NAGY ANDOR

Az automatizált tartalomelemzés lehetőségei

Az Index és az Origo összehasonlításán keresztül szemléltetve

A különféle webes hírportálok senki előtt sem ismeretlenek. Mindenkinek vannak egyéni benyomásai a tartalmukat, minőségüket és stílusukat illetően. Ezek szubjektív észrevételek, nehéz tényleges, általános érvényű különbségeket felsorolni két hasonló tematikájú hírportál között. Kíváncsi voltam, hogy az informatikát miképpen lehetne arra fölhasználni, hogy elfogulatlanul vizsgáljuk meg az egyes médiumokat, és azokról úgy tehessünk megállapításokat, hogy bizonyítottan megállják a helyüket. Az alábbiakban kutatásom eredményét ismertetem, melynek távlati célja, hogy a kidolgozott módszert további finomítások után még eredményesebben lehessen felhasználni automatizált tartalomelemzésre. Ezúttal *a szavak előfordulási gyakoriságának* különböző szempontok szerinti elemzését ismertetem. A módszer – jelenlegi formájában – még nem alkalmas arra, hogy egy-egy orgánum igazi sajátosságait teljesen és hitelesen képes legyen feltárni, ehhez a szavak kontextusát, mellékjelentését (pl. iróniáját) is vizsgálni kellene. Egy későbbi kutatásomban ehhez próbálok közelebb kerülni a jelenlegi módszerem továbbfejlesztése révén.

A kutatásom során két nagy webes hírportál cikkeit vettem alapul, az *Index*ét és az *Origó*ét.

Az *első fázisban* PHP és MySQL programnyelven egy olyan programot írtam, amely *2014. október 14-től november 29-ig*, tehát *47* napon keresztül SQL adatbázisba mentette a két vizsgált hírportál valamennyi cikkét a címmel, tartalommal, publikálási dátummal és rovatnévvel együtt. A következő lépcsőben a program szavakra bontotta az összes cikk teljes tartalmi részét, és a helyesiras.mta.hu valamint a szotar.mokk.bme.hu/szoszablya szóelemző oldalaknak egyenként elküldte a szavakat, majd megvizsgálta a kapott eredményeket, és adatbázisba mentette. Mindkét weboldalt a *Magyar Tudományos Akadémia* gondozza, és helyesírás-, valamint szófajelemzésre alkalmas. A program és a két weboldal segítségével megkaptam a helyesírási hibával leírt szavak számát, valamint azt, hogy melyik szófajú szóból mennyit használtak az egyes hírportálok. A program körülbelül hat óra alatt futott le egy ma középkategóriásnak mondható számítógépen kialakított webszerveren.

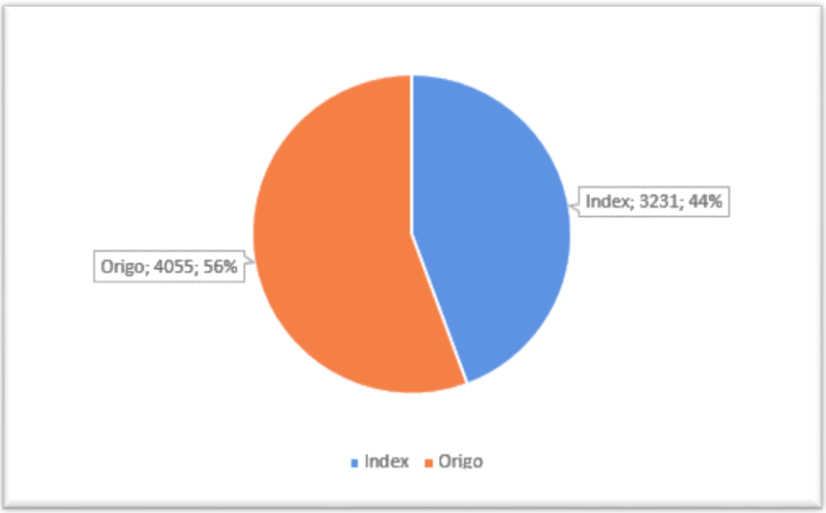
A *következő fázisban* pontos kimutatásokat készítettem, melyhez MySQL és Excel függvényeket hoztam létre.

Több dolgot vizsgáltam:

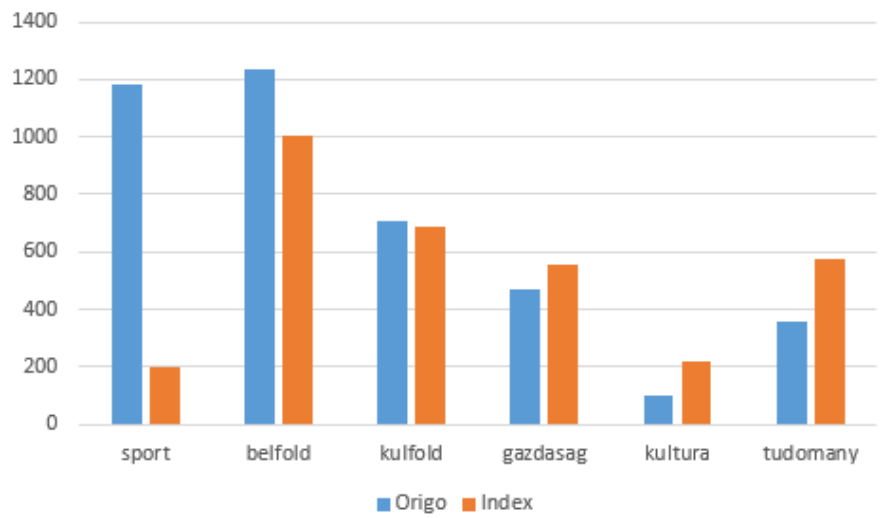
- a leggyakoribb szavakat hírportálra és cikk-kategóriákra lebontva;
- a helytelenül leírt szavak számát, hírportálra és cikk-kategóriákra lebontva;
- a különböző szófajú szavak számát, hírportálra és cikk-kategóriákra lebontva;
- hírportálonként és cikk-kategóriánként az összes szó számát;
- a publikált cikkek számát hírportálonként, napokra lebontva;
- az egyes cikk-kategóriákban megjelent cikkek számának megoszlását hírportálonként, az összes cikkszámot figyelembe véve.

A *harmadik fázisban* a kapott eredmények elemzése, majd a konklúziók levonása következett.

A 47 vizsgált nap alatt az *Index* és *Origo* hírportálokon összesen *7286* cikket publikáltak hat különböző rovatban. A rovatokat a két hírportál nem azonos néven nevezi meg (pl. a belföldi cikk-kategóriának megnevezésére az *Index* a „belföld", az *Origo* az „itthon" kifejezést használja), ezért az egyes rovatoknak a tartalomnak megfelelő nevet adtam, és a későbbiekben a következő megnevezéseket fogom használni: *belföld*, *külföld*, *kultúra*, *gazdaság*, *sport*, *tudomány*. Tehát ez az a hat kategória, amelyekbe az *Index* és az *Origo* is besorolja a cikkeit. A *7286* cikk az ábrán látható módon oszlik meg a két weboldal között:



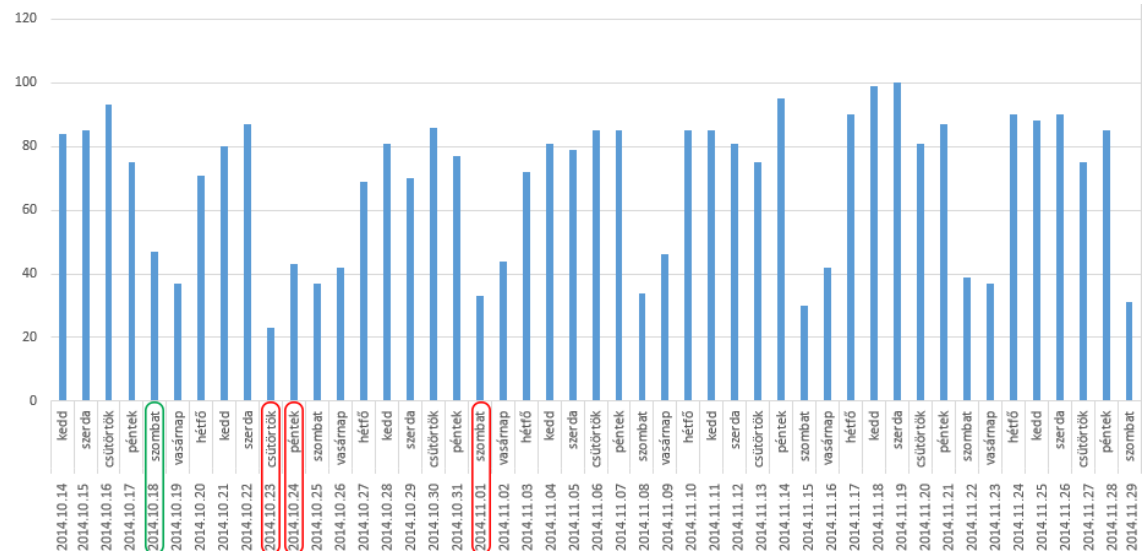
Ahogyan az a diagramon is látszik, az *Origón* *824*-el több cikket publikáltak a *47* nap alatt. Kíváncsi voltam, hogy van-e olyan rovat, amelyben jelentősen eltér a két hírportálon publikált cikkek száma, ezért elkészítettem a rovatok szerinti megoszlás statisztikáját. A következő eredményt kaptam:



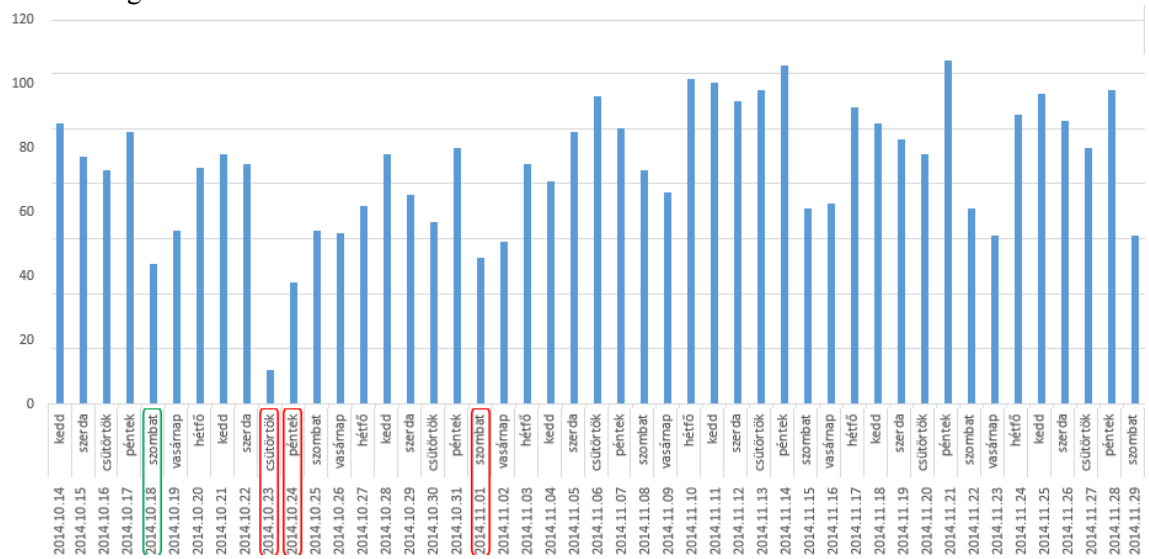
Szembetűnő, hogy az *Origó*n jóval több a sport témájú cikk, szám szerint *987*-tel. A belföldi rovatban publikált cikkek száma is valamivel több, a külföldi cikkek aránya nagyjából egyenlő, ám a gazdasági, kulturális és tudományos jellegű publikációk az *Index*en képviseltetik magukat nagyobb számmal. Hogy ha a sportrovatot figyelmen kívül hagyjuk, akkor az *Origó*n *2872*, az *Index*en pedig *3035* cikk jelent meg a vizsgált időtartam alatt, tehát látszik, hogy az *Origo* ennek az egy, kiemelkedően aktív rovatnak köszönheti, hogy összességében több cikk jelenik meg a weboldalán, mint az *Index*en.

A következő szempont, amire kíváncsi voltam, az az, hogy a napok függvényében hogyan változik a publikált cikkek száma, és a munkaszüneti, illetve valamely munkaszüneti nap helyett ledolgozandó munkanapoknak milyen hatása van. *Október 14.* és *november 29.* között összesen *3* munkaszüneti és egy plusz munkanap volt. A következő ábrán a cikkek napok szerinti megoszlása látható, pirossal a munkaszüneti napokat, zölddel a plusz munkanapot jelöltem:

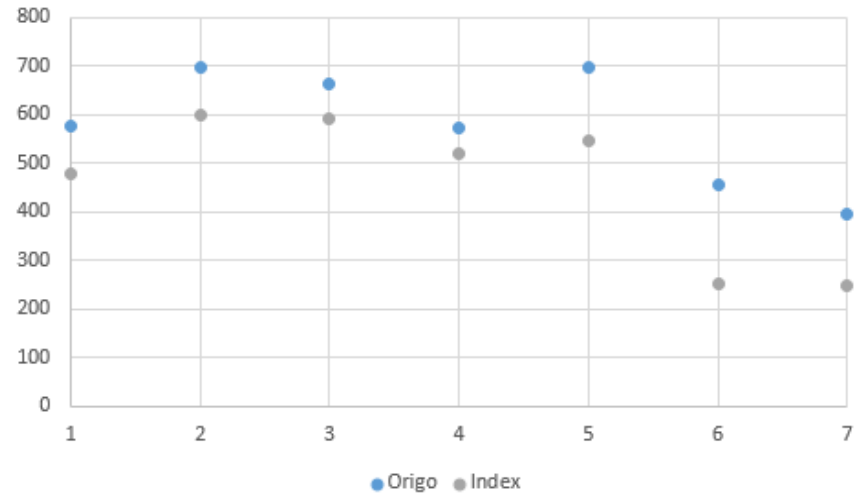
Az Indexen:



Az Origón:



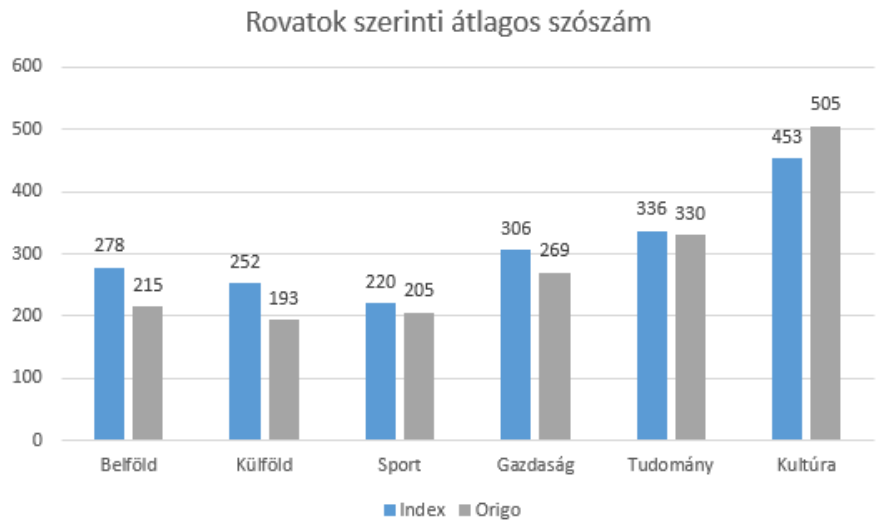
Látszik, hogy egy viszonylag szabályos minta ismétlődik a hetek során (ez az *Index*nél jobban kitűnik), de ez talán a következő grafikonból még tisztábban kivehető (a kék pontok az *Origó*é, a szürkék az *Index*é):



A grafikon vízszintes tengelyén a hét napjai vannak megszámozva, a függőleges tengelyen pedig az adott napon publikált cikkek száma látható. A csúcspont mindkét hírportál esetében keddenként van, utána egy egyenletes csökkenés figyelhető meg a kiadott cikkek számában, majd hétvége előtt van egy ugrás, és hétvégén mindkét hírportálon lecsökken a publikációk száma. A legtöbb internetes hírportálon az a gyakorlat, hogy hétvégén olyan cikkek jelennek meg, amelyeket hét közben írtak meg, és ilyenkor automatikusan publikálják. Így a szerzők megkaphatják pihenőnapnak a szombatot és vasárnapot. és a látogatók sem maradnak új cikkek nélkül. Hétvégén általában a kevésbé fontos, inkább érdekességnek szánt írások dominálnak, azok, amelyeket a szerző hét közben ír meg, és nem jelent problémát, ha csak néhány nappal később jut el az olvasókhoz.

A munkaszüneti napok közül az *október 23-i* éreztette legerőteljesebben a hatását mind az *Index*en, mind pedig az *Origó*n. Ezen a napon a töredékére esett vissza a publikációk száma. Másnap még mindig munkaszüneti nap volt, de ekkor már kevesebb újságírónak lehetett szabadsága, mert mindkét hírportálon megnőtt a publikációk száma. Mindenszentek napja szombatra esett, tehát nem mutatkozott változás a cikkek számában, hiszen szombaton egyébként is a hét közben megírt cikkeket adják ki.

Az *Index*en a 47 nap leforgása alatt összesen *957 618* szót gépeltek le szerzők. Az *Origó*n ezzel szemben *941 551*-et. Ez annak fényében, hogy ez utóbbi hírportálon *824*-el több cikket publikáltak, beszédes adat. Azt jelenti, hogy az *Index*en hosszabbak a cikkek. Kíváncsi voltam, hogy átlagosan hány szóból állnak az egyes cikkek, ezért a szavak számát elosztottam a cikkek számával, így az jött ki, hogy az *Index*en átlagosan *296* szóból tevődik össze egy cikk, míg az *Origó*n *232*-ből. Arra is kíváncsi voltam, hogy a cikkek hosszúsága hogy néz ki az egyes rovatokat vizsgálva. A következő ábrán ez figyelhető meg:

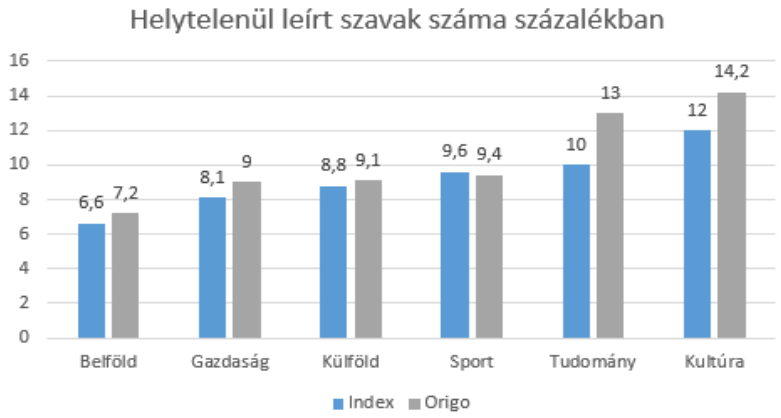


Jól látszik, hogy az *Index*en – egy kivétellel – minden rovatban valamivel hosszabb cikkek jelennek meg. A kivétel a kultúrarovat, ahol az *Origo* dominál, legalábbis a cikkek hosszát tekintve, mert cikkszámban az *Index* előrébb van.

A következő vizsgálati szempontom a helyesírás. Mivel csak a szavak helyesírásának ellenőrzését tudtam automatizálni, ezért a más típusú helyesírási hibákról (pl. a központozás helytelen alkalmazása) nincs adatom. Ennek ellenére úgy gondolom, hogy az összes cikk összes szavának ellenőrzése után kifejezetten pontos képet kaptam arról, hogy a két hírportál közül melyik az, ahol jobban odafigyelnek a helyesírásra.

Az *Index*en *81 933* szóban, az *Origó*n pedig *96 378*-ben volt helyesírási hiba. Ez azt jelenti, hogy az *Index*en a szavak *8,5* százalékát, az *Origó*n pedig a *10,2* százalékát írták helytelenül. Persze ez nem teljesen pontos érték, mert előfordulhat, hogy némely szót azért jelezte hibásnak a Magyar Tudományos Akadémia rendszere, mert olyan szleng, amely nincs az adatbázisában, esetleg idegen nyelvű, vagy szándékosan nem a magyar helyesírás szabályainak megfelelően leírt szó. Mindenesetre ez egy viszonylag pontosnak mondható arány, hiszen a vizsgálati halmaz hatalmas volt. Tehát ha általánosítunk, akkor elmondhatjuk, hogy az *Index*en valamivel kevesebb helyesírási hiba jelenik meg, mint az *Origó*n, de véleményem szerint a helyesírási hibák számát mindkét hírportálnak érdemes lenne csökkentenie, mert a jelenlegi értékeket túl magasnak találom, hiszen a két portál – népszerűsége révén – káros hatást gyakorolhat anyanyelvi kultúránkra.

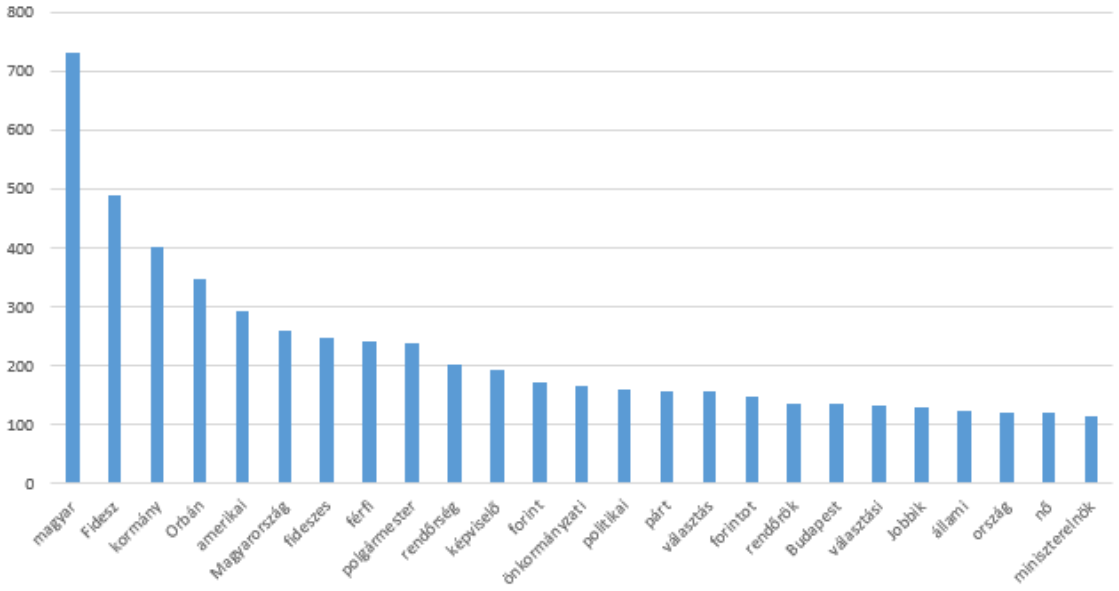
Azt is megvizsgáltam, hogy a helytelenül leírt szavak hogyan oszlanak meg az egyes rovatok között. A grafikonon az látható, hány százaléka helytelen az egyes rovatok szavainak:



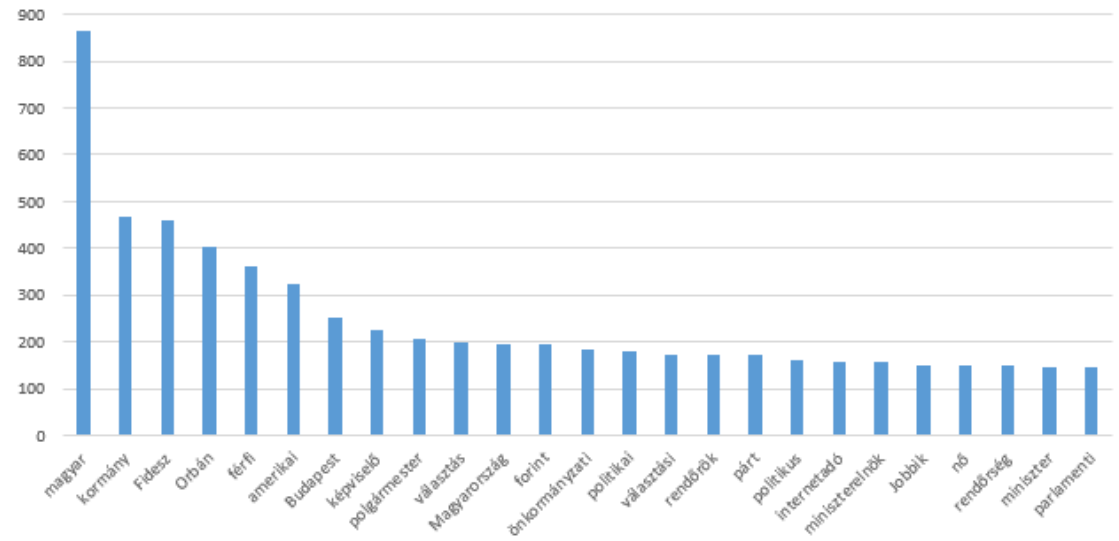
Érdekes módon mindkét oldal esetében a belföldi cikkekben volt a legkevesebb helyesírási hiba, és számomra meglepő módon a kultúrarovat cikkeiben a legtöbb. Véleményem szerint ennek két oka lehet. Az egyik, hogy a kulturális témájú cikkekben több az olyan szó, amely ugyan helyesírási hibától mentes, de a magyar helyesírási szabályzat szótári részében nem található meg. Ilyenek lehetnek például a külföldről átvett divatos zenei irányzatok megnevezései vagy a kultúra egyes területein használt szakzsargonok. A kapott eredmény másik oka az lehet, hogy a két hírportál főoldalára többségében a belföld, gazdaság és külföld rovatból válogatnak cikkeket, így elképzelhető, hogy ezek külön lektoráláson esnek át, így javítva az említett három rovat helyesírási statisztikáját.

A következőkben arra kerestem választ, hogy egyes hírportálok leggyakoribb szavai alapján meg lehet-e mondani, hogy mely témák a leggyakoribbak az adott oldalon. Persze messzemenő következtetéseket ezen adatokból nem lehet levonni. Ahogy azt a cikkem elején írtam, a statisztikát előállító programom további fejlesztése után a szavak jelentésének és konnotációinak vizsgálata egy önálló kutatás témája lehetne, most csak azt vizsgáltam meg, hogy a belföldi rovatban fellelhető önálló tartalommal bíró leggyakoribb 25 szó milyen cikktémákat feltételez, és e leggyakoribb szavak között van-e átfedés az *Index* és *Origo* hírportálokon.

Az *Index* „belföld” rovatának önálló tartalommal bíró 25 leggyakoribb szava:

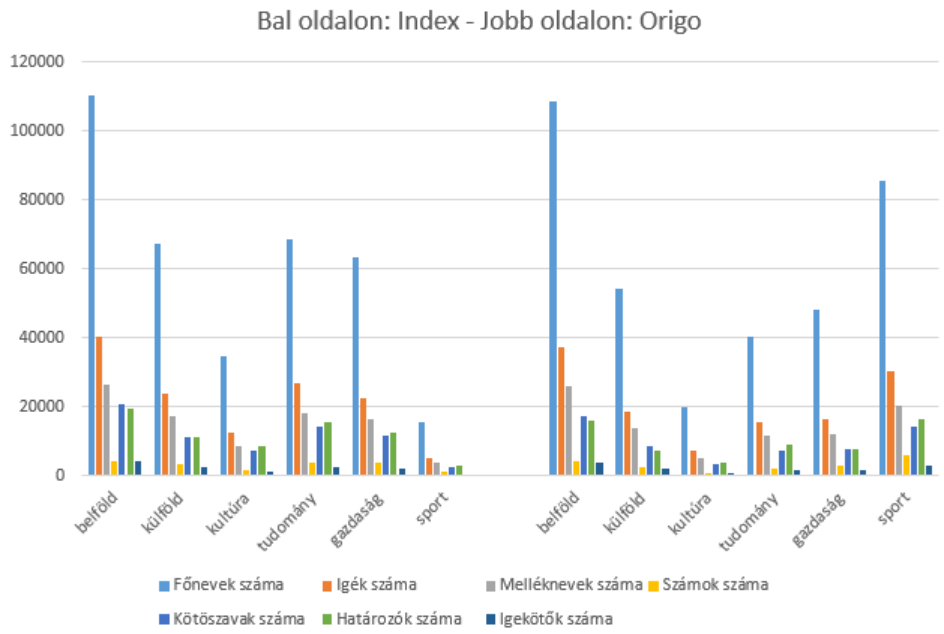


Az *Origo* „belföld” rovatának önálló tartalommal bíró 25 leggyakoribb szava:



Mint látható, a legtöbb gyakori szó és azok száma megegyezik, tehát kizárólag ezt a statisztikát alapul véve azt állapíthatjuk meg, hogy a két hírportál belföld rovatában leggyakrabban az aktuálpolitikáról esik szó, azon belül is a kormányról és a legfelkapottabb történésekről. Biztos vagyok benne, hogy egy mélyebb kutatás eredményeképp számos további érdekes és tartalmibb megállapítást tehetnénk a szavak gyakoriságának alapján.

Végezetül azon szófajok szerint is megvizsgáltam a cikkek szavait, amelyek vizsgálatát technikailag meg tudtam valósítani. Ezek a következők: főnév, ige, melléknév, számnév, kötőszó, határozó, igekötő. Az egyes rovatokban ilyen a szófajok eloszlása:



Mindkét hírportál minden rovatában a főnevek dominálnak, látványos különbség nincs az *Index* és *Origo* között, tehát a szófajok vizsgálatából kiindulva nem tudtam konkrét megállapításokat tenni pl. a cikkek nyelvezetére vagy a megfogalmazások módjára, azaz a stílusára. Valószínűsítem, hogy akármilyen szöveget vizsgálnánk, a fentihez hasonló eredményt kapnánk, egészen egyszerűen a magyar nyelvtan sajátosságai miatt.

Végezetül csak annyit, hogy ez a fajta vizsgálati módszer, tehát a cikkek automatizált kiértékelése és a kapott adatok elemzése jó megoldás lehet a weben fellelhető tartalmak objektív kiértékelésére és összehasonlítására. Persze a fentieknél átfogóbb, mélyebb, „tartalmibb", akár szembesítő-értékelő elemzésre csak egy jóval kiterjedtebb kutatás keretében lehetne vállalkozni, de megítélésem szerint már pusztán a szóállomány statisztikai vizsgálata is érdekes következtetésekre ad lehetőséget.

Felhasznált irodalom:

Antal László: A tartalomelemzés alapjai. Budapest, Magvető Kiadó, 1976.

Krippendorff, Klaus: A tartalomelemzés módszertanának alapjai. Budapest, Balassi Kiadó, 1995.

Riff, Daniel – Lacy, Stephen – Fico, Frederick: Analyzing Media Messages. Using Quantitative Content Analysis in Research. London, Routledge, 2005.

Tikk Domonkos: Szövegbányászat. Budapest, Typotex Kiadó, 2007.

Varga Katalin: Szöveg és tartalom az információs társadalomban. Módszerek és lehetőségek az információ minőségi szelektálására. Pécs, Pécsi Tudományegyetem Felnőttképzési és Emberi Erőforrás Fejlesztési Kara, 2005. (Humán szervező (munkaügyi) menedzser sorozat)