapul. Ezt a korpuszt eredetileg Mikolovék publikálták a word2vec-kel együtt, ezért ennek használata elterjedt (Kutuzov – Kuzmenko 2017). Például a Turkui Egyetem NLP csoportja által fejlesztett online elérhető angol nyelvű szóbeágyazási modell is erre alapul (Turku University word embedding demo 2024). A Google hír gyűjteményen alapuló korpusz összesen 100 milliárd tokent és 2,9

milliárd szót, illetve kifejezést tartalmaz, tehát ez a legnagyobb a négy használt korpusz közül (WebVectors Models 2024). Előfeldolgozottságában eltér a korábban említett korpuszoktól abban, hogy nincs lemmatizálva. Csak tokenizálva van és ki vannak jelölve benne a kollokációk, illetve jobb összehasonlíthatóság érdekében Kutuzovék utólag PoS-címkékkel látták el a szavakat (Kutuzov – Kuzmenko 2017).

A nemi sztereotípiák vizsgálatára már több társadalomtudományos kutatás is készült word embedding modellekkel. Léteznek olyanok, melyek kisebb, konkrétabb korpuszokon vizsgálják a kérdést (például Adukia és társai (2022) a nemi szerepeket vizsgálták egy gyermekkönyvekből álló korpuszon), illetve vannak olyan kutatások, melyek nagy általános korpuszokon teszik ugyanezt. A saját pilot kutatásomban azt vizsgálom, hogy a WebVectors oldalon elérhető négy szóbeágyazási modell segítségével mennyire lehet kutatni a témát. A korábban mások által megfigyelt eredmények megjelennek-e és vizsgálhatóak-e ezeken a viszonylag egyszerű felületeken keresztül releváns társadalomtudományos témák.

Mini kutatásomat Chaloner és Maldonado (2019) munkája inspirálta. Ők WEAT hipotézis tesztelési módszer segítségével kutatták a nemi előítéleteket. A WEAT (Word Embedding Association Test) egy statisztikai teszt, ami a szóbeágyazási modellekben lévő torzításokat észleli a koszinusz hasonlóságok és átlagok vizsgálatát párosítva a hipotézis teszteléssel. Vizsgálatukban négy különböző jellegű korpuszon betanított modellt alkalmaztak, egészen pontosan egy híreken alapuló (Google News), egy közösségi médián alapuló (Twitter), egy biomedikális (PubMed) és egy Wikipédián alapuló, nemek tekintetében kiegyensúlyozott (GAP), korpuszt használtak a modellekhez. Ők olyan korábbi kutatások eredményeiből indultak ki, melyek azt mondták, hogy a nemi előítéletek leginkább öt kategóriában érhetőek tetten a szövegekben. Az öt kategória a következő: karrier és család, matematika és művészet, tudomány és művészet, intelligencia és megjelenés, erősség és gyengeség. Arra az eredményre jutottak, hogy a Google News korpuszon mind az öt előítéletességi kategóriát megtalálták, a másik három korpuszban viszont kevésbé mutatkoztak meg ezek, legkevésbé a biomedikális korpuszra volt jellemző az ilyen jellegű nemi sztereotípiák megléte (Chaloner – Maldonado 2019).

A kutatásom során a WEAT-nél sokkal egyszerűbb szóbeágyazási módszerek álltak rendelkezésemre a WebVectors felületén, illetve a számomra elérhető korpuszok sem annyira sokrétűek, mint amit Chalonerék használtak. Azt szerettem volna megvizsgálni, hogy a

számomra elérhető módszerek segítségével is megfigyelhető-e mind az öt nemi előítéletességi kategória Google News korpuszban.

A Chaloner és Maldonado (2019) tesztjükhöz létrehoztak úgynevezett „cél szavak”-ból álló listákat, ezek olyan szavakat tartalmaznak melyekről feltételezhető, hogy előítéleteket hordoznak az egyik vagy a másik nem irányába. Ezen kívül létrehoztak még egy listát amin egyértelműen férfi vagy női fogalmakat tüntettek fel. Kutatásomban én is ezekből a listákból indultam ki (M1), a nemekre vonatkozó szavak közül kiválasztottam kettő-kettőt, a „férfi”-t („man”) és a „nő”-t („woman”), illetve az „apa”-t („father”) és az „anya”-t („mother”). A WebVectoton elérhető szópárok egymáshoz való hasonlóságát megmutató funkciót választottam a kutatásomhoz. Az egyes előítéletességi kategóriák szavainak és az imént említett négy nemekre vonatkozó szónak az egymáshoz való közelségét vizsgáltam a Google News korpuszon tanított modell segítségével.

Létrehoztam egy Excel fájlt (M2), melynek táblázataiba rögzítettem az eredményeket. Minden szó esetében jelöltem, hogy a „man”-„woman”, illetve a „father”-„mother” párok melyik tagjához vannak közelebb. Ezt követően kiszámítottam az adott szó „férfi” és „nő” szavakhoz való távolságának különbségét és ezt megismételtem minden szó esetében az „apa” és az „anya” szavakkal is. Ha ennek a különbségnek az abszolút értéke nagyobb volt, mint 0,05 és kisebb, mint 0,1 akkor világoskék színnel jelöltem, ha 0,1-nél is nagyobb volt akkor sötéttel. Az alapján, hogy egy adott kategóriában (például karrier) lévő szavak közül hány állt a nőkhöz, illetve a férfiakhoz közelebb, illetve az alapján, hogy a jelentősebb különbségek melyik nem javára szóltak megállapítottam, hogy az egyes kategóriákhoz tartozó nemi sztereotípiák mennyire érzékelhetőek az általam használt módszer segítségével a Google News korpuszban.

A karrier és család kategóriájában a karrierhez tartozó szavak közül ugyan annyi állt közelebb a „nő” szóhoz, mint a „férfi”-hoz és a közelségek közti eltérések sem voltak nagyon jelentősek. Amikor viszont a „nő” és „férfi” szavak helyett az „anya” és az „apa” szavakkal való távolságokat vizsgáltam, akkor azt láttam, hogy a szavak nagy része (nyolc szóból hat) az apához van közelebb és a különbségek abszolútértéke is viszonylag nagy volt a „hivatásos” („professional”) (0,083) és főleg a „karrier” („career”) (0,12) szavak esetében, mind a két szó az „apa”-hoz állt közelebb. A családhoz tartozó szavakhoz alapvetően közelebb álltak a nemeket képviselő szavak, mint a karriert képviselőkhöz. A „nő”-„férfi”, mind az „apa”-„anya”

szavak esetében is az volt megfigyelhető, hogy a család szavainak nagy része közelebb áll a női szavakhoz. A közelségek közti különbségek a „nő”-„férfi” szavaknál három esetben is kifejezetten nagyok voltak, a „szülő” („parent”) (0,127), a „gyerek” („child”) (0,159) és a „házasság” „marriage” (0,114) szavak esetében úgy, hogy mindegyik szó a „nő”-höz volt közelebb. Tehát elmondható, hogy a karrier-család kategóriába tartozó nemi sztereotípiák megjelennek a korpuszban, hiszen a karrierrel kapcsolatos szavak valamivel közelebb voltak a férfiakhoz kapcsolódó szavakhoz és a családhoz tartozó szavak határozottan közelebb voltak a női szavakhoz.

A matematikához kapcsolódó szavak hasonló közelségben voltak a női és férfi szavakhoz egyaránt, az egyes közelségek különbségei közt sem voltak nagyon kiugró értékek. A művészet esetében mind a nyolc szó a nőhöz tartozott, a szülőknél hét szó volt közelebb az „anya”-hoz, és csak egy az „apa”-hoz. A tudomány esetében a szavak a „férfi”-„nő” összehasonlítás esetén két esetben voltak közelebb a „nő”-höz, öt esetben a „férfi”-hoz, egy szó pedig ugyanannyira volt közel mindkettőhöz. Az „apa”-„anya” szópár esetében hasonló volt a helyzet, az „apa”-hoz öt, az „anya”-hoz három szó volt közelebb. A matematika és művészet kategóriában, tehát részben megfigyelhetőek a nemi sztereotípiák, hiszen a művészetek egyértelműen közelebb vannak a női szavakhoz, viszont a matematikához kapcsolódó szavak hasonló közelségben vannak a mind a két nemmel. A tudomány és művészet kategóriában valamivel jobban érzékelhetőek a nemi előítéletek, hiszen a tudományokkal kapcsolatos szavak nagy része közelebb van a férfi szavakhoz.

Az erősség kategóriában mind a „nő”-„férfi”, mind az „apa”-„anya" szópárok esetében két szó tartozott a női és tizenkettő a férfi szavakhoz, ezen kívül azokban az esetekben, mikor nagy különbség volt a két közelség között, mindig a férfiakhoz álltak közelebb a szavak. A gyengeség esetében már nem volt ilyen egyértelmű a helyzet, mivel a „nő” szóhoz hat, a „férfi” szóhoz nyolc, illetve az „anya” szóhoz kilenc, az „apa” szóhoz pedig öt szó volt közelebb. A nagyobb különbségek esetében, legtöbbször a női szavak voltak közelebb a gyengeséggel kapcsolatos szavakhoz. Ebben az esetben elmondható tehát, hogy az erősséghez kötődő szavak esetében egyértelműen látszik a nemi előítélet, de a gyengeséggel kapcsolatban ez már nem feltétlenül van így, hiszen csak kevéssel vannak közelebb az ilyen jellegű szavak a nőiekhez.

Az intelligencia és megjelenés kategóriában az intelligenciához tartozó szavak egyértelműen közelebb voltak a „férfi” és az „apa” szavakhoz, huszonkettő és huszonhárom

szó volt hozzájuk közelebb, a nőkhöz tartozó három, illetve két szó ellenében. A megjelenéssel kapcsolatos szavak alapvetően közelebb vannak a női szavakhoz, bár a „férfi” és „nő” szavakhoz ugyanúgy tizenegy-tizenegy szó van közelebb. A közelségek közti nagy különbségeket vizsgálva azt látjuk, hogy a „csábos” („alluring”) (0,118), a „kéjes” („voluptuous”) (0,158), az „érzéki” („sensual”) (0,137) és a „divatos” („fashionable”) (0,125) szavak sokkal közelebb vannak a „nő”-höz, mint a „férfi”-hoz. Az „apa” és az „anya” esetében még jobban megfigyelhetőek az eltérő közelségek, míg az előbbihez öt szó volt közelebb, addig az utóbbihoz tizenhét. Tehát ebben az esetben jól megfigyelhetőek a nemi sztereotípiák.

Összességében elmondható, hogy valamilyen szintű nemi sztereotípiákat minden kategória esetében meg lehet figyelni a Google News korpuszán egy egyszerű, online szóbeágyazási modell segítségével is.

Konklúzió

Jelenleg egy digitális korban élünk, mely magával hozta új eszközök és technológiák kifejlődését a társadalomkutatásban is. Ezek közé az új módszerek közé tartozik a számítógépes szövegelemzés is, aminek az egyik módszere a szóbeágyazási (word embedding) modellek használata. Ezek a modellek neurális hálók segítségével több dimenziós vektortérbe helyezik a korpuszuk szavait, ebben a vektortérben a szavak úgy helyezkednek el, hogy a hasonló jelentésű szavak egymáshoz közel, míg a különbözőek egymástól távol vannak (Németh – Katona – Kmetty 2020). Ennek a típusú modellnek az elterjedésében Mikolov és társai (2013) munkája jelentett áttörést, mert gyorsabbá és több ember számára elérhetővé tette a módszert.

Kmetty (2022) már írt korábban a szóbeágyazási vektortérmodellek társadalomtudományi alkalmazásáról, ő azt állapította meg, hogy lehetséges társadalmi jelenségek vizsgálatára alkalmazni a módszert, de ehhez nagyon fontos a modell és a korpusz megfelelő kiválasztása. A társadalomkutatásban nem feltétlenül számít előnyősnek a nagy, általános korpuszok használata, és inkább a célra szabott, szelektált, akár kifejezetten kis méretű korpuszok használatát javasolja, illetve az előkészített vektorterek alkalmazását sem tartja célszerűnek.

Kmetty (2022) három fajta társadalomtudományos felhasználási módjáról ír a word embedding modelleknek. A nyelvi modelleken alapuló algoritmusok kritikus vizsgálata az egyik ilyen felhasználási mód, ezt alkalmazzák például Bolukbasi és társai (2016). Egy másik mód a technikai felhasználás valamilyen klasszifikációs modellben, ezt alkalmazza például Yang, Macdonald és Ounis (2018). A harmadik típusú felhasználási mód a tartalmi felhasználás, ebben az esetben a kutatók a vektorok által kimutatott társadalmi összefüggéseket vizsgálják. Az ilyen típusú kutatásokban Kozlowski és társai (2018) kutatása hozott áttörést, ők egy saját kérdőíves kutatással is alátámasztották eredményeiket.

Tartalmi felhasználásra több módon is lehet alkalmazni szóbeágyazási modelleket. Az elemzések kiindulhatnak szimplán a szavak távolságainak vizsgálatából, ilyen például Kmetty, Koltai és Rudas (2021) foglalkozási struktúrákról szóló kutatása. Klaszterelemzéshez is használhatóak a word embedding modellek, így alkalmazza a technikát például Comito, Forestiero és Pizzuti (2019). Tengelyeket alkalmazva is lehet vizsgálni a modellek eredményeit. Ebben is Kozlowski és társai (2018) munkája tekinthető úttörőnek, akik ezen az újításon kívül még történeti összevetést is alkalmaztak. Magyar nyelvű szövegekből is készülnek kutatások word embedding módszerrel. Ilyen például Szabó és társai (2020) kutatása, akik szintén alkalmaztak időbeli összevetést is a munkájukban.

Tehát napjainkban már egész sok témában és sokféle módon alkalmazzák a szóbeágyazási modelleket a kutatók. Saját, nemi előítéletekről szóló pilot kutatásom segítségével pedig azt is szemléltettem, hogy a programozni nem tudó kutatók is tudják alkalmazni a módszert a WebVectors-hoz hasonló, online elérhető word embedding modellek segítségével.

Ugyanakkor megvannak az online elérhető, előre tanított modelleknek is a hátrányai. Az egyik legnagyobb hátrányuk az, hogy nem adható meg saját korpusz a használatukhoz, így kénytelenek vagyunk az elérhető nagy méretű korpuszok közül választani, melyek nem mindig a legideálisabbak a társadalomkutatáshoz, valamint nem tudjuk kombinálni a kutatásunk során a módszert kvalitatív módszerekkel, mert nincs lehetőségünk visszamenni az eredeti korpuszba megvizsgálni azt. Harmadik hátrányuk az, hogy ezek az online modellek (még) nem érhetők el magyar nyelven, ez lekorlátozza az általuk kutatható témákat.

Irodalomjegyzék

Adukia, Anjali – Chiril, Patricia – Christ, Callista – Das, Anjali – Eble, Alex – Harrison, Emileigh – Runesha, Hakizumwami Birali (2022): Tales and tropes: Gender roles from word embeddings in a century of children's books. In Calzolari, Nicoletta – Huang, Chu-Ren – Kim, Hansaem – Pustejovsky, James – Wanner, Leo – Choi, Key-Sun – Ryu, Pum-Mo – Chen, Hsin-Hsi – Donatelli, Lucia – Ji, Heng – Kurohashi, Sadao – Paggio, Patrizia – Xue, Nianwen – Kim, Seokhwan – Hahm, Younggyun – He, Zhong – Lee, Tony Kyungil – Santus, Enrico – Bond, Francis – Na, Seung-Hoon (eds.): *Proceedings of the 29th International Conference on Computational Linguistics.* Kjongdzsu: International Committee on Computational Linguistics, 3086–3097.

Babbie, Earl (1999): *A társadalomtudományi kutatás gyakorlata*. Budapest: Balassi Kiadó.

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

Bolukbasi, Tolga – Chang, Kai-Wei – Zou, James – Saligrama, Venkatesh – Kalai, Adam (2016): Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: Lee, Daniel D. – von Luxburg, Ulrike – Garnett, Roman – Sugiyama, Masashi – Guyon, Isabelle (eds.): *Proceedings of the 30th International Conference on Neural Information Processing Systems*. Curran Associates Inc. Red Hook, 4356–4364.

British National Corpus. (http://www.natcorp.ox.ac.uk/) (Utolsó megtekintés: 2024. április 7.)

Carnegie Mellon Universty Word Embedding Demo (2022). (https://www.cs.cmu.edu/~dst/WordEmbeddingDemo/index.html) (Utolsó megtekintés: 2024. április 7.)

Chaloner, Kaytlin – Maldonado, Alfredo (2019): Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. In Costa-jussà, Marta R. – Hardmeier, Christian – Radford, Will – Webster, Kellie (eds.): *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Firenze: Association for Computational Linguistics, 25–32.

Comito, Carmela – Forestiero, Agostino – Pizzuti, Clara (2019): Word embedding based clustering to detect topics in social media. In: *IEEE/WIC/ACM International Conference on Web Intelligence*. Szaloniki: Institute of Electrical and Electronics Engineers, 192–199.

English Gigaword Fifth Edition. (https://catalog.ldc.upenn.edu/LDC2011T07) (Utolsó megtekintés: 2024. április 6.)

Fares, Murhaf – Kutuzov, Andrei – Oepen, Stephan – Velldal, Erik (2017): Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In Tiedemann, Jörg (ed.): *Proceedings of the 21st Nordic Conference on Computational Linguistics (NoDaLiDa).* Göteborg: Linköping University Electronic Press, 271–276.

Giddens, Anthony (2008): *Szociológia*. Budapest: Osiris Kiadó.

Indig Balázs (2018): Közös crawlnak is egy korpusz a vége – Korpuszépítés a CommonCrawl.hu domainjaiból. In: Vincze Veronika (szerk.): *XIV. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged: Szegedi Tudományegyetem, 125–134.

Király Gábor – Dén-Nagy Ildikó – Géring Zsuzsanna – Nagy Beáta (2014): Kevert módszertani megközelítések. Elméleti és módszertani alapok. *Kultúra és közösség,* 5 (2): 95–104.

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

Kmetty Zoltán (2022): Szóbeágyazási vektortérmodellek társadalomtudományi alkalmazása. *Statisztikai Szemle*, 100 (2): 105–136 https://real.mtak.hu/138404/1/2022_02_105.pdf

Kmetty Zoltán – Koltai Júlia – Rudas Tamás (2021): The presence of occupational structure in online texts based on word embedding NLP models. *EPJ Data Science,* 10 (55): 1-20

Kozlowski, Austin C. – Taddy, Matt – Evans, James A. (2018): The Geometry of Culture: Analyzing Meaning through Word Embeddings. *American Sociological Review*, 84 (5): 905–949.

Kutuzov, Andrey – Kuzmenko, Elizaveta (2017). Building WebInterfaces for Vector Semantic Models with the WebVectors Toolkit. In Peñas, Anselmo – Martins, Andre (eds.): *Proceedings of the Demonstrations at the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Valencia: Association for Computational Linguistics (ACL), 99–103.

Laki László János (2018): Mesterséges intelligencia a gépi fordításban. In Tolcsvai Nagy Gábor (szerk.): *A humán tudományok és a gépi intelligencia*. Budapest: Gondolat Kiadó, 156–183.

Liu, Anthony Word2Vec JS Demo. (https://turbomaze.github.io/word2vecjson/) (Utolsó megtekintés: 2024. április 1.)

Mikolov, Tomas – Sutskever, Ilya – Chen, Kai – Corrado, Greg S. – Dean, Jeff (2013): Distributed Representations of Words and Phrases and their Compositionality. Burges, Christopher J.C. – Bottou, Leon – Welling, Max – Ghahramani, Zoubin – Weinberger, Kilian Q. (eds.): *Proceedings of the Conference on Advances in Neural Information Processing Systems 26 (NIPS).* La Jolla: Neural Information Processing Systems Foundation, 3136–3144.

Németh Renáta (2015): A számok tényleg magukért beszélnek?. Hozzászólás Dessewffy Tibor és Láng László írásához. *Replika,* 92–93: 203–208.

Németh Renáta – Katona Eszter Rita – Kmetty Zoltán (2020): Az automatizált szövegelemzés perspektívája a társadalomtudományokban. *Szociológiai Szemle*, 30 (1): 44–62.

Németh Renáta – Sik Domonkos – Zaboretzky Bendegúz – Katona Eszter (2023): Depression in times of a pandemic–the impact of COVID-19 on the lay discourses of e-mental health communities. *Information, Communication & Society,* 27 (3): 1–23.

Salganik, Matthew J. (2019) *Bit by bit: Social research in the digital age*. Princeton University Press.

Sallay Viola (2015): A kvalitatív megközelítés és a Grounded Theory szerepe a társadalomtudományi kutatásokban. Előszó a magyar kiadáshoz. In: Corbin, Juliet – Strauss, Anselm: *A kvalitatív kutatás alapjai.* Budapest: L'Harmattan – SE Mentálhigiéné Intézet – Sage, 9–22.

Sik Domonkos – Németh Renáta – Katona Eszter (2021): Topic modelling online depression forums: beyond narratives of self-objectification and self-blaming. *Journal of Mental Health,* 32 (2): 386–395.

Sulyok Hedvig – Juhász Valéria – Erdei Tamás (2019): *Beszéd-és nyelvelemző szoftverek a versenyképességért és az esélyegyenlőségért: HunCLARIN korpuszok és nyelvtechnológiai eszközök a bölcsészet-és társadalomtudományokban*. Szeged: SZTE JGYPK Magyar és Alkalmazott Nyelvészeti Tanszék.

Szabó Martina Katalin – Ring Orsolya – Nagy Balázs – Kiss László – Koltai Júlia – Berend Gábor – Vidács László – Gulyás Attila – Kmetty Zoltán (2020): Exploring the dynamic changes of key concepts of the Hungarian socialist era with natural language processing methods. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 54 (1): 1–13. http://doi.org/10.1080/01615440.2020.1823289

Szántó Zsolt – Vincze Veronika – Farkas Richárd (2017): Magyar nyelvű szó- és karakterszintű szóbeágyazások. In: Vincze Veronika (szerk.): *XIII. Magyar Számítógépes Nyelvészeti Konferencia.* Szeged: Szegedi Tudományegyetem, 323–328.

Tóbiás Dániel (2020): *A nemi diszkrimináció megjelenésének elemzése Twitch.tv csatornákon szövegbányászati módszerek segítségével*. Szociológia MA szakdolgozat. Kézirat. ELTE TÁTK, Budapest.

Tóth Olga (2001): Az állami díjas szövőnőktől a tenyérjóslásig. A Nők Lapja 1989-től 1999-ig. *Szociológiai Szemle*, 11(1): 3–21.

Turku University word embedding demo. (http://epsilon-it.utu.fi/wv_demo) (Utolsó megtekintés: 2024. április 6.)

Virágh Enikő Anna – Balázs Szepesi (2022): Vállalkozók reprezentációja a főáramú online médiában Magyarországon. *Szociológiai Szemle,* 32 (3): 24–56.

WebVectors Models. (http://vectors.nlpl.eu/explore/embeddings/en/models) (Utolsó megtekintés: 2024. április 10.)

Yang, Xiao – Macdonald, Craig – Ounis, Iadh (2018): Using word embeddings in Twitter election classification. *Information Retrieval Journal*, 21: 183–207.

Zsibrita János – Vincze Veronika – Farkas Richárd (2013): magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In: Mitkov, Ruslan – Angelova, Galia – Bontcheva, Kalina (eds.): *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013.* Várna: INCOMA Ltd, 763–771.

Melléklet

M1: A pilot kutatáshoz használt szavak

career vs family	career	executive, management, professional, corporation, salary, office, business, career
	family	home, parent, child, family, cousin, marriage, wedding, relative
maths vs arts	maths	math, algebra, geometry, calculus, equation, computation, numbers, addition
	arts	poetry, art, Shakespeare, dance, literature, novel, symphony, drama
science vs arts	science	science, technology, physics, chemistry, Einstein, NASA, experiment, astronomy
	arts	poetry, art, Shakespeare, dance, literature, novel, symphony, drama
intelligence vs appearance	intelligence	precocious, resourceful, inquisitive, genius, inventive, astute, adaptable, reflective, discerning, intuitive, inquiring, judicious, analytical, apt, venerable, imaginative,shrewd, thoughtful, wise, smart, ingenious, clever, brilliant, logical, intelligent
	appearance	alluring, voluptuous, blushing, homely, plump, sensual, gorgeous, slim, bald,athletic, fashionable, stout, ugly, muscular, slender, feeble, handsome, healthy, attractive, fat, weak, thin, pretty, beautiful, strong
strength vs weakness	strength	power, strong, confident, dominant, potent, command, assert, loud, bold, succeed, triumph, leader, dynamic, winner
	weakness	weak, surrender, timid, vulnerable, weakness, wispy, withdraw, yield, failure, shy, follow, lose, fragile, afraid, loser

M2: Az elemzéshez használt táblázatok

GN	man	woman	különbség	father	mother	különbség
executive	-0,025	0,025	-0,050	0,033	0,026	0,007
management	0,022	-0,023	0,045	0,000	0,011	-0,011
professional	0,062	0,031	0,031	0,103	0,020	0,083
corporation	0,133	0,140	-0,007	0,061	0,060	0,001
salary	0,061	0,031	0,030	0,108	0,067	0,041
office	0,079	0,151	-0,072	0,129	0,129	0,000
business	0,036	0,053	-0,017	0,082	0,073	0,009
career	0,152	0,090	0,062	0,243	0,123	0,120

GN	man	woman	különbség	father	mother	különbség
home	0,193	0,187	0,006	0,290	0,358	-0,068
parent	0,065	0,192	-0,127	0,257	0,396	-0,139
child	0,316	0,475	-0,159	0,451	0,618	-0,167
family	0,212	0,264	-0,052	0,572	0,607	-0,035
cousin	0,385	0,343	0,042	0,699	0,656	0,043
marriage	0,140	0,254	-0,114	0,311	0,348	-0,037
wedding	0,110	0,200	-0,090	0,260	0,283	-0,023
relative	nem ismeri	nem ismeri		nem ismeri	nem ismeri	

GN	man	woman	különbség	father	mother	különbség
math	0,018	0,027	-0,009	0,078	0,133	-0,055
algebra	-0,002	0,025	-0,027	0,092	0,160	-0,068
geometry	-0,002	-0,004	0,002	0,074	0,063	0,011
calculus	0,059	0,050	0,009	0,137	0,134	0,003
equation	0,073	0,012	0,061	0,034	0,031	0,003
computation	-0,023	-0,051	0,028	0,004	0,015	-0,011
number	0,102	0,099	0,003	-0,015	-0,006	-0,009
addition	-0,075	-0,073	-0,002	-0,064	-0,064	0,000

GN	man	woman	különbség	father	mother	különbség
poetry	0,122	0,173	-0,051	0,160	0,230	-0,070
art	0,035	0,073	-0,038	0,053	0,089	-0,036
Shakespeare	0,064	0,067	-0,003	-0,015	0,058	-0,073
dance	0,057	0,139	-0,082	0,064	0,143	-0,079
literature	0,067	0,179	-0,112	0,093	0,159	-0,066
novel	0,094	0,095	-0,001	0,061	0,098	-0,037
symphony	0,085	0,111	-0,026	0,032	0,064	-0,032
drama	0,136	0,147	-0,011	0,122	0,120	0,002

GN	man	woman	különbség	father	mother	különbség
science	0,026	0,047	-0,021	0,068	0,081	-0,013
technology	0,008	0,049	-0,041	0,021	0,004	0,017
physics	0,075	0,029	0,046	0,169	0,102	0,067
chemistry	0,057	-0,003	0,060	0,105	0,090	0,015
Einstein	0,181	0,135	0,046	0,166	0,222	-0,056
NASA	-0,008	-0,008	0,000	-0,010	0,037	-0,047
experiment	0,090	0,052	0,038	0,047	0,033	0,014
astronomy	0,050	0,011	0,039	0,064	0,056	0,008

GN	man	woman	különbség	father	mother	különbség
power	0,100	0,043	0,057	0,060	0,017	0,043
strong	0,071	0,045	0,026	0,059	0,034	0,025
confident	0,019	-0,016	0,035	0,012	-0,031	0,043
dominant	0,062	0,040	0,022	0,094	0,020	0,074
potent	0,086	0,006	0,080	0,015	0,005	0,010
command	0,077	0,007	0,070	0,079	0,009	0,070
assert	0,058	0,050	0,008	0,034	0,018	0,016
loud	0,123	0,132	-0,009	0,089	0,120	-0,031
bold	0,117	0,120	-0,003	0,033	0,062	-0,029
succeed	0,025	-0,005	0,030	0,144	0,026	0,118
triumph	0,175	0,130	0,045	0,147	0,098	0,049
leader	0,194	0,122	0,072	0,146	0,038	0,108
dynamic	0,036	0,010	0,026	-0,007	-0,053	0,046
winner	0,173	0,165	0,008	0,105	0,087	0,018

GN	man	woman	különbség	father	mother	különbség
weak	0,058	0,042	0,016	0,077	0,095	-0,018
surrender	0,170	0,078	0,092	0,149	0,123	0,026
timid	0,111	0,092	0,019	0,087	0,119	-0,032
vulnerable	0,038	0,059	-0,021	-0,005	0,042	-0,047
weakness	0,055	0,033	0,022	0,018	0,028	-0,010
wispy	0,100	0,127	-0,027	0,081	0,151	-0,070
withdraw	0,031	0,066	-0,035	0,100	0,083	0,017
yield	nem ismeri	nem ismeri		nem ismeri	nem ismeri	
failure	0,110	0,080	0,030	0,089	0,103	-0,014
shy	0,112	0,084	0,028	0,130	0,136	-0,006
follow	-0,043	-0,029	-0,014	0,091	0,053	0,038
lose	0,079	0,041	0,038	0,075	0,072	0,003
fragile	0,073	0,124	-0,051	0,067	0,157	-0,090
afraid	0,135	0,151	-0,016	0,177	0,205	-0,028
loser	0,205	0,166	0,039	0,106	0,103	0,003

GN	man	woman	különbség	father	mother	különbség
precocious	0,193	0,158	0,035	0,340	0,333	0,007
resourceful	0,099	0,059	0,040	0,095	0,075	0,020
inquisitive	0,123	0,113	0,010	0,162	0,190	-0,028
genius	0,236	0,051	0,185	0,265	0,140	0,125
inventive	0,050	-0,010	0,060	0,018	0,012	0,006
astute	0,101	0,007	0,094	0,095	-0,012	0,107
adaptable	0,052	0,005	0,047	-0,004	-0,008	0,004
reflective	-0,022	-0,028	0,006	-0,048	-0,054	0,006
discerning	0,055	0,084	-0,029	-0,035	-0,030	-0,005
intuitive	0,014	0,018	-0,004	0,033	0,029	0,004
inquiring	0,088	0,104	-0,016	0,113	0,099	0,014
judicious	0,033	-0,027	0,060	0,033	0,006	0,027
analytical	0,017	-0,023	0,040	0,062	0,052	0,010
apt	0,034	0,031	0,003	0,019	-0,017	0,036
venerable	0,071	0,022	0,049	0,078	-0,042	0,120
imaginative	0,058	0,029	0,029	0,060	0,058	0,002
shrewd	0,146	0,058	0,088	0,135	0,059	0,076
thoughtful	0,117	0,095	0,022	0,128	0,121	0,007
wise	0,100	0,027	0,073	0,052	-0,004	0,056
smart	0,092	0,050	0,042	0,059	0,042	0,017
ingenious	0,073	0,018	0,055	0,079	0,039	0,040

GN	man	woman	különbség	father	mother	különbség
alluring	0,037	0,155	-0,118	0,018	0,096	-0,078
voluptuous	0,119	0,277	-0,158	0,058	0,210	-0,152
blush	0,046	0,078	-0,032	0,020	0,041	-0,021
homely	nem ismeri	nem ismeri		nem ismeri	nem ismeri	
plump	0,095	0,140	-0,045	0,040	0,144	-0,104
sensual	0,103	0,240	-0,137	0,041	0,166	-0,125
gorgeous	0,154	0,225	-0,071	0,070	0,191	-0,121
slim	0,105	0,124	-0,019	0,040	0,060	-0,020
bald	0,267	0,197	0,070	0,210	0,176	0,034
athletic	0,114	0,047	0,067	0,067	0,030	0,037
fashionable	0,008	0,133	-0,125	-0,079	0,028	-0,107
stout	nem ismeri	nem ismeri		nem ismeri	nem ismeri	
ugly	0,162	0,157	0,005	0,086	0,130	-0,044
muscular	0,191	0,133	0,058	0,137	0,094	0,043
slender	0,169	0,170	-0,001	0,104	0,148	-0,044
feeble	0,140	0,062	0,078	0,081	0,083	-0,002
handsome	0,228	0,184	0,044	0,229	0,138	0,091
healthy	0,031	0,023	0,008	0,034	0,079	-0,045
attractive	0,001	0,077	-0,076	-0,064	-0,020	-0,044
fat	nem ismeri	nem ismeri		nem ismeri	nem ismeri	
weak	0,058	0,042	0,016	0,077	0,095	-0,018