

TDK-dolgozat

Szekeres Péter

2012

Polaritásmérés magyar nyelvű webes szövegekben

Budapesti Corvinus Egyetem, Gazdálkodástudományi Kar, Információrendszerek Tanszék

Készítette: Szekeres Péter (BCE-GTK, Gazdaságinformatika MSc, II. évf.)

Konzulens: Dr. Kő Andrea (BCE Információrendszerek Tanszék, egyetemi docens)

Budapest, 2012.03.26.

Tartalomjegyzék

1	Vezetői összefoglaló	4
2	Bevezetés	6
3	A Sentiment Analysis üzleti alkalmazásai.....	8
3.1	Brand monitoring	8
3.2	Közösségi média elemzés.....	9
4	Magyar nyelvű szövegek automatikus feldolgozásának nehézségei	11
5	Raglevágás.....	15
5.1	A kivétel alapú raglevágó algoritmus bemutatása.....	15
5.1.1	További toldaléklevágó szabályok definiálása	15
5.1.2	Szabályok kiegészítése kivételekkel	16
5.2	A raglevágó pontosságának tesztelése.....	17
6	Automatikus polaritásmeghatározás	20
6.1	Szótárak a polaritásmeghatározáshoz.....	21
6.2	A pontosság tesztelése	22
7	Összefoglalás.....	25
8	Melléklet	27
8.1	Raglevágás teszteléshez használt url lista	27
9	Felhasznált források	30

1 Vezetői összefoglaló

I. Téma ismertetése

A természetes nyelvek feldolgozása témakörbe tartozik minden olyan terület, amikor emberek által beszélt/írt nyelvet dolgozunk fel, jellemzően számítógéppel. A **véleményelemzés vagy sentiment analysis célja olyan algoritmusok kialakítása, melyek számszerűsítik különböző szövegek véleménypolaritását valamilyen pozitív-negatív skálán.**

A dolgozatban rávilágítok arra, hogy az ismert magyar véleményelemző megoldások pontatlanok a nyelvi struktúrák teljes figyelmen kívül hagyása miatt, és bemutatom az általam kifejlesztett nyelvi véleményelemző algoritmust. **Módszeremben a véleménymérést a szópolaritás meghatározásán túl a vélemény erősségével is kiegészítem, valamint jellemző idiómák, hasonlatok, metaforák és kifejezések kezelésével próbálom javítani a pontosságot, illetve kezelni például a jellemző ironikus megnyilvánulásokat.**

II. Vizsgált hipotézisek

A dolgozatban két hipotézist vizsgálók:

1. hipotézis: a magyar nyelv algoritmikus raglevágásának hatékony módja, ha kivételek segítségével pontosítunk általános raglevágó szabályokat.

2. hipotézis: a magyar nyelvű webes mondatok véleménypolaritásának számítógépes meghatározása a megfelelő algortimussal elérheti az emberi ítélőképesség határát, azaz közel 82%-ban egyezhet egy ember által elvégzett polaritás annotációval.

III. Tesztelés menete

A raglevágó algoritmus teszteléséhez 2012.03.11-én 107 véletlenszerűen kiválasztott weboldal teljes szövegén futtatattam az algoritmust. (A 107 weboldal szövegében összesen 16966 különböző szó szerepelt.)

A polaritásmérést 458 darab véletlenszerűen kiválasztott Index.hu hírportál cikkeiből származó mondatokon teszteltem. A mondatokat magam annotáltam: minden mondat pozitív, semleges vagy negatív címkét kapott. Ezt követően futtattam ezeken a polaritásmeghatározó algoritmusomat.

IV. A kutatás eredményei

89,67%-os pontosságú raglevágó algoritmus felhasználásával, kivételek meghatározásával és többdimenziós szofisztikált szótárrendszer alkalmazásával 75,33%-os

pontosság¹ véleményelemző megoldást sikerült építeni, melyet ha összevetünk ismert véleményelemzési tesztelések adataival (lásd: Pang et al. (2004) illetve Miháltz (2010) munkáit), akkor látható, hogy igen tekintélyes eredménynek számít.

A dolgozatban sikerült mindkét megfogalmazott hipotézist igazolni.

¹ A pontosság itt az accuracy mutatót jelenti.

2 Bevezetés

A természetes nyelvek feldolgozása, rövidítve NLP², témakörbe tartozik minden olyan terület, amikor emberek által beszélt/írt nyelvet dolgozunk fel, jellemzően számítógéppel. **A véleményelemzés vagy sentiment analysis célja olyan algoritmusok kialakítása, melyek számszerűsítik különböző szövegek véleménypolaritását valamilyen pozitív-negatív skálán.**

A számszerűsítéssel lehetővé válik nagyszámú véleményhalmaz hatékony és gyors kezelése, például:

- szűrhetőek a negatív vélemények,
- vagy összesíthetőek a szöveges reakciók egy adott eseményre vagy termékre.

A felhasználói tartalomgenerálás erősödésével a weben és a különböző közösségi platformokon megjelenő szövegek automatizált véleménymérése **üzleti szükségletté vált a brand és PR fókuszú szektorokban.**

Míg angol nyelven több hatékony – vagy leglábbis hatékonynak mondott - megoldás is elérhető, addig magyar nyelven alig található véleménymérő megoldás, ráadásul ezek pontossága erősen megkérdőjelezhető... A véleményelemzés szakirodalmával számos kutató foglalkozik (Miháltz (2010), Berend és Farkas (2008), Tikk (2007)), ennek a dolgozatnak a részletesebb áttekintés nem célja, **mindazonáltal magyar nyelven végzett véleményelemzési kutatással csak Berend és Farkas (2008) munkájában találkozhatunk**, akik fórumhozzászólások alapján próbálták a kettősállampolgárságról szóló magyar népszavazás eredményét megjósolni.

A dolgozatban rávilágítok arra, hogy az ismert magyar véleményelemző megoldások által használt szózsák modell³ a robusztussága ellenére meglehetősen pontatlan a nyelvi struktúrák teljes figyelmen kívül hagyása miatt, és bemutatom az általam kifejlesztett nyelvi véleményelemző algoritmust. **Módszeremben a véleménymérést a szópolaritás meghatározásán túl a vélemény erősségével is kiegészítem, valamint jellemző idiómák,**

² Natural language processing

³ A szózsák (bag-of-words) modell lényege, hogy egy szöveg jellemzőit a benne előforduló szavak vagy szótövek alapján határozzák meg, függetlenül a szövegben vagy mondatban elfoglalt pozíciójuktól. A magyar nyelv elemzésekor kedvelik ezt a módszert, mivel a magyar mondat szerkezeti szabályokat rendkívül bonyolult kezelni.

hasonlatok, metaforák és kifejezések kezelésével próbálom javítani a pontosságot, illetve kezelni például a jellemző ironikus megnyilvánulásokat.

A véleményelemző algoritmusomhoz szükséges volt az agglutináló⁴ magyar nyelvi nehézségek kezelésére **saját raglevágó megoldást készítenem, mely Tordai Anna Light2 szótövezőjén (Tordai, 2005) alapszik, és mely kimagasló 89,67%-os pontosságot ért el.**

A raglevágás bemutatása után ismertetem a polaritásmérési algoritmusomat, illetve az ahhoz használt értékelő szótárak felépítését és azok összeállításának menetét. A dolgozat végén tesztelem a polaritásmérő megoldásom pontosságát, és összevetem különböző rendelkezésre álló referencia adatokkal.

A célom a kutatás során az volt, hogy az előállított megoldás pontossága minél jobban megközelítse a 82%-ot, mely egy általam kijelölt elméleti határ: a jelenlegi kutatási tapasztalatok alapján két különböző ember ugyanazon korpuszon végzett véleményértékelési annotációi legfeljebb ilyen mértékben egyeznek meg. (Grimes, 2011)⁵

⁴ A magyar úgynevezett agglutináló – ragasztó – nyelv, mely a gyakorlatban azt jelenti, hogy nagyon széles toldalékkészlettel (képzők, jelek, ragok) rendelkezik.

⁵ Berend és Farkas (2008) által magyar szövegre készített klasszifikációs modell építése során a 2 annotátor véleményértékelései a tanítóadatok létrehozása során csak 72,94%-ban egyeztek.

3 A Sentiment Analysis üzleti alkalmazásai

Jelen fejezet célja, hogy bemutassa a véleményelemzés tipikus üzleti alkalmazásait, és igazolja a probléma üzleti relevanciáját.

3.1 *Brand monitoring*

A brand napjaink vállalatainak egyik legfontosabb értéke, a marketing és azon belül a PR, hírnév és brand management terület egyre nagyobb hangsúlyt kap. A hírnév (márka) erőssége nehezen meghatározható, nehezen követhető nyomon, és emiatt nehezen menedzselhető volt, azonban a webes tartalomgenerálás fejlődésével új lehetőségek nyíltak meg: megjelentek a webes véleményelemző és különböző brand monitoring eszközök, melyek feladata, hogy felfedezzék és értékeljék a különböző cégekről/brandekről/termékekről szóló szöveges véleményeket a weben, majd jelentések formájában segítsék a vállalati döntéshozatalt elsősorban kommunikációs, de akár termékfejlesztési területen.

A márkaépítés illetve márkafigyelés előtérbe kerülését jól mutatja még például a personal branding irányzatok terjedése, illetve a magyarországi SuperBrands 2011 és MagyarBrands 2011 díjazottak⁶ széles köre.

A brand monitor termékek piacát áttekintve több angol nyelvűre készített megoldással, illetve magyar nyelvűre is készített megoldásokkal is találkozunk:

Angol nyelvű megoldások:

- Radian6 (vezető brand monitoring szolgáltatás angol nyelven)
- OpinionCrawl⁷ (ingyenes online véleményelemző szolgáltatás angol nyelvűre)
- Thismoment brand monitor⁸
- Position2 Brand Monitor
- CSC: Brand monitoring services⁹

⁶ Lásd bővebben: www.superbrands.hu

⁷ <http://www.opinioncrawl.com>

⁸ <http://brandmonitor.thismoment.com>

⁹ <https://www.cscglobal.com/global/web/csc/brand-monitoring.html>

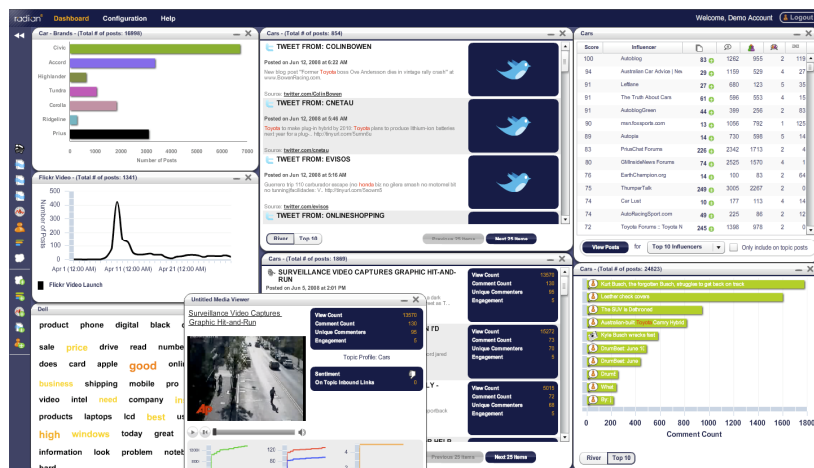
- TIQSentiVal¹⁰

Magyar nyelvű megoldások:

- BuzzMirror (már nem működő szolgáltatás)
- NewZeus (már nem működő szolgáltatás)
- BrandMonitor¹¹
- OpinHu¹²

3.2 Közösségi média elemzés

A közösségi média elemzéseknek – SMA¹³ – szintén fontos eszköze a véleménymérés. Az ilyen elemzések célja, hogy a webkettes portálok (Facebook, Twitter, LinkedIn), blogterek, de sokszor akár a teljes weben megjelenő tartalmakat elemezzék különböző szempontok szerint.



1. ábra: Radian6, a piacvezető SMA megoldás felülete

A közösségi média elemzés üzleti sikerét a márkaérték és az üzleti/fogyasztói reputáció felértékelődése hívta életre. Az SMA jelentőségének és üzleti felhasználásának demonstrálására jó példa az a kutatás, amely a Twitter bejegyzések elemzése alapján

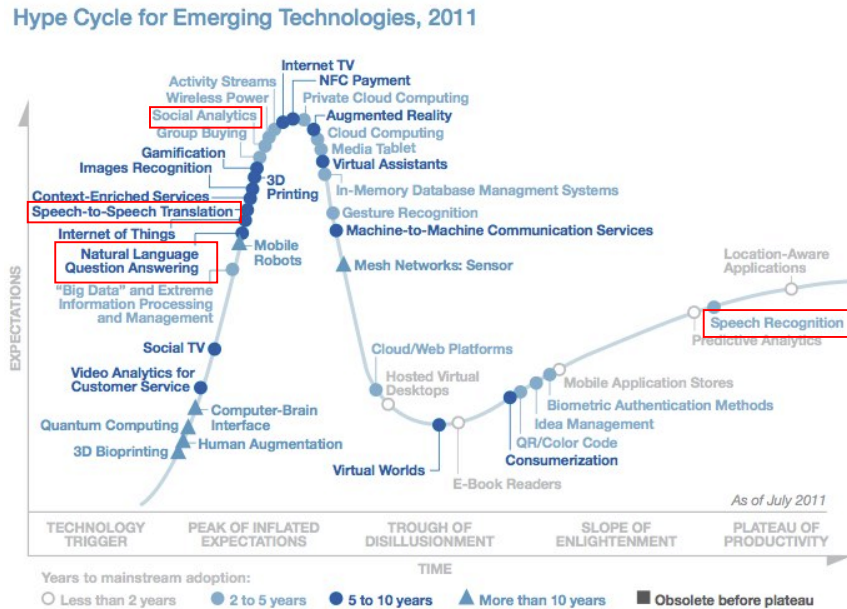
¹⁰ <http://www.tiq-solutions.de/display/loesungen/TIQSentiVal>

¹¹ <http://brandmonitor.hu>

¹² <http://opinhu.com>

¹³ Social Media Analytics

86,7%-os eredménnyel meg tudta jósolni a Dow Jones Ipari Átlag változását. (Grossmann, 2010)



2. ábra: Gartner Hype Görbe (2011)

A 2011-es Gartner görbe alapján kijelenthetjük, hogy technológiamenedzsment szempontból a közösségi elemzési lehetőségek 2011 végére már érett technológiának tekinthetők, üzleti szempontból pedig reális fogyasztói (felhasználói) elvárásokkal, illetve reális hatékonysággal és teljesítménnyel találkozhatunk a piacokon.

4 Magyar nyelvű szövegek automatikus feldolgozásának nehézségei

Angol nyelvű szövegek esetén a véleményelemzők legjobbjai 86% körüli pontossággal bírnak (Miháltz, 2010), a magyar nyelvre készültek ettől valószínűleg jelentősen elmaradnak a nyelvi nehézségek miatt.¹⁴ Jelen fejezet a magyar nyelvű webes szövegek automatikus feldolgozásának problémáival foglalkozik.

Utóragozás kezelése: a strukturalista nyelvtan nézetet leváltó Noam Chomsky által bevezetett generatív nyelvelmélet új megközelítést hozott a nyelvek kezelésébe. A generatív nyelvelmélet jelentősége a nyelvek és a szövegképzés matematikai megragadásában – algoritmizálásában - rejlik: Chomsky modellje alapján egy nyelvnek végtelen variációja létezhet az alap szókészlet (jelkészlet) és a szó- és szövegképzési szabályok alapján. (Chomsky, 1965) A generatív megközeleítés különösen igaz a magyar és a többi agglutináló nyelvre.

A toldalékoló és szabad struktúrájú nyelvek automatikus feldolgozására, mint például a magyar, a szózsák modell a leggyakrabban alkalmazott módszer. Ugyanakkor a szövegek szózsák modell alapú feldolgozásához szükséges, hogy az azonos jelentésűnek tekintett szavak azonos alakra kerüljenek, vagyis a gyakorlatban az utóragokat és bizonyos jeleket le kell vágni a szóvégekről. A levágások algoritmizálásánál képezhetőek szabályok, azonban ezeknek figyelembe kell venniük az összeolvadási, részleges vagy éppen teljes hasonulási nyelvtani szabályokat, valamint az olyan szótöveket, melyek valamilyen utóragra végződnek (például: *sisak*, *reggel*). Külön probléma azon esetek kezelése, amikor a ragozott és a ragnélküli alak is értelmes szótő, például: *folyamod*.

A problémát jól mutatja, hogy egy igének az igei toldalékolás teljes repertoárját felhasználva 216 formája lehet a magyar nyelvben:

¹⁴ A kutatásaim során nem találtam nyilvánosan elérhető magyar nyelvre végrehajtott véleményelemzésről szóló teszteredményeket, így a továbbiakban is az angol nyelven elért eredményekhez viszonyítok, feltételezve, hogy a magyar nyelvre vonatkozó hasonló megoldások pontossága ezeknél nem jobb a nyelvtani bonyolultság miatt.

Igeidő: 3

Igemód: 3

Személy: 3

Szám: 2

Igenem: 2

Alanyi/Tárgyas ragozás: 2

Összesen: $3 \times 3 \times 3 \times 2 \times 2 \times 2 = 216$

Hunglish kifejezések: a webes szövegekben, különösen szakmai szövegekben vagy éppen olyan informális médiumokban, mint a közösségi oldalak, fórumok vagy blogok, igen elterjedt az angol nyelvű szavak használata magyar mondatokban, például:

trademark, outsourcing, trading, like, system, event, share

A webes nyelvi uniformizálódás és mutálódás egy másik jelensége, mikor angol nyelvű szavakat „magyarorsítanak” a nyelvhasználók, azaz a kiejtés elve szerint írják le, és a magyar nyelvtan szabályai szerint ragozzák az idegen eredetű szavakat, például:

júzerrel, lájkol, szisztem, deployol

Szleng és webszleng: amikor webes szövegeket elemzünk, figyelembe kell vennünk, hogy a web nyelvezete modern és sajátos, és általánosak a szleng és webes szleng kifejezések, valamint az emotikonok használata. Különösen igaz mindez érzelmi megnyilvánulásokra:

lol, lolz, omg, sz@r, f@ck u, mindl, h

Beviteli eszköz miatti sajátosságok: a hordozható eszközökről (tipikusan okostelefonokról) történő bevitelnek köszönhetően az informális közösségi oldalakon sokszor találkozhatunk egészen sajátos szintaktikai és szemantikai szabályokat követő mondatokkal és félmondatokkal:

- Nincsenek ékezetek.
- Nincsenek nagybetűk
- Nincsenek írásjelek a szövegben, vagy csak mondatvégi írásjelek használ a szerző.

- A szavak helyesírása lazán kezelt.
- Jellemzők az elgépelések.

Ha számítógéppel akarjuk feldolgozni ezeket a kommenteket és hozzászólásokat, akkor az említett jelenségek tömeges előfordulására fel kell készíteni a rendszert.

Karakterkódolási nehézségek: kevésbé triviális, de magyar nyelvű webes szövegek számítógépes feldolgozásánál tipikus probléma a karakterkódolás megfelelő kezelése:

- HTML oldalak letöltésél illetve szövegek adatbázisba vagy fájlba exportálásánál figyelni kell, hogy a magyar ékezetes betűk (azon belül is elsősorban az ő és ű betűk) megjelenítése megfelelő-e: UTF-8 vagy Latin2 karakterkódolás ajánlott.
- Szintén HTML oldalak feldolgozásakor lehet probléma, ha az adott kiszolgáló az ékezetes és egyéb különleges karaktereket a szövegben HTML karakterkóddal (például: ´ -> á) jelöli a forrásfájlban. Ekkor gondoskodni kell a dekódolásról feldolgozás előtt.

Kontextus- és szakterületi függőség: vélemények szózsák modell alapú elemzésénél probléma lehet, hogy egy adott szó vagy kifejezés kontextusonként eltérő jelentéssel bírhat: egy kifejezés pozitív egy adott kontextusban, míg ellentétes irányú egy másik kontextusban.

Tagadások sokfélesége: a magyar nyelvben a tagadás többféle módon lehetséges:

- Tagadószavakkal és tiltószavakkal (például: *sem, nem, se, ne, senem, sincs, nincs, sincsen, nincsen*), melyek elhelyezkedhetnek közvetlenül a vagy akár 1 pozícióval a tagadás tárgya előtt illetve után.
- Fosztóképzők (például: *-atlan, -etlen, -mentes*) segítségével

Egy véleményelemző rendszernél alapkövetelmény, hogy a rendszernek kezelnie kell a különböző tagadásokat (a polaritás negálását), eltérő

esetben drasztikusan romlik a véleményelemzés pontossága, mivel pontosan ellentétesen értelmezi az adott véleményt.

Az eredményes számítógépes véleményelemzéshez tehát egy jól működő raglevágó algoritmus alapvető szükséglet, mely minden olyan toldalékot levág a szavak végéről, melyek nem módosítják a kívánt mértéken túl a jelentését, és semmiképpen sem a polaritását.

A fentiek alapján két hipotézist fogalmazok meg, melyek teljesülését vizsgálom a továbbiakban:

1. hipotézis: a magyar nyelv algoritmikus raglevágásának hatékony módja, ha kivételek segítségével pontosítunk általános raglevágó szabályokat.

2. hipotézis: a magyar nyelvű webes mondatok véleménypolaritásának számítógépes meghatározása a megfelelő algortimussal elérheti az emberi ítélőképesség határát, azaz közel 82%-ban egyezhet egy ember által elvégzett polaritás annotációval.

5 Ragleválás

Jelen fejezet célja az 1. hipotézis igazolásának bemutatása:

1. hipotézis: a magyar nyelv algoritmikus ragleválásának hatékony módja, ha kivételek segítségével pontosítunk általános ragleváló szabályokat.

5.1 A kivétel alapú ragleváló algoritmus bemutatása

Tordai Anna 4 féle szótövező algoritmust alakított ki magyar nyelvre (Tordai, 2005). Ezek jellemzői:

1. LIGHT1: *„csak a leggyakoribb 14 főnévi esetet kezeli. Ennek ellenére már ez is jelentékenyen javíthatja a keresési hatékonyságot.”* (Tikk, 2008)
2. LIGHT2: *„21 esetet kezel, valamint a LIGHT1 által figyelmen kívül hagyott egykarakteres toldalékok közül az akkuzativusz (tárgyrag -t) és szuperesszívusz (-n) toldalékokat is levágja. Mindkét szótövező figyelembe veszi a szótőjelölt hosszát és, hogy tartalmaz-e érvényes mássalhangzó–magánhangzó kombinációt.”* (Tikk, 2008)
3. MEDIUM: *„12 gyakori főnévi esetet kezel, a birtokos és birtokok, valamint a személyek számát is figyelembe véve. Ezenkívül kezeli a leggyakoribb igealakokat (idő, szám, személy), a melléknevek fokozását, valamint a számneveknél a törtszámnév és sorszámnév toldalékait.”* (Tikk, 2008)
4. HEAVY: *„mind a 21 esetet és az összes igealakot figyelembe veszi.”* (Tikk, 2008)

Tikk (2008) alapján a LIGHT2 rendelkezik a legkedvezőbb tulajdonságokkal gyakorlati tapasztalatok alapján, így ebből szabályrendszerből indultam ki.

A LIGHT2-t két fő irány mentén fejlesztve hoztam létre saját ragleváló algoritmusomat:

1. További toldaléklevágó szabályok definiálása
2. Levágási szabályok kiegészítése kivételekkel

5.1.1 További toldaléklevágó szabályok definiálása

Tordai szótövezőjének szabályrendszerét kibővítettem az igei toldalékok és egyéb toldalékok kezelésével. Ezenkívül a levágások sorrendjét is módosítottam elsősorban empirikus tapasztalatok alapján, hogy bizonyos végződés levágások ne vágjanak le toldalékrészeket, hanem lehetőleg az adott toldalékelem egészben lekerüljön, például:

- $-ett/-ott$ levágás mielőtt $-t$ levágás történik
- többesszám esetén: előbb $-ak, -ek, -ok$ levágás mielőtt $-k$ levágás történik

Így összesen 34 logikai blokkból és **236 toldaléklevágási lépésből** áll a szabályrendszer.

5.1.2 Szabályok kiegészítése kivételekkel

A magas pontosság eléréséhez elendeghetetlen volt, hogy bizonyos levágások ne hajtódjanak végre az adott szón, mivel ekkor a szótő sérülne, vagy egyéb kivételes eset is fennállhat. Ezért minden egyes toldaléklevágási lépés egy függvényhívást jelent, mely függvény bemenetei a következők:

- **Szó:** a szó, amin végre kell hajtani a levágást
- **Levágandó végződés:** a levágandó toldalék, például: $-t$
- **Visszaillesztendő végződés:** bizonyos levágások esetén a végződés módosul, például mert a toldalékolás miatt a szóvégi e $é$ -re módosult korábban
- **Kivétel szótövek listája:** olyan szótövek, melyekről nem szabad levágni az adott végződést, például $-t$ tárgyrag levágása esetén: zrt
- **Kivétel végzõdés listája:** olyan szóvégzõdés, melyekről nem szabad levágni az adott végzõdést, például $-t$ tárgyrag levágása esetén: $ozgat$ (A végzõdés kivételek segítségével bizonyos végzõdés minták megadhatóak, így nem kell minden szótõ egyesével megadni a szótõvek listájában.)

A fent bemutatott kivétel listáknak (szótõ és szóvégzõdés listáknak) összesen **6904 eleme** van. (Mind a 236 toldaléklevágási lépésre együtt.)

A kivételek közé tartoznak még úgynevezett egyedi végzõdés cserék, amikor az algoritmus végén bizonyos levágott alakokat egyedileg módosítani kell, hogy önmagában is értelmezhetõ szótõvet kapjunk, például: $tork \rightarrow torok$, $hav \rightarrow hó$. Ezekbõl a szabályokból 167 db van.

A kivétel szótõvek és szótárak összeállítása 500 db véletlenül kiválasztott weboldal szavai, és a 2003-ban a BME MOKK munkatársai által összeállított 100.000 leggyakoribb webes magyar szó¹⁵ alapján történt félmanuális eszközökkel.

¹⁵ A BME MOKK Hungarian Webcorpus szógyûjteménye elérhetõ a következõ linken: <ftp://ftp.mokk.bme.hu/Language/Hungarian/Freq/Web2.2/web2.2-freq-sorted.top100k.nofreqs.txt>

5.2 A raglevágó pontosságának tesztelése

A raglevágó algoritmus teszteléséhez 2012.03.11-én 107 véletlenszerűen kiválasztott weboldal (URL-eket lásd a mellékletben: 8.1 fejezet) teljes szövegén futtatattam az algoritmust. Minden egyes weboldal szövegét regurális kifejezések segítségével tokenizáltam.

Tulajdonnevek nem szerepelnek a raglevágó kivételkészletében, mivel ezek hatékony kezelése speciális névelemszótárak segítségével történhet, ami nem része jelen dolgozatnak. A szövegfelbontás során használt REGEX alapján a nagy kezdőbetűjű szavak nem kerültek be az elemzésbe, így tulajdonnevekre nem futott le a raglevágás, sem pedig mondatkezdő szavakra. Ugyanakkor feltételezhető, hogy az így kihagyott mondatkezdő kezdő szavak jelentős része előfordul a korpuszban más mondatpozícióban, és így bekerült az elemzésbe.

A weboldalak szövegének tokenizálása után az egyes szavak és azok raglevágott alakja bekerült egy adatbázis táblába. Minden raglevágás (eredeti szóalak – raglevágott alak páros) pontosan egyszer szerepel a táblában.

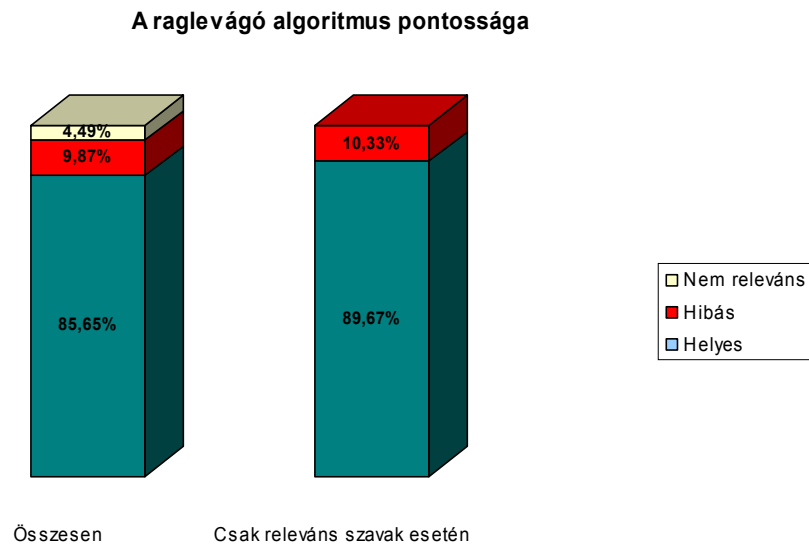
A 107 weboldal szövegében összesen 16966 különböző szó szerepelt, ezek feldoglozása után 10774 raglevágott alak keletkezett.

Miután manuálisan ellenőriztem az összes eredeti szó – raglevágott alak párost, a következő eredményre jutottam:

- **761 db nem releváns elem** volt a listában (nem relevánsnak tekintettem egy bejegyzést, ha tulajdonnév toldalékolt alakja vagy angol nyelvű szó volt a kiindulási szó)
- **14531 db jó raglevágás történt**

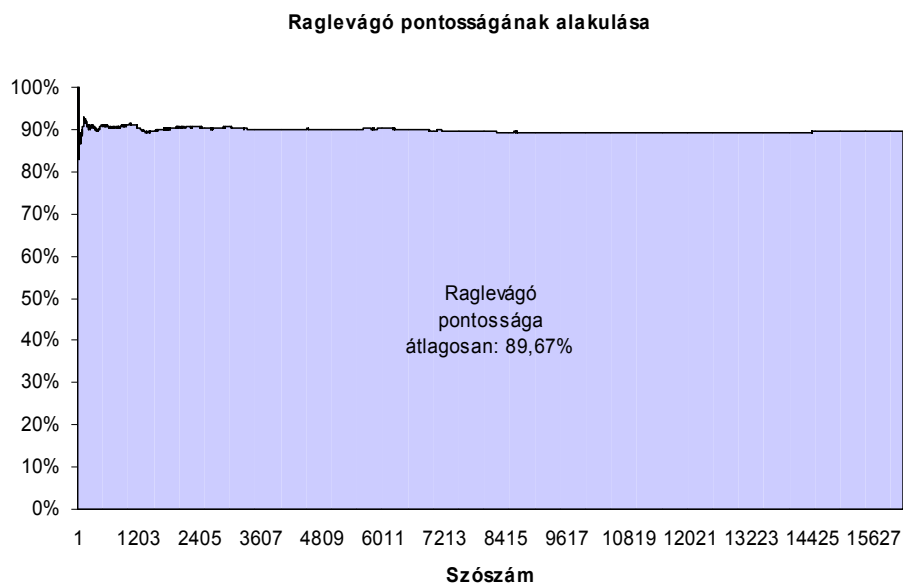
Ezek alapján a mintán futtatott raglevágás

- **Átlagos pontossága: 89,67%**
- **A pontossága szórása: 0,637% (0,637 százalékpont)**



3. ábra: A raglevágó algoritmus pontossága a vizsgált mintán

Tehát a vizsgált esetek 90%-ában helyesen működött az algoritmus, mely egy kritikus szintet elérve (a 4. ábra alapján kb. 1500 szó) gyakorlatilag konstansan, szignifikáns ingadozások nélkül 90%-os pontosságot nyújtott. (lásd: 4. ábra)



4. ábra: A raglevágó algoritmus pontosságának alakulása a szószám növekedésével

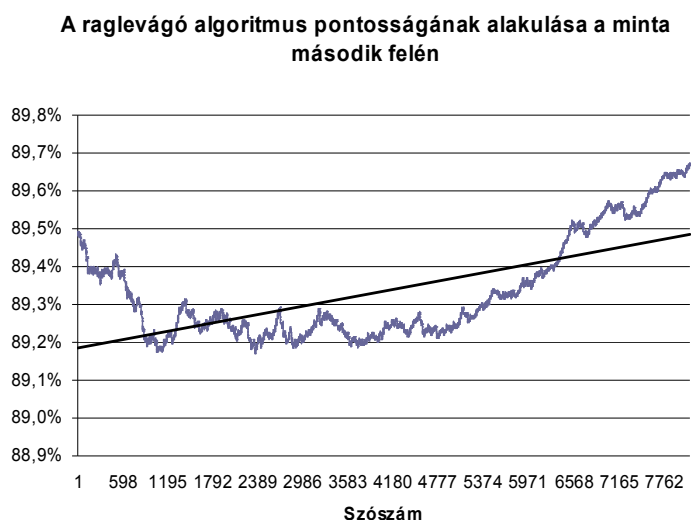
A fenti ábrán látható, hogy a minta feldolgozásával, ahogy új szavak kerültek a listába, hogyan változott a raglevágó átlagos jósága. Amennyiben lineáris trendet illesztünk a kapott adatsorra, a következő képletet kapjuk:

$$y = -0,000000830698427x + 0,904047547089899$$

Ez egy nagyon csekély mértékben, de a feldolgozott szószám növekedésével folyamatosan csökkenő pontosságot mutat. Azonban látható például, hogy a pontosság 100%-ról indul, így ha a minta elején lévő ingadozásokat nem vesszük figyelembe, és a releváns minta második részétől (a 8102. értéktől) vizsgáljuk a pontosság változását, akkor a következő lineáris trendet kapjuk (lásd: 5. ábra):

$$y = 0,000000367228480x + 0,891859283527337$$

A fenti trend pedig már egy növekvő trend.



5. ábra: A raglevágó algoritmus pontosságának alakulása a minta második felén

Becslésem szerint a jelenlegi szabályrendszerrel és raglevágási metodológiával akár 95% fölött pontosság is elérhető elsősorban a kivétel szótövek és szóvégződések további bővítésével. A fentiek alapján az 1. hipotézist igazoltnak tekintem.

6 Automatikus polaritásmeghatározás

Jelen fejezet célja a 2. hipotézis igazolásának bemutatása:

2. hipotézis: a magyar nyelvű webes mondatok véleménypolaritásának számítógépes meghatározása a megfelelő algortimussal elérheti az emberi ítélőképesség határát, azaz közel 82%-ban egyezhet egy ember által elvégzett polaritás annotációval.

Az általam definiált polaritáselemző mondatokra bontja a szöveget, és ezekben azonosít kifejezéseket, majd a mondatban lévő kifejezésekre kapott polarítások összegzése adja a mondat polarítását, míg a mondatok polaritásösszege adja a teljes szöveg polarítását. Hasonlóan működik Wilson et al. (2005) OpinionFinder megoldása illetve a magyar Geox Kft. OpinHu (Miháltz, 2010) megoldása.

Az algoritmus működése a következő:

1. Weboldal HTML kódból a szöveg kinyerése
2. Szöveg feldarabolása mondatokra szintaktikai szabályok alapján
3. Mondatok feldarabolása szavakra
4. Szavak kisebetűssé alakítása
5. Szavak végéről ragok levágása
6. Jellemző – szótárban is tárolt - kifejezések keresése a mondatban, pontszám adása találat esetén
7. Az előző lépésben azonosított kifejezésekben nem szereplő szavak között polaritással bíró szavak azonosítása a mondatban
8. Polaritás erősítő/módosító szavak azonosítása a polaritással bíró szavakat megelőző szavak vizsgálatával
9. Negáló szavak azonosítása
10. A 7., 8. és 9. lépésben felépített kifejezések pontszámainak kiszámítása
11. Mondatok pontszámainak kiszámítása a pontszámok összegzésével.
12. Ugrás a következő a mondatra és 3. lépésre, ha van még feldolgozatlan mondat
13. Mondatok pontszámainak összegzése adja a szöveg összegét.

Itt jegyezném meg, hogy jelen megoldásnál a cél nem a szubjektív vélemények azonosítása és polaritáselemzése, hanem a teljes szöveg illetve szövegrészek polaritásának meghatározása, feltételezve, hogy a szövegben leírtak a szerző szubjektív gondolatai vagy pedig olyan egyértelműen pozitív/negatív események megfogalmazása, melyek együttesen adják az adott szöveg polaritását, hangulatát.

6.1 Szótárak a polaritásmeghatározáshoz

A polarítások meghatározásához különböző szótárakat kellett létrehozni.

Véleményszó szótár

A véleményező szótár 9754 szónak tartalmazza a polaritását egy -3 és 3 közé eső skálán:

- -3: nagyon negatív
- -2: negatív
- -1: enyhén negatív
- 1: enyhén pozitív
- 2: pozitív
- 3: nagyon pozitív

A szótár 4000 véletlenszerűen kiválasztott magyar weboldal tartalma alapján, 23 szleng kifejezésekkel foglalkozó oldal tartalma alapján és a 2003-ban a BME MOKK munkatársai által összeállított 100.000 leggyakoribb webes magyar szó¹⁶ alapján, illetve ezek manuális és automatikus kibővítésével¹⁷ történt.

Vélemény kifejezés szótár

A kifejezés szótárban olyan kifejezések kerültek szintén -3 és 3 közé eső polaritásértékkel, melyek polaritása pusztán a bennük lévő szavakból lenne megállapítható. Idetartoznak jellemző ironikus kifejezések is. (Például: *pórusul jár*). A kifejezés szótár 2012.03.25-én 90 kifejezésből áll.

¹⁶ A BME MOKK Hungarian Webcorpus szógyűjteménye elérhető a következő linken: <ftp://ftp.mokk.bme.hu/Language/Hungarian/Freq/Web2.2/web2.2-freq-sorted.top100k.nofreqs.txt>

¹⁷ Automatikus kiegészítésre példa a középfokú mellénevek fokainak automatikus kiegészítése felső és legfelső fokra.

A szótárat egy robot segítségével illetve manuálisan építettem fel. A dolgozat készítése során a robot 2650 weboldalról gyűjtötte ki 2-3-4-5 szavas kifejezéseket, és számolta az előfordulásukat. Az így megtalált leggyakrabban előforduló többszavas kifejezések építettem fel a szótárat.

Véleményerősség szótár

A véleményerősség szótár olyan szavakat tartalmaz, melyek erősítik vagy gyengítik, esetleg módosítják, például: **bámulatosan** jó.

Véleménymódosító szótár

A véleménymódosító szó olyan szavakból áll, melyek nem csak gyengítik vagy erősítik a polaritás szavak vagy kifejezések polaritását, de módosítják is annak irányát a saját irányukra, függetlenül, hogy milyen volt a polaritás szó eredeti polaritása, például:

- jó -> **aggasztóan** jó
- rossz -> **aggasztóan** rossz

A fentiekben tagadószavak és úgynevezett beékelődő szavakból is készült egy kis méretű szótár. Ezek összállítása manuálisan történt.

Családi **problémái** voltak az öregnek, de még így is **nehezen bocsátottuk meg**, hogy 45 percig intézte az ügyeit

	család	probléma	van	az	öreg	de	még	így	is	nehezen	bocsát	meg	hogy	45	perc	inté	az	ügy
NegWord	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
PhrasePolarity	.	.	.	.	.	.	.	.	.	-2	.	.	.	.	.	.	.	.
WordPolarity	.	-2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
WordPower	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
WordModifier	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

6. ábra: Példa elemzésre (mondat forrása:

http://sportgeza.hu/sport/2012/03/24/eroszak_a_sportban/)

6.2 A pontosság tesztelése

A polaritásmérést 458 véletlenszerűen kiválasztott Index.hu hírportál cikkeiből származó mondatokon teszteltem. A mondatokat magam annotáltam: minden mondat pozitív, semleges vagy negatív címkét kapott. Ezt követően futtattam ezeken a polaritásmeghatározó algoritmusomat, az eredményeket a következő két táblázat tartalmazza:

	True poz.	True sem.	True neg.
Pred. poz.	42	18	1
Pred. sem.	23	217	14
Pred. neg.	11	46	86

1. Táblázat: A polaritásmérés összefoglalása

	Menny.	Precision ¹⁸	Recall ¹⁹	F mérték ²⁰
Poz.	76	68,85%	55,26%	61,31%
Sem.	281	85,43%	77,22%	83,25%
Neg.	101	60,14%	85,15%	85,15%
Súlyozott átlag		77,10%	75,33%	80,03%

2. Táblázat: A polaritásmérés hatékonysága

A tesztelés **accuracy²¹** (magyarul a precision-hoz hasonlóan pontosság) értéke **75,33%**.

Ha összevetjük más, angol szövegekre végzett véleményelemzésekkel a fenti eredményeket, akkor azt látjuk, hogy a legjobb eredmények közé tartozik, különösen, ha figyelembe vesszük, hogy a magyar nyelvű szövegek véleményelemzése mennyivel nehezebb, mint az angol szövegeké:

Ki	Hogyan	Nyelv	Accuracy
Pang et al (2002)	SVM algoritmussal unigramok felhasználásával IMDB weboldal filmkritikáinak elemzése	angol	82,90%
Pang et al (2004)	Szubjektivitás osztályozással majd Naiv Bayes módszerrel megvalósított hierarchikus osztályozással IMDB weboldal filmkritikáinak elemzése	angol	86,40%
Jelen dolgozat	Szótár alapú módszerrel weboldalak szövegének véleményelemzése	magyar	75,33%

3. Táblázat: Összevetés a legjobb angol nyelvű megoldásokkal

Az általam vizsgált irodalomban a legjobb pontosságú megoldás a Pang et al. (2004) megoldása. Ettől mindössze 11%-kal marad el a jelen dolgozatban bemutatott megoldás pontossága. Ha azonban a Miháltz [2010] által is hivatkozott 2007-es SemEval angol nyelvű szövegeken végzett véleményelemzési verseny résztvevőinek eredményeivel vetem össze jelen

¹⁸ Precision (magyarul pontosság, de nem összetévesztendő az accuracy értékkel): $TP / (TP + FP)$

¹⁹ Recall (magyarul felidézés vagy megbízhatóság): $TP / (TP + FN)$

²⁰ Az F mérték számításához használt együttható: $\alpha=0,5$

²¹ Accuracy: $(TP + TN) / N$

eredményeket, akkor egyértelmű fölényt mutat a dolgozatban bemutatott módszer. (Az összevetés csak elméleti, mivel eltérő korpuszon és nyelven végrehajtott tesztekéről van szó, így illusztrációként használok a SemEval 2007 verseny eredményeit.)

Rendszer	Accuracy	Precision	Recall	F mérték
ClaC	55,10%	61,42%	9,20%	16,00%
UPAR7	55,00%	57,54%	8,78%	15,24%
SWAT	53,20%	45,71%	3,42%	6,36%
ClaC-NVB	31,20%	31,18%	66,38%	42,43%
SICS	29,00%	28,41%	60,17%	38,6%
OPINHU	55,20%	90,25%	51,95%	65,94%
Jelen rendszer	75,33%	77,10%	75,33%	80,03%

4. Táblázat: Összevetés más angol nyelvű megoldásokkal

Magyar nyelvű szövegekre mindössze néhány hasonló munka készült, ezért a módszerem pontosságát Berend és Farkas (2008) véleményelemzési kutatásának pontosságával vetem össze. Szintén referenciaként szolgál azaz alapérték pontosság (Baseline), amit úgy kapnánk, ha tesztelnénk minden egyedet a leggyakoribb, semleges címkével látnánk el.

Ki	Hogyan	Nyelv	Accuracy
Berend és Farkas (2008)	C4.5 döntési fa algoritmussal kombinált fórum válaszolási gráfmegoldással népszavazásról alkotott vélemények osztályozása	magyar	71,76%
Baseline pontosság	Ha minden mondatot, a leggyakoribb, a semleges kategóriába sorolunk	magyar	61,35%
Jelen dolgozat	Szótár alapú módszerrel weboldalak szövegének véleményelemzése	magyar	75,33%

5. Táblázat: Összevetés más magyar szövegosztályozások pontosságával

A fentiek alapján az 2. hipotézist igazoltnak tekintem.

7 Összefoglalás

A kutatásom alapján igazoltnak tekintem, hogy a magyar nyelvű szövegeken is megvalósítható a legjobb, angol nyelvre készített polaritásmérési megoldások pontossága. A magyar és a többi ragasztó nyelv (például: baszk, török) számítógépes feldolgozásának alapkövetelménye a megfelelő szótövezési algoritmus alkalmazása.

89,67%-os pontosságú raglevágó algoritmus felhasználásával, kivételek meghatározásával és többdimenziós szofisztikált szótárrendszer alkalmazásával 75,33%-os pontosságú²² véleményelemző megoldást sikerült építeni, melyet ha összevetünk ismert véleményelemzési tesztelesek adataival (lásd: Pang et al. (2004) illetve Miháltz (2010) munkáit), akkor látható, hogy igen tekintélyes eredménynek számít.

A dolgozatban sikerült mindkét megfogalmazott hipotézist igazolni:

- **1. hipotézis:** a magyar nyelv algoritmikus raglevágásának hatékony módját sikerült megvalósítani kivételek segítségével pontosított általános raglevágó szabályokkal.
- **2. hipotézis:** a magyar nyelvű webes mondatok véleménypolaritásának számítógépes meghatározása során megközelítettem az emberi ítélőképesség elméleti határát (82%), azaz 75,33%-ban egyezett az automatikus polaritásmeghatározás az általam elvégzett polaritás annotációval.

A kutatás teljes értékeléséhez hozzátartozik, hogy érdemes lenne nagyobb számú mondathalmazon is elvégezni a tesztelést, valamint jelen esetben az emberi annotáció 1 személy által történt, míg hasonló kutatásokban (lásd: Berend és Farkas (2008)) sokszor 2-3 szakértő által annotált mondatokon szoktak tesztelni, ahol is a tesztmondatok közé csak a minden annotáló által azonosan megítélt mondatokat veszik be.

A bemutatott polaritáselemzés pontossága a későbbiekben tovább növelhető még különféle módszerek segítségével:

- **Szinoníma szótár kialakítása és integrálása:** a szinoníma szótárban szereplő de a vélemény szótárban nem szereplő szavak polaritására és/vagy erősségére becslés adható a vélemény szótárban szereplő szinonímák értékeinek átlagolásával.

²² A pontosság itt az accuracy mutatót jelenti.

- **Kifejezések bővítése:** a kifejezés szótár bővítésével jelentősen javítható a rendszer pontossága, képessé válik a leggyakoribb metaforák illetve ironikus kifejezések polaritásának helyes megállapítására.
- **Névelemek felismerése:** tulajdonnév adatbázis kialakításával elkérülhetők olyan hibák például, hogy a `Jobbik` szót félreértelmezze a rendszer.

A fenti lépésekkel bővített megoldás kialakítása és tesztelése nem része jelen dolgozatnak, egy későbbi munkában kerül majd ismertetésre.

8 Melléklet

8.1 Raglevágás teszteléshez használt url lista

Letöltve: 2012.03.11-én

ID	URL
1	http://belseg.blog.hu/2012/01/19/az_ev_autos_reklamja
2	http://belseg.blog.hu/2012/02/01/ezt_a_suzukit_inkabb_ne_vedd_meg
3	http://belseg.blog.hu/2012/02/01/tudjak_mi_a_furcsa
4	http://belseg.blog.hu/2012/02/01/tudjak_mi_a_furcsa?utm_source=bloghu_megosztas&utm_medium=buzz&utm_campaign=blhshare
5	http://belseg.blog.hu/2012/02/01/tudjak_mi_a_furcsa?utm_source=bloghu_megosztas&utm_medium=facebook&utm_campaign=blhshare
6	http://belseg.blog.hu/2012/02/01/tudjak_mi_a_furcsa?utm_source=bloghu_megosztas&utm_medium=iwiw&utm_campaign=blhshare
7	http://belseg.blog.hu/2012/02/02/darth_vader_visszavag?utm_source=bloghu_megosztas&utm_medium=buzz&utm_campaign=blhshare
8	http://belseg.blog.hu/2012/02/02/darth_vader_visszavag?utm_source=bloghu_megosztas&utm_medium=facebook&utm_campaign=blhshare
9	http://belseg.blog.hu/2012/02/02/darth_vader_visszavag?utm_source=bloghu_megosztas&utm_medium=iwiw&utm_campaign=blhshare
10	http://belseg.blog.hu/2012/02/03/ferrari_ff_ahogy_nem_ismered?utm_source=bloghu_megosztas&utm_medium=buzz&utm_campaign=blhshare
11	http://belseg.blog.hu/2012/02/03/ferrari_ff_ahogy_nem_ismered?utm_source=bloghu_megosztas&utm_medium=facebook&utm_campaign=blhshare
12	http://belseg.blog.hu/2012/02/03/ferrari_ff_ahogy_nem_ismered?utm_source=bloghu_megosztas&utm_medium=iwiw&utm_campaign=blhshare
13	http://belseg.blog.hu/authors/ommm
14	http://belseg.blog.hu/tags/911
15	http://belseg.blog.hu/tags/ fiat
16	http://belseg.blog.hu/tags/ opel
17	http://belseg.blog.hu/tags/ porsche
18	http://belseg.blog.hu/tags/ starwars
19	http://belseg.blog.hu/tags/ volkswagen
20	http://belseg.blog.hu/tags/ walter
21	http://blog.hu/user/126323
22	http://blog.hu/user/160607
23	http://blog.hu/user/173221
24	http://blog.hu/user/183045
25	http://blog.hu/user/20224
26	http://blog.hu/user/244269
27	http://blog.hu/user/257496
28	http://blog.hu/user/261439
29	http://blog.hu/user/353857
30	http://blog.hu/user/361059
31	http://blog.hu/user/419166/tab/data

ID	URL
32	http://blog.hu/user/648363
33	http://blog.hu/user/655065
34	http://blog.hu/user/88713
35	http://blog.hu/user/93791
36	http://cemp.blog.hu/2011
37	http://cemp.blog.hu/2011/02
38	http://cemp.blog.hu/2011/02/21
39	http://cemp.blog.hu/2011/02/21/elesben_az_elso_tvo_peugeot_508
40	http://mailto:sales@cemp.hu
41	http://nol.hu/archivum/orban_es_az_eu_brusszelben_mar_az_anyagaszal_tartanak
42	http://nol.hu/archivum/paul_lendvai_azert_meghagyna_orbant_miniszterelnoknek
43	http://nol.hu/kulfold/_az_egyeduli_fenypont_az_hogy_magyarország_csak_egy_kis_krumpli
44	http://nol.hu/kulfold/bbc_magyarország_feladta
45	http://nol.hu/kulfold/cimkek/lapszemle
46	http://nol.hu/kulfold/magyarország_megkapta_az_ultimatumot
47	http://nol.hu/moderalasi_alapelvek
48	http://totalcar.hu/24ora/?tol=2011-02-04&ig=2012-02-04&cimke=ausztria
49	http://totalcar.hu/24ora/?tol=2011-02-04&ig=2012-02-04&cimke=bonusmalus
50	http://totalcar.hu/24ora/?tol=2011-02-04&ig=2012-02-04&cimke=daewoo
51	http://totalcar.hu/24ora/?tol=2011-02-04&ig=2012-02-04&cimke=elgurult
52	http://totalcar.hu/24ora/?tol=2011-02-04&ig=2012-02-04&cimke=hitel
53	http://totalcar.hu/24ora/?tol=2011-02-04&ig=2012-02-04&cimke=nissan
54	http://totalcar.hu/24ora/?tol=2011-02-04&ig=2012-02-04&cimke=suzuki
55	http://totalcar.hu/magazin/hirek/2011/02/03/ettol_igazi_darth_vader_leszel
56	http://totalcar.hu/tanacsok/karszakerto/?p=0
57	http://totalcar.hu/tanacsok/karszakerto/?p=1
58	http://totalcar.hu/tanacsok/karszakerto/?p=2
59	http://totalcar.hu/tanacsok/karszakerto/2011/10/14/behuzott_kezifek_mellett_lecsuszott_a_kocsi
60	http://totalcar.hu/tanacsok/karszakerto/2011/10/18/biztositas_ket_biztositonal
61	http://totalcar.hu/tanacsok/karszakerto/2011/10/25/totalkaros_a_suzukim
62	http://totalcar.hu/tanacsok/karszakerto/2011/10/26/munkaba_jaras_akadalyozasa
63	http://totalcar.hu/tanacsok/karszakerto/2011/10/30/elgurult_a_daewoo
64	http://totalcar.hu/tanacsok/karszakerto/2011/11/07/magyar_bonus_ausztriaban
65	http://totalcar.hu/tanacsok/karszakerto/2011/11/08/ceges_auto_eladasa
66	http://totalcar.hu/tanacsok/karszakerto/2011/11/11/kotelezo_mas_nevere
67	http://totalcar.hu/tanacsok/karszakerto/2011/11/15/ervenyes_a_biztositas_vagy_sem
68	http://totalcar.hu/tanacsok/karszakerto/2011/11/18/valtozhat-e_az_evkozi_kotes_ideje_ha_most_ujat_kotok
69	http://totalcar.hu/tanacsok/karszakerto/2011/11/22/cserbenhagytak_mit_tehetek
70	http://totalcar.hu/tanacsok/karszakerto/2011/11/23/fenyezés_utani_ertekcsokkenes
71	http://totalcar.hu/tanacsok/karszakerto/2011/11/28/biztositovaletas
72	http://totalcar.hu/tesztek/2012/02/01/lekested-e_az_utolso_igazi_porschet
73	http://totalcar.hu/tesztek/2012/02/03/ma_mar_mentem_haromszazzal
74	http://valodisafrany.blog.hu/2011
75	http://valodisafrany.blog.hu/2011/10

ID	URL
76	http://valodisafrany.blog.hu/2011/10/19/valodi_francia_pekseg_cukraszat_budapest_n
77	http://valodisafrany.blog.hu/2011/11
78	http://valodisafrany.blog.hu/2011/11/23/fuszeres_sutoben_sult_halfile_mandulas_ha_gymas_rizzsel_es_fugevel
79	http://valodisafrany.blog.hu/2011/12
80	http://valodisafrany.blog.hu/2011/12/05/vietnami_tavaszi_tekercs_azaz_a_nem
81	http://valodisafrany.blog.hu/2011/12/29/tojasfelho_leves
82	http://valodisafrany.blog.hu/2011/3
83	http://valodisafrany.blog.hu/2011/4
84	http://valodisafrany.blog.hu/2011/5
85	http://valodisafrany.blog.hu/2011/6
86	http://valodisafrany.blog.hu/2011/7
87	http://valodisafrany.blog.hu/2011/8
88	http://valodisafrany.blog.hu/2011/9
89	http://valodisafrany.blog.hu/2012/01
90	http://valodisafrany.blog.hu/2012/01/22/lumaconi_gigante_orias_csigateszta_lazacca_l_es_brokkolival
91	http://valodisafrany.blog.hu/2012/01/31/paradicsomos_oriasbab
92	http://valodisafrany.blog.hu/2012/01/31/paradicsomos_oriasbab?utm_source=bloghu_megosztas&utm_medium=buzz&utm_campaign=blhshare
93	http://valodisafrany.blog.hu/2012/01/31/paradicsomos_oriasbab?utm_source=bloghu_megosztas&utm_medium=facebook&utm_campaign=blhshare
94	http://valodisafrany.blog.hu/2012/02/04/kijevi_torta_az_edes_tevedes?utm_source=bloghu_megosztas&utm_medium=buzz&utm_campaign=blhshare
95	http://valodisafrany.blog.hu/2012/02/04/kijevi_torta_az_edes_tevedes?utm_source=bloghu_megosztas&utm_medium=facebook&utm_campaign=blhshare
96	http://valodisafrany.blog.hu/2012/1
97	http://valodisafrany.blog.hu/2012/2
98	http://valodisafrany.blog.hu/archive
99	http://valodisafrany.blog.hu/tags/bab
100	http://valodisafrany.blog.hu/tags/kijevi_torta
101	http://www.noltv.hu/csatorna/1/legfrissebb
102	http://www.noltv.hu/csatorna/1/legvitatottabb
103	http://www.noltv.hu/video/2853.html
104	http://www.noltv.hu/video/3262.html
105	http://www.noltv.hu/video/4108.html
106	http://www.noltv.hu/video/4121.html
107	http://www.noltv.hu/video/4133.html

6. Táblázat: A raglevágás tesztelés használt szövegek forrás URL-jei

9 Felhasznált források

Baccianella, S., Esuli, A. és Sebastiani, F. [2010]: SENTIWORDNET 3.0: An enhanced lexical resource for sentiment analysis and opinion mining

Bárházi, E., Héder, M. [2010]: Panaszlevelek szerkezetének gépi felismerése, Magyar Számítógépes Nyelvészeti Konferencia 2010.

Berend, G., Farkas, R. [2008]: Opinion Mining in Hungarian based on textual and graphical clues, in Proceedings of the 4th Intern. Symposium on Data Mining and Intelligent Information Processing, Santander, 2008.

Chomsky, N. [1965]: Aspects of the Theory of Syntax, MIT Press

Fajsz, B [2010]: Üzleti haszon az adatok mélyén – Az adatbányászat mindennapjai, Alinea Kiadó – IQSYS Informatikai és Tanácsadó Zrt.

Grimes, S. [2011]: Text Analytics and Sentiment Analysis (forrás: <http://www.b-eye-network.com/view/15276>, utoljára letöltve: 2012.03.12-én)

Grossman, L. [2010]: Twitter can predict the stock market (forrás: <http://www.wired.com/wiredscience/2010/10/twitter-crystal-ball/>, utoljára letöltve: 2012.03.12-én)

Halácsy, P., Kornai, A., Németh, L., Rung, A., Szakadát, I., Trón, V. [2004]: Creating open language resources for Hungarian, In Proceedings of the 4th international conference on Language Resources and Evaluation (LREC2004), 2004

Kornai, A, Halácsy, P, Nagy, V, Oravecz, Cs, Trón, V, and Varga, D [2006]: Web-based frequency dictionaries for medium density languages, In: Proceedings of the 2nd International Workshop on Web as Corpus, edited by Adam Kilgariff, Marco Baroni ACL-06, pages 1--9.

Miháltz Márton [2010]: OpinHu: online szövegek többnyelvű véleményelemzése, VII. Magyar Számítógépes Nyelvészeti Konferencia

Pang, B., Lee, L., Vaithyanathan, S. [2002]: Thumbs up? Sentiment Classification using Machine Learning Techniques. In: Proceedings of EMNLP-2002 (2002) 79–86

- Pang, B., Lee, L. [2004]:** A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In: Proceedings of ACL-2004 (2004)
- Pang B. és Lee L. [2008]:** Opinion-mining nad sentiment analysis, In: Foundations and TrendsR, Information Retrieval Vol. 2, Nos. 1–2 (2008) 1–135. pp.
- Sántáné-Tóth [2008]:** Döntéstámogató rendszerek, Panem kiadó, Budapest
- Szekeres, P. [2008]:** A gépi tanulás intelligens módszerei a döntéstámogatásban -A legemberközelibb technikák összehasonlítása, BCE Gazdálkodáskodástudományi Kar, Tudományos Diákköri dolgozat
- Szekeres, P. [2009]:** Az információszerzés új dimenziói – az adatbányászat, a szövegbányászat és a webbányászat üzleti előnyei, BCE Gazdálkodáskodástudományi Kar, Tudományos Diákköri dolgozat
- Szekeres, P. [2011]:** A tőzsdei kisbefektetői hangulat elemzése webes információk alapján, BCE Gazdálkodáskodástudományi Kar, Tudományos Diákköri dolgozat
- Tikk, D. [2007]:** Szövegbányászat, Typotex Elektronikus Kiadó Kft, Budapest
- Tordai, A. és Rijkee, M. [2005]:** Hungarian monolingual retrieval at CLEF 2005
- Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., Patwardhan, S. [2005]:** OpinionFinder: A system for subjectivity analysis. In: Proceedings of HLT/EMNLP 2005 Interactive Demonstrations