Nagy Andor
PTE KPVK

Automatizált tartalomelemzés

Tagozat: Könyvtártudomány
Témavezető: Dr. habil. Varga Katalin

1. Bevezetés

1.1. A kutatásom előzményeiről

Évek óta a napi rutinom része néhány magyar hírportál cikkeinek gyors átolvasása. Megfigyeltem, hogy sok esetben hemzsegnek a helyesírási hibáktól, és vannak olyan tévesztések, amelyek alapján meg lehet mondani, hogy ki írta a konkrét cikket. Sőt, az egyes szerzők fogalmazási módja is jól megkülönböztethető, pl. van, aki távolságtartóan fogalmaz, de olyan is van, akinek a közvetlen hangvételű, sok jelzővel megtűzdelt cikkei a védjegyévé váltak.

2014-ben belekezdtem egy olyan kutatásba, amely a jelen tanulmányom előzményének is tekinthető. Kíváncsi voltam, hogy a cikkek egyedi jellemzőinek módszeres vizsgálatával milyen összefüggések tárhatók fel. Például feltettem a kérdést, hogy van-e olyan befolyásoló tényező, amely hatással van az egyes szerzők szóhasználatára, helyesírására, fogalmazási módjára. Ilyen tényező lehet pl. a cikk megírásának napja, a különböző ünnepek közelsége, a cikk rovata, a szerző hozzáállása az adott témához stb.

Rögtön két problémával is szembesültem. Az egyik, hogy az általam vizsgált két hírportálon, az *Index*en[1] és az *Origó*n[2] naponta közel 100-100 cikket publikálnak, amely akkora adathalmazt jelent, hogy egy embernek túl nagy feladat lenne a módszeres áttekintése. A másik problémát a saját szubjektivitásom jelentette, mert attól tartottam, hogy akármennyire is próbálnék objektív maradni, az eredmények kiértékelésében megjelenne a saját nézőpontom is.

Ezeket figyelembe véve döntöttem úgy, hogy kidolgozok egy olyan automatizált megoldást, amely megkönnyíti a dolgomat, és mellette objektív maradhatok. A kutatásom eredményeiről 2015-ben publikáltam (Nagy 2015). Az automatizált tartalomelemzés kifejezés létezéséről ekkor még csak hallomásból volt tudomásom, és csak akkor tudatosult bennem, hogy valójában automatizált tartalomelemzést végeztem, amikor elkészültem a kutatással, és az egyik oktatóm elolvasta a tanulmányomat. Ezt követően

[1] Index http://index.hu [2016. 10. 01.]
[2] Origo http://origo.hu [2016. 10. 01.]

kezdtem bele abba a kutatásba, amelynek eredményeiről a jelen tanulmányban számolok be.

1.2. A tartalomelemzésről

A *content analysis* (tartalomelemzés) kifejezést először a 20. században használták (Waples, Berelson, Bradshaw 1940), de a szövegek szisztematikus elemzésével már az egyházak is foglalkoztak a 17. században, mivel ekkoriban kezdtek elterjedni a nem egyház által közreadott nyomtatványok, és erre a folyamatra aggódva tekintettek az egyházi vezetők. Először a 17. századi Svédországban végeztek olyan tartalomelemzést, amelynek módszerei hasonlítanak a mai tartalomelemző megoldásokra (Bexter, Babbie 2003). A *Cion énekei* című, 90 tételes, ismeretlen szerzőtől származó pietista zsoltárgyűjteményt elemezték abból a célból, hogy bebizonyítsák: olyan szektás elemeket tartalmaz, amelyek veszedelmesek a hivatalos papságra. Először csak a szimbólumokat számolták össze, ám ezek számában nem találtak jelentős eltérést a jóváhagyott daloskönyvekhez képest, de amikor a szimbólumokat a kontextusukkal együtt elemezték, akkor már igen.

A ma is használatos technikák csak jóval később, az első világháborút követően alakultak ki, és ebben az időben indult ugrásszerű fejlődésnek a számítástechnika is. A tartalomelemzés és a számítástechnika együttes fejlődése nem véletlen. A háborúban szemben álló felek igyekeztek olyan információkat kinyerni egymás kommunikációjából, amelyek feltárják a szöveg elsődleges jelentésén túli látens tartalmat, és ennek egyik legfőbb eszköze a számítógépek által automatizált tartalomelemzés.

1.3. Az automatizált tartalomelemzésről

Fontosnak tartom kiemelni az *automatizált* jelzőt, mivel úgy gondolom, hogy az automatizált tartalomelemzés nem a tartalomelemzési technikák egy újabb eszköze, hanem egy egészen más megközelítést igénylő tevékenység. Az évtizedek alatt kidolgozott tartalomelemzési módszertanok még nem számolhattak olyan számítási teljesítménnyel, mint amilyenekre a mai számítógépek képesek. Korábban ugyan léteztek olyan tartalomelemzési technikák, amelyeket kifejezetten az automatizálás céljából hoztak létre, de a legtöbb az emberi érzékszervekre, az emberi gondolkodásra épít, így csak jelentős változtatásokkal ültethetők át modern számítástechnikai környezetbe.

Amikor ma egy nagy teljesítményű számítógépen automatizáltan elemzünk egy szöveget, akkor olyan új megoldásokat alkalmazhatunk, amelyek korábban szóba se jöhettek, mert értelmetlen lett volna kidolgozni egy olyan algoritmust, amely évek alatt futott volna le. Persze az automatizált tartalomelemzési technikák egy része a hagyományos tartalomelemzési

technikákat veszi alapul, de sok, korábban alapvetésnek tartott megállapítás már nem állja meg a helyét, ha a mai értelemben vett automatizált tartalomelemzésről beszélünk.

2. Az automatizált tartalomelemzés lehetséges alkalmazási területei

A számítástechnika fejlődésének köszönhetően olyan új alkalmazási területek előtt nyílt meg a tartalomelemzés lehetősége, amelyeket korábban azért nem vettek számításba, mert még nem születtek meg a szükséges technikai vívmányok. Pl. a nagyfelbontású fényképezőgépeknek hála a számítógép ma már a festményeken megjelenő ecsetvonásokat is képes kiértékelni és összehasonlítani akár több ezer másik festmény ecsetvonásaival. Ezek alapján egy algoritmus képes lehet összepárosítani azokat a festményeket, amelyek egyazon ecsettel készültek, vagy ugyanazon művész munkái, sőt, az ecset nyomvonalának mélységéből és irányából akár még az alkotó pillanatnyi lelkiállapotára is lehet következtetni. Mindezt természetesen teljesen automatizált módon, egy számítógépes algoritmus segítségével.

Az értelmezésem alapján tartalomelemzésnek nevezhetünk minden olyan tudatos vagy tudattalan cselekedetet, amely a minket körülvevő világ makro- vagy mikroszintű elemzésére irányul. Tartalomelemzést végezhetünk minden olyan emberi cselekedeten, szóbeli vagy írásos megnyilatkozáson, emberi vagy természeti alkotáson, amely mögött valamilyen tartalmat vélünk felfedezni, vagy amelyekben a felszínen nem látható összefüggésekre szeretnénk rávilágítani.

3. Az automatizált tartalomelemzés módszerei

Amit ma értünk automatizált tartalomelemzésen, annak nincs egy olyan szakirodalma sem, amely teljes egészében körüljárná a diszciplínát, és részletesen bemutatná az összes automatizált tartalomelemző módszert. Még az ezzel foglalkozó kutatók is csak tapogatóznak, nincs olyan oktatóanyag, amely egyértelműen leírja, hogy milyen feladatra milyen algoritmust érdemes alkalmazni. Az automatizált tartalomelemzés – ahogyan azt korábban említettem – az első számítógépek megjelenésével párhuzamosan kezdett el kialakulni, de az igazi fejlődés csak a kétezres évektől kezdődött el, ugyanis a számítógépek teljesítménye akkor érte el azt a szintet, hogy a korábban napokig tartó kiértékelési folyamat néhány órára redukálódott. Ezzel együtt át kellett értékelni a korábban kidolgozott módszertanokat, hiszen nagyon sok új megközelítés vált lehetővé.

A kutatásomban az írott szövegek tartalomelemzésének folyamatára fókuszálok, ezen belül is a hírportálokon megjelenő tartalmak elemzésére. Teszem ezt azért, mert amatőr programozóként csak az írott szövegek tartalomelemzésének menetét tudom hitelesen bemutatni. A feltárt technikák alkalmazhatók más típusú szövegek elemzéséhez is, de az automatizált tartalomelemzés sajátossága, hogy nem létezik olyan univerzálisan hasznosítható megoldás, amely minden műfajú szöveg elemzésekor hatékonyan működik. Az alapelv hasonló egy hír és egy irodalmi szöveg elemzésénél, de az alkalmazott technikákat mindig a célnak megfelelően kell módosítani. Léteznek kereskedelmi forgalomban kapható szoftverek, amelyek képesek – elsősorban statisztikai módszerekkel – az automatizált tartalomelemzés egyes formáira, de mivel nem tudjuk, hogy ezek pontosan hogy működnek, nem látjuk a működésük közben kiszámított részeredményeket, ezért nincs lehetőségünk olyan alapos ellenőrzésre, mint amikor egy konkrét adatbázishoz szabott programot használunk.

4. A kutatásom célja és hipotéziseim

A kutatásom célja a szövegek esetében alkalmazható legelterjedtebb automatizált tartalomelemző módszerek feltérképezése és a működésük bemutatása. Kézzelfogható segítséget szeretnék nyújtani azoknak a kutatóknak, akik ebben fantáziát látnak és szeretnék bővíteni a kutatási módszereik eszköztárát, valamint szeretném felhívni a könyvtáros szakemberek figyelmét is az automatizált tartalomelemzésben rejlő lehetőségekre. Lehetetlen lenne valamennyi területre kitérni, ahol az automatizált tartalomelemzés alkalmazható, ezért én kifejezetten a webes hírportálok tartalomelemzésére fókuszáltam, mivel ez az a terület, ahol rengeteg írott szöveges tartalom található meg digitális formában

Amikor belekezdtem a kutatásomba, azt a hipotézist fogalmaztam meg, hogy viszonylag egyszerű algoritmusok segítségével is elvégezhetők olyan tartalomelemzések, amelyek az egyszerűségükhöz képest igen komoly eredménnyel járhatnak. A hipotézis bizonyítása vagy megcáfolása érdekében a tanulmányomban részletezett módszerek egy részét a gyakorlatba ültettem és egy igen nagy mintán végeztem automatizált tartalomelemzést.

5. A szövegek tartalomelemzésének a folyamata

A következőkben azt a folyamatot fogom részletesen bemutatni, amelyen mindenképpen végig kell mennie annak a kutatónak, aki valamiféle automatizált tartalomelemzést kíván végezni szöveges tartalmakon. A folyamat leírásához összesen három olyan munkát (Varga et al. 2015) vettem alapul, amelyek jó alapot szolgáltattak a munkámhoz, de igyekeztem gya-

korlatiasabb megközelítésből bemutatni a lépéseket, úgy, hogy azok kézzelfogható segítséget nyújtsanak egy automatizált tartalomelemző szoftver elkészítéséhez. A folyamatleírást a saját szoftvereim elkészítésével párhuzamosan alakítottam ki, támaszkodva a szakirodalomra és a saját gyakorlati tapasztalataimra, észrevételeimre.

5.1. A forrás kiválasztása

– Ez lehet valamilyen digitális szöveg, pl. a hírportálok cikkei, vagy akár egy korpusz/kifejezésgyűjtemény, amilyen az MTA Nyelvtudományi Intézetének a Magyar Nemzeti Szövegtára.
– Lehet nyomtatott formában létező szöveg, ezeket optikai karakterfelismerő szoftverrel lehet digitalizálni.
– Akár hangos szöveg is lehet az automatizált tartalomelemzés tárgya, ugyanis léteznek olyan megoldások, amelyek elég nagy pontossággal képesek írott szöveggé alakítani a szóban elhangzott tartalmat, de használhatunk feliratsávot is, pl. a tévéhíradókhoz a teletexten[3] általában elérhető a hallássérülteknek szánt feliratsáv.

A tanulmányom fontos részét képezi az a próbaként elvégzett automatizált tartalomelemzés, amelyet azért folytattam le, hogy a megismert automatizált tartalomelemző módszereket a gyakorlatban is kipróbáljam, és kézzelfoghatóbbá tegyem.

A forrásom 3 hírportál több hónapos cikktermése volt, ez összesen 40.540 cikket jelent. Ezeket a cikkeket automatizált módon mentettem adatbázisba úgy, hogy időzítve megnyitottam a hírportálok RSS-csatornáját, és onnan egy script segítségével adatbázisba mentettem minden tartalmat, dátummal, címmel ellátva. Áprilistól júniusig csak az *Index*[4] és az *Origo*[5] cikkeit mentettem le, majd júliustól elkezdtem lementeni a *Kuruc.info*[6] nevű hírportál cikkeit is, mert úgy gondoltam, hogy érdekes ellentétekre derülhet fény a másik két hírportállal szemben. Az adatgyűjtést április 1-jén kezdtem, és október 31-én fejeztem be.

5.2. A szöveg megtisztítása és a számítógép által értelmezhetővé tétele

– Bármilyen írott szöveg tartalomelemzését végezzük, biztos, hogy olyan elemeket is tartalmazni fog, amelyek nem a szöveghez tartoznak. Egy

[3] PCMag. http://www.pcmag.com/encyclopedia/term/52714/teletext [2016. 12. 27.]
[4] Index. http://index.hu
[5] Origo. http://origo.hu
[6] Kuruc.info. http://kuruc.info

könyvben ilyen lehet az oldalszám, a fej- és lábléc, a hivatkozások; egy hírportál cikkeinél pedig a reklámok, a szerzők nevei és a weboldal más elemei.
– Nagyon kell figyelni a szöveg kódolására[7], mert azonos nyelven belül is sokféle kódolást használnak az elektronikus szövegek megjelenítésére, és ez problémát jelenthet több különböző forrásból származó szövegek együttes elemzésénél. Ilyen kódolások pl. az *UTF-8* vagy az *ISO-8859*. Különböző kódolásoknál egységesítés után kell felvinni a szövegeket az adatbázisba.
– Miután már csak a tényleges tartalom jelenik meg az adatbázisban, ki kell szűrni azokat a szavakat, karaktereket, amelyek nem relevánsak a kutatás szempontjából. Ezek lehetnek pl. a névelők és a központozás, de mindig a konkrétan elérni kívánt cél határozza meg azt, hogy mit tekinthetünk irrelevánsnak.
– Ha a tartalomelemzés szempontjából releváns, akkor azt is meg kell adni a szoftver számára, hogy mit tekintsen teljes mondatnak, mondatkezdő nagybetűnek vagy éppen tulajdonnévnek. Ez azért jelent nehézséget, mert a számítógép nem értheti, hogy egy szövegrészletben azért van pont, mert lezár egy mondatot, vagy azért, mert egy rövidítést jelöl. Egy lehetséges megoldás, hogy a pontot akkor tekintjük mondatzáró pontnak, ha utána nagybetű vagy már semmi nem szerepel, de ilyenkor sem biztos a siker, hiszen a pont utáni nagybetű jelölhet mondaton belüli tulajdonnevet is.

5.3. Milyen kérdésekre keressük a választ az automatizált tartalomelemzés segítségével, milyen információkat szeretnénk kinyerni a szövegből?

Egy olyan témát választottam a szemléltetésképpen elvégzett automatizált tartalomelemzésemhez, amelyről tudtam, hogy sok cikk jelenik meg róla a három hírportálon, és várhatóan az adatgyűjtésem teljes 7 hónapja alatt fognak róla publikálni, vagyis lesz miből dolgoznom. Ez a téma a *bevándorlási hullám*. Arra voltam kíváncsi, hogy az áprilistól októberig terjedő időszakban hogyan változik az események tálalása, milyen kontextusban írnak a bevándorlásról, és van-e valamilyen különbség abban, ahogy a három hírportál számol be a fejleményekről.

[7] TechTerms. http://techterms.com/definition/characterencoding [2016. 10. 20.]

5.4. A szóba jöhető automatizált tartalomelemzési módszerek kiválasztása

Ehhez kapcsolódóan nem létezik egy olyan komplex lista, amely felsorolja az összes lehetséges módszert, mert az automatizált tartalomelemzés ugyan átvette a legtöbb klasszikus tartalomelemző metódust, de a számítógépek alkalmazásával szinte végtelenné vált a lehetséges megoldások tárháza. Kutatásomban a legelterjedtebb, leginkább kézzelfogható módszerekkel foglalkoztam. Ahogy azt korábban kifejtettem, szerettem volna elkerülni, hogy a tanulmányom egy száraz technikai ismertető legyen, ehelyett igyekeztem úgy bemutatni az eredményeimet, hogy azok kézzelfoghatók, a gyakorlatba is könnyen átültethetők legyenek. A következőkben a legelterjedtebb automatizált tartalomelemzési módszereket fogom ismertetni. Ezek közül többet felhasználtam a saját magam által elvégzett tartalomelemzésekben is, ezek eredményeiről az egyes módszerek ismertetéséről szóló fejezetekben számolok be.

5.4.1. Gyakoriság-elemzés

Ez a legegyszerűbb módszer, szám szerint megvizsgáljuk, hogy egy meghatározott adat hányszor fordul elő a szövegben. „Ez talán az egyik legegyszerűbb, a legkönnyebben gépesíthető módszer. »Minden olyan esetben, amikor a megfigyelt gyakoriságok arányai 'meglepő'-ek, arra van szükség, hogy explicitté tegyük azt a megoszlást, amellyel ezeket a megfigyeléseket összevetettük, s hogy maga az összevetés valóban igazolható legyen.« (Krippendorff 1995). Vagyis mindig ellenőrizni kell, hogy a gyakoriság mihez képest nagy vagy kicsi. Ennek érvényesnek kell lennie a kulcsszó-meghatározásra is." (Varga 2005: 53)

5.4.2. A változók közötti relációk elemzése

Ezzel a módszerrel bizonyos adatok együttes előfordulását térképezhetjük fel, így pl. megállapítható, hogy egy adott témáról hogyan vélekedik két különböző ember, mégpedig úgy, hogy a választott témára jellemző szavak környezetét vizsgáljuk külön az egyik és külön a másik ember esetében.

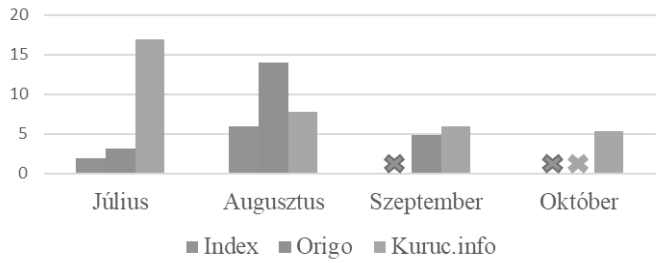
Ezt a módszert arra használtam, hogy vettem a „*bevándorló*" kifejezést és annak összes lehetséges szinonimáját, és ezeket azonosként kezelve feltérképeztem azokat a kifejezéseket, amelyekkel rendszeresen együtt szerepelnek. Ezeket számszerűsítettem, és két csoportot alkottam belőlük:

– az egyik az országok és nemzetek neveit tartalmazza;
– a másik pedig a kifejezetten negatív érzelmeket kiváltó kifejezéseket (pl. „*terrorista*", „*öngyilkos merénylő*" stb.).

Hónapokra és hírportálokra lebontva megnéztem azt a 25 kifejezést, amelyek leggyakrabban szerepelnek együtt a „*migráns*" kifejezéssel és annak

szinonimáival, ezek közül a leggyakrabban előfordulók: *terrorista, terrorizmus, Németország, Magyarország, Törökország, kormány, befogadás, unió, kvóta*.

Ezután kiválasztottam azokat, amelyek kifejezetten negatív tartalmat hordoznak, és diagramon ábrázoltam az eredményeket. Azért júliustól és nem áprilistól indulnak a diagramok, mert júliustól kezdve volt adatom mind a három hírportálról.



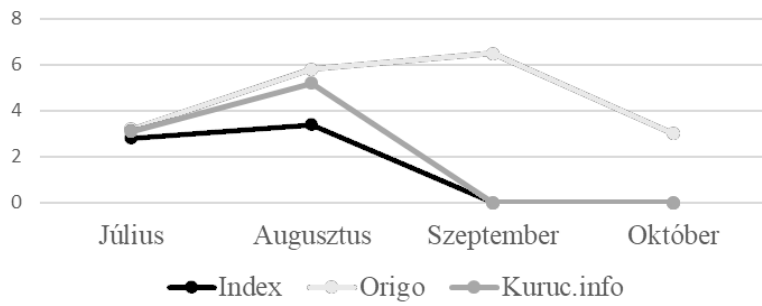
1. ábra Negatív érzelmeket kiváltó kifejezések használata (%)

A diagramon (1. ábra) a hírportálok százalékos értékei nem egymáshoz viszonyulnak, hanem azt mutatják, hogy egyenként az egyes hírportálokon a „*migráns*" és szinonimáival gyakran együtt használt kifejezések hány százaléka egyértelműen negatív.

A diagramon az látszik, hogy augusztus kivételével mindig messze a *Kuruc.infón* használták a legtöbb negatív érzelmet kiváltó kifejezést a migránsok kapcsán, a sorban a második az *Origo*, az *Indexen* pedig nagyon ritkán írtak negatív kontextusban a migránsokról. Szeptemberben és októberben már egy negatív kifejezés sem jelent meg a „*migráns*" és szinonimáinak szövegkörnyezetében. Augusztusban tetőzött az ilyen jellegű kifejezések használata, ekkor volt a riói olimpia, így feltehetően a biztonsági intézkedések kapcsán használták sokat a „*terrorista*", „*terrorizmus*", „*szükségállapot*" kifejezéseket.

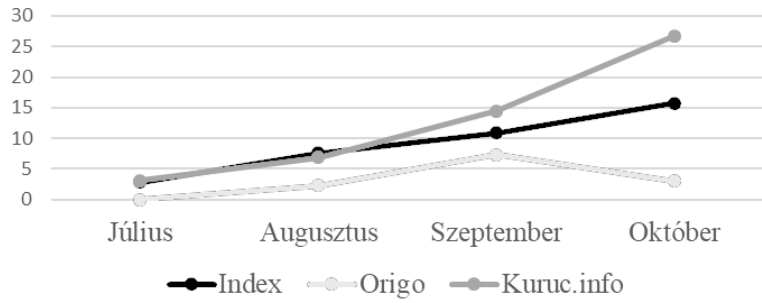
Az általam alkotott második csoport az országok és népek neveit tartalmazza, de az egyszerűség kedvéért nem tettem különbséget az országok és a hozzájuk tartozó népek említésénél, tehát pl. a „*Magyarország*" és a „*magyar*" kifejezéseket azonosként kezeltem. Összesen 7 ország (és nép) nevét említik a cikkek egy szövegkörnyezetben a „bevándorló" és annak rokon értelmű kifejezéseivel. A gyakoriságvizsgálat után a következő eredményre jutottam (2. ábra):

2. ábra Az egyes országok említésének gyakorisága. A legsötétebb színnel jelölt országot a legtöbbször, a legvilágosabb színnel jelölt országot a legkevesebbszer hozták szóba a hírportálok



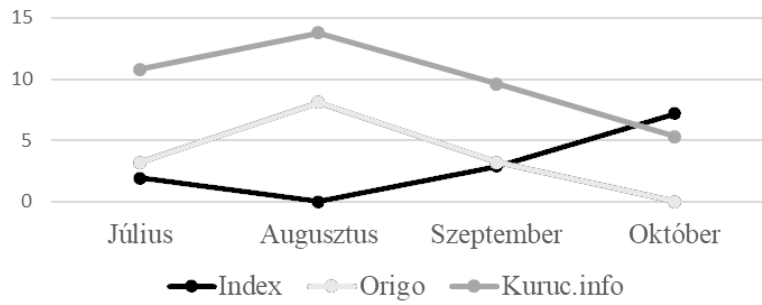
3. ábra Törökország említésének gyakorisága (súlyozás utáni darabszám)

A Törökországról való diskurzus az *Indexen* és a *Kuruc.infón* augusztusban tetőzött, az *Origo* cikkeiben viszont még szeptemberben is sokszor megemlítették az országot, és ellentétben a másik két hírportállal, októberben is több alkalommal szóba hozták (3. ábra).



4. ábra Magyarország említésének gyakorisága (súlyozás utáni darabszám)

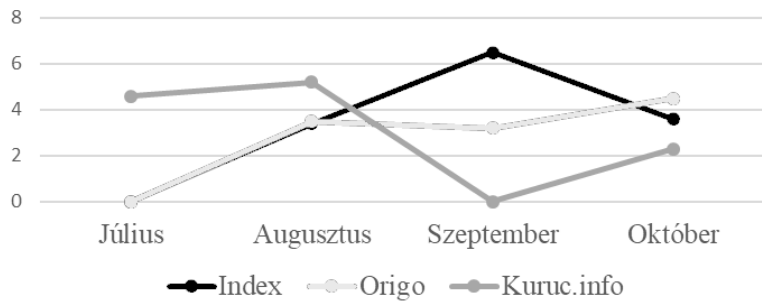
A „*Magyarország/magyar*" kifejezések előfordulását vizsgálva ellentétes irány figyelhető meg az előző diagramhoz képest, ez feltehetően az októberi népszavazás[8] következménye; a sajtót inkább a helyi események érdekelték (4. ábra).



5. ábra Németország említésének gyakorisága (súlyozás utáni darabszám)

A Németországgal kapcsolatos kifejezéseket az *Origo* és a *Kuruc.info* szinte mindig ugyanakkor használta, az Index viszont teljesen szembe ment ezzel a tendenciával. Ennek több oka lehet, pl. elképzelhető, hogy a *Kuruc.info* és az *Origo* egymástól vagy egy közös hírforrásból vettek át híreket, esetleg az *Index* kevésbé tartotta lényegesnek a németországi eseményeket (5. ábra).

[8] Magyarország Kormánya http://nepszavazas2016.kormany.hu [2016. 11. 20.]

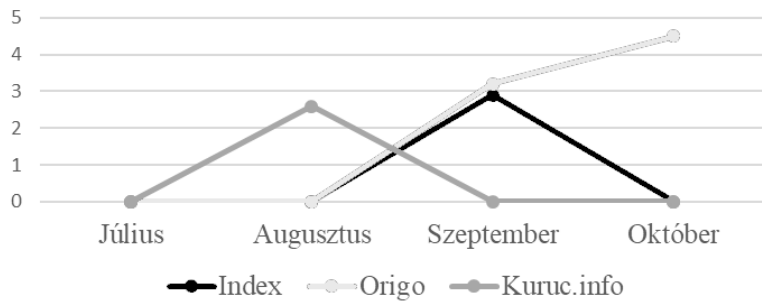


6. ábra Szíria említésének gyakorisága (súlyozás utáni darabszám)

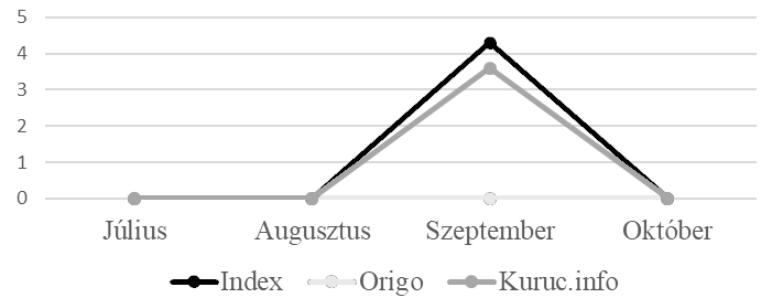
Úgy tűnik, hogy Szíria említésére nagyon különböző időpontokban került sor, de augusztusban és októberben mindhárom hírportál közel ugyananynyi alkalommal említette meg az országot. Utánanéztem, hogy mi lehet ennek az oka. Augusztusban többek között arról lehetett olvasni mindhárom hírportálon, hogy a szírek és törökök több települést is visszafoglaltak az Iszlám Államtól [9], októberben pedig egy szír bevándorló megkísérelt felrobbantani egy bombát egy német repülőtéren[10], illetve a magyar határnál határőrre támadt egy szír menekült[11]. (6. ábra).

Görögországot viszonylag kevés alkalommal említették az elmúlt 4 hónapban a bevándorlók kapcsán, de szeptemberben az *Index* és *Origo* szinte ugyanannyiszor írt az országról. Ennek az egyik görög menekülttáborban történt gyújtogatás lehet az oka[12]. (7. ábra).

[9] Index http://index.hu/kulfold/2016/08/24/meg_egy_varost_felszabaditottak_a_sziriai_felkelok [2016. 11. 5.]
[10] Origo http://www.origo.hu/nagyvilag/20161008-nagyszabasu-rendori-akcio-a-nemetorszagi-chemnitzben-robbantasra-keszulhettek.html [2016. 11. 5.]
[11] Index http://index.hu/belfold/2016/10/16/karoval_tamadt_egy_hatarvadaszra_a_szir_menekult [2016. 11. 5.]
[12] Index http://index.hu/kulfold/2016/09/19/tuz_miatt_kiuritettek_egy_gorog_menekulttabort [2016. 11. 5.]



7. ábra Görögország említésének gyakorisága (súlyozás utáni darabszám)



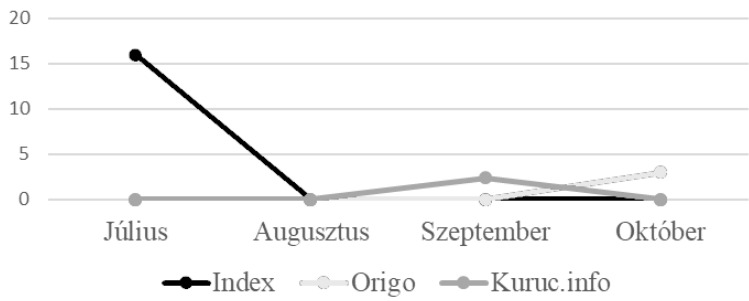
8. ábra Irak említésének gyakorisága (súlyozás utáni darabszám)

Irak említését az *Origo* szinte teljesen mellőzte az elmúlt 4 hónapban (legalábbis a „bevándorló" és annak szinonim kifejezései kontextusában), de az *Indexnél* és *Kuruc.infónál* is csak szeptemberben írtak az országról. Egy kis kutatómunka után úgy tűnik, hogy az amerikai és iraki katonai csapatokról publikált[13] olyan híreket az Index és a Kuruc.info, amelyet az Origo kevésbé tartott fontosnak (8. ábra).

Afganisztánról nem sokat lehet olvasni a magyar sajtóban a bevándorlók kontextusában, de a júliusi kiugró értékéknek utánanéztem. Úgy tűnik,

[13] Kuruc.info https://kuruc.info/r/4/163910 [2016. 11. 10.]

hogy a kiugró értéket az afgán menekültek helyzetéről szóló hosszú, helyzetelemző cikkel[14] érte el az *Index*. Nem konkrét időponthoz kötődő eseményről számol be, tehát a másik két hírportálon ezért nem említették Afganisztánt a bevándorlók vonatkozásában (9. ábra).



9. ábra Afganisztán említésének gyakorisága (súlyozás utáni darabszám)

5.4.3. Kontingenciaelemzés

Ezzel a módszerrel az együttesen előforduló kifejezések közötti kapcsolatot vizsgálhatjuk meg, pl. megnézhetjük, hogy egy-egy párban álló kifejezés milyen gyakran jelenik meg egy megadott szövegrészen belül, vagy az együttes megjelenésük kizárja-e más szövegpárok felbukkanását.

Magam is készítettem egy olyan algoritmust, amelynek segítségével kilistáztam az összes együttes előfordulást az *Index* és az *Origó* teljes áprilisi és októberi cikktermésében, illetve a *Kuruc.info* júliusi és októberi cikkeiben. Reméltem, hogy így kiderül, hogy az egyes hírportálok mely kifejezéseket szeretik más kifejezésekkel együtt alkalmazni, pl. a különböző közszereplők neveit milyen kontextusban használják.

Sajnos az algoritmus elkészítése, majd lefuttatása és az eredmények kiértékelése után azt tapasztaltam, hogy az eredményeimből nem olvasható ki semmilyen összefüggés vagy mélyebb tartalom. Persze lehet, hogy ez az algoritmusom hibája, de az is lehet, hogy a módszer ebben a formában nem célravezető. Bár azt sajnálom, hogy kontingenciaelemzéssel nem tudtam feltárni semmilyen hasznos információt, de így legalább egy tévutat is be tudok mutatni, és látszik az is, hogy milyen buktatók lehetnek az automatizált tartalomelemzés folyamatában. Az egyik ilyen, hogy csak az algoritmus megírás után derül ki, hogy érdemes volt-e elkészíteni a programot,

[14] Index http://index.hu/kulfold/2016/07/27/europa_teljesen_felkeszuletlen_pedig_ujabb_menekulthullam_johet [2016. 11. 10.]

vagyis kimutatható-e az elemezni kívánt szövegből bármilyen összefüggés, látens tartalom. Nincs rá garancia, hogy a program megírásába fektetett idő és pénz végül megtérül, ezzel számolni kell.

5.4.4. Klaszterálás

Amikor már túl sok együttes előfordulást találtunk, akkor értelmezhetetlen adathalmazt kapunk. Ilyenkor jön képbe a klaszterálás. A folyamat során az összetartozó, hasonló jelentéssel bíró csoportokat összevonjuk, és a későbbiekben együttesen kezeljük őket. Nem vizsgáljuk meg mindig külön-külön, hogy milyen más szópárokkal állnak kapcsolatban, hanem a nagyon hasonló szópárokat egyszerűen azonosnak tekintjük. Ez kicsit hasonlít a könyvtári osztályozásra.

5.4.5. Kontextuális osztályozás

Ez az a módszer, amikor a kifejezések szövegkörnyezetét vizsgáljuk, és annak alapján próbálunk rokon értelmű szövegrészeket, kifejezéseket találni, hogy mennyi közös vonás van az egyes kifejezések nyelvi környezetében. Nyilván minél több a közös vonás két különböző szó nyelvi környezetében, annál inkább értékelhetjük szinonimaként azokat. Ha pl. a „*kutya*" és az „*eb*" kifejezés is rendszeresen a „*kutyaház*" kifejezéssel szerepel együtt, akkor beállíthatjuk a tartalomelemző szoftverben azt, hogy amennyiben ez sokszor megismétlődik, akkor kutyát és az ebet kezdje el azonos kifejezésként kezelni.

5.4.6. Szótár alapú tartalomelemzés

Ez egy viszonylag egyszerű módszer, de a számítógép szempontjából eléggé erőforrás-igényes. A lényege, hogy veszünk valamilyen tematikus szótárat, amely egy meghatározott témakört fed le, és az ebben szereplő összes szót összehasonlítjuk a tartalomelemzés tárgyát képező szövegekkel.

Én ezt arra használtam, hogy megállapítsam, hogy a három hírportálon milyen megoszlásban jelennek meg pozitív és negatív tartalmú cikkek. Pozitívon azt értem, amikor pl. egy zsiráfbébi születéséről írnak, negatívon pedig azt, amikor pl. bankrablásról.

Az első lépéshez az ingyenesen használható *Szószablya*[15] című weboldalt hívtam segítségül. Ezt a Budapesti Műszaki és Gazdaságtudományi Egyetemen fejlesztik, a lényege, hogy a weboldal felületén be lehet írni egy szót, amire az oldal válaszként néhány másik adat kíséretében kiírja, hogy mi a

[15] Szószablya http://szotar.mokk.bme.hu/szoszablya/searchw.php [2016. 12. 05.]

szó szótöve és a szófaja. Egy script[16] segítségével összesen 9 millió 831 ezer szót küldtem be a 100-as tömbökben[17], és így egy olyan adatbázist építettem fel, amelyben szerepel minden egyes szó szótöve és szófaja.

Erre a melléknévszótáram elkészítéséhez volt szükség. Kézi módszerrel kiválasztottam 500 olyan melléknevet, amelyről viszonylag egyértelműen megállapítható, hogy pozitív vagy negatív tartalmat hordoz. Ilyen pl. a „*díjnyertes*", „*hűséges*", „*megbízható*", a másik oldalon pedig a „*radioaktív*", „*igazságtalan*", „*fasiszta*". Persze szövegkörnyezettől függően egy-egy pozitívnak vagy negatívnak vélt melléknév jelenthet egészen mást is, mint amire a szótár összeállításánál gondoltam, de nem is azt vártam, hogy egy ilyen melléknévszótárral teljes biztonsággal tudom majd kategorizálni a cikkeket pozitív és negatív tartalom szerint, hanem kísérletként tekintettem rá, kíváncsi voltam, hogy mennyire működőképes egy ilyen megoldás.

Tehát miután elkészítettem a pozitív/negatív melléknévszótárat, összehasonlítottam a 40 ezer cikk 9 millió szavával, így megkaptam, hogy mely cikkben hány pozitív és hány negatív tartalmú melléknév jelenik meg. Azt a cikket ítéltem negatív vagy pozitív tartalmúnak, amelyben legalább 5 szó felbukkant a melléknévszótáramból, és legalább 60%-kal több pozitív melléknevet tartalmazott, mint negatívat, vagy fordítva.

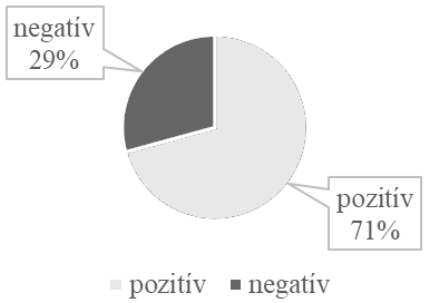
Meglepő módon a módszert nagyon sikeresnek bizonyult, az algoritmus nagyon sok cikkről pontosan meg tudta állapítani, hogy pozitív-e vagy negatív, a hibaarány mindössze 5%-os. Az viszont kiderült, hogy az 500 szavas melléknévszótár nagyon kicsi ahhoz, hogy ilyen sok cikkre alkalmazható legyen, így csak néhány száz olyan cikket talált a programom, amelyben szerepeltek kifejezések a melléknévszótáramból (10. ábra).

Ez nem meglepő eredmény, a statisztikában a tudományos, kulturális és sportrovatok cikkei is szerepelnek, ha csak a politikai és gazdasági híreket néznénk, akkor valószínűleg máshogy nézne ki a diagram.

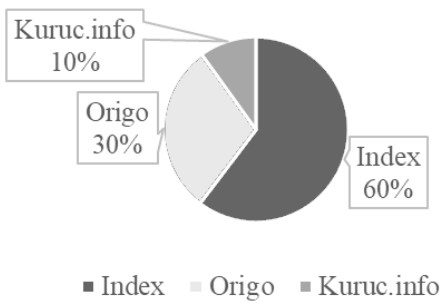
Mivel az *Index* és *Origo* publikált tartalmait áprilistól gyűjtöttem, a *Kuruc.infó*ét pedig csak júliustól, ráadásul nem is azonos számú cikk jelent meg az egyes hírportálokon, ezért valahogy meg kellett oldanom, hogy összehasonlíthatók legyenek az adatok. Ehhez kiszámoltam, hogy az egyes

[16] TechNetwork https://pcforum.hu/szotar/?term=script&tm=miaz [2016. 11. 5.]
[17] Carnegie Mellon University https://www.cs.cmu.edu/~adamchik/15-121/lectures/Arrays/arrays.html [2016. 10. 29.]

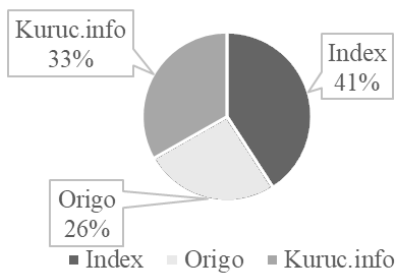
hírportálokon havonta átlagosan hány cikk jelenik meg, majd úgy súlyoztam az értékeket, hogy a pozitív és negatív melléknévszámok összehasonlíthatóak legyenek (11. ábra).



10. ábra A pozitív és negatív töltetű cikkek megoszlása a három hírportálon, összesítve



11. ábra A pozitív töltetű cikkek megoszlása a három hírportál között (súlyozás után)



12. ábra A negatív töltetű cikkek megoszlása a három hírportál között (súlyozás után)

Érdekes, hogy az *Index* mind a pozitív, mint a negatív melléknevekben vezet, ez valószínűleg azért van, mert vannak rovatai, amelyekben szinte csak pozitív hangvételű cikkek jelennek meg, és vannak olyanok, amelyekben nagyon sok a borús képet festő cikk, ezek valószínűleg a *Belföld* és a *Külföld* rovatban foglalnak helyet (12. ábra).

Az Origón fele annyi pozitív tartalmú cikk jelent meg, mint az Indexen, viszont negatív sincs sok, ebből arra következtetek, hogy sok olyan cikket publikált, amelyek semleges hangvételűek, vagy a melléknévszótáram viszonylag kis mérete miatt nem kerültek be a statisztikába.

A *Kuruc.infó*n mindössze 10% a másik két hírportálhoz viszonyított pozitív cikkek aránya, az összes többi vagy negatív hangvételű, vagy semleges.

5.4.7. Több módszer összefűzése

Fentebb leírtam, hogy melyek azok a klasszikus tartalomelemző módszerek, amelyeket részben az automatizált tartalomelemzés is átvett. Amikor automatizált módszerekkel szeretnénk különféle összefüggésekre fényt deríteni és feltárni egy-egy szöveg látens tartalmát, akkor szinte soha nem elég csak egyféle módszert alkalmazni. Erre jó példa az, hogy a fentebb ismertetett tartalomelemzések során minden esetben éltem a *gyakoriságelemzés* módszerével, még akkor is, ha nem konkrétan a gyakoriságelemzés módszertanát szerettem volna szemléltetni.

Továbbá, egy olyan programot is készítettem, amely egyszerre támaszkodik a *gyakoriságelemzés* módszerére, a *szótár alapú tartalomelemzésre* és a *klaszterálásra*, pedig a program „mindössze" azt vizsgálja meg, hogy 80 nap leforgása alatt hány olyan esemény volt, amelyről nagyon hasonló módon írt cikket az *Index* és az *Origo*.

Mivel a számítógép, amelyen a tartalomelemzést végeztem, korlátozott erőforrásokkal rendelkezik, ezért csak a cikkek egynegyedével, 10.000 cikkel dolgoztam, ez a 2016. március 21. és május 31. közötti időszakot öleli fel.

A két hírportál rendszeres olvasójaként az volt a hipotézisem, hogy naponta több olyan cikk is megjelenik a két weboldalon, amelyek közel azonosak, nemcsak a téma, amelyről írnak, hanem a szóhasználatuk tekintetében is. Ennek több oka is lehet, pl. az, hogy közös hírforrást használnak (pl. Magyar Távirati Iroda[18]), de egymástól is vehetnek át híreket.

A programom működési mechanizmusa a következőképpen néz ki:

[18] Magyar Távirati Iroda http://www.mti.hu/mti/Default.aspx [2016. 11. 10.]

1. A korábban már említett Szószablya nevű weboldal segítségével megállapítottam a 10.000 cikk összes szavának a szótövét, és kiszűrtem mindazokat, amelyeket az elemzés szempontjából nem tartottam relevánsnak (névelők és speciális karakterek).
2. Eltávolítottam a cikkeken belül ismétlődő szóelőfordulásokat. Ezt azért tettem meg, mert megfigyeltem, hogy a kezdeti próbálkozásaim sikertelensége arra vezethető vissza, hogy a két hírportál közül az egyiken sok esetben terjengősebben számoltak be az egyes eseményekről, és így hiába egyezett meg szinte teljes egészében két cikk, a szóismétlések miatt a tartalomelemző algoritmusom két cikket nem tekintett egyezőnek, ha az egyik sokkal bővebben számolt be az adott eseményről. Többféleképpen korrigálhattam volna az algoritmuson, de végül néhány próbaelemzés után arra jutottam, hogy az ismétlődő szóelőfordulások eltávolításával tudom elérni a legpontosabb eredményt.
3. A fenti lépések után gyakorlati szempontból mind a 10.000 cikk egy-egy szótárrá vált. Ezután már csak össze kellett hasonlítanom az összes „szótár" szókészletét a többi „szótáréval".
4. Ezt követően meg kellett határoznom, hogy mikor tekinthető egyezőnek két „szótár" (cikk). Ezt egyszerű kísérletezéssel próbáltam megállapítani. Első körben akkor tekintettem egyezőnek két cikket, ha azok szókészlete legalább 30%-ban megegyezett. Így sok olyan cikket is egyezőnek tekintett az algoritmus, amelyek valójában teljesen másról szólnak, ezért addig növeltem a határt 1-1%-kal, mire közel teljesen pontos eredményeket kaptam. Ez 49%-nál következett be, tehát elmondható, hogyha pusztán a cikkek szókészlete összehasonlításának útján szeretnénk összepárosítani a közel azonos tartalmú cikkeket, akkor legalább a kifejezéseik 49%-ának kell megegyezniük ahhoz, hogy összetartozónak tekinthessük azokat.

Ezekre az eredményekre jutottam:

Összesen 10.000 cikkel dolgoztam, ez a 2016. március 21. és május 31. közötti időszakot tölti ki. A cikkek az *Index* és az *Origo* hírportálokról származnak, de nem teljesen egyenlő az eloszlásuk, a vizsgált időszakban 5425 cikket publikáltak az Indexen, 4575-öt az Origón.

A 10.000 cikk között 1383 volt olyan, amelyet a tartalomelemző programom egyezőnek tekintett. Ez azt jelenti, hogy az *Index* 5425 cikkének 25,5%-a hasonló formában megtalálható az *Origón* is. Az *Origo* 4575 cikke

között 30,2% volt azon cikkek aránya, amelyek nagyon hasonlítottak az *Index* cikkeire.

Természetesen ezek az eredmények nem tekinthetők teljesen pontosnak. Az is elképzelhető, hogy több, de az is, hogy valamivel kevesebb a szinte teljesen egyező cikkek aránya a két hírportálon. Az első 200 egyező cikk alapos szemrevételezése után a 200-ból 14 esetben találtam olyan párt, amelyet nem tartok egyezőnek. Ez 7%-os hibaarány, és bár ez nem akkora minta, hogy általánosíthassunk, arra alkalmas, hogy lássuk: már egy viszonylag egyszerű algoritmus is aránylag nagy pontossággal képes összepárosítani az összetartozó szövegeket.

Pontosabb eredményeket csak egy jóval kifinomultabb automatizált tartalomelemzési algoritmus segítségével lehetne elérni. A tanulmányomnak nem az volt a célja, hogy tökéletesen pontos adatokkal szolgáljak a magyar hírportálokról; hiszen a programokat azért készítettem el, hogy hitelesen és kézzelfogható, közérthető módon tudjak írni az automatizált tartalomelemzés módszereiről, lépéseiről, és mindezek hasznosságát konkrét példákon keresztül szemléltessem.

5.5. Az elkészített algoritmusok végigfuttatása az összes szövegen, és az eredmények eltárolása az adatbázisban

Figyelni kell arra, hogy az eredményeket úgy tároljuk el, hogy azok később számszerűsítve is felhasználhatók legyenek, pl. a felismert főnevek legyenek 1-essel jelölve az adatbázisban, a melléknevek 2-essel és így tovább. Számokkal mindig egyszerűbb dolgozni, pl., ha átvisszük a kapott adatokat Excelbe, akkor ott is sokkal könnyebb diagramokat, kimutatásokat készíteni, ha minden tulajdonságot egy meghatározott számmal látunk el.

5.6. Ellenőrzés

Ez az egyik legfontosabb lépés az automatizált tartalomelemzés során, és nem szabad abban bízni, hogyha egy algoritmus működött egy konkrét szövegcsoporton, akkor egy másikon is ugyanúgy működni fog. Az automatikusan kiszámolt eredmények közül néhányat érdemes manuálisan is átszámolni, illetve mindig meg kell nézni, hogy mi az oka a kiugró értékeknek.

6. Konklúzió

A tanulmányom megírása során az jelentette a legnagyobb nehézséget, hogy úgy fogalmazzak meg megállapításokat az automatizált tartalomelemzés módszereiről és folyamatáról, hogy közben tisztában voltam azzal, hogy bármit is írok le, minden kijelentésemet kezdhetném úgy, hogy „például”, vagy úgy, hogy „jelen pillanatban”. Annyira dinamikusan változó tudományterületről van szó, hogy mire valaki alaposan kidolgoz egy

automatizált tartalomelemzési módszert, elkészíti a dokumentációját, megírja a szükséges algoritmust és publikálja az eredményeit, addigra lehet, hogy már el is veszítette az aktualitását. A programozási nyelvek folytonos átalakulása, a számítógépek kapacitásának exponenciális fejlődése és az új technikai vívmányok megjelenése következtében újra és újra át kell értékelnünk azokat a megállapításokat az automatizált tartalomelemzés módszereit illetően, amelyeket korábban tényként kezeltünk. Persze mindig lesznek olyan alapvető módszerek, amelyeket valamilyen formában a jövőben is használni fognak a szövegek automatizált elemzése során, de ezek technikai megvalósítása az informatikával együtt fog fejlődni, átalakulni.

A hipotézisem az volt, hogy viszonylag egyszerű algoritmusok segítségével is olyan – hasznos – eredményekhez juthatunk, amelyeket nem lehetett volna elérni a szövegek „manuális" elemzése során. Úgy gondolom, hogy a hipotézis igazolást nyert, hiszen sikerült olyan összefüggéseket feltárnom az *Index*, az *Origo* és a *Kuruc.info* nevű hírportálok tartalmaiban, amelyek számszerűek, pontosak és objektívek. A bemutatott tartalomelemzési módszerekkel és technikai megoldásokkal természetesen nemcsak webes hírportálok automatizált tartalomelemzése valósítható meg, hanem átültethetők bármilyen más területre is, ahol szövegekkel dolgoznak.

Nagyon fontosnak tartom hangsúlyozni, hogy a tanulmányomban ismertetett módszerek csupán egy kis részét fedik le mindannak, amit automatizált tartalomelemzésnek tekinthetünk. A kutatásomban arra kerestem a választ, hogy egyszerű algoritmusok segítségével hogyan segíthetnénk a tudományos munkát. Továbbá a kézzelfogható példák bemutatásával szeretném felkelteni az automatizált tartalomelemzés iránti érdeklődést, elsősorban könyvtári vagy a könyvtárhoz közel álló területeken. Úgy gondolom, hogy az automatizált tartalomelemzés alkalmazása a könyvtári szolgáltatások minőségi javulását eredményezné.

Irodalom

Antal László (1976): *A tartalomelemzés alapjai*. Magvető Könyvkiadó, Budapest.

Baxter, L. A., Babbie, E. R. (2003): *The Basics of Communication Research*. Wadsworth Publishing, Belmont.

Bengtsson, M. (2016): How to plan and perform a qualitative study using content analysis. *NursingPlus Open*, 2016. 2. sz. 8–4. http://www.sciencedirect.com/science/article/pii/S2352900816000029 [2016. 12. 24.]

Ehmann Bea, Balázs László, László János, Gushin, V. (2012, szerk.): Izolált kiscsoportok pszichodinamikája: a Mars-500 űranalóg szimuláció legénységi kommunikációjának tartalomelemzéses vizsgálata. In: Vargha András (szerk.): *A tudomány emberi arca: A Magyar Pszichológiai Társaság XXI. Országos Tudományos Nagygyűlése: Kivonatkötet*; Magyar Pszichológiai Társaság, Budapest.

Géring Zsuzsanna (2014): Tartalomelemzés: A virtuális és a valós világ határán: Egy vállalati honlap-elemzés bemutatása. *Kultúra és Közösség*, 5. 1. sz http://publikaciotar.repozitorium.bgf.hu/611/1/Gering_tartalomelemzes_2014.pdf [2016. 12. 29.]

Graneheim, U. H., Lundman, B. (2004): Qualitative content analysis in nursing research: concepts, procedures and measure to achieve trustworthiness. *Nurse Education Today*, 24. 2. sz. 105–112 http://www.nurseeducationtoday.com/article/S0260-6917(03)00151-5/abstract [2016. 12. 24.]

Hsieh, H.F.; Shannon, S.E. (2005): Three approaches to qualitative content analysis. *Qualitative Health Research*. 15. 9. sz. 1277–1288.

Krippendorff, K. (1995): *A tartalomelemzés módszertanának alapjai*. Balassi Kiadó, Budapest. 119.

Krippendorff, K. (2013): *Content Analysis: An Introduction to Its Methodology*. SAGE Publishing, London.

Krippendorff, Klaus (1967): *An Examination of Content Analysis: A Proposal for a Framework and an Information Calculus for Message Analytic Situations* (egyetemi doktori disszertáció). Urbana, University of Illinois.

László János; Ehmann Bea (2003, szerk.): LAS Verticum: Egy szó feletti tartalomelemző szoftver. In: *Magyar Számítógépes Nyelvészeti Konferencia*, MSZNY, Szeged.

Lowe, W. (2008): Understanding wordscores. *Political Analysis*. 16. 4. sz. 356–71.

Nagy Andor (2015): Az automatizált tartalomelemzés lehetőségei. *Tudásmenedzsment*, 16. 1. sz. 132–139. http://epa.oszk.hu/02700/02750/00037/pdf/EPA02750_tudasmenedzsment_2015_01_132-139.pdf [2017. 01. 01.)]

Neuendorf, K. A. (2016): *The Content Analysis Guidebook*. SAGE Publishing, London.

Tesch, R. (1990): *Qualitative Research: Analysis Types & Software Tools*. Falmer Press, Bristol.

Varga Katalin (2005): *Szöveg és tartalom az információs társadalomban*. Pécsi Tudományegyetem Felnőttképzési és Emberi Erőforrás Fejlesztési Kar, Pécs.

Waples, D., Berelson, B., Bradshaw, F. R. (1940): *What Reading Does to People: A Summary of Evidence on the Social Effects of Reading and a Statement of Problems for Research*. University of Chicago Press, Chicago. 9.

Weber, R.P. (1990): *Basic Content Analysis*. Sage Publications, Newbury Park.