Trend Miner

Large-scale Cross-lingual Trend Mining Summarization of Real-time Media Streams

TrendMiner
(Politikai témájú SM üzenetek (szociál)pszichológiai vizsgálata)

Miháltz Márton

NYELVTUDOMÁNYI INTÉZET
MAGYAR TUDOMÁNYOS AKADÉMIA

Hungarian Natural Language Processing Meetup, 2014. szeptember 25.

TrendMiner Projekt (FP7)

- Közösségi média streamek valós idejű, nagyléptékű trendfigyelése és összegzése
- 2011-2014:
  - DFKI (de), U. Sheffield, U. Southampton (uk), OntoText (bg), Sora (at), Internet Memory (fr), Eurokleis (it)
- 2013-2014:
  - MTA NYTI (hu), U. Madrid, Daedalus (es), IPIPAN (pl)

TrendMiner 2011-2014

- Twitter streamek valós idejű követése
  - Felhőalapú architektúra; nem tárol minden üzenetet
- Entitások felismerése és ontológiához kapcsolása, több nyelven
  - „ Wikification” / egyértelműsítés (Dbpedia)
- Összegzés (summarization)
  - Többnyelvű spektrális klaszterezés, reprezentáns választása
- Trendfigyelés
  - Gépi tanulás: SM üzenetek + valós idősorozatok modellezése
- Vizualizáció (UI)
- Use Cases
  - Pénzügyi folyamatok
  - Politikai események
  - Gyógyszerek és hatóanyagok a SM-ban

Trend Miner

TrendMiner 2013-2014 (MTA NYTI)

- Milyen **(szociál)pszichológiai** jelenségek, trendek figyelhetők meg a **magyar** Fb hozzászólók **politikai** témákra reagáló üzeneteiben?
- Politikusok, politikai szerveztek publikus **Facebook** posztjaira érkezett publikus **kommentek** letöltése és elemzése
  - Alapszintű **NLP** (tokenizálás, PoS, szótövesítés) -- domain-adaptáció
  - **Entitások**: politikai szereplők (személyek, pártok, szervezetek)
  - Sentiment (**érzelempolaritás**)
  - **Szociálpszichológiai** dimenziók: *közösségiség-ágencia, individualizmus-kollektivizmus, optimizmus-pesszimizmus, elsődleges -másodlagos gondolkodási folyamatok*
- Együttműködés
  - MTA TTK Kognitív Idegtudományi és Pszichológiai Intézet Narratív Pszichológiai Kutatócsoport

Adatok

- Jelenleg: **1.9M komment** 141K poszthoz
  - 2013.10.01 – 2014.09.22 időszak
  - Facebook Graph API, folyamatos letöltés
- **1344 fb** oldalról
  - Szervezetek: politikai pártok, tagszervezeteik
  - Személyek: országgyűlési képviselők és –jelöltek (OEVK)
  - Hivatalos és nem hivatalos oldalak
- 3 kategóriában
  - Magyar országgyűlés 2010-2014
  - Magyar országgyűlés 2014-2018 + országgyűlési választások 2014
  - Magyar EP képviselők 2014-2019 + magyar EP választások 2014
- Források (entitások): valasztas.hu, wikipedia.hu

Adatbázis

- Fb források (oldalak)
  - Oldal címe, tárgy típusa és neve (személy vagy párt), affiliációja (jelölés & tagság: hu2010, hu2014, eu2014)
- Fb posztok, kommentek
  - Oldal id, felhasználó id, időpont
- Annotációk
  - Entitások, pszichológiai tartalmak a kommentekben
  - Komment id, mondat- és tokenpozíció, annotált szöveg, szöveg szótövesített fejjel, annotációs címke
- Pontértékek
  - Minden kommenthez: komment hossza + 14 mutató (sentiment, valency_pos/neg, agency_pos/neg, …)
- SQL adatbázis

Feldolgozó pipeline

- Letöltés fb Graph API-val, DB-ba töltés
- **Tokenizálás** (huntoken)
- **PoS**-taggelés (hunmorph)
- **Morfológiai** elemzés (hunmorph)
- Szótő és morf. elemzés **egyértelműsítése**
- **Entitások** azonosítása, **tartalomelemzés** (NooJ)
- Annotációk, kiszámított mutatók DB-ban tárolása

Trend Miner

Domain-adaptáció

- A meglévő NLP eszközöket eltérő szövegdomainre fejlesztették ki
  - „sztenderd” nyelv (hírek)
  - Közösségi média szövegeken rosszul teljesítenek
- Adaptáció előkészítése korpuszvizsgálattal
  - 1.25M fb komment, 29M token
  - 2.25M ismeretlen token (694K típus)
  - Gyakorisági lista: f > 15 manuális átnézése
  - Gyakori problématípusok, releváns és gyakori ismeretlen szavak kigyűjtése stb.

Trend Miner

Domainadaptáció: tokenizálás

- Huntoken eszköz
- Gyakori problémák
  - **Hiányzó szóközök** írásjelek után
    Első mondat vége.Következő mondat eleje
  - Többször **ismételt írásjelek**
    első rész……. második rész
    => [„első”, „rész.”, „…”, „…”, „második”, „rész”]
  - **URL-ek** felbontása
  - **Emotikonok** felbontása
    :D => [„:”, „D”]
  - Nagy **számoknál** ezres csoportok felbontása
    125 000 => [„125”, „000”]
  - Ismeretlen **rövidítések**

Domainadaptáció: szótövesítés/PoS

- Hunpos + hunmorph + egyértelműsítő szkript
- Gyakori problémák
  - Gyakori és fontos i**smeretlen** szavak (nincs szótő, elemzés): hozzáadás az elemzőhöz analóg, ismert szavakkal
  - Ismeretlen **rövidítések**, **betűszavak** hozzáadása (normalizált alakkal)
  - Gyakori **elírt** szóalakok: javítás helyes alakra
  - Nem sztenderd **kisbetű-/nagybetűhasználat** pl. *CSUPA NAGYBETŰS MONDATOK*
  - Hiányzó **ékezetek**
  - Ismeretlen **szleng** szavak
  - Szóvégi mássalhangzók **többszörözése** pl. *pfffffffffff, uffffffffff, ejjjjjjjjjjj*
  - Ismeretlen **emotikonok**

NooJ

- **NooJ** (GUI): véges állapotú automaták fejlesztése és futtatása szövegeken
  - Lexikai, morfológiai és/vagy szintaktikai elemzés, konkordanciák
  - Annotációk importja és exportja
  - .NET, Java
  - Letölthető: http://nooj4nlp.net
- **NooJ-cmd**:
  - Nyílt forrású **Java NooJ** API-ra épül
  - Minden NooJ GUI feature => parancssori megfelelő
  - Szabadon hozzáférhető: https://github.com/tkb-/nooj-cmd

Trend Miner

TrendMiner NooJ tartalomelemzés

- NooJ nyelvtanok (automaták) annotációhoz:
  - **Szereplők** (entitások)
  - **Érzelmi** valencia (sentiment)
  - **Regresszív** képzeleti szótár
  - **Közösségiség-ágencia**
  - **Optimizmus-pesszimizmus**
  - **Individualizmus-kollektivizmus**

NooJ nyelvtanok fejlesztése

- Együttműködés MTA TTK PI **szociálpszichológus** kutatóival
  - Pólya Tibor, Fülöp Éva, Csertő István, Kővágó Pál
- Fejlesztői korpusz
  - 176K minta fb komment 570 fb oldalról (4.9M token)
  - NLP annotáció
  - Gyakorisági listák: szótő, szótő+szófaj, szót ő+morfológiai elemzés stb.

Trend Miner

1. Politikai szereplők (NE-k)

- Maxent NER eszköz (huntag): alacsony teljesítmény ezen a domain-en
  - Híroldalak szövegein tanult (sztenderd nyelv)
  - Kategóriahibák, hamis pozitív entity-k, entityhatár -felismerési problémák stb.
- NooJ lexikon és nyelvtan
  - **Személynevek**:
    családnév (+ utónév_szótő), becenevek
  - **Szervezetek nevei**:
    Hivatalos név, rövidített/betűszavas változat, becenevek
  - Automatikusan DB-ból + kézzel (korpuszban gyakori becenevek, névváltozatok)

Trend Miner

2. Érzelmi valencia

- Érzelmek **pozitív** vagy **negatív** polaritással
  - Főnevek, melléknevek, igék, határozószók, emotikonok, többszavas kifejezések
  - 500 pozitív, 420 negatív elem
  - Kontextusfüggő polaritás: pl. **negáció** felismerése egyszerű szabályokkal
- Fejlesztése:
  - f > 100 tartalmas szavak a fejlesztői korpuszból (3500 típus)
  - 6 független annotátor: pozitív, negatív, semleges
  - >= 4 annotátor egyetért: végső ellenőrzés és döntés
  - Lexikonok, szabályok: NooJ nyelvtanba szerkesztés

Trend Miner

3. Regresszív képzeleti szótár

- Martindale (1975, 1990): szövegben tükröződő pszichológiai folyamatok felfedése
- 2 szint a gondolkodási folyamatokban:
  - **Elsődleges**: asszociatív, konkrét, a realitáshoz kevésbé kapcsolódó (fantázia, ábrándozás, álmok)
  - **Másodlagos**: absztrakt, logikus, realitásközpontú és problémamegoldásra fókuszáló
- 29+14 további kategória (közösségi viselkedés, megismerés, érzékelés, érzelmek stb.)
- Magyar változat: Pólya--Szász 2013
- 3000+ kifejezés

Trend Miner

4. Közösségiség-ágencia

- 2 alapvető dimenzió a **társas értékelésben**:
  - **Közösségiség**: az egyén **másokhoz/csoporthoz** való viszonyát jellemzi
    - **morális** szempontból (pl. *együttműködés, becsületesség, hűség, őszinteség, önfeláldozás)*
    - **érzelmi** szempontból (pl. *barátságosság, szeretet, ragaszkodás, tisztelet*)
  - **Ágencia**: az egyént a **célkövető viselkedés hatékonysága** szempontjából jellemzi
    - **motiváció** (pl. *ambiciózus, elszánt, céltudatos, akarat*)
    - **kompetencia** (pl. *intelligens, ügyes, ravasz, szakértelem*)
    - **kontroll** (pl. *önérvényesítő, sikeres, győztes, hatalom*)
- Pozitív és negatív értékek mindkét dimenzióban
  - Köntextusfüggő lehet (pl. tagadás)
- Fejlesztés:
  - f > 100 tartalmas szavak a fejlesztői korpuszból (3500 típus)
  - 3+3 annotátor: minden szó: +/0/- közösségiség/ágencia szempontból
  - 7. annotátor: egyetértés <100% => végső ítélet
- 640 elem

Trend Miner

5. Optimizmus-pesszimizmus

- Események **idejének** szerepe az egyéni gondolkodásban
  - **Múlt** dominál: a személy megváltoztathatatlanak gondolja a világot
  - **Jelen** dominál: reálisan megvalósítható feladatok fontossága
  - **Jövő** dominál: nyitott lehetőségek megjelenése
- Szófaji, morfológiai elemzésre + időkifejezések felismerésére alapul
- 2 mutató:
  1. | jövő idejű igealakok | / | múlt idejű igealakok |
  2. | jelen vagy jövő idejű igealakok | / | múlt idejű igealakok |
- Mindkettő: minél magasabb, annál magasabb szintű optimizmus

Trend Miner

6. Individualizmus-kollektivizmus

- **Individualizmus**: mennyire fontos a személy kategóriája a világról való gondolkodásban
- **Személyes névmások** használata:
  - **Gyakori**: a személy kategóriája van az előtérben (fontos), individualizmus szintje **magas**
  - **Ritkább**: a személy kategóriája a háttérben (környezet az előtérben), individualizmus szintje **alacsonyabb**
- Szófaji, morfológiai elemzésre alapul
- 1 mutató:
  |személyes névmások| /
  (|személyragos igealakok| +
  |birtokos személyragos főnévi alakok|)
- Magasabb érték magasabb szintű individualizmusra utal

Trend Miner

Trendminer Politikai Ontológia

- OWL politikai témájú ontológia
- Fogalmi osztályok, tulajdonságok, axiómák
- Lengyel, magyar, osztrák adatok
- Individuumok: személyek, pártok, események (választások stb.), jelölések stb.
  - 1300 magyar politikus és párt + kapcsolataik
  - 2010, 2014 magyar és 2014 EP választások
- Szabadon hozzáférhető lesz

Jelenlegi munka: kiértékelés

- Pszichológiai annotációk pontosságának mérése
  - Minta komment korpusz: 500 komment (10K szó), minden forrásból
  - Humán annotáció (gold standard) 2+1
- NLP eszközök domain-adaptációjának hatékonysága
  - Ismeretlen szavak arányának csökkenése

Trend Miner

Jelenlegi munka: vizsgálatok

- Ingroup-outgroup
  - Saját csoport vs. másik csoport (párt, párttömb)
  - kikről és hogyan folyik a diskurzus
- Időbeli trendek
  - Választási időszakok, választások
- Korreláció közvélemény-kutatások eredményeivel
- Vizualizációs felület (OntoText)
  - Sentiment, entitások, kulcsszavak
  - Grafikonok, táblázatok, együtt-előfordulási mátrixok, térkép
  - Topikfelhők
- **Open Data**
  - **Eszközök** egy része **szabadon** hozzáférhető lesz (github)
  - **Adatbázis** (2013-2014, 1.9M komment+annotációk): **szabadon** hozzáférhető!
  - Elemző **partnereket** szívesen látunk!

Trend Miner

Köszönöm a figyelmet!