

TDK-dolgozat

Szekeres Péter

2011

A tőzsdei kisbefektetői hangulat elemzése webes információk alapján

Budapesti Corvinus Egyetem, Gazdálkodástudományi Kar, Információrendszerek Tanszék

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

Konzulens: Vas Réka (BCE Információrendszerek Tanszék, adjunktus)

Budapest, 2011.03.25.

Tartalomjegyzék

1	Vezetői összefoglaló.....	4
2	Bevezetés.....	5
3	Az online tőzsdei magánbefektetők döntési modellje	6
4	Az opinion-mining alapjai	11
5	A SIMA alapműködése	12
5.1	A SIMA architektúrája.....	13
5.2	A SIMA elemző algoritmus.....	14
6	A web bejárása	17
7	A letöltött dokumentumok szövegbányászati elemzése.....	21
7.1	Szövegelemzési szempontok modellezése	22
7.1.1	A SIMA szótárainak kezelése.....	22
7.1.2	Relevanciamutató.....	23
7.1.3	Véleménymutató	25
7.1.4	A tartalom prioritása	27
7.2	Az elemzés kimenete	27
8	A kutatás bemutatása	29
8.1	A vizsgálat menete.....	29
8.2	A futtatások eredményeinek elemzése	30
8.2.1	Az eddigi eredmények összefoglalása.....	30
8.2.2	Az eredmények statisztikai vizsgálata	31
9	Összefoglalás	33
10	Melléklet	35
11	Felhasznált források.....	36

1 Vezetői összefoglaló

I. Téma ismertetése

Egy saját fejlesztésű Java és MySQL alapú szoftverrendszerrel, a SIMA-val vizsgáltam a magyar internetes részvénykereskedésre ható vélemények és az OTP részvényárfolyam változásának kapcsolatát.

II. A SIMA eszköz

A SIMA két fő komponensből áll:

- Crawler: feladata a web bejárása és a linkek összegyűjtése a weblapokról
- Opinion processor: feladata a releváns tartalmak azonosítása, és a vélemények számszerűsítése

III. Feldolgozási és értékelési modell

A releváns weblapok kiválasztása a meta adatok, egy URL tiltólista és a relevanciát azonosító gazdasági szótárak alapján, a releváns tartalmak véleményeinek számszerűsítése a pozitív és negatív szótár szavainak előfordulási aránya alapján történik. A futások végén a megvizsgált véleményeket/befektetői hangulatot a SIMA egyetlen -1 és 1 közé eső indexben foglalja össze.

IV. A kutatás eredményei

A kutatás során 10 kereskedési nap futási eredményeit és árfolyamváltozását vetettem össze. Ezek alapján pozitív sztochasztikus kapcsolatot azonosítottam az OTP részvények valós árfolyamváltozása és a releváns webes információk megállapított hangulata/véleménye között.

A kovariancia 0,652 értékű, azonban csak nagyon gyenge lineáris kapcsolatot sikerült azonosítani az eddigi eredmények alapján.

V. Következtetések

A SIMA hatékonyan tömöríti a webes véleményeket egyetlen, könnyen értelmezhető mutatóba, azonban a szótárak szavainak tanulási algoritmusokkal történő finomítása illetve súlyozása jelentős hatékonyság javulást eredményezhet.

2 Bevezetés

Az e-kereskedelmi rendszerek lehetővé tették a 2000-es évekre, hogy a brókerházak és befektetési alapok vagy egyéb közvetítő szereplők kihagyásával tulajdonképpen közvetlenül – csupán a pénzintézetek online platformjait igénybe véve – kereskedhessenek a magánbefektetők részvényekkel a magyar tőzsdepiacra:

- KBC Equitas: WebBroker rendszer
- CIB Bank: eBroker rendszer
- Portfolio.hu: Portfolio Alap és PRO rendszer
- Gold Forex Trading: DealBook 360

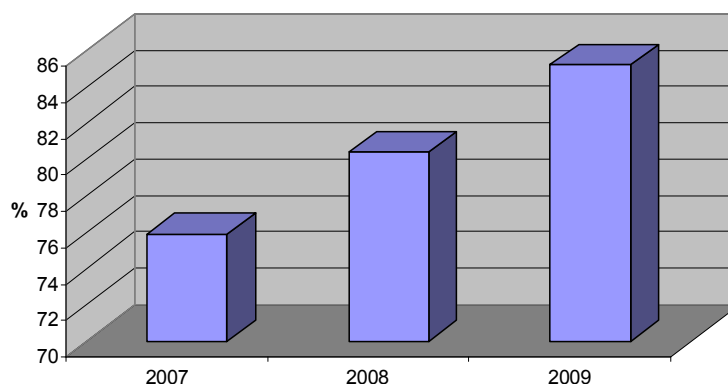
A fenti, illetve az ezekhez hasonló rendszerekben bárki nyithat értékpapírszámlát, és kezdhet részvénykereskedésbe. A belépéshez csak minimális tőke szükséges, és a tranzakciós díjak is igen alacsonyak. Az online tőzsdei rendszerek felhasználói jellemzően árfolyam spekuláns magánbefektetők, egyben rendszeres internetfelhasználók, akik a kereskedési döntéseik alapjául szolgáló információkat egyrészt a használt rendszerben elérhető statisztikai adatokból, árfolyamadatokból és elemzésekből nyerik, másrészt különböző internetes oldalak: hírportálok, tematikus fórumok, blogok befolyásolják kereskedésüket.

A dolgozatom célja, hogy bebizonyítsam, hogy a kisbefektetők befektetési hangulata reprezentálható a weben megjelenő, részvénykereskedelemhez köthető információkkal, és ezen hangulat pozitív kapcsolatban áll az egyes részvények árfolyam változásával.

Egy általam fejlesztett szoftver rendszer, a **SIMA** (*Small Investor Mood Analyzer*) segítségével elemzem a magyar weboldalakon található információkat. A rendszer minden futás során megállapít egy az adott részvényre vonatkozó mutatószámot, mely tömörítve jellemzi a részvényre vonatkozó tartalmak optimista/pesszimista voltát (**AIM** – *Aggregated Investor Mood* index). A dolgozatban a SIMA számítási modelljének, illetve a modell kialakításához vezető hipotézisek bemutatása után ismertetem egy 10 kereskedési napot érintő megfigyelés eredményeit, mely során azt vizsgáltam és elemeztem statisztikai kapcsolati mutatók segítségével, hogy az OTP részvény árfolyamának valós változása milyen statisztikailag értelmezhető kapcsolatba hozható a megelőző napon a részvényre számított AIM értékekkel.

3 Az online tőzsdei magánbefektetők döntési modellje

Az internet terjedése és egyben az e-kereskedelmi szolgáltatások iránti felhasználói bizalom erősödése stabil – 2007 és 2009 között majdnem évi 5%-os - növekedést eredményezett a magyar internetes banki és pénzügyi szolgáltatások igénybevételének területén.



1. ábra: A magyarországi internethasználók között a banki és pénzügyi szolgáltatások igénybevevőinek aránya a KSH adatai alapján¹

Az online tőzsdei kereskedés magyarországi megjelenésének köszönhetően kisbefektetők ezreinek magántőkéje növeli a Budapesti Értéktőzsde forgalmát. A platformok, melyeken keresztül a felhasználók kereskedhetnek a következő főbb szolgáltatásokat nyújtják:

- Késleltetett (15 perccel korábbi) tőzsdei adatok²
- Részvénykereskedelem
- Határidős kereskedelem
- Állampapír kereskedelem

¹ A KSH internet igénybevételének célját tartalmazó összefoglaló táblázat alapján, letölthető: http://portal.ksh.hu/pls/ksh/docs/hun/xstadat/xstadat_eves/i_oni011.html weboldaltól, utoljára letöltve: 2011.03.19-én

² Az azonnali információkhoz való hozzáférés jellemzően külön fizetős szolgáltatásként kínálják.

- Daytrade ügyletek: napon belüli kereskedés jellemzően alacsonyabb jutalékkal, mint normál kereskedés esetén
- Technikai daytrade ügyletek: tőkeáttételes napon belüli kereskedés

A magyar magánbefektetők elsősorban a magyar blue chippekkel (OTP, MOL, TELEKOM, RICHTER) kereskednek, ezek közül is kiemelkedik az OTP, mely köztudottan volatilis árfolyama miatt igen népszerű a spekuláns magánbefektetők között. Az MTI adatai szerint 2010-ben a BÉT részvényforgalmának 97,5%-át tette ki az 5 legnagyobb papír forgalma: **az OTP papírok a teljes forgalom 67,2%-át tették ki.** A teljes részvényforgalom 25-30%-át adta a mintegy 15 ezer magánbefektető.³ Mindez azt jelenti, hogy egy magánbefektető átlagosan 2010-ben az 5522 milliárd Ft összforgalomból 110,44 millió Ft-tal vette ki részét éves szinten.⁴

Mindezek mellett napjainkra kijelenthető, hogy **az online kereskedés a magánbefektetők legnépszerűbb kereskedési formája** a közvetlen döntési lehetőségek, az alacsony jutalékok valamint a gyors és egyszerű üzletkötés miatt.⁵

A továbbiakban a vizsgált probléma formalizálásának elősegítése miatt felteszem, hogy az online tőzsdézők egy részvénnel kapcsolatban kétféle döntést hozhatnak: vesznek vagy eladnak.⁶ Az így kialakított félig-strukturált problémának könnyen megadható az egyszerűsített leíró döntési modellje (*descriptive decision model*):⁷

**Előkészítés +
Tervezés**

1. A magánbefektető kiválasztja a számára érdekes részvényt
2. Elemzi és értékeli a rendelkezésére álló releváns információkat

³http://www.mfor.hu/cikkek/Tartosan_nott_a_hazai_maganbefektetok_aktivitasa_a_BET_en.html weboldal adatai alapján, utoljára letöltve 2011.03.19-én

⁴ A részvényforgalom 30%-ával számolva.

⁵http://bet.hu/topmenu/befektetok/tozsde_lepesrol_lepesre/azonnali_piacismeretek/hogyan_kereskedj_unk_a_tozsden/tozsdei_kereskedo_ceg_kival weboldal adatai alapján, utoljára letöltve 2011.03.19-én

⁶ A valóságban egy magánbefektető számtalan egyedi alternatíva (shortolás, longolás, kockázatcsökkentés, FIFO/LIFO értékesítés, Stop-loss lehetőségek, portfólió diverzifikáció, day-trading, szavazati jogok, osztalékvárakozások előnyben részesítése, stb.) és azok kombinációi közül választhat.

⁷ Az egyes lépésekhez Turban általános döntési folyamatának lépései (Sántáné-Tóth [2008]) is hozzá lettek rendelve.

Választás**3. Eldönti, hogy elad-e vagy vesz-e részvényt****Megvalósítás****4. Az online platformon keresztül kiadja ajánlatát**

Bár a valóságban előfordulhat, hogy egy adott esemény, illetve arról szóló információ indukálja a vételt vagy eladást, a továbbiakban a fenti általános modellel dolgozom. A SIMA feladata elsősorban a 2. lépés támogatása, ennek megfelelően lehetőséget ad a program a megfelelő részvény kiválasztására (1. lépés). A 3. lépés a 2. lépés eredményeinek értékelése alapján történik racionális esetben, míg a 4. lépés az erre dedikált kereskedési rendszerben történik.

A döntést befolyásoló tényezők – információk és adatok – igen sokfélék lehetnek:

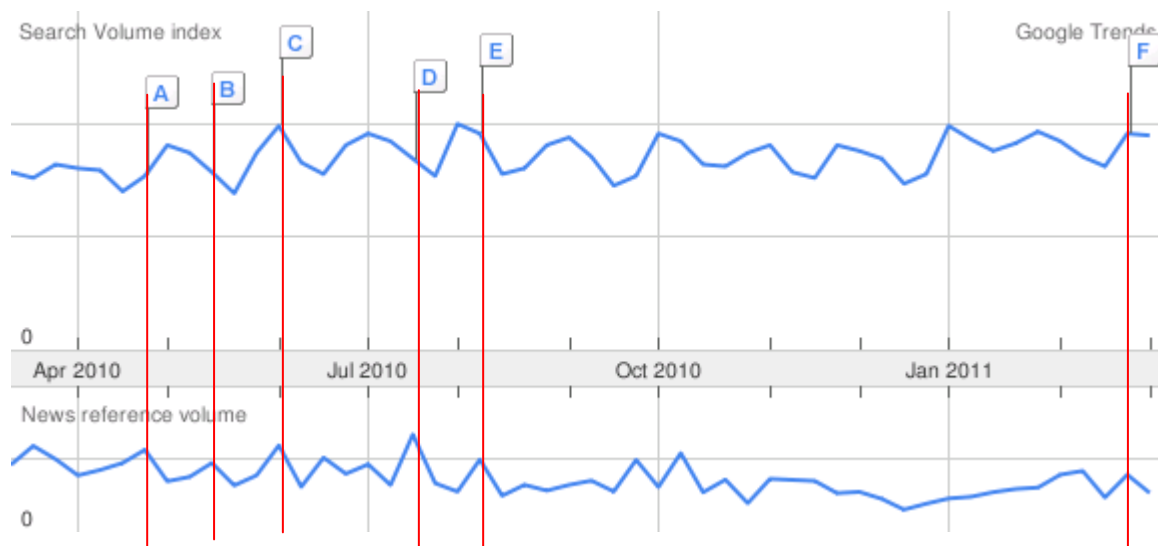
Döntési tényező	Webes előfordulás
1. Statisztikai adatok, elemzések, trendek	1. A legtöbb online kereskedési platform biztosít technikai elemzéseket 2. Az üzleti/gazdasági/hír portálok szöveges összefoglalói a technikai elemzések eredményeiről, trendekről
2. Gazdasági események, hírek	1. Hírportálok 2. Blogok (a fontosabb gazdasági/üzleti eseményekről jellemzően megjelennek bejegyzések) 3. Fórumok (a fontosabb gazdasági/üzleti eseményekről jellemzően megjelennek bejegyzések)
3. Hivatalos szakértői nyilatkozatok, előrejelzések, pénzügyi elemzések	1. Gazdasági témájú internetes újságok 2. Elemző cégek weboldalai
4. Egyéni vélemények, fórumok, blogok	1. Fórumok 2. Hír kommentek

Döntési tényező	Webes előfordulás
	3. Blog bejegyzések
5. Belső információk	1. Nem jellemző
6. Belső megérzések	1. Nem jellemző
7. Mások befektetési stratégiája	1. Blogok 2. Fórumok 3. Tőzsdei közösségi vagy egyéb közösségi oldalak
8. Saját befektetési stratégia	1. Nem jellemző

1. Táblázat: A részvénykereskedésre vonatkozó döntési tényezők és webes megjelenési formájuk

A fenti táblázatból látható, hogy a kereskedési döntést meghatározó legtöbb tényezőnek létezik releváns webes megjelenési formája. A különböző gazdasági hírportálok, blogok, fórumok, kommentek, internetes újságok, kiadványok tartalmainak elemzésével a legtöbb tényezővel kapcsolatban szerezhethetünk információkat. Hasonló kutatást végzett Berend és Farkas [2008], mely során magyar nyelvű fórum vélemények alapján készítettek előrejelző rendszert vizsgálva az egyes fórumozók egymással való interakcióit.

A webes információk aktualitását és relevanciáját támasztja alá, hogy a fontosabb – kereskedést is befolyásoló – gazdasági, üzleti eseményeknél jellemzően megnövekszik a céggel/részvénnel vonatkozó internetes hírek/hivatkozások száma. A Google Trends szolgáltatásának 2011.03.19-i adatai alapján például látható (2. ábra), hogy a Google által kiemelt OTP események (A-F) egybeesnek az OTP-hez kapcsolódó hírek mennyiségének lokális maximumhelyeivel.



2. ábra: Az OTP szó keresési és hivatkozási adatai Google Trends alapján⁸

A SIMA célja tehát, hogy a célrészvénnnyel (kibocsátó céggel) illetve a gazdasági témákkal foglalkozó webes szöveges információkat felkutassa és értékelje, valamint ez alapján elősegítse a vételi/eladási döntés meghozását.

⁸ A <http://trends.google.com/trends?q=otp&date=ytd&geo=hu&ctab=0&sort=0&sa=N> weboldal adatai alapján, utoljára letöltve: 2011.03.19-én

4 Az opinion-mining alapjai

A részvényekhez köthető tartalmak elemzése ún. *opinion-mining* elveken és megközelítésen alapszik. Az *opinion-mining* vagy *sentiment analysis* célja, hogy a weben található strukturálatlan szöveges adatokat elemezze, és a vizsgálatok eredményeként tömörítse vagy csoportosítsa az információkat.

Amíg az internetes keresők elsősorban tényadatokat vizsgálnak kulcsszavak segítségével, addig az *opinion-mining* kevésbé egzakt számítási és feldolgozási algoritmusokra épül (Liu [2006]). Fontos szerep jut annak, hogy mit vizsgálunk, mi a vizsgálat célja (információ/tudás kivonatolás, csoportosítás, vélemények osztályozása, termékek összehasonlítása, stb.), milyen nyelven történik a vizsgálat, és annak a nyelvnek milyen nyelvtani és morfológiai szabályai vannak. **Ahogy Pang és Lee [2008] is hangsúlyozza, a szubjektív vélemények feldolgozása egyre több üzleti és nem üzleti rendszerben kap szerepet.**

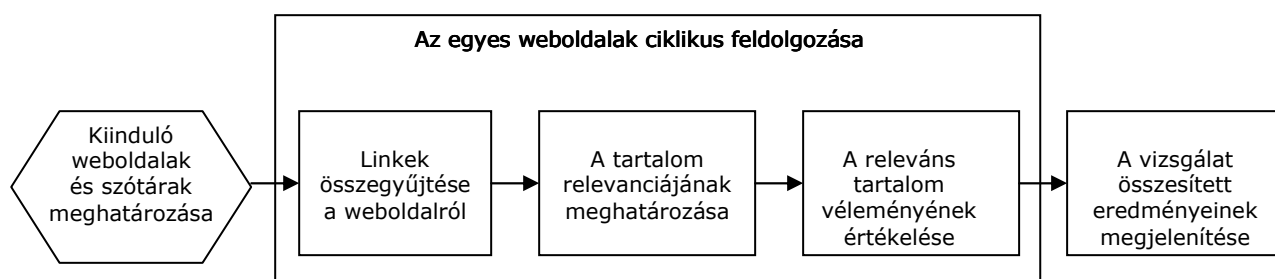
Az *opinion-mining* tipikus alkalmazási lehetőségei (Liu [2006]):

- **Üzlet és szervezetek:** fogyasztói szokások, vélemények gyors és automatikus feldolgozása és elemzése
- **Egyének:** politikai vélemények tömörítése, termékekről információszerzés, egyéb döntési helyzetekben információgyűjtés
- **Hirdetések elhelyezése:** felhasználó által generált tartalmak mellé tematikusan kapcsolódó hirdetések automatikus elhelyezése
- **Véleménykeresés:** egyéb, általános célú véleménykutatás

Jelen dolgozat témája az első kategóriába tartozik, a cél az részvénnel kapcsolatos közvetlen és közvetett (általános gazdasági) tartalmak automatikus feldolgozása.

5 A SIMA alpműködése

A SIMA célja, hogy az egy futtatás alatt talált releváns webes szöveges tartalmakat összetömörítse egyetlen -1 és 1 közé eső mutatószámmá (**AIM** index). A mutatószám képzése során minden egyes megvizsgált weboldaltól ellenőrzi a szoftver, hogy a tartalom releváns-e a vizsgált részvények szempontjából vagy egyáltalán gazdasági szempontból, és ha releváns a tartalom, akkor az adott szempontra – részvény vagy általános gazdasági hangulat - véleménymutatót számít.



3. ábra: A feldolgozás lépései

Tehát egy weboldal feldolgozásának 3 különböző eredménye lehet:

- **A vizsgált részvényre vonatkozó SSIM (*Single Stock Investor Mood*) index:** a weboldal részvennyel kapcsolatos véleményét, hangulatát tükrözi. A futtatás végén SSIM értékek átlaga adja a **SIM** mutatót (*Stock Investor Mood*).
- **Általános gazdasági véleményt tükröző SGIM (*Single General Investor Mood*) index:**⁹ az adott weboldal általános gazdasági helyzettel, körülményekkel kapcsolatos véleményét, hangulatát tükrözi.¹⁰ A futtatás végén SGIM értékek átlaga adja a **GIM** mutatót (*General Investor Mood*).

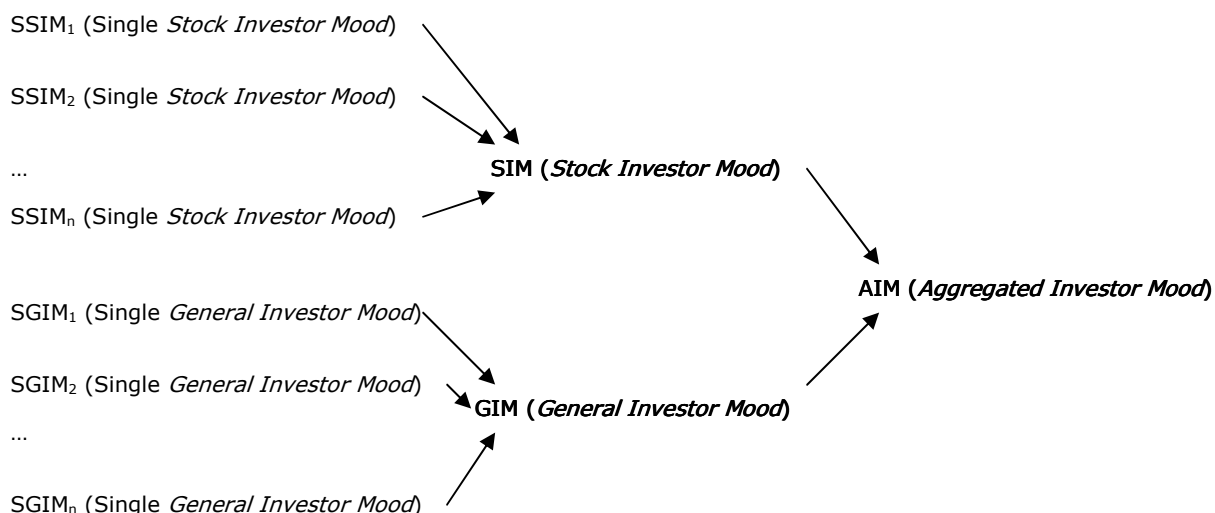
⁹ Az egyes indexek még finomodnak az oldal fontosságát meghatározó prioritással való súlyozással (lásd később)

¹⁰ Az értékelő modell több feltételezéssel is él:

- A magyar nyelvű gazdasági cikkek többsége a magyar gazdasági helyzettel foglalkozik.

- **Az oldal tartalma nem releváns**

A futás befejezésekor a rendszer az addig feldolgozott részvény (SIM mutató) és általános gazdasági hangulatot (GIM mutató) átlagolja,¹¹ mivel a modell a vélemények összesítésénél abból indul ki, hogy az általános befektetési (gazdasági) hangulat is hasonló mértékben határozza meg a kisbefektetői tőzsdei döntéshozatalt, mint a célrészvényre vonatkozó információk.



4. ábra: A SIMA mutatószámrendszere

5.1 A SIMA architektúrája

Architektúráját tekintve a rendszer három fő részre bontható:

- 1 **SIMA adatbázis:** MySQL adatbázis, mely tartalmazza a vizsgálat során használt tiltott URL címeket és sztringeket, tartalmazza a szövegbányászati funkcióknál használt szótárakat, valamint historikusan tárolja az egyes futások paramétereit.
- 2 **SIMA alkalmazás:** grafikus felhasználói felülettel rendelkező Java desktop alkalmazás, mely a web bejárása során párhuzamosan kigyűjti a weboldalon található linkeket, illetve elemzi az adott weboldal szövegét.

-
- A magyar befektetői hangulattal, tehát közvetve a magyar részvényárakkal a gazdasági témájú írások hangulata pozitív sztochasztikus kapcsolatban áll.

¹¹ Tehát $AIM = (SIM + GIM) / 2$

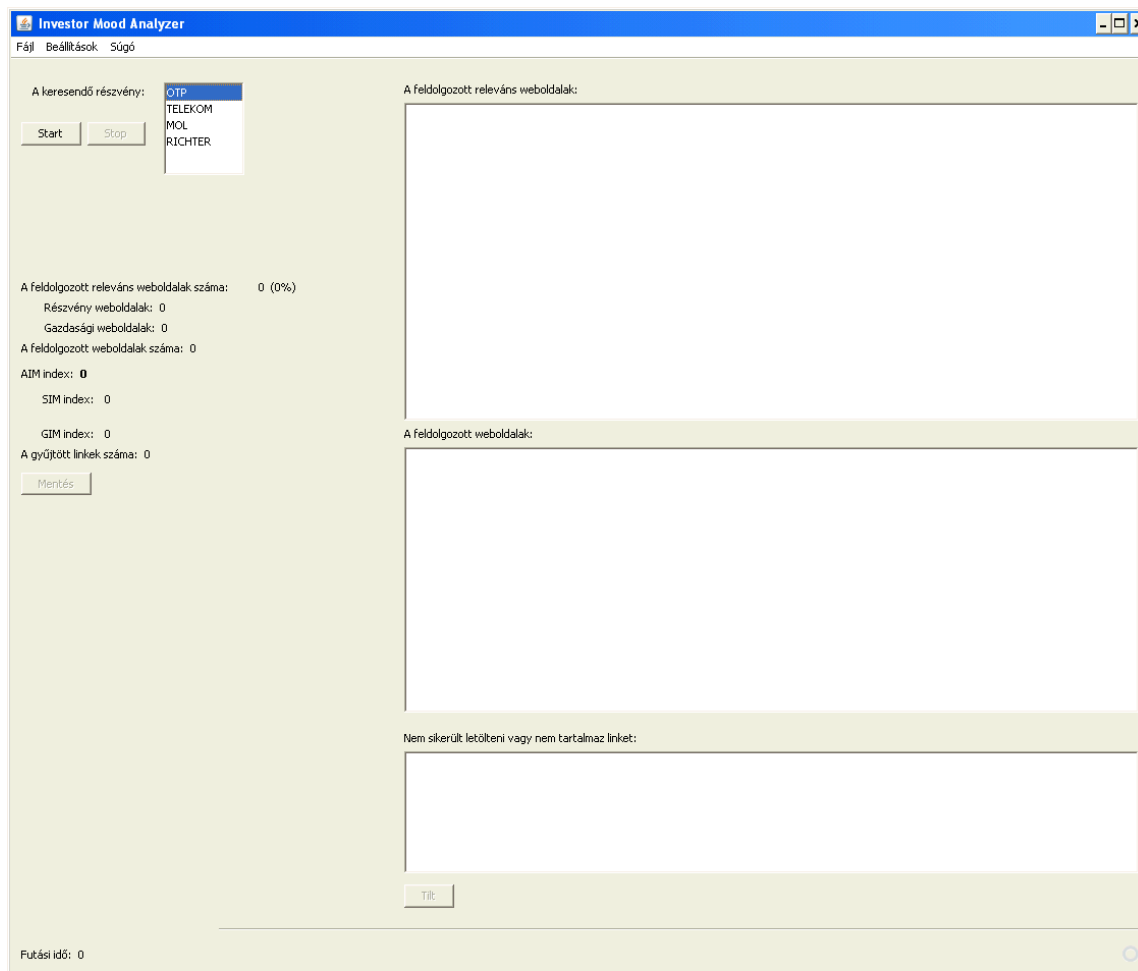
3 **Weben található tartalmak:** a rendszer inputjául szolgáló tartalmak.



5. ábra: A SIMA architektúrája

5.2 A SIMA elemző algoritmus

A SIMA futtatása előtt be kell jelölnünk, hogy mely részvényt vizsgáljuk.



6. ábra: A SIMA főképernyője

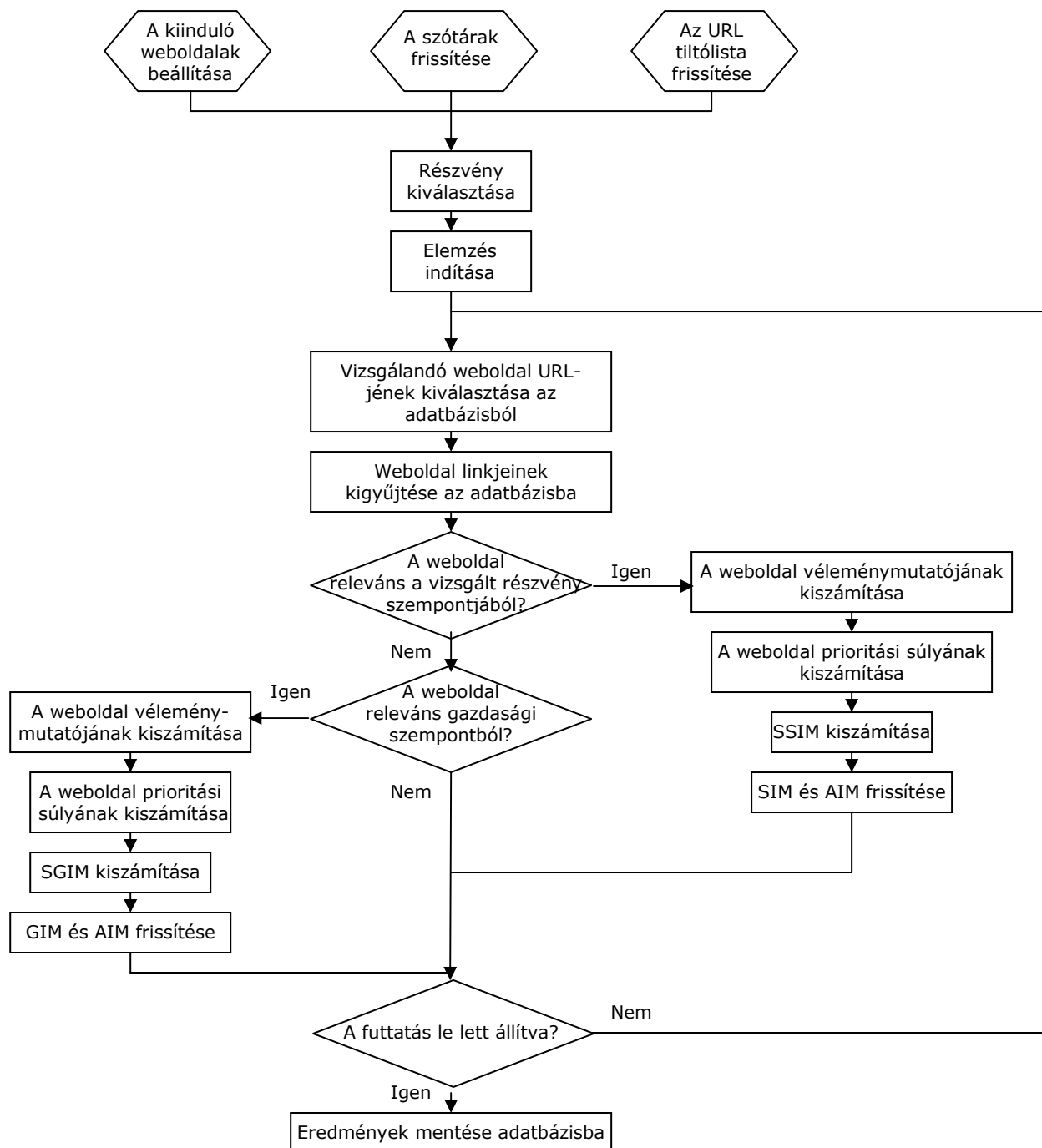
A weblapok elemzése során a SIMA először begyűjti az oldalon található linkeket a memóriájába (LinksBuffer sztring tömbbe), majd eldönti a weboldal meta információi és a domain név alapján, hogy az adott weboldal releváns-e gazdasági vagy az adott részvény szempontjából, vagyis kapcsolódik-e a vizsgálat tárgyához a tartalom, érdemes-e részletesen elemezni és értékelni.¹² **A relevancia megállapítása tulajdonképpen egy bináris szöveg osztályozási feladat, mely során a szózsákmodellre¹³ és klaszterhipotézisre¹⁴ (Tikk [2007]) alapozva a SIMA a gazdasági és részvény szótárak szavainak előfordulásai alapján eldönti, hogy a tartalom releváns-e vagy sem.**

Amennyiben releváns a weboldal a vizsgált részvény vagy általános gazdasági szempontból, akkor a rendszer feldolgozza a teljes szöveget, és megállapítja az egyedi véleménymutatót az adott szempontra (O_s vagy O_g index) a pozitív és negatív szótárak szavainak (véleményszavak) előfordulásainak segítségével. Ezt követően a mutatót súlyozza az oldal egy úgynevezett prioritásmutatóval. A futás – vagyis a weboldalak feldolgozása – bármikor leállítható, és megtekinthetőek a kiválasztott részvényre vonatkozó eredmények. A futtatás elméletileg akár végtelen ideig folytatódhat.

¹² A begyűjtött linkeket nem jelöli meg a rendszer, hogy releváns oldalról származnak-e vagy sem, mivel a rendszer épít a nem tematikus kapcsolatokra, a portál jellegű oldalak hivatkozásaira.

¹³ A szózsákmodell lényege, hogy egy szöveget a benne előforduló szavak, kifejezések előfordulásai alapján vizsgálunk, mintha bedobnánk a szöveg a szavait egy zsákba, a szavak szövegen belüli sorrendje nem számít.

¹⁴ A klaszterhipotézis feltételezi, hogy a szövegek osztályozása során az azonos kategóriába tartozó dokumentumok hasonló szavakat tartalmaznak.



7. ábra: A SIMA folyamatmodellje

A következő fejezetek részletesen bemutatják az egyes funkciók működését.

6 A web bejárása

Az internetes keresőmotorokra jellemző tipikus crawler és webtérképkezelési funkciók (Tikk [2007]) a SIMA-ban egy modulban kerültek megvalósítása. A szoftver egy előre megadott URL halmazból (kiinduló weblapok) kezdi el a felderítést. Futás közben a LinksBuffer tárolóba begyűjtött URL címeit dolgozza fel a szoftver egymás után sorban, a talált új linkek pedig folyamatosan a lista végére kerülnek (FIFO feldolgozás).¹⁵

A fentiek mellett minden egyes új URL cím esetén lefut egy egyszerű úgynevezett **arató vagy csonkoló metódus** (Tikk [2007]), mely egy adott URL címet feldarabol a „/” jelek mentén visszafelé haladva, és az így képzett további URL címeket is beilleszti az adatbázisba. (A SIMA mindig csak az első „/” jelig csonkol, a csonkolás a már csonkolt címek feldolgozásakor folytatódik.)

A kiinduló weboldalak összeállítása manuálisan történt, a legismertebb magyar hírportálok és blogterek alapján. A következő táblázat mutatja a kiinduló weboldalak listáját és típusát:

Sorszám	Weboldal	Típus
1	origo.hu*	Hírportál
2	index.hu*	Hírportál
3	portfolio.hu	Gazdasági hírportál
4	webisztan.blog	Blogtér
5	www.startlap.hu	Portál
6	blog.hu*	Blogtér
7	blogger.hu	Blogtér
8	freeblog.hu*	Blogtér

¹⁵ Amennyiben a szoftver érvénytelen URL címen található tartalmat akar elemezni, naplózza az eseményt, és folytatja a futást az adatbázisban található következő URL címmel.

Sorszám	Weboldal	Típus
9	hirkereso.hu*	Hírportál
10	vilaggazdasag.hu	Gazdasági hírportál
11	fn.hu	Gazdasági hírportál

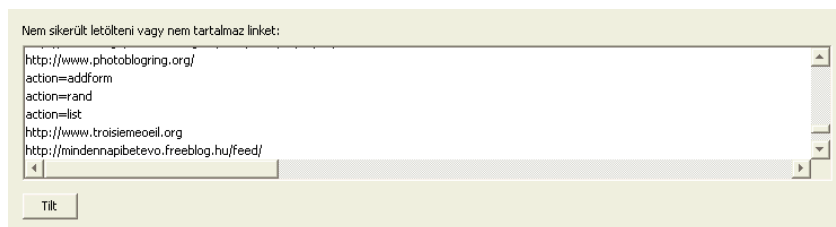
2. Táblázat: A kiinduló weboldalak listája

A táblázatban csillaggal van jelölve, ha a site szerepel a 2009-es leglátogatottabb weboldalak listájában.¹⁶

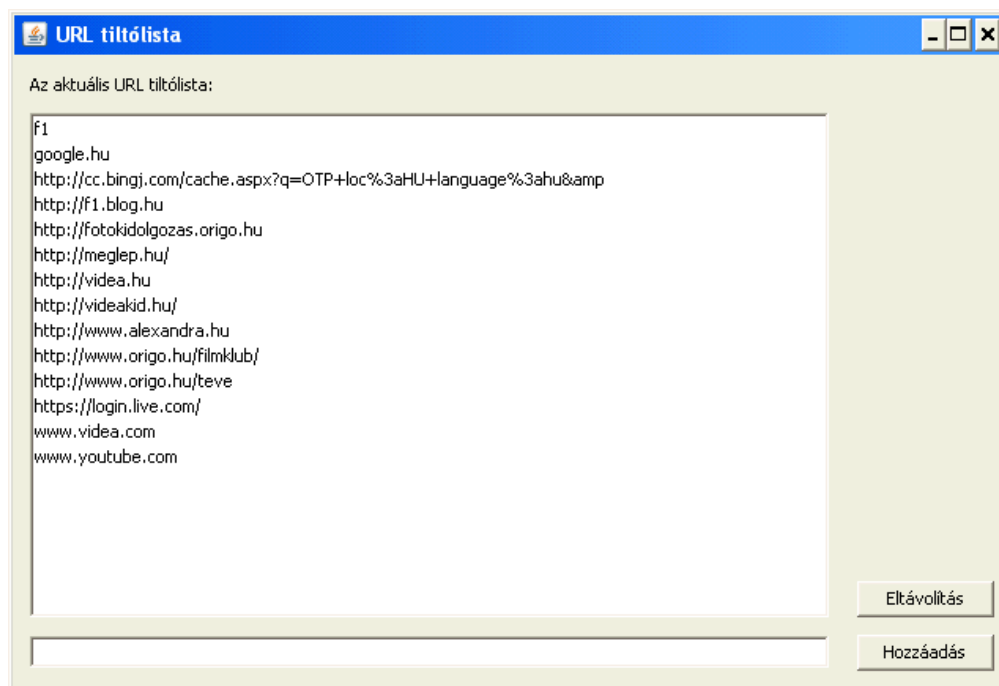
Mivel a szoftver futása bármikor megállítható, ha úgy gondoljuk, hogy elengedő weblapot dolgoztunk fel, ezért a bejárt weboldalak tartalmának elemzése párhuzamosan zajlik az adott oldalon található linkek feltárásával és kigyűjtésével. A webes információk gyors fluktuációja miatt minden futáskor újrakezdődik a bejárás és az elemzés.

A feldolgozási idő javítása érdekében a SIMA mielőtt elkezdené egy URL cím tartalmát letölteni (az eddigi tapasztalatok alapján a SIMA ezen művelete a legidőigényesebb), ellenőrzi, hogy az URL cím, mint sztring tartalmaz-e tiltólistabeli elemet. A tiltólista elemei lehetnek konkrét URL-ek, de szerepelnek bizonyos tiltószavak is: például sport, baba, film, stb. melyek gyakorlatilag kizárják, hogy egy weboldal tartalma releváns, így nincs is értelme letölteni a tartalmát. A SIMA minden egyes futás után összegyűjti a főfelületén azokat a web címeket, amelyeknek tartalmát nem sikerült letölteni illetve, amelyek nem tartalmaznak linkeket, és a 'Tilt' gomb megnyomásával hozzáadhatjuk a teljes címet vagy egy részét a tiltó listához, hogy a program a következő futás során ne is foglalkozzon az adott oldallal. A tiltólista funkció kihasználja, hogy a legtöbb URL címnek része a tartalom címe (ékezetes karakterek nélkül).

¹⁶ A <http://dotcomm.hu/blog/wp-content/uploads/2009/07/A-legl%C3%A1togatottabb-magyar-weboldalak.pdf> weboldal alapján, utoljára letöltve: 2011.03.19-én



8. ábra: Tiltólista javaslatok a főfelületen futás után



9. ábra: A SIMA URL tiltólistájának kezelőfelülete

A fejezet összefoglalásaként illetve a következő fejezetek bevezetéseként a webrobotok viselkedését befolyásoló tényezők (Tikk[2007]) az alábbi módon jelennek meg a SIMA-ban:

- **Kiválasztási elv** (*selection policy*): a letöltendő dokumentumok kiválasztásának szabályai. A SIMA-ban csak a kiinduló weboldalak manuális megadásával lehet befolyásolni a letöltendő oldalakat. Az URL címen található tartalom elemzése előtt azonban a SIMA eldönti a metaadatok és az URL cím alapján, hogy releváns-e tartalom, megéri-e részletesen elemezni a teljes tartalmat.
- **Újralátogatási elv** (*re-visit policy*): meghatározza az indexfrissítési paramétereket. A SIMA minden egyes futtatáskor „előlről kezd” a bejárást és az elemzést is.

- **Udvariassági elv** (*politeness policy*): a webszervereken okozható többletterhelést szabályozza. A SIMA esetében nem releváns: nem okoz nagyobb terhelést, mint egy emberi látogató.
- **Párhuzamos feldolgozási elv** (*parallelization policy*): a párhuzamosan dolgozó robotok összehangolása. A SIMA-nál jelenleg nem releváns.

A következő fejezetek a weboldalak tartalmának elemzési módját ismertetik.

7 A letöltött dokumentumok szövegbányászati elemzése

A bejárás során letöltött szövegeket két szempont szerint elemzi a rendszer:

- A szöveg alapján megállapítható vélemény az adott részvényről (cégről)
- A szövegből származtatható általános gazdasági vélemény

A SIMA értékelő modellje abból indul ki, hogy az általános gazdasági hangulat körülbelül ugyanolyan mértékben befolyásolja a kisbefektetők döntéseit, mint az adott céggel kapcsolatos információk.

A vélemények tömörítésére és számszerűsítésére képzett mutatók:

- **SSIM** = $O_s * R * P$, ahol
 - **SSIM** (*Single Stock Investor Mood*): az adott tartalomra számított egyedi részvény véleménymutató
 - **O_s**: az adott webes tartalom (írójának) véleménye a részvényről (lásd bővebben: 7.1.3 fejezet)
 - **R**: az oldal relevanciája (lásd bővebben: 7.1.2 fejezet)
 - **P**: az oldal prioritása (lásd bővebben: 7.1.4 fejezet)
- **SGIM** = $O_g * R * P$, ahol
 - **SGIM** (*Single General Investor Mood*): általános gazdasági véleménymutató
 - **O_g**: az adott webes tartalom (írójának) általános gazdasági/befektetési hangulata (lásd bővebben: 7.1.3 fejezet)
 - **R**: az oldal relevanciája (lásd bővebben: 7.1.2 fejezet)
 - **P**: Az oldal prioritása (lásd bővebben: 7.1.4 fejezet)

Egy releváns tartalom értékelése után a kapott gazdasági és részvénytmutató értékek átlagát (SIM és GIM), valamint az AIM mutatót frissíti a rendszer a vonatkozó releváns weboldalak számának arányában. (A mutatórendszert a 4. ábra mutatja be az 5. fejezetben.)

7.1 Szövegelemzési szempontok modellezése

A fentieknek megfelelően négyféle szótárat használ a szoftver:

- **Célrészvény szótárak:** a rendszer által vizsgálandó részvényekhez köthető szavak, jellemzően tulajdonnevek. (Például Telekom esetén: Telekom, T-Home, Magyar Telekom)
- **Általános gazdasági szótár:** olyan szavak halmaza, melyek előfordulása az adott cikk gazdasági témájára utal: gazdaság, üzlet, pénz, árfolyam, bank, stb.
- **Pozitív (gazdasági) szótár:** pozitív illetve pozitív gazdasági szavak, és azok szinonimái, valamint a negatív szótár szavainak a „nem” szó hozzáadásával képzett ellentétjei (Például: veszteséges -> nem veszteséges).
- **Negatív (gazdasági) szótár:** negatív illetve negatív gazdasági szavak, és azok szinonimái valamint a pozitív szótár szavainak a „nem” szó hozzáadásával képzett ellentétjei (Például: nyereséges -> nem nyereséges).

Bizonyos szótáraknak vannak közös elemei: például általános gazdasági és a pozitív illetve negatív szótáraknak vannak közös szavaik, azonban a pozitív és negatív szótárak diszjunktak. A célrészvény szótár jellemzően csak az adott részvény nevét, a kibocsátó cég nevét, illetve adott esetben anyavállalat/leányvállalat valamint főbb termékek/márkák nevét tartalmazza, ennek megfelelően a részvény szótárak diszjunktak a többi szótárral.

7.1.1 A SIMA szótárainak kezelése

Az egyes szótárak összeállításának alapjai a felhasználó által manuálisan megadható/módosítható - szakértői - alapszavak listája, melyek alapján egy online szinonima szótár (<http://www.szinonimaszotar.hu>) segítségével kibővíti a szoftver a szótáradatbázisát, majd a felesleges szavak eltávolíthatóak. Külső webes szolgáltatások igénybevétele jellemző az *opinion-mining* illetve *text-mining* területre, Berend és Farkas [2009] például a Wikipédiát, mint tudásbázist használta fel egy címkézési feladat során.

A szótárak kezelése egy inicializáló feladatnak tekinthető, a megfelelő szótárak összeállítása után nem szükséges a módosításuk. A frissítést egy a szoftver menürendszeréből elérhető funkció valósítja meg.

A szótárak frissítésekor módosíthatjuk az alapszavak listáját, majd lefut a SIMA szinonimakereső metódusa minden tárolt szóra, és kigyűjti a szavak szinonimáit. Minden talált szó kap egy prioritás értéket, mely rendre 1-gyel nagyobb a kiinduló szó prioritásánál.

(Az alapszavak prioritása 1.) Így megadható, hogy hány szintig keressen a SIMA szinonimákat. (Például egy 6-os szintű frissítés már igen sok felesleges szót eredményezhet a szótárba, valamint figyelembe kell venni, hogy a szótárak növekedésével exponenciálisan romlik a rendszer feldolgozási sebessége.) Az automatikus szótárbővítés után manuálisan eltávolíthatóak a mégsem szükséges szavak.

A szótárak szavai a SIMA adatbázis Lexicon táblájában tárolódnak. A tábla attribútumai:

- **ID:** a rekord egyedi azonosítója
- **LEXICON:** a szótár azonosítója
- **WORD:** a szó maga
- **STOCKWORD:** jelzi, hogy az adott szó részvénnel kapcsolatos szó-e
- **WORDPRIORITY:** a szó fontossága, lehetséges értékei: egész számok (1 jelenti a legfontosabb kategóriát)

Szótárak kezelése			
Szótár	Szó	Részvényező?	Prioritás
gazdaság	konjunktúra	☐	1
gazdaság	gazdaság	☐	1
pozitív	jó	☐	1
negatív	zuhan	☐	1
negatív	recesszió	☐	1
negatív	rossz	☐	1
pozitív	pozitív	☐	1
pozitív	siker	☐	1
pozitív	konjunktúra	☐	1
gazdaság	profit	☐	1
gazdaság	bevétel	☐	1
gazdaság	üzleti eredmény	☐	1
gazdaság	pénz	☐	1
gazdaság	üzlet	☐	1
gazdaság	bank	☐	1
OTP	Országos Takarékp...	☒	1
OTP	OTP	☒	1
TELEKOM	t-systems	☒	1
TELEKOM	t-home	☒	1
TELEKOM	MATÁV	☒	1
TELEKOM	telekom	☒	1
pozitív	dübörög	☐	1
pozitív	szárnyal	☐	1
gazdaság	felhasználó	☐	1
gazdaság	vásárló	☐	1

Aktuális szótár statisztikák

Alapszavak száma: 365

Összes ismert szó száma: 365

Gazdasági szótár: 176

Pozitív szótár: 86

Negatív szótár: 86

Részvényezőszótárak:

OTP: 3

TELEKOM: 7

MOL: 4

RICHTER: 3

Szinonimaszint: 0

Szótár frissítése

10. ábra: A SIMA szótár modulja

A szótárak tartalma alapján a korábban említett szószakmodellre és klaszterhipotézis elveire építve elemzi a SIMA a szövegek szóhalmazait és meta információit.

7.1.2 Relevanciamutató

Relevancia esetében a SIMA az úgynevezett lekérdezéstől függő – azaz témaspecifikus - jellemzőit vizsgálja a szövegnek (Tikk [2007]). A szöveg relevanciamutatója (R) diszkrét,

bináris változó, értéke 0 vagy 1 lehet. (0, ha nem releváns a szöveg a vizsgálat szempontjából, és 1, ha releváns.) A szoftver nem skálázza tovább a relevancia fogalmát, ehelyett egy másik mutatóval, a weboldal prioritásával súlyozza az adott szöveg által számszerűsített véleményét (lásd: 7.1.4 fejezet).

Ebben a fázisban a szoftver csak a html oldal head részében található adatokat, valamint az URL címet vizsgálja. A relevancia megállapítása egyben egyfajta tartalomszűrő a teljes szöveg feldolgozása előtt.

A dokumentum relevanciáját meghatározó faktorok:¹⁷

- A dokumentum keletkezési ideje: jó esetben meghatározható a meta tagek elemzésével, minél régebbi a dokumentum, annál kevésbé releváns. (Jelenleg nem része a SIMA működésének)
- A gazdasági/célrészcím szótárak szavai (továbbiakban témaszavak) mekkora arányát teszik a teljes szószámnak
- A témaszavak mekkora aránya szerepel a szövegben

Az azonosított faktorok alapján a következő relevanciakorlátokat használja a SIMA:

1. Szerepel témaszó az URL címben
2. Szerepel témaszó az oldal címében

HTML kód az azonosításhoz:

```
<title>...</title>
```

3. Szerepel témaszó a kulcsszavak között

HTML kód az azonosításhoz:

```
<meta name="keywords" content="..." />
```

4. Szerepel témaszó a leírásban

HTML kód az azonosításhoz:

¹⁷ A weboldal linkjét tartalmazó kiinduló weboldalak relevanciája nem számít a relevancia meghatározásakor. A weboldalak keletkezési idejének elemzése jelenleg szintén nem része a relevancia meghatározásának, mivel gyakorlati tapasztalatok alapján a feldolgozott releváns weboldalak kb. 1%-a volt időben nem aktuális, ugyanakkor a megfelelő metaadat hiánya esetén a tartalom korának meghatározása jelentősen növeli a futási időt.

<meta name="description" content="..." />

Bár blogok esetében nem mindig vannak a metaadatok definiálva az informális jellegük miatt, elemzésükkor a szöveg címének és az URL címnek a vizsgálatával kellően jó eredményt kapunk.

7.1.3 Véleménymutató

Liu [2010] szerint a vélemény a következő öt változóval fejezhető ki:

- o_j : **célobjektum**: célrészvény vagy általános gazdaság
- f_{jk} : **a célobjektum egy tulajdonsága vagy komponense**: a SIMA-ban a gazdasági és a célrészvény szótárak szavainak felel meg.
- so_{ijkl} : **a vélemény értéke**: a SIMA mutatót számol a pozitív és negatív szótárak szavainak előfordulásai alapján.
- h_i : **a célobjektum tulajdonságának vagy komponensének vélemény alkotója**: a SIMA-ban nem kezelt.
- t_i : **vélemény ideje**: a SIMA-ban nem kezelt.

A tartalomról következő befektetői hangulatot csak akkor elemzi a SIMA, ha releváns az adott szöveg, hogy csökkentse a felesleges feldolgozások számát.

Egy releváns szöveg véleménymutatójának számításakor két dolgot dönt el a SIMA: hogy mire vonatkozik a véleménymutató (célrészvényre vagy általános gazdasági hangulatra) illetve, hogy mennyi a véleménymutató értéke. Amennyiben egy releváns tartalom célrészvényre vonatkozik, akkor a szoftver nem vizsgálja általános gazdasági hangulatra a szöveget. Egy tartalom akkor vonatkozik egy célrészvényre, ha a részvénytársaság legalább 1 eleme szerepel a szövegben. Mivel a részvénytársaságok tulajdonnevek, jelenlétük egyértelműen jelzi, hogy a szöveg köthető a részvényhez. Amennyiben a relevánsnak ítélt tartalomban nincs részvénytársaság, úgy a megállapított véleménymutató az általános gazdasági hangulatra fog vonatkozni.

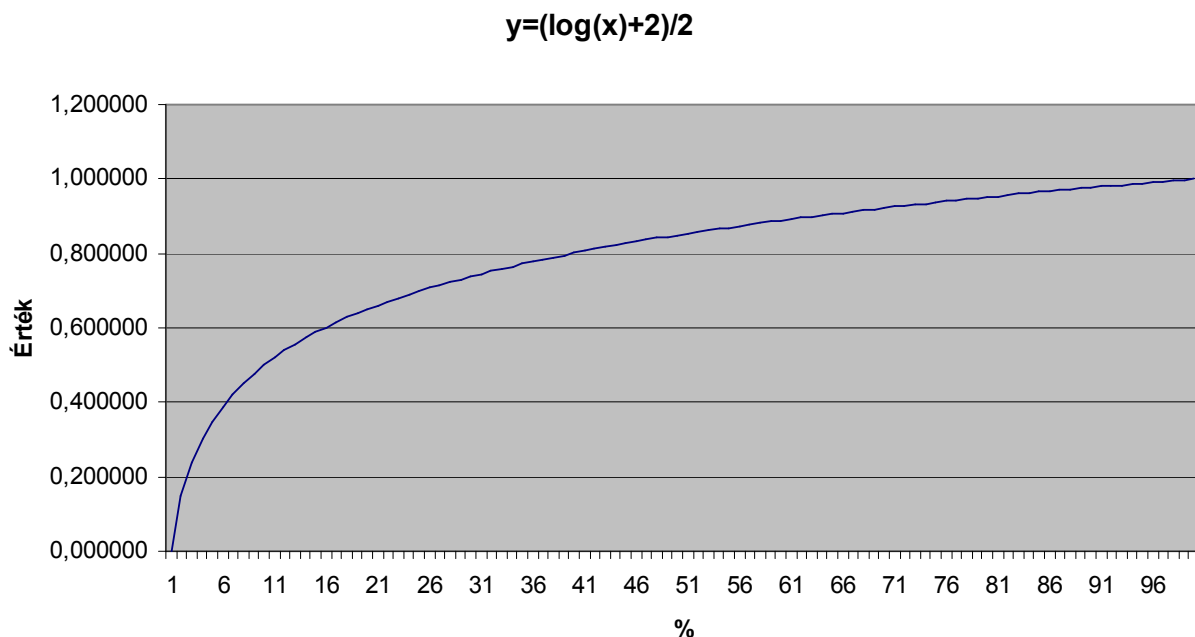
A véleménymutató megállapítása jelenti a tényleges szövegelemzést, mely során mindig egy dokumentumot vizsgálunk, az eredmény pedig egy -1 és 1 közé eső mutató. Más kutatások jellemzően nem egyetlen folytonos intervallumon számszerűsítik a véleményt, például Baccianella, Esuli és Sebastiani [2010] SentiWordNet 3.0 megoldása külön pontozza az adott tartalom pozitív, negatív és neutrális voltát.

A SIMA a következőképpen számolja a mutatót:

1. Megszámolja, hogy a pozitív szótár szavainak hány százaléka szerepel a szövegben.
2. A kapott százalékos értéket (0 és 1 közé eső számot), behelyettesíti a következő képletbe: $(\log(x) + 2) / 2$
3. Megszámolja, hogy a negatív szótár szavainak hány százaléka szerepel a szövegben.
4. A kapott százalékos értéket (0 és 1 közé eső számot), behelyettesíti a következő képletbe: $(\log(y) + 2) / 2$
5. Kiszámolja a véleménymutatót a két érték különbségeként:

$$O = [(\log(x) + 2) / 2] - [(\log(y) + 2) / 2]$$

A számítások során használt függvény jellegzetessége, hogy már alacsony 2%-os találati arány esetén is viszonylag nagy értéket, 0,15-öt ad eredményül, míg 30%-os találat fölött viszonylag alacsony a különbség az egyes értékek között.



11. ábra: A vélemények megállapításához használt függvény

Az elemzés során a html fejlécekben (<hx> </hx> tagek között) megtalált szavak duplán számítanak, ugyanakkor egy vélemény szó megtalálásakor a szoftver azonnal ellenőrzi, hogy szerepel-e előtte a „nem” szó.

A vélemény meghatározásának eredménye tehát egy O_s vagy O_g -1 és 1 közé eső index, attól függően, hogy részvényvéleményt vagy általános befektetési hangulatot számolt a program.

A SIMA a vélemény és a prioritás meghatározásakor jelenleg a teljes html forráskódot elemzi, azonban a tényleges és lényeges html tartalom megtalálásának vannak érdekes alternatívái, Hsu és Weninger [2008] például egy forráskód sor szöveges karaktereinek és tagjeinek arányából (*Text-to-Tag Ratio* - *TTR*) állapítják meg, hogy a forráskód, mely része tartalmazza a szövegfeldolgozás szempontjából legfontosabb információkat.

7.1.4 A tartalom prioritása

A tartalom prioritása (P) egy 1 és 2 közé eső súly, melynek célja, hogy a végeredményben (AIM indexben) megjelenjen a tartalom „szakmaisága”. A SIMA feltételezi, hogy minél több gazdasági szótárszó szerepel egy szövegben, az annál „szakmaibb”, fontosabb. P számítása hasonló O-hoz, a gazdasági szótár szavainak előfordulási arányát a következő képletbe helyettesíti:

$$P = 1 + (\log(x) + 2) / 2$$

Ezúttal is a html fejlécekben megtalált szavak duplán számítanak. A prioritás és a véleménymutató számítása párhuzamosan történik. (Mivel P maximum értéke 2, matematikai elméleti szempontból az AIM index 1 fölé is emelkedhetne, azonban könnyen belátható, hogy az alkalmazott modellel még 0,4 fölé is csekély eséllyel emelkednek az indexek, mivel a szövegek szavainak viszonylag kis arányát teszik ki a szótárszavak, tekintve.)

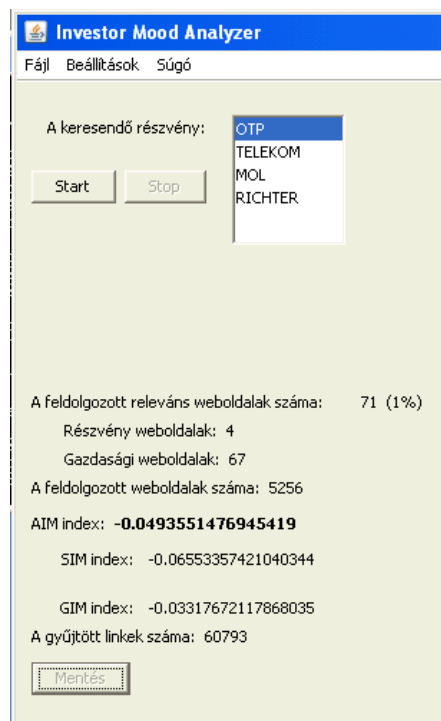
7.2 Az elemzés kimenete

Minden egyes átvizsgált szöveg után a szoftver frissíti a SIM (átlagos SSIM) vagy GIM (átlagos SGIM) mutató értékét. Majd a kettő - nem súlyozott! - átlaga alapján az AIM indexet.¹⁸

¹⁸ Az AIM index számításakor nem súlyozza a szoftver, hogy hány weboldalt elemzett a SIM illetve GIM index számításához, a két mutató egyenértékűnek számít a SIMA modelljében.

A futtatás eredményeit a program az adatbázis Results táblájában tárolja részvényenként:

- **DATE:** A futás dátuma
- **STOCK:** A vizsgált részvény
- **AIM:** aggregált kis befektetői hangulat index
- **SIM:** a részvényre vonatkozó átlagos hangulat index
- **GIM:** az általános gazdasági hangulatra vonatkozó index
- **SIM_COUNTER:** A SIM kiszámításához talált releváns weboldalak száma
- **GIM_COUNTER:** A GIM kiszámításához talált releváns weboldalak száma



12. ábra: A SIMA futtatásának eredményei

8 A kutatás bemutatása

8.1 A vizsgálat menete

Jelen fejezet célja, hogy ismertesse a 10 kereskedési napot érintő kezdeti vizsgálat menetét, mely során minden nap lefutott a SIMA elemezve az OTP részvénnel kapcsolatos webes véleményeket. A vizsgálatban érintett napokon körülbelül este 8 és 10 órakor indult el, és másnap reggel 7 óra körül lett le állítva a futás. Egy futás során jellemzően 35.000-40.000 weboldalt elemzett a SIMA, 200-300 releváns weboldalt talált és körülbelül 250.000 linket gyűjtött össze.¹⁹ A vizsgálat célja, hogy elemezze a számított értékek, és futás befejezésének napján történt OTP részvényárfolyam változás közötti kapcsolatot. Jelenleg a futás befejezésével egyező napi árfolyamváltozást vizsgálom, azonban kellően nagy megvizsgált nap után érdemes lehet megvizsgálni, hogy az információkban levő vélemény esetleg csak néhány nap elteltével épül be igazán az árfolyamba, mivel például Bollen, Mao és Zeng [2010] arra jutottak a Twitter vélemények és a Dow Jones ipari átlag közötti összefüggések vizsgálata során, hogy bejegyzésekben található hangulat 3-4 nappal később realizálódik az indexben.

¹⁹ A futási teljesítményt erősen befolyásolta a rendelkezésre álló lassú internet kapcsolat (max. 15 mbit/s) illetve, hogy a SIMA egy 1,66 GHz-es PC lett futtatva.

8.2 A futtatások eredményeinek elemzése

8.2.1 Az eddigi eredmények összefoglalása

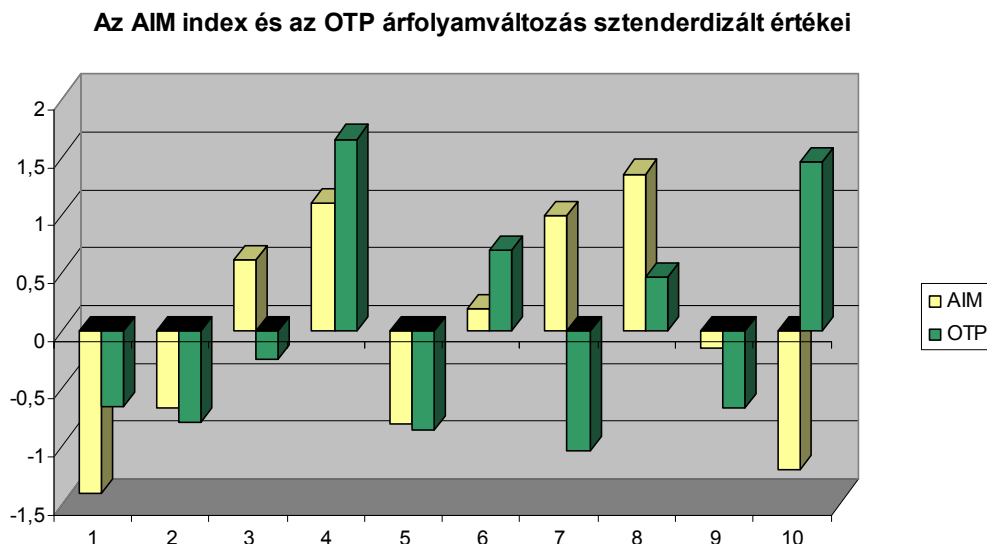
Az alábbi ábra összefoglalja a futások és az OTP árfolyamváltozását a vizsgált időszakra:

1. hét	OTP	N/A	-77	-93	-32	176
	AIM	N/A	-0,099693	-0,076407	-0,036195	-0,020772
		2011.03.07	2011.03.08	2011.03.09	2011.03.10	2011.03.11
		H	K	Sz	Cs	P
2. hét	OTP	N/A	N/A	-100	71	-120
	AIM	N/A	N/A	-0,080586	-0,049355	-0,024195
		2011.03.14	2011.03.15	2011.03.16	2011.03.17	2011.03.18
		H	K	Sz	Cs	P
3. hét	OTP	45	-79	N/A	155	N/A
	AIM	-0,012747	-0,060224	N/A	-0,093016	N/A
		2011.03.21	2011.03.22	2011.03.24	2011.03.24	2011.03.25

13. ábra: A futtatások eredményei

Az ábrán heti naptári bontásban láthatóak az értékek, a felső sorokban az OTP részvények adott napi árfolyamváltozása, míg az alsókban az adott nap reggelén befejezett 8-10 órás futtatások által számolt AIM index értékek 6 tizedesjegy pontossággal.²⁰ A SIMA a kísérlet során 10 esetből 4-szer tévedett az árfolyamváltozás irányát illetően. (Megvizsgálandó a jelenség, hogy minden futásra negatív érték született az AIM indexre, elképzelhető, hogy módosítani kell a modellen.)

²⁰ A mezők színezése jelzi az adott érték kategóriáját: minél sötétebb kék a háttér, annál pozitívabb az eredmény, minél barnább, annál negatívabb az érték.



14. ábra: Az OTP árfolyamváltozás és az AIM index sztenderdizált értékeinek összehasonlítása

Ha a változók sztenderdizált²¹ értékeit vizualizáljuk, akkor látható, hogy a SIMA a 10 futásból kétszer, a 7. (2011.03.18-i) és a 10. (2011.03.24-i) futás során tévedett „nagyot”, a többi esetben közel esik (3. futtatás) vagy megegyezik a két érték előjele.

8.2.2 Az eredmények statisztikai vizsgálata

Jelen fejezet célja, hogy néhány ismert statisztikai kapcsolati mutató kiszámításával értelmezze az AIM mutató és az árfolyamváltozás közötti kapcsolatot.

Kovariancia

A kovariancia az együttingadozást összefoglaló mérőszám. Az értéke a két változó közötti kapcsolat tényét és irányát képes azonosítani (Hunyadi-Vita [2002]). A SIMA esetén $C=0,652175$ vagyis a mutató értelmében pozitív kapcsolat van a két változó között.

Determinációs együttható

A korrelációs kapcsolat elemzésére determinációs együttható jobban megfelel, mint a korrelációs együttható,²² mivel ez százalékos formában is értelmezhető: azt mutatja, hogy

²¹ A sztenderdizálás lényege, hogy két különböző mértékegységű változó előfordulásait össze lehessen hasonlítani. A sztenderdizálás során az adott értékből kivonjuk a változó átlagértékét, majd elosztjuk mindezt a változó értékeinek szórásával.

az AIM indexek ismerete hány százalékkal csökkenti az OTP részvényárfolyam változás becslésekor elkövetett hibát, ha a becslés lineáris analitikus regressziófüggvény segítségével történik. A SIMA esetében $r^2=3,6\%$, tehát az eddigi megfigyelések alapján az AIM index ismerete 3,6%-kal csökkenti az árfolyamváltozással kapcsolatos bizonytalanságot.

A fentiek alapján a két változó között pozitív sztochasztikus kapcsolat van, és gyenge a lineáris korrelációs kapcsolat. Az eredmények nem tekinthetők reprezentatívnak az alacsony mintaadat mennyiség miatt, de érdemes folytatni a megfigyeléseket, és újraszámolni a kapcsolati mutatókat egy nagyobb időintervallumon, illetve javítani a modellt.

²² A determinációs együttható a korrelációs együttható négyzete.

9 Összefoglalás

Az eddigi kutatás bebizonyította, hogy érdemes tovább folytatni a szoftver fejlesztését/finomítását, javítani az értékelő és előrejelző képességét.

A szoftver fő erősségeinek és ismert hiányosságainak összefoglalása:

Erősség	Hiányosság
▪ Egyedi véleményekre nem túl érzékeny a számítási modell ▪ A szótárak egyszerű kezelhetősége ▪ Az eredmények könnyű és egyértelmű értelmezése ▪ Tiltólisták alkalmazása ▪ Konkrét részvény és általános gazdasági vélemények egyenrangúsága	▪ Toldalékolás problémájának kezelése ▪ A SIMA nem tud PDF dokumentumokat megnyitni és elemezni ▪ A feldolgozás sebessége jelenleg lassúnak mondható ▪ A program statikus, nem tanul automatikusan a korábbi futások tapasztalataiból (szótárbővítés, URL tiltás)

3. Táblázat: A SIMA erősségei és hiányosságai

A szoftver legfontosabb azonosított fejlesztési lehetőségei/irányai:

- A szótár szavain szótövező és toldalékoló algoritmusok alkalmazása
- Korábbi vizsgálatok/futások eredményeinek felhasználása a teljesítmény javításához
- A modellben szavazó bizottság létrehozása a relevancia-, vélemény- és prioritásindexek meghatározására Bayes valószínűségekre épülő számítási modell illetve neurális hálók alkalmazásával
- Az online tőzsdéző szörfözésének modellezése a linkek szövegének illetve a linkek kontextusának elemzésével. Az így érintett weblapok tartamának szövegelemzése.

További fejlesztési javaslatok:

- Kliens-szerver architektúra kialakítása, hogy párhuzamosan több gép bevonásával lehessen végezni az elemzést

- Feldolgozási idő javítása további szűrésekkel, rendezésekkel, Db és sztring műveletek minimalizálásával, teljesítménytesztekkel
- Nyelvhelyességi rugalmasság biztosítása elírások helyes kezelése a tartalmakban
- Angol nyelvű cikkek is tudjon értelmezni: a szótárak szavainak lefordítása egy online szótár segítségével,
- Kiinduló weboldalak módosítása egy a célközönség internet szörfözési szokásairól készült kérdőíves felmérés alapján.
- A relevánsnak talált weboldalakat automatikus kivonatolásával részvényjelentések generálása.

Az egyes jövőbeli fejlesztéseknél mindenképpen figyelembe kell venni, hogy az értékelési/osztályozási hatékonyság növelése mennyivel növeli a futási időt, hiszen az átlagolt értékeknek köszönhetően a rendszer nem érzékeny az egyedi értékekre, és sok esetben a gyors kezdeti becslés többet ér 99%-os pontosságú, de lassú becslésnél...

10 Melléklet

A következő táblázat tartalmazza részvényárfolyam változásokat és a SIMA által számított értékeket (a számított index értékek 6 tizedesjegy pontosságúak):

Dátum	SIM	GIM	AIM	Árfolyamváltozás
2011.03.08.	-0.140828	-0.058558	-0.099693	-77 Ft
2011.03.09.	-0.103117	-0.049698	-0.076407	-93 Ft
2011.03.10.	-0.041873	-0.030517	-0.036195	-32 Ft
2011.03.11.	-0.002040	-0.039504	-0.020772	176 Ft
2011.03.16.	-0.194559	0.033385	-0.080586	-100 Ft
2011.03.17.	-0.065533	-0.033176	-0.049355	71 Ft
2011.03.18.	-0.022042	-0.026349	-0.024195	-120 Ft
2011.03.21.	-0.027634	0.002139	-0.012747	45 Ft
2011.03.22.	-0.043590	-0.076857	-0.060224	-79 Ft
2011.03.24.	-0.089367	-0.096664	-0.093016	155 Ft

4. Táblázat: A kutatás eredményeinek összefoglaló táblázata

11 Felhasznált források

- Baccianella, S., Esuli, A. és Sebastiani, F. [2010]:** SENTIWORDNET 3.0: An enhanced lexical resource for sentiment analysis and opinion mining
- Berend G. és 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.
- Berend és Farkas [2009]:** A Wikipédia felhasználása az absztrakt címkézési feladatban, VI. Magyar Számítógépes Nyelvészeti Konferencia - MSZNY 2009.
- Bollen, J., Mao, H. és Zeng, X. [2010]:** Twitter mood predicts the stock market
- Hunyadi L. és Vita L. [2002]:** Statisztika közgazdászoknak, kiadta: Központi Statisztikai Hivatal, Budapest
- Hsu, W. H. és Weninger T. [2008]:** Text extraction from the web via Text-to-Tag Ratio
- Liu, B. [2006]:** Opinion Mining and Search, Google meghívásra tartott előadás, Pittsburgh, Sept 29, 2006
- Liu, B. [2010]:** Opinion Mining and Sentiment Analysis: NLP Meets Social Sciences. Invited Workshop on Social Theory and Social Computing workshopon tartott előadás, Honolulu, Hawaii, 2010.05.22-23.
- MTI [2011]:** Tartósan nőtt a hazai magánbefektetők aktivitása a BÉT-en, letölthető: http://www.mfor.hu/cikkek/Tartosan_nott_a_hazai_maganbefektetok_aktivitasa_a_BET_en.html weboldaltól, utoljára letöltve: 2011.03.19-én
- Pang B. és Lee L. [2008]:** Opinion-mining and 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
- Tikk, D. [2007]:** Szövegbányászat, Typotex Elektronikus Kiadó Kft, Budapest