Katona Eszter
Eötvös Loránd Tudományegyetem, Társadalomtudomány Kar
PhD hallgató

Knap Árpád
Eötvös Loránd Tudományegyetem, Társadalomtudomány Kar
PhD hallgató

A Natural Language Processing és módszereinek felhasználhatósága és terjedése a társadalomtudományban

Absztrakt

Tanulmányunkban áttekintő jelleggel bemutatjuk a tartalomelemzési technikák klasszikus, illetve az utóbbi években rendkívüli fejlődésen átesett, általunk „újfajta" jelzővel kategorizált módszereit. Ezek ismertetését követően olyan kutatási módszertan kidolgozására teszünk kísérletet, melynek alkalmazásával detektálható, hogy a társadalomtudományokon belül születő tudományos közleményekben mekkora részaránnyal bírnak a Natural Language Processing körébe tartozó eljárások. Kutatásunk alapját a JSTOR adatbázisa jelenti. A JSTOR-ról kikért metaadatok elemzését, valamint a platform működésének bemutatását követően interaktív vizualizációs technikákkal mutatjuk be a kapott eredményeket, végül ezek interpretációja után lehetséges irányokat jelölünk ki a témával kapcsolatos további vizsgálatok számára.

Kulcsszavak: tartalomelemzés, szövegelemzés, szociológia, társadalomtudomány, természetes nyelvfeldolgozás, Natural Language Processing, NLP, JSTOR.

Tartalom

Hagyományos tartalomelemzési módszerek áttekintése

„A tartalomelemzés egy olyan kutatási technika, amely szövegekből (és egyéb, jelentéssel bíró alapanyagokból) megismételhető és érvényes következtetéseket von le azok használatának kontextusára vonatkozóan" (Krippendorff 2004).

A kvalitatív tartalomelemzés egy, a szöveges adatok szubjektív értelmezésének és az adatokban rejlő, elsősorban látens tartalmak kutatására alkalmas módszer, mely empirikusan és módszertanilag ellenőrzött eszközöket alkalmaz. Segítségével a szövegeket kommunikációs kontextusukban elemezhetjük, bennük mintázatokat kereshetünk, és az adatokat klasszifikálhatjuk – anélkül, hogy szövegeinket kvantifikálnánk. A módszer célja konzisztenciák és jelentések azonosítása a szövegekben (Hsieh & Shannon 2005, Mayring 2000, Patton 2002).

A kvantitatív tartalomelemzés ezzel szemben a szövegekben előforduló szavak vagy témák gyakoriságának vizsgálatára alkalmas. A szavak együttes előfordulása alapján elsősorban manifeszt, – de ezen túlmenően bizonyos esetekben látens – tartalmakat kívánnak felderíteni, előre meghatározott kategóriák szerint. A kategorizálást megelőzően kódolást végeznek a kutatók, majd ezek mennyiségi elemzése következik (Antal 1976.).

Herring (2010) 5 fázisra osztja a kvantitatív kutatás menetét, melyek a következők:

1. a kutatási kérdés és/vagy hipotézis megfogalmazása,
2. a mintaválasztás,
3. az elemzési egységek és a kódolási kategóriák definiálása,
4. a kódolók képzését követő kódolás a megbízhatóság ellenőrzésével, valamint
5. a kódolás eredményeképpen előállt adatok elemzése, interpretálása.

A kvalitatív és kvantitatív tartalomelemzési módszerek összehasonlításával a két paradigma külön-külön is könnyebben megérthetővé válik. Az alábbi összehasonlítás alapja Zhang és Wildermuth (2005) cikke.

A kvantitatív elemzéseket gyakran alkalmazzák elsősorban a tömegkommunikációban arra, hogy „megszámolják" az egyes manifeszt szövegelemek előfordulási gyakoriságát. Ezt az eljárást sokan kritizálják a szövegben rejlő szintaktikai és szemantikai információk teljes mellőzése miatt. Ezzel szemben a kvalitatív megközelítés az antropológia, a szociológia és a pszichológia területéről ered, és a szövegben rejlő valós jelentéseket kívánja feltérképezni, tehát nem magának a szövegnek a számokkal leírható tulajdonságait ismerteti.

Szintén elmondható a kvantitatív megközelítésről, hogy működése alapvetően deduktív: hipotézisek és korábban, megelőző felmérések eredményei és már létező ismeretek alapján megfogalmazott kérdések megválaszolására, elméletek alátámasztására vagy megcáfolására

törekszik. Ettől eltérően a kvalitatív elemzés induktív, célja alapvetően a témák azonosítása, emiatt sokkal inkább alkalmas elméletek létrehozásához, mint azok alátámasztására.

Nyilvánvaló, hogy a mintavételi eljárás megválasztása teljesen különböző szempontok szerint történik a két esetben. A kvantitatív elemzések véletlen mintavételezésen vagy más, valószínűségszámításon alapuló eljáráson nyugszanak annak érdekében, hogy statisztikailag érvényes megállapítások levonására legyenek alkalmasak. Ezzel szemben a kvalitatív módszerek gyakran alkalmaznak szakértői, önkényes mintavételezést, amelynek során szándékosan választanak ki a kutatás szempontjából releváns szövegeket, tartalmakat.

Végezetül fontos megemlíteni, hogy a két módszer alkalmazásának eredményeképpen előálló produktumok is különböznek. A kvantitatív eljárások statisztikai módszerekkel elemezhető, számszerűsített adatok létrehozására alkalmasak. Ezzel szemben a kvalitatív tartalomelemzés során leírások, tipológiák születnek, valamint szükségszerűen szubjektív, részletes beszámolók az adott szövegről, amelyek bemutatják a vizsgált jelenség jelentésének árnyalatait (Zhang és Wildermuth 2005).

Ami általánosságban is elmondható a két módszerről, az ebben a specifikus esetben is igaz: a kvalitatív módszerek érvényessége magasnak, megbízhatósága alacsonynak mondható, míg a kvantitatív eljárások alacsony érvényességű, de magas megbízhatóságú eredményt szolgáltatnak.

Ahogyan azt a tanulmány későbbi részében is láttatni kívánjuk, a „Natural Language Processing" tárgykörén belül alkalmazott módszerek mindkét megközelítésből merítenek, tehát nem csupán számszerű adatok leírására tesznek kísérletet, hanem a mögöttes jelentéstartalom feltérképezését is célul tűzik ki.

Új szövegelemzési módszerek: Natural Language Processing

Az NLP általánosságban

A tanulmány következő szakaszában Tikk Domonkos 2007-es, valamint Sebők et al. 2016-os átfogó, szövegbányászattal foglalkozó műveit vettük alapul.

A számítógépes szövegfeldolgozás, más néven Natural Language Processing adathalmazainak alapját legtöbbször a „Big Data" kifejezéssel illethető források jelentik. A Big Data kifejezést sokszor használják az adat méretére vonatkozóan, de a nagy mennyiségű adatok gyűjtésére, tárolására, feldolgozására és elemzésére vonatkozó tevékenységek összefoglaló megnevezéseként is (Karin van Es 2017, 13). A Big Data definiálásakor hagyományosan a „három V betűt" szokták használni, amely megjelenik egyebek mellett Zikopoulos et al. könyvében (2012, 5-9). E szerint az ilyen típusú adatok három legfontosabb jellemzője az adatok nagy mérete

(volume), nagyfokú változatossága (variety; a strukturált adatok mellett nagy mennyiségben megtalálható félig-strukturált és strukturálatlan adat is1), valamint az adatok sebessége és elérhetősége (velocity). Az utóbbi szemponthoz kiemelendő, hogy nem csak az adatok mérete nőtt meg, hanem a keletkezésük sebessége is, ezzel párhuzamosan pedig lecsökkent a „szavatossági idejük". Az adatokban lévő potenciált az tudja a leghatékonyabban kihasználni, aki gyorsan, ideális esetben valós időben képes feldolgozni az adathalmazt. A fenti definíciós kísérlet kiterjesztéseként értelmezhető Kitchin (2014) munkája, aki kiemeli, hogy szintén jellemző a Big Data-ra az, hogy teljes populációkat fed(het) le (de legalábbis sokkal nagyobb mintaméreteket alkalmaz a korábban megszokottaknál); nagy felbontással, részletezettséggel rendelkezik; relációs jellegű (a különböző adatforrásokból származó adatok gyakran összekapcsolhatók); valamint rugalmas és skálázható.

A Big Data adatforrásait Lazer és Radford (2017, 21-23) három nagy kategóriába sorolja.

Digitális élet: Facebook, Twitter, Wikipedia – a tevékenység alapvetően online, de esetenként kapcsolatban van offline dolgokkal is. Ezekre az adatforrásokra tekinthetünk úgy, mint a társadalom „mikrokozmoszaira", amelyekben bizonyos össztársadalmi szinten is meglévő problémákat vizsgálhatunk speciális környezetekben2; vagy mint az élethez kapcsolódó fontos, ugyanakkor elkülöníthető területekre, amelyeknek önmagukban is vizsgálatra méltók3.

Digitális lábnyomok: ebbe a kategóriába főleg a cégek és a bürokratikus intézmények által felhalmozott adatok tartoznak. A lábnyomok, avagy lenyomatok – ellentétben az előző csoportban taglalt adatokkal – nem magát a tevékenységet jelentik, hanem annak valamilyen digitális reprezentációját. Példaként említhetők a digitálisan tárolt hívásadatok vagy a választói adatbázisok.

Digitalizált élet: ide tartozik minden olyan adat, amely az offline életünkkel kapcsolatos, arra vonatkozik, azonban digitális formában van eltárolva. A Bluetooth-szal ellátott eszközök képesek regisztrálni egymás távolságát, ezáltal elemezhető az ezeket birtokló személyek pozíciója és közelsége; a városi kamerák videófelvételeit vizsgálva következtetéseket vonhatunk le az emberek interakcióiról; a Google Books milliói könyvet tartalmazó korpusza pedig izgalmas szövegelemzéseknek adhat teret.

---

1 Lásd még: Csepeli (2015, 173).

2 Például Edelman és Luca (2014) tanulmánya, amely az Airbnb platformon tapasztalható „digitális diszkriminációt" kutatja.

3 Erre példa Bakshy, Messing és Adamic (2015) írása, amelyben azt vizsgálták, hogy a Facebook „információs burokba" zárja-e a felhasználókat, amelynek révén csak a világnézetükkel kompatibilis hírekkel találkoznak a platformon.

„*...Míg 2002-ben a világon összesen 27 exabyte-nyi adat keletkezett, addig ma 7 nap alatt jön létre ennyi információ. Ha a pontos számok kérdésesek is lehetnek, az exponenciális növekedés ténye vitathatatlan*" (Dessewffy – Láng, 2015). A szociológia szempontjából nem feltétlenül az adatok mennyisége, sokkal inkább az adatok online hozzáférhetősége fontos. Ahogy Ságvári (2017) is írja, az új típusú, online termelődő adatok megjelenése lehetőséget ad arra, hogy olyan kérdéseket vegyünk górcső alá, melyek elemzésére korábban, az adatok hiánya miatt nem volt lehetőségünk. A különböző forrásokból származó adatok összekapcsolása rengeteg komplex elemzési lehetőséget rejt magában, ám ezekhez el kell szakadnunk a megszokott, rendezett adatstruktúráktól, teret kell engednünk újfajta módszerek alkalmazásának.

Megkülönböztetünk egymástól strukturált és strukturálatlan szövegeket. Strukturált szövegek alatt korábban már feldolgozott dokumentumokat értünk, melyekről különböző információkkal, metaadatokkal rendelkezünk (lehetnek ezek a szövegből kinyert információk, mint a szavak száma, vagy a szöveg érzelmi töltetére vonatkozó információ), melyek az adatainkat bizonyos szempont szerint jellemzik. Strukturálatlannak tekintjük az összes, hétköznapokban termelődő szöveget, melyek nincsenek elemzési célra előkészítve. A nyelvtani szerkezeteknek, a szöveg különböző jellemzőinek köszönhetően a tanulmányokat, újságcikkeket, weboldalakat azonban túlzás lenne strukturálatlan adatnak hívni, az ilyen tartalmakat inkább szabadformátumúnak, gyengén vagy félig strukturáltnak szoktuk nevezni, de ez nem jelenti azt, hogy az ilyen szövegeink már elemzésre alkalmas állapotban vannak. A szövegek elemezhető formára hozása az NLP egyik fontos feladata. Az előfeldolgozás egyes lépései egymással kölcsönhatásba lépnek és befolyásolják a modell teljesítményét, emellett a hibák javítása később gyakran nagyon költséges lehet, így megéri az elemzés előtt kezelni őket (Rehurek, 2011).

A következőkben az általunk legfontosabbnak tartott lépéseket mutatjuk be. Az első lépés a Named Entity Recognition vagy Named Entity Extraction tulajdonképpen nem más, mint a névelemek felismerése. A folyamat során megkeressük a szövegben szereplő tulajdonnevek, földrajzi nevek, intézmények vagy szervezetek neveinek különböző előfordulásait, és ezeket egységes alakra hozzuk. A következő lépés a szövegek szótövezése és szófaj szerinti szűrése. A morfológia szabályainak megfelelően a szavak alakjai eltérőek lehetnek, emiatt szükséges szótövezést végezni, ami egyfajta normalizálás, mellyel egységes alakra hozzuk a kifejezéseinket. A szófaji szűrés során megadhatjuk, hogy mely szófaji kategóriába sorolt kifejezéseket szeretnénk a korpuszunkban viszontlátni. Alapvetően a főnév, melléknév és ismeretlen kategóriák szoktak a korpuszba kerülni, az igék kiírása változó. Megkülönböztetjük egymástól a tartalomszavakat (content words pl. főnév, melléknév) a funkciószavaktól (function words például a kötőszavak). A stopszavazással megszabadulunk azoktól a kifejezésektől, melyek tartalmi jelentéssel nem rendelkeznek (Tikk et al., 2007). A következő feladat az előfeldolgozás során a kollokációk

vizsgálata, vagyis a szignifikáns bigramok megkeresése. A bigram egy gyakran együtt előforduló szópár (vagy akár szótöbbes, azaz n-gram). A magyar lexikográfiában állandósult szókapcsolatként, idiómaként szoktak rá hivatkozni (Reményi 2010). Az előfeldolgozás utolsó lépésében azokat a kifejezéseket, amik nagyon kevés, vagy nagyon sok dokumentumban szerepelnek kizárjuk az elemzésből. A szóeloszlás egy hatványfüggvény, amely Zipf-eloszlást követ, emiatt a leggyakoribb szavak a legtöbb dokumentumban megtalálhatóak (Tikk et al., 2007). Az emberi tudás, valamint nyelvi képességeink lehetővé teszik a keletkező strukturálatlan szöveghalmazok feldolgozását, azonban ilyen tevékenységet csak limitált mennyiségben, és viszonylag lassan vagyunk képesek végezni. Ezzel szemben a számítógép gyors és hatékony munkát végez, ugyanakkor nem képes egyebek mellett a különböző grammatikai struktúrák, vagy a kontextus megértésére. A probléma kezelésére, a két képesség ötvözésével a természetes nyelvfeldolgozás (Natural Language Processing, NLP) tesz kísérletet. Ez az a tudomány, amely segítséget nyújt a gyorsan termelődő, nagy mennyiségű szöveges adatok hatékony feldolgozásában. A természetes nyelvfeldolgozás a számítógéptudomány (computer science), a mesterséges intelligencia (Artifical Intelligence, AI) és a nyelvészet közös területe (bővebben a témáról: Liddy, 2001).

Az NLP arra tesz kísérletet, hogy a jelentés minél teljesebb reprezentációját nyerje ki a szövegből, feltárja a benne rejlő összefüggéseket, továbbá segítséget jelent az adatok kereshetőségének kialakításában és ezáltal a bennük történő eligazodásban is. A természetes nyelvfeldolgozás során az alacsony mértékben strukturált szöveghalmazainkat kisebb egységekre bontjuk, és kapcsolatokat keresünk a szöveg különböző elemei között. A Natural Language Processing körébe tartozó módszerek célja tehát a nem triviális tudás feltárása strukturáltalan szövegekből (Kao – Poteet, 2007:1). A tudományterülethez számos eltérő céllal alkalmazott és különböző logikát alkalmazó eljárás tartozik, amelyek közül a leggyakrabban használt típusokat tanulmányunk későbbi részében ismertetjük.

Az NLP felhasználása az üzleti életben igen elterjedt, a gépi fordítás alkalmazása, vagy a chatbotok jelenléte ma már mindennaposnak mondható. E két felhasználási mód nagyban épít a mesterséges és az üzleti intelligenciára egyaránt. Az üzleti intelligencia eleinte, a kifejezés megjelenésekor - a 80-as, 90-es években - leginkább strukturált, számszerű adatok elemzését jelentette a jobb algoritmizálhatóság miatt. Mára azonban a szövegelemzés szerves részét képezi az üzleti intelligencia területének: a hírszerzéstől kezdve a biomedikális összefüggések feltárásán és az ajánló algoritmusokon át a terroristák felismeréséig számos területen alkalmazzák.

NLP a társadalomtudományban

Az NLP felhasználásának elterjedtségéhez tehát nem fér kétség az üzleti élet területén, azonban a társadalomtudósok is egyre gyakrabban fordulnak a módszer eszközeihez.

Kmetty et al. (2017) az öngyilkosság mintázatát vizsgálták egy, a Twitterről származó korpuszon, és azt találták, hogy az öngyilkosságok időbeli eloszlásának heti ingadozása illeszkedik a rossz hangulatú üzenetek arányához a Twitteren, bár ahhoz képest, amit az öngyilkossági adatok mutatnak, a tweetekben intenzívebben látható a felhasználók hangulatának romlása vasárnaponként. A parlamenti felszólalások is termékeny talajt adnak az újfajta szövegelemzési módszereknek. A Precognox[4] és a K-Monitor[5] is felhasználta ezen tartalmakat: előbbi a beszédek nemi töltetét vizsgálta szóbeágyazási modell segítségével, a K-Monitor pedig topikmodellezést végzett, hogy jobban megismerjük a parlament nyelvezetét, és annak sajátosságait.

Az NLP a hagyományos kutatási módszerek kiegészítésére is alkalmas. Ahogy Roberts, Stewart, Tingley és Airoldi (2013) tanulmányukban írják, a társadalomtudományban is egyre közkedveltebb eszköz az NLP látens nyelvi, politikai és pszichológiai kutatásokban. Ők például arról írnak, hogy a hagyományos kérdőíves módszert alkalmazó kutatóknak gyakran nehéz dönteni, hogy kérdőívükben alkalmazzanak-e nyílt kérdéseket, hiszen ezek elemzése nem triviális. A feladat megkönnyítésére létrehozták a Structural Topic Model nevű módszert, amit a tanulmányunk későbbi részében bemutatásra kerülő látens Dirichlet allokáció (LDA) kiterjesztésével készítettek el.

Az NLP segítségével továbbá lehetségessé válik az online diskurzusok elemzése, parlamenti felszólalások, vagy a social media különböző platformjainak vizsgálata. Ezen kívül fórumok és hírportálok is gyakran képezik alapját az NLP módszertanával történő vizsgálódásoknak.

A következőkben áttekintjük az NLP gyakran alkalmazott eszközeit, majd pedig a módszer terjedésének dinamikáját vizsgáljuk meg szociológiai folyóiratokban.

Kapcsolódó módszerek áttekintése

Felügyelt vs. nem felügyelt módszerek

A gépi tanulás legfontosabb célkitűzése olyan hatékony és robusztus algoritmusok előállítása, amelyek segítségével lehetőség nyílik ismeretlen elemek előrejelzésére, vagy új esetek

[4] https://zolizoli.github.io/prezis/gender2.sozi.html#frame9546
[5] https://k.blog.hu/2017/12/04/a_parlament_nyelve

kategóriákba sorolására. A modellalkotás során megkülönböztetünk felügyelt (supervised), és nem-felügyelt (unsupervised) módszereket.

Felügyelt esetben előzetesen ismert kategóriákkal rendelkezünk, és a cél az, hogy új elemeket tudjunk beilleszteni a már rendelkezésre álló kategóriarendszerbe. Ilyenkor mindig tudjuk elemeink legalább egy részéről, hogy melyik csoportba tartoznak, ezt nevezzük tanítóhalmaznak. Felügyelt esetben tudjuk, hogy milyen struktúrába szeretnénk az adatainkat kódolni, és ezt tanítjuk meg az algoritmusnak. A felügyelt módszerek közé tartozik a lineáris és logisztikus regresszió, vagy az SVM, azaz a Support Vector Machine módszere.

Például egy e-mail fiók esetében a spam felismerése a feladatunk. Készíthetünk egy automatikus osztályozót a szűrésre, úgy, hogy kész példákat használunk a tanuló algoritmusunk tanítására. Ebben az esetben a tanuló mintánk a felcímkézett levelekből áll, és egy teszt mintán értékeljük ki az algoritmusunk teljesítményét. A teszt adatokkal az algoritmus a tanulás során nem találkozik. A spam-szűrőnk esetében a spam címkét az algoritmusnak kell prediktálnia. Ahhoz, hogy a predikció pontosságát mérni tudjuk, az algoritmus által adott címkéket összehasonlítjuk a tesztminta címkéivel.

A nem felügyelt módszereket a kognitív tudományok területén fejlesztették ki, de egyre nagyobb teret nyert az utóbbi időben a statisztikán belül is. A nem felügyelt módszereket hagyományosan szinonimák, névelemek felderítésére használják, ám számos modell is támaszkodik a módszerre, így például látens topikok, témák felismerésére is gyakran használnak ilyen eljárásokat (Sebők et al 2016). Nem felügyelt módszerek esetén nem rendelkezünk semmilyen előzetes kategóriarendszerrel, nincsenek előzetes feltevéseink arra nézve, hogy milyen témák rejlenek a szövegeinkben. Ilyenkor tanítóhalmazzal sem rendelkezünk, hiszen a modellünkre bízzuk, hogy különböző statisztikai feltevések alapján valamilyen, háttérben meghúzódó struktúrát találjanak a szövegekben.

Klasszifikáció

Akkor használunk klasszifikációt, ha szövegeinket előre meghatározott osztályokba szeretnénk sorolni. Például a fent említett spam-szűrő elkészítése is tipikusan klasszifikációs feladat. Ebben az esetben van egy, a modell tanítására előkészített, felcímkézett részhalmaz a szövegeinkből, ezt neveztük az előbb tanítóhalmaznak. Ezen a halmazon modellünk megtanulja, hogy szövegeink bizonyos tulajdonságai melyik osztályba tartozást valószínűsítik, majd a modell feladata az lesz, hogy ismeretlen elemeken, a tanult módon klasszifikálja a szövegeinket.

Felügyelt klasszifikáció során fontos lépés a tanítóhalmaz előállításának módja. Ez kézzel, kódolók segítségével történik, nem pedig automatikus módon. A tesztelésre szánt adatok is előre

be vannak kategorizálva, de a tesztelés során a modell a valós címkéket nem látja, a modellnek kell megtippelni a csoportbatartozást. Ha a tanító halmazunk elkészült, a tesztelésre szánt mintaelemeinken értékeljük ki a modellünk teljesítményét.

A dokumentumszett (korpusz) elemeit az előbb bemutatott felügyelt tanulási módszer mellett szótár alapon is klasszifikálhatjuk. Ebben az esetben először egy szótárt, szólistát készítünk el, majd a listánk alapján soroljuk osztályokba a szövegeinket, annak megfelelően, hogy mely kifejezések kerültek elő az egyes dokumentumokban, és melyek nem (Sebők et al 2016).

Látens tartalom felderítése: klaszterelemzés, topikmodellek

Tanulmányunkban nagyobb hangsúlyt fektetünk a topikmodellezés bemutatására, hiszen a többi eljárás a szövegelemzésen kívül a numerikus adatok elemzése során gyakran alkalmazott, így sokkal többet tárgyalt, sokkal nagyobb ismeretnek örvendő módszer.

A klaszterelemzés során a szövegeket bizonyos jellemzőik szerint, hasonlósági alapon rendezik csoportokba. Sem a topikmodellezés, sem pedig a klaszterelemzés során létrehozandó csoportok száma nem ismert előzetesen. A klaszterelemzés nem felügyelt módszer, célja a korpuszban rejlő látens struktúra feltárása, a korpusz szövegeinek csoportosítása.

A valószínűségi topikmodellek nagyméretű szöveges korpuszok megértésében nyújtanak segítséget: az adataink, korpuszunk felderítésére, az adatokban található rejtett struktúrák feltárására alkalmasak (Blei, 2011). Segítenek megtalálni a látens topikokat, melyek a korpusz főbb témáit képezik. Blei és Lafferty (2009) szerint a topikmodellek olyan kapcsolatokat tárhatnak fel a dokumentumok között és a dokumentumokon belül, melyek nem nyilvánvalóak, és amelyekre a priori nem számítunk.

Képzeljük el, hogy rendelkezésünkre áll egy nagyobb szöveges korpusz, például online sajtó összegyűjtött anyagai egy (vagy akár több) portálról. Ebben az esetben topikmodellezéssel vizsgálhatjuk azt, hogy a szövegek milyen témák köré csoportosulnak, és levonhatjuk olyan következtetést, hogy az adott topikok alapján milyen csoportokkal szembeni ellenszenv jelenik meg: rasszizmusról, antiszemitizmusról, idegenellenességről vagy épp nőgyűlöletről szólnak-e a cikkek inkább.

A topikmodellek az utóbbi években gyors fejlődésen estek át. A topikmodellek a látens szemantikus indexelésből (latent semantic indexing, LSI) alakultak ki (Deerwester et al., 1990), bár az LSI még nem nevezhető valószínűségi modellnek. Hoffmann (2001) az LSI-t továbbgondolva, egy valószínűségi generatív folyamattal kiegészítve alakította ki a valószínűségi látens szemantikus indexelést (probabilistic latent semantic indexing, PLSI), amit az információkinyerés és a klaszterezés során alkalmaznak leginkább.

Ezt követően Blei, Ng és Jordan (2003) írt először a látens Dirichlet allokációról (Latent Dirichlet Allocation, LDA). Ez a modell azt feltételezi, hogy a szavak erős szemantikai információkkal rendelkeznek, és a hasonló témákkal foglalkozó dokumentumok hasonló szavak csoportjait használják. A látens témákat tehát a korpusz dokumentumaiban gyakran együtt szereplő szavak csoportjainak azonosításával fedezik fel.

Előzetesen csak a témák számát határozzák meg, és csak egy megfigyelt változó van: a dokumentumok szavai[6]. Ezekhez az inputokhoz az LDA visszaadja a dokumentumok topikjainak megoszlását, valamint az egyes topikok szóeloszlását.

Az LDA-nak több kiterjesztése létezik, ezek közé tartozik például a korábban említett Structural Topic modell, ami a kérdőívek nyílt kérdéseinek feldolgozását segíti (Roberts et al 2013), a korrelált topikmodell (correlated topic models), ami a topikok közötti interakciót modellezi (Blei – Lafferty 2007), a dinamikus topikmodell (dynamic topic models) ami a topikok időbeli változását vizsgálja (Blei – Lafferty 2006), valamint a szerző-topikmodell (author topik model), ami szerzőségi információval egészíti ki az LDA-t (Rosen-Zvi – Chemudugunta – Griffiths – Smyth – Steyvers 2008).

Szóbeágyazási modellek

Ahogy Garg et. al (2018) írja, a szóbeágyazási modellek (word embedding) a gépi tanulás módszerébe tartoznak. Minden szót egy vektor képvisel. A vektorok közötti geometriai kapcsolat szavak közötti szemantikai kapcsolatnak feleltethető meg.

Vegyünk két mondatot, melyeknek jelentése nagyon hasonló: „Legyen szép napod.”, és „Legyen gyönyörű napod.”. Ha ezt a két mondatot szótár (jelölése: V) formában szeretnénk reprezentálni, a következőt kapnánk: V = {legyen, szép, gyönyörű, napod}.

A mondat szavai vektor reprezentációban:

```
legyen =    [1,0,0,0];
szép =      [0,1,0,0];
gyönyörű =  [0,0,1,0];
napod =     [0,0,0,1].
```

A szóvektorainkból kialakíthatunk egy négydimenziós vektorteret, melyben az egyes szavak egy-egy dimenziót képeznek. Ebben a térben a „szép” és a „gyönyörű” kifejezések ugyanolyan

[6] Dinamikus modell esetén az időponttal, az Author Topic Model esetén pedig a szerzővel egészül ki a változók listája.

különbözőnek tűnnének, mint a „legyen” és a „napod”, pedig ez nem igaz. A célunk ezért az, hogy a hasonló szavak térben egymáshoz közel legyenek.

A szavakhoz úgy szeretnénk vektort rendelni, hogy ezáltal a vektortér struktúrájáról tehessünk állításokat. Nem lenne jó, ha a „szép” és „gyönyörű” szavak teljesen különböző beágyazással rendelkeznének, hiszen a legtöbb mondatban felcserélhetők. A szóvektorok közötti geometriai kapcsolatoknak tükrözniük kell a szavak szemantikai kapcsolatait. A szóbeágyazási modellek feladata az emberi nyelv geometriai terének leképezése. Egy vektortérben a szinonimák hasonló szóvektorokba történő beágyazódását várjuk, tehát azt, hogy a nagyon különböző dolgokat jelentő szavak egymástól távoli pontokba rendeződjenek, míg a hasonló jelentésű szavak közelebb kerüljenek egymáshoz. Egy gyakran példaként bemutatott geometriai transzformáció a vektortérben, hogy ha a „király” vektorból kivonjuk a „férfi” vektort, és hozzáadjuk a „nő” vektort, akkor a „királynő” vektort kapjuk vissza[7].

Szentimentelemzés, emócióelemzés

E módszer célja, hogy a szövegen belül a negatív és a pozitív hangulat mértékét vizsgáljuk. Thelwall és Wilkinson (2010) a MySpace-ről gyűjtött adatokon a nemek közötti nyelvhasználati különbségeket határozták meg szentimentelemzéssel. Ahktar és Soria (2009) Facebook adatokat elemeztek érzelmi töltetük alapján, Thelwall et al. (2011) pedig valós események utáni pozitív és negatív hangulatokat tártak fel Twitter bejegyzésekből.

Az érzelmek felismerésére szöveges korpuszokban sokféle módszert alkalmazhatunk, egyfelől tekinthetünk úgy a problémára, mint klasszifikációs feladatra: ebben az esetben osztályokba szeretnénk sorolni a szövegeinket. Használhatunk különböző klasszifikációs módszereket (Support Vector Machines, Random forest), melyek teljesítményének összehasonlításáról Kinnunen (2017) írásában olvashatunk. Ennél a módszernél nagyon fontos az előzetes annotálás, hiszen ilyenkor az előzetesen besorolt szövegeken tanítjuk be (traineljük) a modellünket. Meghatározhatunk különböző számú kategóriákat, különböző címkékkel, például pozitív-negatív, vagy pozitív-negatív-semleges / nagyon negatív-negatív-semleges láthatjuk el azokat.

Szentimentelemzést szótár alapon is végezhetünk - a magyar nyelvre is rendelkezésre áll szentimentszótár. Ebben az esetben az egyes szavak érzelmi töltetéből indulunk ki, és a szövegek szavait keressük a szótárunkban, majd ezeket társítjuk az előre meghatározott érzelemhez.

A legújabb és legizgalmasabb lehetőség, ha deep learning módszerek segítségével igyekszünk felderíteni, hogy milyen érzelmeket rejtenek szövegeink. Gatti és dos Santos (2014) konvolúciós

[7] Forrás: https://jjallaire.github.io/deep-learning-with-r-notebooks/notebooks/6.1-using-word-embeddings.nb.html

neurális hálót (convolutional neural network, CNN) használ, melynek előnye, hogy nem csak a szavak szintjén vizsgálódik, hanem a karakterek és mondatok szintjét is figyelembe veszi az elemzés során. Az algoritmus tehát különböző szinteken (layers) tanul a szövegek jellemzőiből, ami például a tagadó szerkezetek felismerésében is sokat segít.

A terjedés dinamikája

Kutatásunkban arra kerestük a választ, hogy a dolgozat korábbi részében bemutatott, általunk „új" szövegelemzési módszereknek nevezett eljárások hogyan terjednek a tudományos közegben. A kérdés fő relevanciája abban rejlik, hogy - mint ahogy korábban is láthattuk - ezek a metódusok nem a tudományos közeg, hanem üzleti vállalkozások által kerültek kifejlesztésre, így egyelőre az akadémiai felhasználásuk csekély.

A társadalmi folyamatokhoz kapcsolódó fogalmaknak, az emberek gondolkodásának és attitűdjeinek, továbbá bizonyos konstrukciók jelentéstartalmának változását a társadalomtudomány hagyományosan kvantitatív kérdőíves kutatásokkal vizsgálja. Azt láthattuk azonban az utóbbi években, hogy a közbeszéd dinamikája – részben a közösségi média platformoknak és az online hírportáloknak köszönhetően – annyira „felpörgött", hogy például egy politikai tárgyú közvéleménykutatás eredménye gyakran néhány nap alatt érdektelenné válik, megállapításai nem érvényesek többé. A természetes szövegfeldolgozással, és ahhoz kapcsolódóan automatizált adatgyűjtési módszerek segítségével lehetséges társadalmi folyamatokat valós időben nyomon követni, és szociológiailag releváns problémákat megválaszolni. Így tehát különösen érdekes megnézni, hogy a társadalomtudományok mennyiben „fedezték fel" ezeket az újfajta eljárásokat.

Módszertan

A kutatásunk célja az volt, hogy adatot gyűjtsünk arra vonatkozóan, hogy a tudományban, és elsősorban a társadalomtudományban születő cikkek milyen arányban tartalmaznak „újfajta" szövegelemzési módszereket. A terjedés időbeli dinamikáját kívántuk vizsgálni, tehát azt, hogy az adott évben megjelent összes cikk közül hány darab alkalmazza az újfajta módszereket.

Google Trends

A terjedés ütemének méréséhez többféle lehetőséget is számba vettünk. A végleges módszertan megtalálásának dilemmáját a következő néhány gondolatban mutatjuk be annak érdekében, hogy

az olvasó ne vádolhassa a cikk szerzőit azzal, hogy nem erőltették meg magukat kellően azért, hogy a lehető legközelebb kerüljenek a valóság leírásához.

A Google Trends[8] által megjelenített adatok a Google keresőben végzett keresések trendjeit tartalmazzák. A rendszer nem ad ki nyers számokat arról, hogy adott időszakban bizonyos kulcsszavakra hányszor kerestek rá, illetve a találatok számáról sem kapunk információt, csupán arról, hogy az adott kulcsszóra vonatkozó „érdeklődés" hogyan alakult, idő és régió szerint. Ezek az adatok tehát csak egymáshoz viszonyítva, és adott időszakon belül értelmezhetőek. Ugyan arra van lehetőség, hogy kiválasszunk egy „kategóriát", ezek a kategóriák a weboldalak üzemeltetői által szabadon megadhatók és megváltoztathatók, így megbízhatóságuk kétséges, másrészt a Google-ben való keresés során mi nem tudunk kategóriát választani, így kérdéses, hogy ez a kategóriaválasztás pontosan hogyan befolyásolja a megjelenített adatokat a Google Trends-en. A legnagyobb probléma az, hogy bár megadhatjuk, hogy csak a „science" kategórián belüli trendeket mutassa a rendszer, arra már nincsen lehetőség, hogy tudományterületre szűrjünk, így tehát a Google Trends használata nem lehetséges ebben a kutatásban.

Dimensions.io

A Dimensions.io nevű platform több mint 100 kutatói szervezettel működik együtt, és kulcsrakész analitikai és vizualizációs megoldásokat kínál a felhasználóknak. Első ránézésre tökéletes eszközül szolgálhatna egy ilyen kutatás megvalósításához, azonban sajnos több olyan limitációval is szembesültünk, amely lehetetlenné tette az eszköz használatát. Ilyen például a keresőkifejezések kezelésének módja: úgy tűnik, hogy a Dimensions-ön, ha egy komplex, több szóból álló kifejezést adunk meg, akkor nem az összes szó együttes meglétét vizsgálja a rendszer, hanem minden olyan tartalom találatnak számít, amelyben legalább az egyik szó megtalálható. Ez a működési mód teljesen értelmezhetetlenné teszi az eredményeket, hiszen így például nincsen értelme rákeresni a „big data" kifejezésre, mivel az összes olyan cikk meg fog jelenni, amiben szerepel a „data" szó - ami meglehetősen sok fals találatot eredményez. Ez a működési mód kiküszöbölhető idézőjelek használatával, de a keresőkifejezéseink jelentős részére így is rendkívül furcsa, már-már megmagyarázhatatlan eredményeket és trendeket kaptunk. További probléma egyrészt, hogy a platform csak 2010-ig enged visszanézni az időben, illetve az, hogy a szolgáltatásról nem találtunk olyan részletes leírást, amely alapján eldönthető lenne, hogy elegendően nagy „merítéssel" bír-e a tudományos publikációk keresésében. Végül akkor vetettük el a platform használatát, amikor szembesültünk vele, hogy a szociológia tudományterületbe

[8] A szolgáltatás a következő linken érhető el: https://trends.google.com/trends/

számos olyan cikk is besorolásra került, amely már első ránézésre sem tartoznak ebbe a kategóriába - feltehetően a Dimension-ön működő algoritmus még nem kellően kiforrott.

Google Scholar, Harzing's Publish and Perish

A Google Scholar rendkívül fejlett keresője nagyban megkönnyítené hasonló kutatások végzését, azonban a platform API[9] hiányában automatizált adatgyűjtésre csak web-scraping módszerekkel alkalmas[10]. Ennél nagyobb, áthidalhatatlan probléma azonban a tudományterületekre való szűkítés lehetőségének hiánya, amely miatt lehetetlenné válik, hogy kizárólag a szociológia témakörébe tartozó cikkeken belüli előfordulást vizsgálhassunk.

Létezik ugyan egy szoftver, a Harzing's Publish and Perish[11], amely képes többféle forrásból - így Google Scholarból is - tudományos publikációkat keresni, a Google Scholar limitációi miatt egy keresésre mindössze a legtöbbet idézett 1000 cikket jeleníti meg[12]. További probléma, hogy a használata rendkívül lassú: ez a program is web-scraping alapon működik, így egyszerre mindössze 10 találatot tölt le, és az automatizált adatgyűjtés ellen bevetett CAPTHA-kódok miatt a kérések kiküldésének sebessége is limitált, így 1000 cikk adatainak legyűjtése akár több órát is igénybe vehet.

A Publish and Perish képes keresni a Crossref adatbázisában is, azonban itt még szorítóbb limitációba futunk, ugyanis maximum 200 találatot lehet letölteni egy kereséshez kapcsolódóan. A szoftver ezen kívül elvileg képes még a viszonylag friss, egyelőre kísérleti fázisban működő Microsoft Academic-en is keresni, azonban ez a funkció nem működik rajta, így a Publish and Perish használatát elvetettük.

A JSTOR scrapelése

Következő lehetőségként a legnagyobb online tudományos adatbázishoz, a JSTOR-hoz fordultunk. Az 1995-ben alapított, non-profit alapítvány által üzemeltetett platform több mint 12 millió tudományos folyóiratcikkhez és könyvhöz biztosít hozzáférést[13], valamint kitűnő keresési felülettel rendelkezik. Kutatók és egyetemi hallgatók regisztrálhatnak a JSTOR Data For Research

[9] Application Programming Interface, vagy alkalmazásprogramozási felület. Számos olyan webes API van, amelyre csatlakozva adatokat szerezhetünk automatizált, programozott módon. Ilyen például a Twitter vagy a Facebook API-ja, de létezik Google Custom Search API is.
[10] Az ún. web-scraping módszerek alkalmazásával weboldalak forráskódjából nyerhetünk ki adatokat. A legtöbb webhelytulajdonos tiltja az ilyen jellegű adatgyűjtést, de a terület jogi szabályozottsága még kialakulóban van.
[11] Forrás: https://harzing.com/blog/2017/11/publish-or-perish-version-6
[12] Forrás: https://harzing.com/popbook/ch13_2_4.htm
[13] Forrás: https://about.jstor.org/

szolgáltatásra, amelyen keresztül letölthetők különböző keresési eredmények, egyszerre max. 25000 találatig.

Ahhoz, hogy elérjük célunkat először arról kellett információt szereznünk, hogy egyes években hány darab tudományos közlemény született a szociológia területén. Dilemmát okozott az a kérdés, hogy biztosan csak a szociológiára szűkítsük-e kutatásunkat, vagy válasszunk más, rokon tudományágakat - például szociális munka, statisztika, feminista- és nőkutatások - is a kereséshez, hiszen a JSTOR-on összesen 74 tudományterület közül lehet választani[14]. Azt találtuk azonban, hogy mivel a kategóriák gyakran fedésben vannak egymással[15] - tehát egy cikk több halmazba is tartozhat - így a számunkra releváns cikkek mindenképpen rendelkeznek a szociológia címkével, míg azok, amelyek például csak a statisztika témakörbe tartoznak, nem jelentik kutatásunk tárgyát. Kizárólag a szociológia körébe tartozó cikkek halmazán dolgoztunk tehát tovább.

JSTOR Data For Research

Az egyes években született publikációk teljes számának kigyűjtéséhez Python-ban írtunk egy programot, amely tetszőlegesen generált évszám-listán (pl. 1886-tól 2018-ig) iterál végig, minden évszám-párhoz (tehát minden egyéves intervallumhoz) keresést végez a JSTOR Data for Research felületén, majd a visszakapott HTML állományból kimenti a találatok számát[16]. A keresést olyan módon végeztük, hogy az összes, a szociológia témakörébe tartozó találat számát adja vissza a rendszer, tehát nem szűkítettünk csupán a könyvekre vagy a folyóiratcikkekre.

Miután előállt az alaphalmazunk, kulcsszavas kereséseket végeztünk a Data For Research felületen. A kulcsszavakat, valamint a hozzájuk tartozó találatok számát az alábbi táblázat ismerteti.

1. táblázat: az egyes keresőkifejezések csoportosítása, valamint a keresésekhez kapcsolódó találatok száma

Témakör	Adathalmaz neve	Kulcsszó	Időintervallum	Találat
Natural Language Processing	sociology-natural_language_processing	natural language processing	-	5 803
Szövegelemzés	sociology-text_analysis_1800-1985	text analysis	1800-1985	22 039
	sociology-text_analysis_1986-2000	text analysis	1986-2000	19 250
	sociology-text_analysis_2001-2010	text analysis	2001-2010	19 942
	sociology-text_analysis_2011-2018	text analysis	2011-2018	18 627
	sociology-qualitative_text_analysis	qualitative text analysis	-	13 412

[14] Lásd a Data For Research keresőjét: https://www.jstor.org/dfr/
[15] Például a Delirious Naples: A Cultural History of the City of the Sun című mű (https://www.jstor.org/stable/j.ctv8bt213) egyszerre tartozik a Sociology, Language & Literature, valamint Art & Art History kategóriákba.
[16] A program forráskódja megtalálható a Függelékben.

	sociology-quantitative_text_analysis	quantitative text analysis	-	12 724
	sociology-text_mining	text mining	-	6 438
Computational Social Science	sociology-computational_social_science	computational social science	-	2 648
Adattudomány	sociology-data_science_1800-1970	data science	1800-1970	19 134
	sociology-data_science_1971-1985	data science	1971-1985	21 371
	sociology-data_science_1986-1998	data science	1986-1998	22 392
	sociology-data_science_1999-2008	data science	1999-2008	22 809
	sociology-data_science_2009-2015	data science	2009-2015	22 632
	sociology-data_science_2016-2018	data science	2016-2018	5 182
Adatbányászat	sociology-data_mining	data mining	-	13 526
Big Data	sociology-big_data_1800-2007	big data	1800-2007	24 263
	sociology-big_data_2008-2018	big data	2008-2018	17 792
NLP módszerek	sociology-topic_model_1800-2005	topic model	1800-2005	19 971
	sociology-topic_model_2006-2018	topic model	2006-2018	17 547
	sociology-topic_segmentation	topic segmentation	-	1 466
	sociology-sentiment_analysis	sentiment analysis	-	21 971
	sociology-word_embedding	word embedding	-	1 327
	sociology-text_classification	text classification	-	12 418
	sociology-text_clustering	text clustering	-	1 888
	sociology-automatic_summarization	automatic summarization	-	37

Az egyes kulcsszavakat az áttekinthetőség érdekében csoportosítottuk. Létrehoztunk egy kategóriát az NLP-ben használt módszereknek, ide soroltuk a topic model, topic segmentation, sentiment analysis, word embedding, text classification, text clustering, automatic summarization kifejezéseket. Összevontuk a hagyományos szöveganalitikára vonatkozó kereső kifejezéseinket, így a qualitative text analysis, a quantitative text analysis és a text analysis egy kategóriába került. A többi kulcsszót önállóan használtuk, továbbá létrehoztunk egy egyéb cikkek kategóriát is az összehasonlíthatóság kedvéért, így összesen nyolc kategóriát kaptunk. A kulcsszó-kategória megfeleltetéseket a fenti táblázat ismerteti.

A táblázatban minden sor egy-egy archívumot takar, amelyek metaadatokat tartalmaznak az adott keresés során megtalált cikkekről és könyvekről. Ezek a metaadat-halmazok hierarchikus struktúrával rendelkező XML fájlok, amelyek információt szolgáltatnak egyebek mellett az adott tartalom típusáról (research article, review article, book, book review, stb.), nyelvéről, a megjelenés helyéről és idejéről, de gyakran tartalmaznak terjedelmet és különböző azonosítószámokat (pl. DOI) is. Ezekből a gazdag adathalmazokból számunkra csupán a megjelenés idejére illetve az alkalmazott kulcsszóra volt szükség.

Miután meggyőződtünk arról, hogy a keresőszavainkra túlnyomórészt valóban releváns találatok érkeznek, kidolgoztunk egy eljárást a duplikátumok kiszűrésére. Mivel a keresőkifejezések között

számos olyan van, amely rokonterületeket, vagy egymásba érő halmazokat reprezentál, ezért számítottunk arra, hogy valami alapján ki kell törölnünk a többszörös találatokat. Szerencsére a fájlnevek tartalmazzák az adott tartalom JSTOR-on belüli egyedi azonosítóját, ezért a fájlnevek alapján végeztük a szűrést: amelyik fájlnév már szerepelt a feldolgozás során, az többször nem kerül be a végső adathalmazba. Így az eredetileg kapott, több mint 366 ezer találatot végül 208 673-ra szűkítettük. Bár az adatokhoz keresőkifejezések segítségével jutottunk hozzá, az elemzést már tágabb témakörönként végeztük, emiatt nem volt lényeges, hogy egy témakörön belül pontosan melyik keresőszavaknál fordult elő az adott cikk: az eredmények szempontjából csupán az az érdekes, hogy beletartozik-e az adott témakörbe.

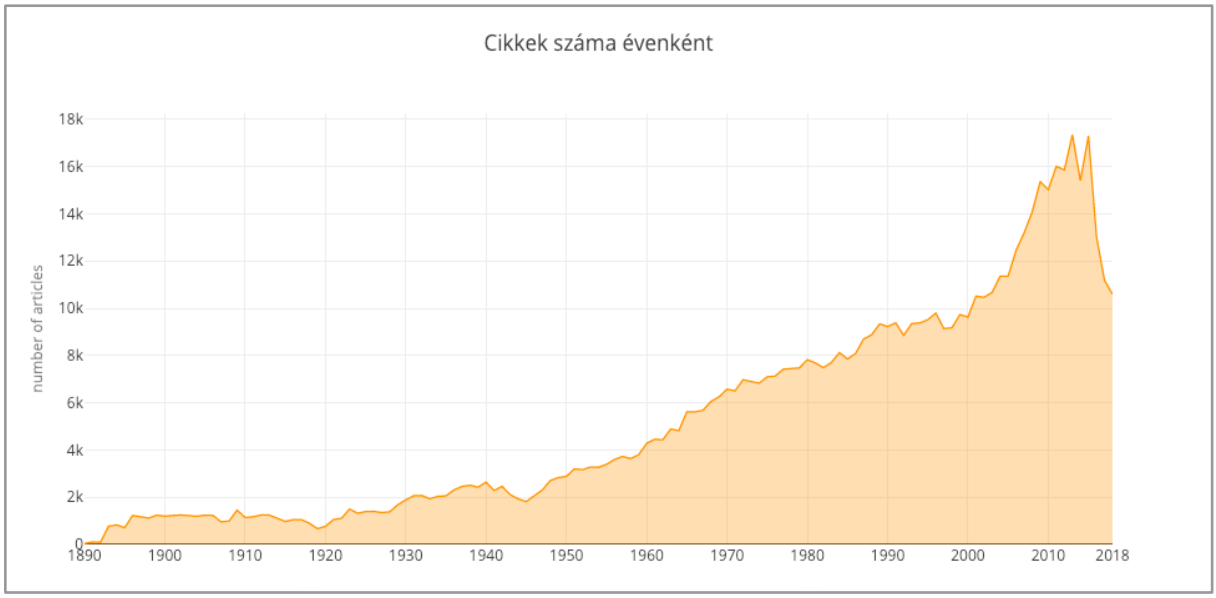
Az adatok feldolgozására megírt függvény tehát végigmegy az összes archívum összes fájlján, azokból pedig kigyűjti a számunkra releváns adatokat: a keresőkifejezést, a megjelenés idejét, publikáció típusát és nyelvét, valamint az egyedi azonosítót. Az adathalmazt ezt követően egy pivot táblába összesítettük az évek és az 1. táblázatban látható témakörök szerint.

Az ismertetett keresést 1886-től[17] 2018-ig futtattuk - ez az idei év kivételével lefedi a JSTOR teljes, szociológia témakörbe tartozó archívumát.

Eredmények

A cikkek számának változása

1. ábra: szociológia témájú cikkek száma évenként, 1890-2018



[17] A legrégebbi, JSTOR-on található szociológiai cikk 1886-os, azonban ezt követően négy évig egyetlen tartalmat sem találtunk, így a jobb vizualizáció érdekében adatainkat 1890-től jelenítjük meg.

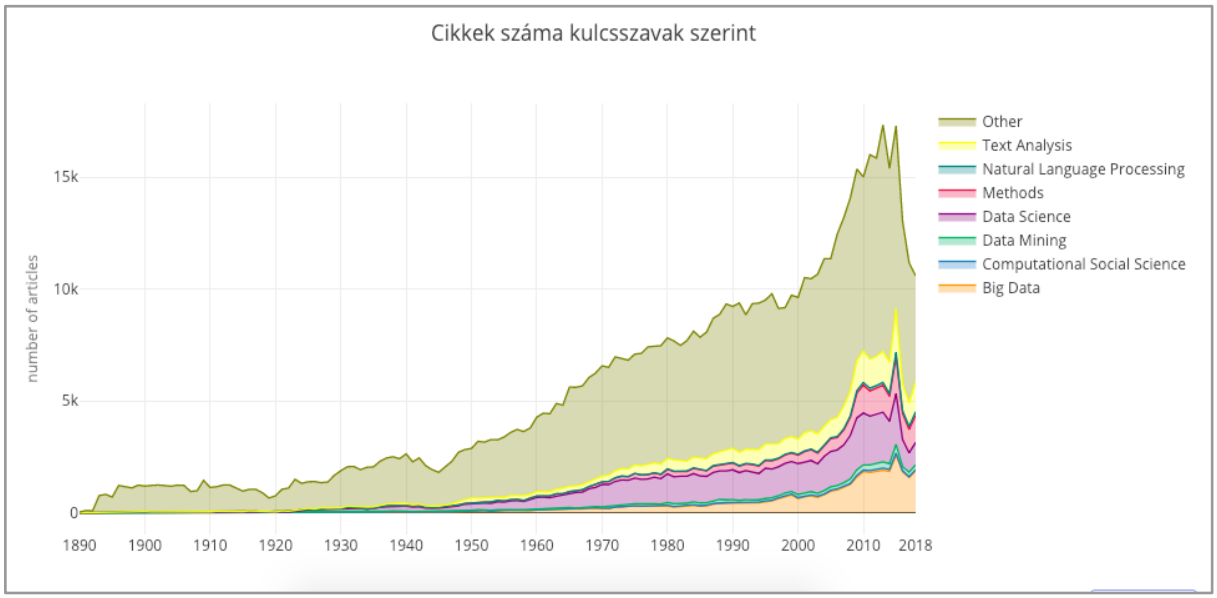
Látható, hogy a cikkek száma dinamikusan emelkedik 1900-tól kezdve. Az első jelentősebb megtorpanás a növekedés sebességében a második világháború környékén látható, illetve az 1980-as években is volt egy stagnáló fázis, majd szintén kisebb visszaesés mutatkozik a '90-es évek végén. A 2000-es évek elejétől ugyanakkor rendkívüli ütemű növekedés látható, amelyet követően 2013-15-ben tetőzik az új publikációk száma. Elgondolkodtató lehet, hogy vajon miért tapasztalunk ekkora visszaesést a cikkek számában az utolsó néhány, három-négy évben. Ennek az az oka, hogy a JSTOR a legtöbb folyóirattal olyan megállapodást köt, amelynek alapja egy „mozgó fal" a folyóirat legfrissebb száma és a JSTOR-on megjelenő legújabb kiadás között. Ez a gyakorlatban azt jelenti, hogy a folyóirat új száma a megjelenéssel nem kerül ki automatikusan a JSTOR-ra, hanem csak 2-3 év múlva lesz elérhető. Ez elsősorban a kiadó érdekét szolgálja, hiszen ha az legújabb szám is szabadon hozzáférhető lenne, akkor senki sem fizetne a folyóirat elolvasásáért. A „mozgó falon" kívül néhány kiadócéggel „fix fal" megállapodása van a JSTOR-nak, az ebben érintett folyóiratok esetében egy adott dátumnál frissebb tartalmak nem jelenhetnek meg a platformon. Kijelenthető tehát, hogy az utolsó néhány év visszaesése nem a szociológiai tudomány hanyatlásának jele, hanem a JSTOR működésének jellegéből fakadó sajátosság, limitáció.[18]

A vizsgált módszerek terjedése

A tárgyalt módszertan alapján előálló adathalmazt szerettük volna minél befogadhatóbb módon megjeleníteni, ezért egy interaktív adatvizualizációs eszköz mellett döntöttünk. Az ábra elkészítéséhez a Python plotly, numpy és pandas csomagjait használtuk. Az alábbiakban a vizualizációk statikus verzióját közöljük - az interaktív változatok megtekinthetők a https://aknap.eu/nlp-szociologia weboldalon.

[18] Forrás: https://about.jstor.org/whats-in-jstor/journals/

2. ábra: szociológia témájú cikkek száma évenként, kulcsszavak szerinti bontásban, 1890-2018



Az általunk vizsgált témákhoz tartozó cikkek arányának növekedése szintén megfigyelhető. Míg az ezredforduló előtti időszakban nem érte el a megnevezett témakörökbe eső tartalmak aránya az összes találat egyharmadát, addig 2015-ben - ami az adatok alapján az utolsó olyan év, amely a JSTOR működési jellegéből fakadóan teljesértékűen elemezhető - meghaladta annak felét is. A azt mondhatjuk tehát, hogy egyértelmű növekedési trend látszik az új típusú szöveg- illetve adatelemzési módszerek terjedésében a szociológián belül, de áttörésről egyelőre nem beszélhetünk.

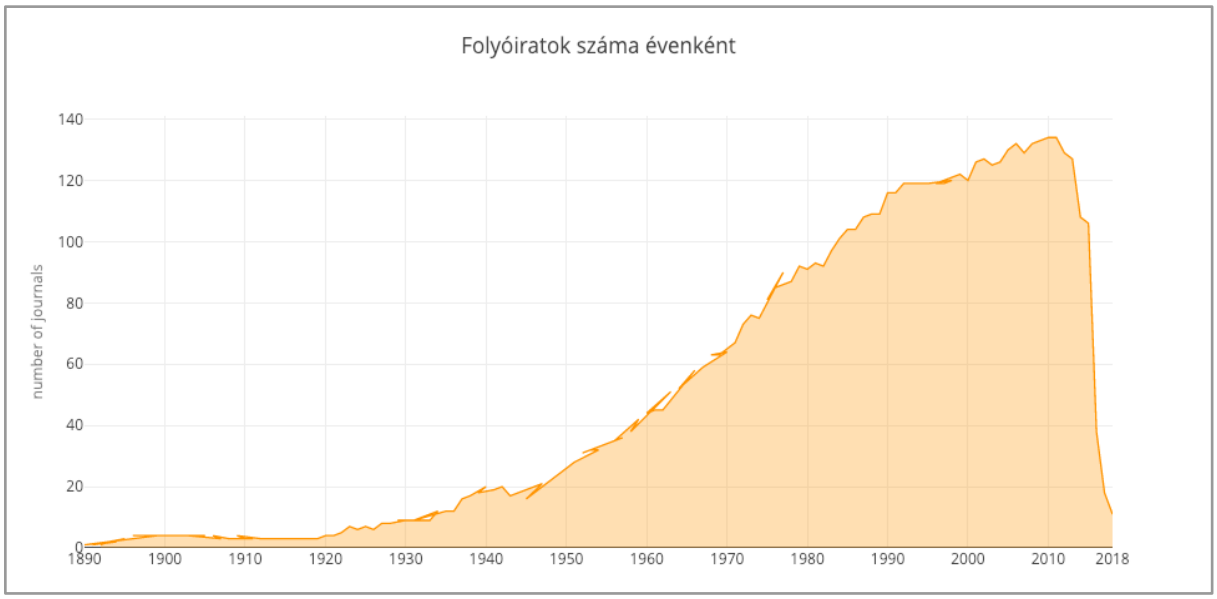
A továbbiakban az egyes témakörökhöz tartozó számokhoz külön-külön is, a teljesség igénye nélkül igyekszünk interpretációt adni. A „big data" kifejezés elterjedése, buzzword-ként történő használata a 2000-es évek elejére datálható, azóta ez a szókapcsolat szinte „mém" lett a tudomány és az üzlet világában, olyan dolog, amiről mindenki hallott már, de pontosan senki sem tudja, hogy mi az (Donoho 2017, 747). Adatainkból látható, hogy a kifejezés elterjedése egybeesik a szakirodalom megállapításaival: a szó „karrierje" 2003-tól ível fölfelé. Az is megállapítható azonban, hogy az ezredforduló előtti időszakban is van évente néhány száz eset. Itt feltehetően fals találatokkal van dolgunk, ami nem különösebben meglepő annak fényében, hogy a kifejezés alkotó szavai külön-külön mennyire gyakran használatosak. A „data science" kifejezés ennél is később, a 2010-es évektől kezdve vált buzzworddé, azonban ezt a szókapcsolatot már korábban is használták, egyebek mellett a „computer science" szóösszetétel helyettesítésére, majd a '70-es évektől kezdve számítógépes adatfeldolgozás kapcsán. Ennek ellenére itt is valószínű, hogy számos hibás találat került az adathalmazunkba. A többi témakör esetében nem láthatunk olyan számokat, amelyek arra engednének következtetni, hogy az egyes keresőkifejezések nem működtek megfelelően. A felsorolt problémák orvoslására egy későbbi kutatás keretében jelen írásunk korábbi részében ismertetett eljárást hívhatnánk segítségül. A JSTOR által szolgáltatott

metaadatfájlok gyakran tartalmaznak absztraktot is, ezek tartalmának klaszterezésével feltehetően kiszűrhetők lennének a téves találatok.

A folyóiratok számának változása

A kulcsszavas kereséseket követően letöltöttük a JSTOR-ról az összes folyóiratcikk metaadatát annak érdekében, hogy az egyes években előforduló folyóiratok számát megtudhassuk. Ez 514 950 darab fájl feldolgozását jelentette. Módszerünk lényege az, hogy a cikkeken végigiterálva a folyóiratok azonosítóját egy, az adott évhez tartozó listába tesszük, de csak abban az esetben, ha még nem szerepel az adott folyóirat az adott évben, majd ezt követően egyszerűen összeszámoljuk, hogy évenként hány darab folyóirat jelent meg az adatokban. Ezzel az egyszerű módszerrel ugyan azt nem tudjuk detektálni, ha egy folyóirat beszünteti működését, azonban a folyóiratok számának alakulását így is figyelemmel kísérhetjük.

3. ábra: a JSTOR-on megjelenő folyóiratok száma a szociológia témájú cikkekben, 1890-2018



Az ismertetett adatfeldolgozási mód miatt ábránkon néhol előfordul, hogy visszaesés mutatkozik egyes évek között. Az azonban világosan látszik, hogy a JSTOR-on fellelhető folyóiratok számának növekedése 1950-től 1990-ig dinamikusnak mondható, ezt követően közel egy évtizedes stagnálás mutatkozik. Ezt a megközelítőleg 120 folyóiratos szintet valamelyest meghaladja a 2000-es évek, majd korábbi adatainkkal összhangban csökkenés mutatkozik a 2010-es évektől kezdődően. Az talált adatok reálisnak tűnnek annak tükrében, hogy a JSTOR-on, összetett keresés definiálásakor 151 folyóirat közül lehet választani a szociológia tudományágban[19].

[19] https://www.jstor.org/action/showAdvancedSearch

További elemzési lehetőségek

A JSTOR által biztosított gazdag adathalmaz rengeteg lehetőséget nyújt a kutatás továbbviteléhez. A Data For Research felületen kiérhető metaadatok segítségével a társszerzőségek hálózati megközelítésű elemzésére nyílik lehetőség. Egy ilyen vizsgálat során felderíthető lenne, hogy mely tudományterületek képviselőinek segítségével terjednek át egyes módszerek a társadalomtudományokba, milyen diffúziós folyamatok játszódtak le a diszciplínák között. A módszerek elterjesztésében kulcsszerepet vállaló szerzők azonosítása is lehetséges lenne.

Egy másik megközelítéssel azonosítani lehetne azokat a folyóiratokat, amelyek úttörő szerepet vállaltak egyes módszerek elterjesztésében. Természetesen ehhez sokkal jobban szükséges specifikálni a kereséseket, mivel az esetleges fals találatok helytelen következtetések levonását eredményezhetik.

Végezetül a JSTOR beépített szövegelemző szolgáltatásával lehetőség van letölteni a cikkekhez tartozó bi-, tri-, illetve unigramokat, amelyek elemzése tovább segíthet a téves találatok felismerésében, klaszterezésükkel pedig lehetőség nyílna a szociológia fókuszpontjainak feltérképezésére a különböző időszakokban.

Hivatkozások listája

Ahktar, J. – Soria, S. (2009). „Sentiment Analysis: Facebook Status Messages." *Stanford University Technical Report*.

Antal, L. (1979). *A tartalomelemzés alapjai*. Magvető Könyvkiadó, Budapest.

Bakshy, E. – Messing, S. – Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science* 348 (6239): 1130-1132.

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

Blei, D. M. (2011). „Introduction to probabilistic topic models". *Communications of the ACM*.

Blei, D. M. and Lafferty, J. D. (2006) „Dynamic topic models". *International Conference on Machine Learning*, New York, NY, USA, pages 113-120

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

Csepeli György. (2015). „A szociológia és a Big Data." *Replika* 3-4 (92–93): 171-176.

Deerwester, S. – Dumais, S.T. – Furnas, G.W. – Landauer, T.K. – Harshman, R.: (1990). „Indexing by latent semantic analysis". In *Journal of the Association for Information Science and Technology*.

Dessewffy, T. – Láng, L. (2015). „Big Data és a társadalomtudományok találkozása a műtőasztalon". *Replika* 92–93, 157–170. http://replika.hu/system/files/archivum/92-93_11_dessewffy_lang.pdf [Letöltve: 2019. 01. 25.]

Donoho, D. (2017). „50 Years of Data Science." *Journal of Computational and Graphical Statistics* (Taylor & Francis) 26 (4): 745-766. doi:10.1080/10618600.2017.1384734.

dos Santos, C. – Gatti M. (2014). „Deep convolutional neural networks for sentiment analysis of short texts". *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, Dublin, Ireland https://www.aclweb.org/anthology/C14-1008 [Letöltve: 2019. 01. 25.]

Edelman, B. – Luca, M. (2014). „Digital Discrimination: The Case of Airbnb.com". *Harvard Business School Working Paper.*

Garg, N., Schiebinger, L., Jurafsky, D., Zou, J. (2018). „Word embeddings quantify 100 years of gender and ethnic stereotypes". *Proceedings of the National Academy of Sciences*, 115 (16). https://www.pnas.org/content/pnas/115/16/E3635.full.pdf [Letöltve: 2019. 01. 25.]

Herring, Susan C. (2010). „Web Content Analysis: Expanding the Paradigm". In Hunsinger, J. – Allen, M. – Klastrup, L. eds. *The International Handbook of Internet Research*. Springer Verlag, Dordrecht. 233-249. https://www.sfu.ca/cmns/courses/2012/801/1-Readings/Herring%20WebCA%202009.pdf [Letöltve: 2019. 01. 25.]

Kao, A. – Poteet, S. R. (2007). *Natural Language Processing and Text Mining*. Springer-Verlag London. http://129.219.222.66/publish/pdf/natural_language_processing_and_text_mining.pdf [Letöltve: 2019. 01. 25.]

Kinnunen, T. (2017). „Hate speech detection". *futurice.com* https://futurice.com/blog/hate-speech-detection [Letöltve: 2019. 01. 25.]

Kmetty, Z. – Koltai, J. – Bokányi, E. – Bozsonyi, K. (2017). „Seasonality Pattern of Suicides in the US – a Comparative Analysis of a Twitter Based Bad-mood Index and Committed Suicides". In *Intersections*. http://intersections.tk.mta.hu/index.php/intersections/article/view/302 [Letöltve: 2019. 01. 25.]

Krippendorff , K (2004) „Content Analysis. An Introduction to its Methodology".

Liddy, E. D. (2001). „Natural Language Processing". In *Encyclopedia of Library and Information Science*, 2nd Ed. NY. Marcel Decker, Inc. https://surface.syr.edu/cgi/viewcontent.cgi?article=1043&context=istpub [Letöltve: 2019. 01. 25.]

Rehurek, R. (2011). *Scalability of Semanic Analysis in Natural Language Processing*. https://radimrehurek.com/phd_rehurek.pdf [Letöltve: 2019. 01. 25.]

Reményi Andrea (2010). „Kollokációk korpuszalapú vizsgálata". In: *Fordítástudomány* XII. 2. szám, 67–95.

Roberts M. E. – Stewart B. M. – Tingley D. – Airoldi E. M. (2013). „The Structural Topic Model and Applied Social Science". In *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*. https://scholar.princeton.edu/files/bstewart/files/stmnips2013.pdf [Letöltve: 2019. 01. 25.]

Ságvári, B. (2017). „Társadalomtudomány a Big Data korában". *Statisztikai Szemle* 95 (5), 491-504. http://real.mtak.hu/54832/1/2017_05_491.pdf [Letöltve: 2019. 01. 25.]

Sebők, M. (szerk,) (2016). *Kvantitatív szövegelemzés és szövegbányászat a politikatudományban*. L'Harmattan Kiadó, Budapest. https://qta.tk.mta.hu/uploads/files/Kvantitativ_szovegelemzes_keszpdf.pdf [Letöltve: 2019. 01. 25.]

T. Hofmann (1999). *Probabilistic latent semantic indexing*. http://ciir.cs.umass.edu/publes/ir-464.pdf [Letöltve: 2019. 01. 25.]

Thelwall, M. – Wilkinson, D. – Uppal, S. (2010). „Data Mining Emotion in Social Network Communication: Gender Differences in MySpace." *Journal of the American Society for Information Science and Technology* 61: 190–99.

Thelwall, M., – Buckley, K. – Paltogou G. (2011). „Sentiment in Twitter Events." *Journal of the American Society for Information Science and Technology* 62: 406–18. https://www.researchgate.net/publication/317300025_Deep_Learning_for_Hate_Speech_Detection_in_Tweets [Letöltve: 2019. 01. 25.]

Tikk, D. – Farkas, R. – Kardkovács, Z. T. – Kovács, L. – Répási T. – Szarvas G. – Szaszkó S. – Vázsonyi M. (2007). *Szövegbányászat*. Typotex, Budapest.

Zhang, Y. – Wildemuth, B. M. (2005). *Qualitative Analysis of Content.* 1 (2):1-12 https://www.ischool.utexas.edu/~yanz/Content_analysis.pdf [Letöltve: 2019. 01. 25.]

Függelék

A tanulmányban megtalálható ábrák interaktív verziói elérhetők a https://aknap.eu/nlp-szociologia oldalon. Szintén ezen az oldalon található az adatfeldolgozás során használt programkód is.