

Dziś u MO Kura i Miś, czyli o analizie polskich tweetów politycznych

Maciej Ogrodniczuk

Arkadiusz Cacko

Mateusz Kopeć

Aleksander Wawer

Krzysztof Kasianiuk

Instytut Podstaw Informatyki
Polskiej Akademii Nauk

Collegium Civitas

Seminarium ZIL, 24.11.2014

O projekcie TrendMiner

Metryczka projektu:

- Tytuł: Large-scale, Cross-lingual Trend Mining and Summarisation of Real-time Media Streams
- Finansowany z programu FP7 oraz wsparcia ministerialnego
- Daty: 2011–2014, z polskim udziałem w ostatnim roku
- Szef projektu: Thierry Declerck (DFKI)

O projekcie TrendMiner

Pozostali partnerzy:

- **University of Sheffield** — UK (Kalina Bontcheva)
 - **Ontotext AD** — Bułgaria (Atanas Kiryakov, Alex Simov)
 - **University of Southampton** — UK (Mahensan Niranjana)
 - **Internet Memory Research** — Francja (France Lafarges)
 - **Eurokleis** — Włochy (Francesco Bellini)
 - **SORA GmbH** — Austria (Christoph Hofinger, Paul Ringler)
 - **Hardik Fintrade** — India (Suresh Aswani)
-
- **Universidad Carlos III de Madrid** (Paloma Martínez)
 - **RILHAS** — Węgry (Tamás Váradi)
 - **IPI PAN**
 - **DAEDALUS SA** — Hiszpania (José Luis Martínez)

O projekcie TrendMiner

Misja:

Opracowanie innowacyjnych, replikowalnych, dostępnych na zasadzie open-source metod wielojęzycznego wydobywania informacji i agregacji dużych ilości danych z mediów strumieniowych przy użyciu zaawansowanych metod lingwistycznych z dziedziny przetwarzania tekstu, wnioskowania, uczenia maszynowego, ekonomii i nauk politycznych.

Misja polska:

- opracowanie ontologii polskiej sceny politycznej
- zestawienie infrastruktury lingwistycznej do przetwarzania tweetów w języku polskim na potrzeby analizy politycznej.

Budowa polskiej ontologii politycznej

Cel:

Uzyskanie możliwie pełnej wiedzy na temat uczestników życia politycznego w Polsce w szerokim zakresie czasowym.

Kryteria doboru danych:

- Do zbioru zostały włączone dane, które dotyczą sfery politycznej, definiowanej instytucjonalnie.
- Podstawowy porządek instytucjonalny w Polsce określa Konstytucja.
- Trzy rodzaje władzy – ustawodawcza, wykonawcza i sądownicza.
- Dodatkowe instytucje (np. NBP, NIK, KRRiTV).
- Zalety: porządek instytucjonalny jest w Polsce dość stabilny – luki w danych mogą być łatwo zidentyfikowane.

Budowa polskiej ontologii politycznej

Efekt:

Dzięki kryterium instytucjonalnemu wiemy:

- na ile zbiór danych jest i będzie pełny – wystarczy śledzić działalność głównych instytucji (obecnie przede wszystkim rządu),
- wiemy, kto pełnił jakie funkcje w różnych instytucjach,
- wiemy, kto pełnił określone instytucje w różnych momentach czasu.

Budowa polskiej ontologii politycznej

Źródła:

Historyczne (1989–2009):

- Baza danych pozyskana z polskiego Sejmu (III-VI kadencji)
- Opracowanie Kalety „Ludzie władzy” („People in Power”) – 1989–2009

Obecne (2009–):

- Biuletyn Informacji Publicznej
- Monitor Polski (dziennik rządowy)
- Strony Sejmu, Senatu i innych instytucji z różnych gałęzi władzy
- Inne (e.g. Wikipedia, DBPedia, doniesienia prasowe, życiorysy w plikach PDF)

Budowa polskiej ontologii politycznej

Kategorie zebranych danych:

- Imię, drugie imię, nazwisko
- Data urodzenia
- Początek i koniec sprawowania funkcji
(Ministrowie – daty dzienne, pozostali – miesięczne)
- Oficjalna nazwa funkcji
- Nazwa instytucji
- Afiliacja polityczna (członkostwo w partii lub partia rekomendująca na stanowisko)
- Źródło informacji

Znajdowanie kont Twittera polityków

Metody poszukiwań oparte o:

- imię i nazwisko
- analizę list obserwujących
- analizę lista obserwowanych
- obserwowanie kont partii politycznych.

Problem ograniczeń API Twittera :(

Budowa ontologii

Ontologia:

- próba przetworzenia zbioru danych na wiedzę
- ustrukturyzowany sposób przedstawiania wiedzy, zawierający zbiór pojęć i relacje między nimi
- specyfikacja: **OWL** (składnia XML, oparty o RDF)
- na podstawie:

<http://www.dfki.de/lt/onto/political.owl>

Podstawowe definicje

Klasa

```
<owl:Class rdf:about="http://www.dfki.de/lt
/onto/political.owl#PoliticalFunction">
  <rdfs:label xml:lang="de">politische Funktion
  </rdfs:label>
  <rdfs:label xml:lang="en">political function
  </rdfs:label>
  <rdfs:label xml:lang="pl">funkcja polityczna
  </rdfs:label>
</owl:Class>
```

Podrzędność

```
<owl:Class rdf:about="http://www.dfki.de/lt
/onto/political.owl#Executive">
  <rdfs:label xml:lang="pl">władza wykonawcza
  </rdfs:label>
  <rdfs:subClassOf rdf:resource="http://www.dfki.de
/lt/onto/political.owl#PoliticalFunction"/>
</owl:Class>
```

Struktura

- Struktura klas jako odwzorowanie obiektowości (dziedziczenie).
- Każdy obiekt może być powiązany z innym za pomocą różnych relacji.



DatatypeProperty

```
<owl:DatatypeProperty rdf:about="http://www.dfki.de
/lt/onto/political.owl#firstName">
  <rdfs:domain rdf:resource="http://www.dfki.de
    /lt/onto/political.owl#Person"/>
  <rdfs:range rdf:resource="http://www.w3.org/2001
    /XMLSchema#string"/>
</owl:DatatypeProperty>
```

ObjectProperty

```
<owl:ObjectProperty rdf:about="http://www.dfki.de
/lt/onto/political.owl#worksFor">
  <rdfs:domain rdf:resource="http://www.dfki.de/lt
    /onto/political.owl#PoliticalActor"/>
  <rdfs:range rdf:resource="http://www.dfki.de/lt
    /onto/political.owl#PoliticalActor"/>
</owl:ObjectProperty>
```

Przykład osoby:

```
<owl:NamedIndividual rdf:about="http://.../onto/
    political.owl#Donald_Franciszek_Tusk__Sejm4">
  <rdf:type rdf:resource="http://.../political.owl
    #MemberOfParliament"/>
  <endReason rdf:datatype="http://...XMLSchema
    #string"></endReason>
  <dateFrom rdf:datatype="http://...XMLSchema
    #dateTime">2001-10-19</dateFrom>
  <isLowerHouse rdf:datatype="http://.../XMLSchema
    #boolean">true</isLowerHouse>
  <occursWith rdf:resource="http://.../onto
    /political.owl#Election_2001_10_19"/>
</owl:NamedIndividual>
```

Łączenie:

```
<owl:NamedIndividual rdf:about="http://political.owl
    #Donald_Franciszek_Tusk">
  <rdf:type rdf:resource="http://political.owl
    #Politician"/>
  <firstName rdf:datatype="http://XMLSchema
    #string">Donald</firstName>
  <hasFunction rdf:resource="http://political.owl
    #Donald_Franciszek_Tusk__Sejm4"/>
  <hasFunction rdf:resource="http://political.owl
    #Donald_Franciszek_Tusk__Sejm5"/>
</owl:NamedIndividual>
```

Przykład partii

```
<owl:NamedIndividual
  rdf:about="http:political.owl#AWS__Sejm3">
  <rdf:type rdf:resource="http:.../political.owl
    #PartySuborganization"/>
  <nickname rdf:datatype="http://www.w3.org/2001
    /XMLSchema#string">AWS</nickname>
  <occursWith rdf:resource="http:.../political.owl
    #Election_1997_10_20"/>
  <name rdf:datatype="http://www.../XMLSchema
    #string">AWS</name>
  <hasMember rdf:resource="http:.../political.owl
    #Adam_Andrzej_Słomka__Sejm3"/>
  <hasMember rdf:resource="http:.../political.owl
    #Adam_Biela__Sejm3"/>
  ...
</owl:NamedIndividual>
```

W podobny sposób zamodelowana praca na rzecz urzędów, wszystkie kadencje, elekcje, przeszłe i terażniejsze funkcje itp.

Przykładowe zapytania

Lista partii w obecnej kadencji Sejmu:

```
party and hasOrganization
  some ('party faction' and hasMember
    some ('member of parliament'
      and isLowerHouse value true
      and occursWith
        some (Election and current value true)))
```

Lista senatorów z Platformy:

```
hasFunction
  some (inverse hasMember
    some (inverse hasOrganization value PO
      and occursWith
        some (Election and current value true)
      and hasMember
        some (isLowerHouse value false)))
```

Garść statystyk

Zebrailiśmy:

- kilkaset kont Twittera polityków
- dane polityków 8 kadencji (1989–2014)
- w sumie ponad 7000 osobo-funkcji
- dane o ponad 40 partiach
- wszystko zawarte w jednym pliku jako ontologia

Twitter



twitter.com

- Twitter (skrót: TT) to darmowy serwis społecznościowy, służący wymianie krótkich wiadomości (tweetów)
- Powstał w 2006 roku
- Ma ponad 271 milionów aktywnych użytkowników

Twitter



twitter.com

- Twitter (skrót: TT) to darmowy serwis społecznościowy, służący wymianie krótkich wiadomości (tweetów)
- Powstał w 2006 roku
- Ma ponad 271 milionów aktywnych użytkowników
- Oficjalne logo Twittera zostało nazwane Larry the Bird, na cześć słynnego koszykarza
- 7 listopada 2013 Twitter zadebiutował na NYSE



Co to jest tweet?

- Zapis chwili lub myśli
- Może zawierać tekst, linki do zdjęć lub filmów
- Może zawierać wzmianki o innych użytkownikach
- Może zawierać określenia tematu tweeta (hashtagi)
- Limit długości to 140 znaki (Normalization Form C)
- Na tweeta można odpowiedzieć, można go podać dalej albo dodać do ulubionych



Ellen DeGeneres

@TheEllenShow



Follow

If only Bradley's arm was longer. Best photo ever. [#oscars](#)
pic.twitter.com/C9U5N0tGap

Reply Retweet Favorite More

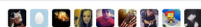


RETWEETS

3,429,266

FAVORITES

2,020,229



7:06 PM - 2 Mar 2014

Flag media

Co to jest oś czasu?

- Jako użytkownik Twittera wybieramy użytkowników, których chcemy obserwować
- Wszystkie tweety tych użytkowników (posortowane chronologicznie) tworzą naszą oś czasu
- Dodatkowo w osi czasu znajdują się tweety zawierające wzmianki o nas

Najliczniej obserwowani:

Użytkownik	Obserwujący
KATY PERRY	57 370 126
Justin Bieber	54 824 536
Barack Obama	46 977 046
YouTube	45 245 639
Taylor Swift	44 188 987

Co to jest oś czasu?

- Jako użytkownik Twittera wybieramy użytkowników, których chcemy obserwować
- Wszystkie tweety tych użytkowników (posortowane chronologicznie) tworzą naszą oś czasu
- Dodatkowo w osi czasu znajdują się tweety zawierające wzmianki o nas

Najliczniej obserwowani:

Użytkownik	Obserwujący
KATY PERRY	57 370 126
Justin Bieber	54 824 536
Barack Obama	46 977 046
YouTube	45 245 639
Taylor Swift	44 188 987

Tweety publiczne i prywatne

Domyślnie wszystkie nasze tweety są publicznie widoczne. Można to zmienić w ustawieniach – wtedy gdy ktoś nas zechce obserwować, musimy wyrazić na to zgodę.

API Twittera

- API – interfejs opisujący interakcję programów
- Twitter posiada dwa rodzaje API:
 - REST API – pobieranie danych na żądanie
 - Streaming API – odbieranie nowych tweetów na bieżąco
- Aby mieć możliwość programistycznego pobierania tweetów konieczne jest założenie specjalnego konta
- Każdy rodzaj pobierania danych jest ograniczony limitami
- Istnieją mniej ograniczone płatne formy dostępu do danych

API Twittera

- API – interfejs opisujący interakcję programów
- Twitter posiada dwa rodzaje API:
 - REST API – pobieranie danych na żądanie
 - Streaming API – odbieranie nowych tweetów na bieżąco
- Aby mieć możliwość programistycznego pobierania tweetów konieczne jest założenie specjalnego konta
- Każdy rodzaj pobierania danych jest ograniczony limitami
- Istnieją mniej ograniczone płatne formy dostępu do danych

Udostępnianie pobranych danych [14.11.2014]

<https://dev.twitter.com/terms/api-terms> 6b. If you provide Content to third parties, including downloadable datasets of Content or an API that returns Content, you will only distribute or allow download of Tweet IDs and/or User IDs.

Przykład danych zwracanych przez API

```
{
  "id" : 436780833140387840,
  "id_str" : "436780833140387840",
  "text" : "#FF dla @aronsem @jackislander @AzraelK @AGozdyra
           @AleksZawisza Polecam obserwacji i Pozdrawiam :)",
  "created_at" : "Fri Feb 21 08:34:00 +0000 2014",
  "geo" : null,
  "coordinates" : null,
  "place" : null,
  "lang" : "pl",
  "entities" : { ... },
  "user" : {
    "id" : 1001972246,
    "id_str" : "1001972246",
    "screen_name" : "MikiWrobelek",
    "name" : "Mikołaj Wróbelek",
    "lang" : "pl",
    "location" : "United States of Europe",
    ...
  },
  ...
}
```

Język polski na Twitterze

Pytania:

- czy da się go przetwarzać naszymi narzędziami lingwistycznymi (Morfeuszem? Nerfem? Spejdem?)
- czy internetowy śmietnik nie objawia się tam w pełnej okazałości?

Jest źle:

```
na 4fun 'chce 398 + pozdrowienia' aatw
bedzie numer 7138. :) ja wyslalam 3
aatw boze jak to mi sie kojarzy
z koncertem jusa
pierwszy raz wypilam kawe bez mleka.
o fuj, never again
```

Język polski na Twitterze

Jest gorzej niż źle:

chyba pojade na 'obietnice' alr dopiero pod koniec miesiaka

because i'm workiiiiing\n(clap along if you feel like masz jeszcze osiem godzin do przepracowania dzis)

awww ten jarajacy mnie i czywiscie wiele wiele innych directionerrek louis jako pilakarz takze na mikowym vimeo zamilska vs rss b0ys vs [mmd] <https://t.co/sgkb9kecso>

kazdy z bap ma insta, tylko dae to wyjatek ccc:

@aytaehyungay masz biasa sunggyu i biaske soonkyu, jak sie z tym czujesz?

Język polskiej polityki na Twitterze

A może u polityków jest lepiej?

Czy dr hab. N. Maliszewski też przewiduje dziś w sejmie sukces PEK? I czy już zarezerwował sobie tę tezę?

@LUKASZKIJEK ps. Takie pytania też tracą populizmem;)

@niepokornadusza Pan naprawdę UWAŻASZ, że prez. Putin potrzebuje mojego podjudzania, aby takim jak Pan i, ja przy okazji, sprawić lanie?

Język polskiej polityki na Twitterze

A może nie jest tak źle?

RT @rafsalak: W najnowszym @PressRedakcja "Warsztat socjal ninji" - świetny, przekrojowy materiał od @DRozanska. Lektura każdego community...

Dziś u MO Kura i Miś. Koń by się uśmieł.

Co my tu mamy?

- całkiem znośny język,
- różne problemy, które wydają się naprawialne (lepiej interpretowalne niż w przypadku ogólnym).

No to sprawdźmy systematyczniej:

W 3000 losowo wybranych tweetach znaleźliśmy dużo nierozpoznanych form:

- skróty: *FB, GW, JVR, NATO, €, prof, dzienn., TVP2*, ale nie: *F16* czy *ZET* w *Radio ZET*,
- diakryty: *mąż, odczytuje, pisałem, pomarzyć*, ale nie: *exposé, Müller*,
- emotikony
- ucięte słowa (na końcu tweetu)
- błędy pisowni: *swrdecznie, członkowstwo, gospodarkama*, także *okeeeeeeej*,
- język obcy: *community, Sammyego, I like it, Ora et labora, Assassin's Creed Identity, slitfocia*
- wielkość liter (utrudniająca rozpoznawanie nazw własnych).

Inne nierozpoznane formy

Dokładniejsza kategoryzacja form „innych”:

- neologizmy: *sorkokorki*, *Kopaczinho*,
- slang: *kminię*, *Bolandzie*, *Lemingradu*, *Łomatko* (często obecny w słownikach slangu)
- słowa nowe: *smartfon*, *audiobook*,
- słowa złożone: *homopropaganda*, *nadredaktor*,
- formy poprawne, ale rzadsze: *zmolestowanego*,
- formy o niestandardowej transkrypcji: *nie-by-wa-łe-go*,
- odmienione formy nazw własnych, szczególnie przymiotnikowe: *palikotowy*,
- formy zniekształcone dla zabiegu stylistycznego: *szłem Świętokrzysko*, *kwiaty pachły*, *a wiatr wiał*, *jeden na jedemu*.

Nierozpoznane formy

Statystyki:

Kategoria	% tweetów z tym problemem
Skróty	26,23
Diakryty	10,93
Emotikony	12,00
Ucięte słowa	6,03
Błędy pisowni	3,37
Język obcy	2,93
Wielkość liter	1,27
Inne	2,77
Dowolny problem	49,40

Co z tym robić?

Można:

- starać się rozwijać skróty,
- diakrytyzować,
- emotikonów pewnie nie ruszać (przydadzą się do analizy sentymentu),
- usuwać ucięte słowa z końca, próbować interpretować pozostałe,
- ew. poprawiać błędy pisowni — ale ostrożnie! dużo większy zysk osiągniemy z poprawy samych diakrytów,
- próbować interpretować słowa angielskie, ale to nietrywialny problem (odmiana?)
- używać słowników slangu, najnowszej polszczyzny, ...

Eksperyment z LT „z pudełka” (LT0)

Przegląd wyników

- *Chantal Biya nazywana jest Lady Gagą* → *Chanta Bila nazywana jest Lady Gągoł*
- *Baracka Obamy* → *Baranka Obawy*
- *pracow.* → *placów.*
- *Yorke* → *Córkę*
- *smartfonów* → *smart fonów*

Sprawdźmy więc warianty działające tylko na słowach, które:

- nie są w całości zapisane wersalikami
≈ poprawiaj wszędzie oprócz skrótów (LT1),
- nie rozpoczynają się wielką literą
≈ poprawiaj wszędzie oprócz nazw własnych (LT2).

Warianty korekty

Ewaluacja w porównaniu z TM (same diakryty)

	LT0	LT1	LT2	TM
Błędy niewykryte	126	164	268	184
Wykryte i poprawione	614	576	472	589
Błędnie poprawione	695	483	178	228

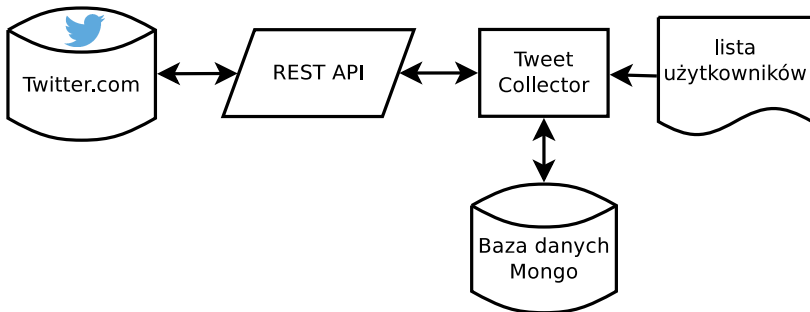
	LT0	LT1	LT2	TM
P	0,47	0,54	0,73	0,72
R	0,83	0,78	0,64	0,76
F1	0,60	0,64	0,68	0,74

Architektura systemu

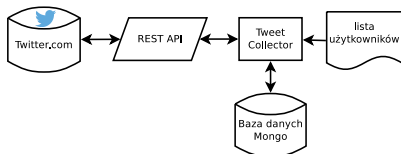
1. Zbieranie

2. Przetwarzanie

3. Wizualizacja



Zbieranie – Tweet Collector

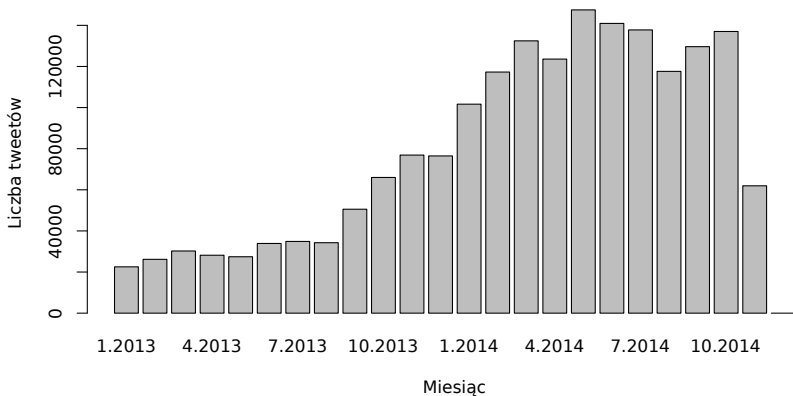


- Zbierane są wszystkie tweety napisane przez wybranych użytkowników Twittera
- Korzystamy z REST API, ponieważ mamy wtedy możliwość pobierania do 3200 tweetów historycznych
- Program zbierający jest uruchamiany cyklicznie, poszukując nowych tweetów do pobrania
- Limit 300 żądań/15 minut, maks. 200 tweetów/żądanie
- „Wybrakowane” tweety są odrzucane
- Dane zapisujemy w bazie MongoDB¹

¹<http://www.mongodb.org/>

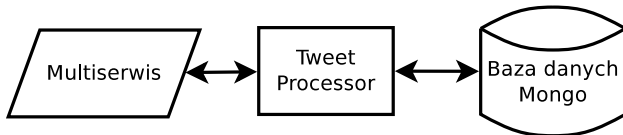
Zbieranie – Tweet Collector

- 2,055,635 tweetów, w tym 1,810,823 po polsku (88%)
- średnio ok. 128,500 tweetów miesięcznie w 2014 roku

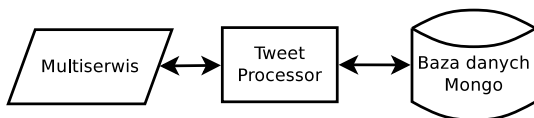


Architektura systemu

1. Zbieranie 2. Przetwarzanie 3. Wizualizacja



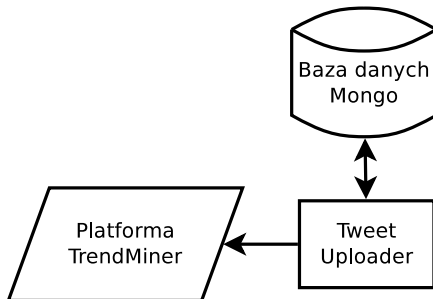
Przetwarzanie – Tweet Processor



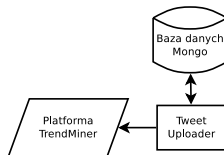
- Program przetwarzający jest uruchamiany cyklicznie, poszukując nowych tweetów do przetworzenia
- Tekst tweeta jest „czyszczony”, a następnie przetwarzany z użyciem efektywnych narzędzi przetwarzania języka:
 - Tagowanie i segmentacja: WCRFT + Polimorf
 - Wykrywanie nazw własnych: Nerf
 - Parsowanie powierzchniowe: Spejd
 - Wykrywanie wystąpień: MentionDetector
- Używamy tych narzędzi jako usługi udostępnianej przez Multiserwis: `glass.ipipan.waw.pl/multiservice`
- Wyniki przetwarzania wraz z dodatkową analizą są zapisywane w bazie danych

Architektura systemu

1. Zbieranie 2. Przetwarzanie 3. Wizualizacja



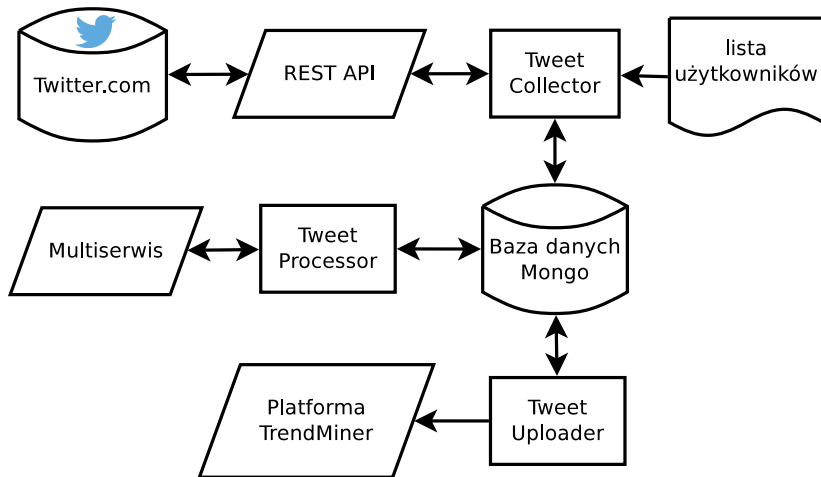
Wizualizacja – Tweet Uploader



- Program zbierający jest uruchamiany cyklicznie, poszukując nowych tweetów do wysłania
- Wizualizacja wykorzystuje następujące wyniki analizy:
 - wydźwięk tweeta jako wartość z przedziału $[-1, 1]$
 - miejsca wspomniane w tweecie (wraz ze współ. GPS)
 - tematy poruszane w tweecie
 - nazwy własne wymienione w tweecie
 - zlematyzowaną treść tweeta
- Partie i politycy wymienione w tweecie są łączone z ontologią

Architektura systemu

1. Zbieranie 2. Przetwarzanie 3. Wizualizacja



<https://github.com/mkopec87/trendminer-pl>

Szczegóły przetwarzania tweetów

Przetwarzanie każdego tweeta przebiega następująco:

- 1 Wykrywamy język, w jakim jest napisany. Jeśli nie jest to polski, tweet nie jest dalej przetwarzany
- 2 Treść tweeta jest czyszczona
- 3 Wyczyszczona treść jest przetwarzana pomocą narzędzi Multiserwisu
- 4 Wykrywane są nazwy własne w treści tweeta
- 5 Lokalizowane są miejsca poruszane w treści tweeta
- 6 Wykrywane są tematy poruszane w treści tweeta
- 7 Wykrywany jest wydźwięk tweeta

„Czyszczenie” treści tweeta

- 1 Normalizacja kodowania znaków spoza UTF-8
- 2 Usunięcie znacznika "podaj dalej":
RT @mkopec: Skandal! → Skandal!
- 3 Rozwinięcie hashtagów: #ProjektPraca w
#Bydgoszcz! → Projekt Praca w Bydgoszcz!
- 4 Rozwinięcie wzmianek o użytkownikach: @mkopec →
Mateusz Kopeć
- 5 Parsowanie znaczników HTML: > → >
- 6 Usunięcie linków i emaili
- 7 Normalizacja emotikon: :)))) → :)
- 8 Normalizację powtórzeń liter: jazdaaaaa → jazda
- 9 Usunięcie „przyciętych” słów: Czytaj więcej:
<http://marszalek.sejm.pl/ms...> → Czytaj
więcej:
- 10 Normalizacja białych znaków

„Czyszczenie” treści tweeta – diakrytyzacja

- ❶ Eksperymenty z użyciem narzędzia LANGUAGE TOOL nie przyniosły dobrych rezultatów – problemem jest nadmiarowe poprawianie słów
- ❷ W naszych danych główny problem pisowni to brakujące diakryty
- ❸ Zaimplementowaliśmy własny, prosty diakrytyzer
- ❹ Jego działanie jest następujące:
 - Tokeny są sprawdzane ze słownikiem POLIMORF
 - Nierozpoznane słowa mają dodane diakryty w celu stworzenia rozpoznanego słowa
 - W wypadku wielu możliwości, wykorzystywana jest frekwencja unigramowa NKJP

Wykrywanie nazw własnych

- 1 Podstawą są nazwy własne wykryte przez Nerfa
- 2 Odrzucamy nazwy własne wykryte < 10 razy
- 3 Odrzucamy zagnieżdżone imiona i nazwiska
- 4 Lematyzujemy pozostałe nazwy własne
- 5 Jeżeli zlematyzowaną nazwę możemy odnaleźć w ontologii, dodajemy stosowny odnośnik do obiektu w ontologii i pełną nazwę
 - Do ontologii jest dodana ręcznie stworzona lista alternatywnych nazw, przezwisk
 - Osoby są wyszukiwane także po samym imieniu/nazwisku
 - Niejednoznaczności rozstrzygamy frekwencją pełnej wersji nazwy

Wykrywanie miejsc

- 1 Podstawą są nazwy geograficzne, które rozpoznał Nerf
(`geogName`, `placeName`)
- 2 Odrzucamy nazwy własne wykryte < 10 razy
- 3 Lematyzujemy pozostałe nazwy własne
- 4 Wyszukujemy nazwy w bazie `geonames.org`
(pełna zgodność nazwy)
- 5 Zwracamy pierwszą lokalizację odnaniezoną w bazie
dla każdej z nazw, wraz ze współrzędnymi GPS

Wykrywanie tematów

- 1 Podstawą są wystąpienia zwrócone przez Mention Detectora
- 2 Odrzucamy wyrażenia wykryte < 10 razy
- 3 Odrzucamy wyrażenia, które zostały zidentyfikowane jako nazwy własne
- 4 Lematyzujemy pozostałe wyrażenia
- 5 Odrzucamy wyrażenia ze stoplisty

Lematyzacja korpusowa

- ❶ Wykrywanie miejsc, tematów i nazw własnych używa lematyzacji wyrażen wielowyrazowych
- ❷ Lematyzacja opiera się na dotychczasowej frekwencji form odmienionych, z uwzględnieniem liczby i przypadku pierwszego tokena (lub głowy skł., jeśli wykryta)
- ❸ Przykład: w tekście pojawia się wyrażenie: Naczelnej Rady Lekarskiej
 - ❶ Lematyzujemy słowo po słowie: naczelny rada lekarski
 - ❷ Sprawdzamy dotychczasowe odmiany z tym samym lematem słowo po słowie:
 - Naczelnej Rady Lekarskiej – 13 wystąpień, 0 wystąpień z głową w mianowniku i liczbie poj.
 - Naczelna Rada Lekarska – 6 wystąpień, 6 wystąpień z głową w mianowniku i liczbie poj.
 - ❸ Wybieramy lemat Naczelna Rada Lekarska

Wykrywanie wydźwięku

Wydźwięk tweeta to wartość rzeczywista:

$$S = \frac{1.5t_{pos} - t_{neg}}{t}$$

gdzie

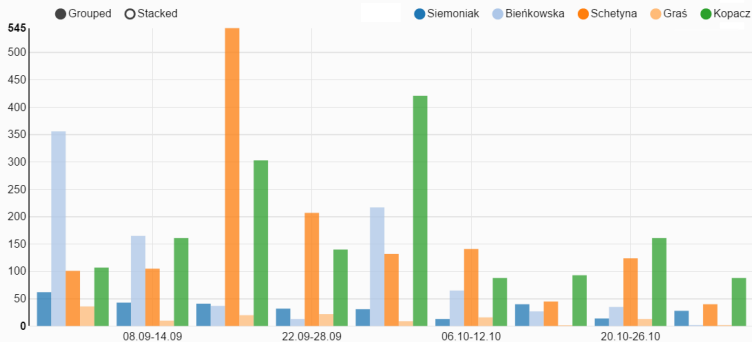
- t to liczba tokenów w treści tweeta
- t_{pos} to liczba tokenów o pozytywnym wydźwięku
- t_{neg} to liczba tokenów o negatywnym wydźwięku

Lista tokenów o określonym wydźwięku ma 3277 pozycji (formy podstawowe).

Przykład wizualizacji

Reakcje Twittera na wybór premiera:

Mentions of the entities for subperiods



Przykład wizualizacji

Reakcje Twittera na wybór premiera (komentarz):

- 30 sierpnia Donald Tusk został nominowany do objęcia stanowiska przewodniczącego RE,
- giełda nazwisk premierów in spe: Tomasz Siemoniak, Elżbieta Bieńkowska, Grzegorz Schetyna, Paweł Graś, Ewa Kopacz,
- kandydatura Bieńkowskiej, o najwyższej pozycji w pierwszym okresie, stała się bezprzedmiotowa w momencie ogłoszenia jej nominacji na komisarza KE (10 września),
- Schetyna powraca z politycznego niebytu i trafia do MSZ,
- 22 września Ewa Kopacz zostaje premierem RP.

Dalsze badania

Łańcuch lingwistyczny:

- Odmiana hashtagów i wzmianek o użytkownikach
- Próba zgadywania uciętych słów
- Rozwijanie skrótów
- Rozpoznawanie wyrażień slangowych
- Rozpoznawanie spolszczeń słów angielskich
- Rozwój algorytmów „czyszczenia”
- Normalizacja wielkości liter

Kierunki aplikacyjne:

- neologism watch na podstawie Twittera?
- wizualizacja „tematów dnia” („politycznych tematów dnia”)?

Dalsze badania

Przykładowe kierunki badań o charakterze politologicznym:

- zmiana polityczna: czy te same osoby są u władzy od 1989 roku?
- kariery polityczne: kto wspiął się, a kto spadł z drabiny politycznej?
- układy władzy: na ile zintegrowane są środowiska i grupy polityczne?

Co dalej?

Dalsze kierunki rozwoju:

- Rozbudowa bazy danych o nowe dane dotyczące innych instytucji (np. na poziomie województw i samorządów).
- Powiązanie:
 - wiedzy o relacjach pomiędzy jednostkami z procesem podejmowania decyzji (na etapie rządowym i legislacyjnym w parlamencie),
 - uzyskanej wiedzy z bazami tematów i treści medialnych,
 - polskich danych (np. medialnych) z danymi na poziomie Europejskim (w ramach UE).
- Budowa narzędzi filtrowania i prezentowania posiadanych danych dla naukowców i społeczeństwa.

Podziękowania

Opisane prace były sfinansowane przez projekt *TrendMiner*,
w którym udział wzięli:

Maciej Ogrodniczuk	IPI PAN (ZIL)
Mateusz Kopeć	
Aleksander Wawer	
Arkadiusz Cacko	IPI PAN (ZPSI)
Dariusz Czerski	
Paweł Łoziński	
Paweł Ziniewicz	
Krzysztof Kasianiuk	Collegium Civitas