



OŚRODEK PRZETWARZANIA INFORMACJI  
PAŃSTWOWY INSTYTUT BADAWCZY

# Eko-system BabelNet

Warszawa 05.04.16

# O produkcji

Babelnet (<http://babelnet.org/>) jest to wielojęzykowa encyklopedia oraz semantyczna sieć łącząca koncepty i nazwy własne w wielki graf relacji.

Składa się z co najmniej 14 mln elementów, zwanych babelnet synsetami. Każdy babel synset reprezentuje znaczenie i zawiera zbiór synonimów (babel sense), które wyrażają to znaczenie w szerokim zakresie różnych języków.

BabelNet pokrywa ponad 270 języków, i powstał w ramach automatycznej integracji: WordNeta, Wikipedii, OmegaWiki, Wiktionary, WikiData itd.

System rozwijany jest przez Uniwersytet Sapienza w Rzymie, oraz dysponuje różnymi interfejsami dostępowymi (m.in. JAVA API, HTTP Rest API, SPARQL).

# Zakres danych

**BabelNet 3.6** covers 271 languages and is obtained from the automatic integration of:

- **WordNet**, a popular computational lexicon of English (version 3.0).
- **Open Multilingual WordNet**, a collection of wordnets available in different languages (downloaded in August 2015).
- **Wikipedia**, the largest collaborative multilingual Web encyclopedia (November 2014 dump).
- **OmegaWiki**, a large collaborative multilingual dictionary (July 2015 dump).
- **Wiktionary**, a collaborative project to produce a free-content multilingual dictionary (August 2014 dump).
- **Wikidata**, a free knowledge base that can be read and edited by humans and machines alike (November 2014 dump).
- **Wikiquote**, a free online compendium of sourced quotations from notable people and creative works in every language (March 2015 dump).
- **VerbNet**, a Class-Based Verb Lexicon (version 3.2).
- **Microsoft Terminology**, a collection of terminologies that can be used to develop localized versions of applications (July 2015 dumps).
- **GeoNames**, a free geographical database covering all countries and containing over eight million placenames (April 2015 dump).
- **WoNeF**, an improved, expanded and evaluated automatic French translation of WordNet (high precision version, downloaded in August 2015).
- **ItalWordNet**, a lexical-semantic database developed in the framework of two different research projects: EuroWordNet and SI-TAL (downloaded in December 2015).
- **ImageNet**, an image database organized according to the WordNet hierarchy (2011 release).

# Zakres danych

- **Multilinguality**: the same concept is expressed in tens of languages
- **Coverage**: 271 languages and 14 million entries!
  - **6M** concepts and **7.7M** named entities
  - **745M** word senses
  - **380M** semantic relations (27 relations per concept on avg.)
  - **11M** images associated with concepts
  - **41M** textual definitions
  - **1.6M** concepts with domains associated



# BabelNet



BabelNet

[LOG IN](#) [REGISTER](#)

ENGLISH

TRANSLATE INTO...

SEARCH

[PREFERENCES](#)

All

Concepts

Named Entities



15 results

● Noun

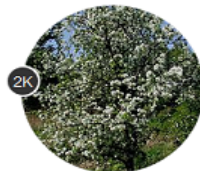
## Noun



apple, apple blossom, apple peel

Fruit with red or yellow or green skin and sweet to tart crisp whitish flesh

ID: 00005054n | Concept



apple, Malus pumila, orchard apple tree

Native Eurasian tree widely cultivated in many varieties for its firm rounded edible fruits

ID: 00005055n | Concept



Apple Inc., Apple (company)

Apple Inc. is an American multinational corporation headquartered in Cupertino, California, that designs, develops, and sells consumer electronics, computer software, online services, and personal computers.

ID: 03739345n | Named Entity



# BabelNet

English

Arabic

Chinese

French

German

Greek

Hebrew

Hindi

Italian

 all preferred languages

- Dictionary
- Images
- Translations
- Sources
- Categories
- Compounds
- Other forms
- External Links

 • bn:00005054n • NOUN • Concept • Categories: Apples, Malus, Plants described in 1803, Plants with sequenced genomes...

**EN** **apple**   • **apple blossom**  • **apple peel**  • **Apples**  • **pomiculture**

English

Fruit with red or yellow or green skin and sweet to tart crisp whitish flesh   More definitions

IS A: **edible fruit** • **pome** • **fruit** • **Fleshy fruit** PART OF: **apple** • **apple tree**

## EXPLORE NETWORK




## RELATED:

apple  
Fuji (apple)  
Cydonia oblonga  
McIntosh  
Malus sieversii  
fire blight  
Jonagold  
pomegranate  
Golden apple  
Malus  
pear

## Translations

**AR** التفاح, *Malus domestica*, أنواع التفاح, التفاح, التفاح شائع, تفاح مسكّن, تفاحة, تجرة تفاح, فواكه التفاح, *domestica malus*, التفاح الأخضر

**ZH** 苹果, 蘋果, *Malus domestica*, 大蘋果市, 泰, 蘋果樹, 苹果, 青苹果

**EN** apple, apple blossom, apple peel, Apples, pomiculture, Green Apples, American Delicious, An apple a day, Apple, Apple-blossom, Apple-blossoms, Apple-tree, Apple/Nutritional information, Apple blossoms, Apple Popularity, Apple production, Apple trees, Appleblossom, Appleblossoms, Apples and teachers, Culture of apple, Dried apple, Epli, Green Apple, *Malus communis*, *Malus domestica*, *Malus domesticus*, Nutritional information about the apple, *Pyrus malus*, World's Largest Apple,  

**FR** pomme, peau de pomme, Pommes, Pomiculture, Pomme de France, Pomme vert, Pommerale, *pomme verte*

**DE** Apfel, Apfelblüte, Apfelschale, green apple, *malus domestica*, *äpfel*

**EL** μήλο, Μηλιά, *malus domestica*, μήλο, πράσινο μήλο

**HE** תפוח, תפוח, *Malus*, גרנד אלכסנדר, גרני סמית, דלישס, ענה, תפוח עץ, *malus domestica*, תפוח ירוק, תפוחים

**HI** सेब, सेब, *malus domestica*, ग्रीन एप्पल

# BabelNet - JSON

## Response example

```
{
  "senses": [
    {
      "lemma": "Apple_System_on_Chips",
      "simpleLemma": "Apple_System_on_Chips",
      "source": "WIKIRED",
      "sensekey": "",
      "sensenum": 0,
      "frequency": 0,
      "position": 1,
      "language": "EN",
      "pos": "NOUN",
      "synsetID": {
        "id": "bn:14792761n",
        "pos": "NOUN",
        "source": "BABELNET"
      },
      "translationInfo": "",
      "pronunciations": {
        "audios": [],
        "transcriptions": []
      }
    },
    {
      "lemma": "Apple_System_on_a_Chip",
      "simpleLemma": "Apple_System_on_a_Chip",
      "source": "WIKIRED",
      "sensekey": "",
      "sensenum": 0,
      "frequency": 0,
```

# BabelNet Synset vs Sense

## **BabelSynset**

A **BabelSynset** is a set of multilingual lexicalizations (**BabelSenses**) that are synonymous expressions of a given concept or named entity. Each **BabelSynset** has its unique ID.

## **BabelSense**

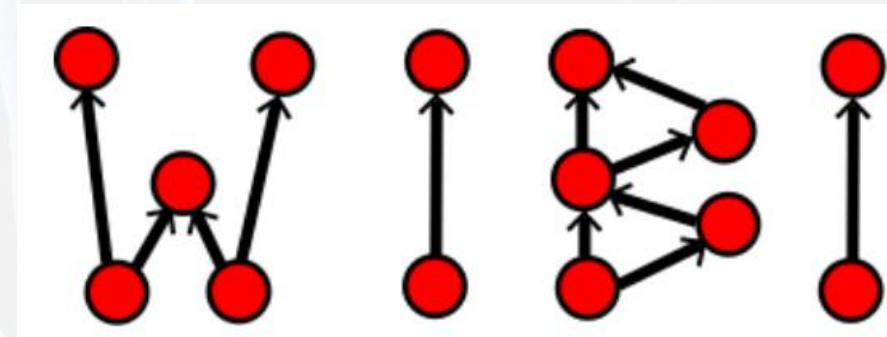
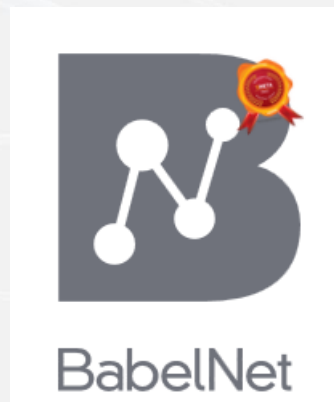
A **BabelSense** is a particular, language-specific lexicalization occurring in a given **BabelSynset**. Each **BabelSense** is tied to a particular source (WordNet, Wikipedia, Wiktionary, automatic translations, etc.).



# Pokłosie BabelNetu

W oparciu o Babelnet (<http://babelnet.org/>) powstały dwa narzędzia:

- 1) Babelfy – wielojęzykowy system ujednoznaczniania
- 2) Wikipedia Bi taksonomia – taksonomia stron Wikipedii dopasowana do taksonomii kategorii.



# Babelfy



Babelfy

this apple is very tasty

Enable partial matches: ☐

ENGLISH

[expanded view](#) | [compact view](#)

this

apple

is very

tasty



apple

Jabłko – jadalny,  
kulisty owoc jabłoni o  
soczystym i chrupkim  
mięszu, spożywany

tasty

Pleasing to the sense  
of taste

# Babelfy



Babelfy

Restauracja urzeka wdziękiem, zapachem, obsługa kelnerska najwyższy poziom, polecamy: gęś, królika i polędwice wołową

Enable partial matches: ☐

POLISH

[expanded view](#) | [compact view](#)

iom , polecamy: gęś , królika i polędwice wołową



cja – w  
wie część  
awarta  
zpośrednio



## gęś

Web-footed  
long-necked typically  
gregarious migratory  
aquatic birds usually



## królika

Królik – polska nazwa  
zwyczajowa kilku  
gatunków zajęczaków  
z rodziny



## polędwice

The portion of the loin  
(especially of beef)  
just in front of the  
rump

# Klasteryzacja wyników wyszukiwania z pomocą BabelNetu

# Wstęp

- Celem jest zweryfikować jak Babelnet/Babelfy pomaga w poprawie jakości grupowania krótkich tekstów.
- Badamy dobrze znane metody klasteryzacji krótkich tekstów wzbogacone semantyczną informacją z Babelnetu.
- Trzy poziome eksperymenty:
  - Porównanie trzech popularnych metod klasteryzacji wyników wyszukiwania
  - Ocena wpływu BabelNetu/Babelfy na jakość klasteryzacji
  - Weryfikacja idei klasteryzacji opartej na samym ujednoznacznieniu zapytania
- Ewaluacja na zbiorze AMBIENT z użyciem czterech popularnych miar: Rand Index (RI), Adjusted Rand Index (ARI), Jaccard Index (JI) and F1 measure



# Przegląd stanu wiedzy

- Klasteryzacja Wyników Wyszukiwania (ang. Search Results Clustering - SRC) jest specyficzną dziedziną w ramach klasteryzacji dokumentów
- Kontekstowy opis dokumentu (tzw. snippet) zwracany przez wyszukiwarke jest krótki, często niekompletny, oraz ograniczony względem zapytania, co powoduje iż określenie miary podobieństwa między dokumentami jest dużym wyzwaniem.
- Podejścia do klasteryzacji wyników wyszukiwania mogą być klasyfikowane jako: data-centric lub description-centric
- Data-centric – Bisecting K-means, HAC
- Description-centric – STC, Lingo, KeySRC

# Rest API

- Babelfy – text disambiguation
  - <https://babelfy.io/v1/disambiguate>
- BabelNet – get categories and glosses for the given synset
  - <https://babelnet.io/v3/getSynset>
- BabelNet – get hypernyms for the given synset
  - <https://babelnet.io/v3/getEdges>

# Pierwszy eksperyment

<b>Algorithm</b>	<b>RI</b>	<b>ARI</b>	<b>JI</b>	<b>F1</b>
Lingo	62.52	18.09	30.76	49.01
STC	66.95	23.05	28.10	53.08
K-means	62.79	7.69	12.83	49.79

# Drugi eksperyment

Improvement	RI	ARI	JI	F1
Lingo	62.52	18.09	30.76	49.01
synsets+	<b>63.52</b>	<b>18.61</b>	29.21	<b>49.76</b>
categories+	<b>63.04</b>	17.01	27.46	<b>49.36</b>
categories+1	61.73	16.48	29.55	48.65
categories+2	62.17	17.44	30.30	48.80
glosses+	<b>62.69</b>	12.27	21.30	47.24
hypernyms+	61.52	16.35	29.44	48.32

# Trzeci eksperyment

<b>Approach</b>	<b>RI</b>	<b>ARI</b>	<b>JI</b>	<b>F1</b>
Lingo	62.52	18.09	30.76	49.01
babelC11	50.60	1.67	26.87	41.53
babelC12	50.44	1.56	27.06	40.41



# Wnioski

- Wprowadziliśmy nowe semantyczne cechy z BabelNet/Babelfy (jak ujednoznacznione synsety, categories/glosses opisujące synsety, czy semantyczne krawędzie) w celu weryfikacji jak one wpływają na wynik klasteryzacji
- Najlepsze usprawnienia dotyczą rozszerzania o synsety i są stosunkowo słabe (co jest zaskoczeniem?).
- Klasteryzacja snippetów tylko z użyciem informacji z Babelnetu daje gorsze wyniki niż Lingo.
- Występuje problem wydajności środowiska badawczego (requests limits ale także czasy odpowiedzi).
- Należy opuścić sferę płaskiego rozszerzania na poczet heurystyk grafowych

# Wnioski

- Duże możliwości są też w modelu word2vec zaadoptowanym do przestrzeni znaczeń
- Projekt o nazwie Sense Embeddings zrealizowany w ramach ekosystemu BabelNet

# Koniec części 1

Ośrodek Przetwarzania Informacji – Państwowy Instytut Badawczy  
al. Niepodległości 188B, 00-608 Warszawa  
tel. 22 570 14 00  
[www.opi.org.pl](http://www.opi.org.pl)

# Whose line is it anyway?

Czyli charakterystyka emocjonalna filmu  
na podstawie napisów

Warszawa 05.04.16

# Problem

- Dla każdej linii dialogu w danym filmie przypisać jeden z ośmu stanów emocjonalnych: *'love'*, *'happiness'*, *'surprise'*, *'emotionless'*, *'sad'*, *'disgust'*, *'anger'* oraz *'fear'*.
- Znaleźć "profil emocjonalny" kultowych filmów
- Aby dodać pikanterii zadanie wykonać dla dwóch języków *'love'*, *'happiness'*, *'surprise'*, *'emotionless'*, *'sad'*, *'disgust'*, *'anger'* and *'fear'*
- Sprawdzić, czy wyniki są stabilne pomiędzy językami

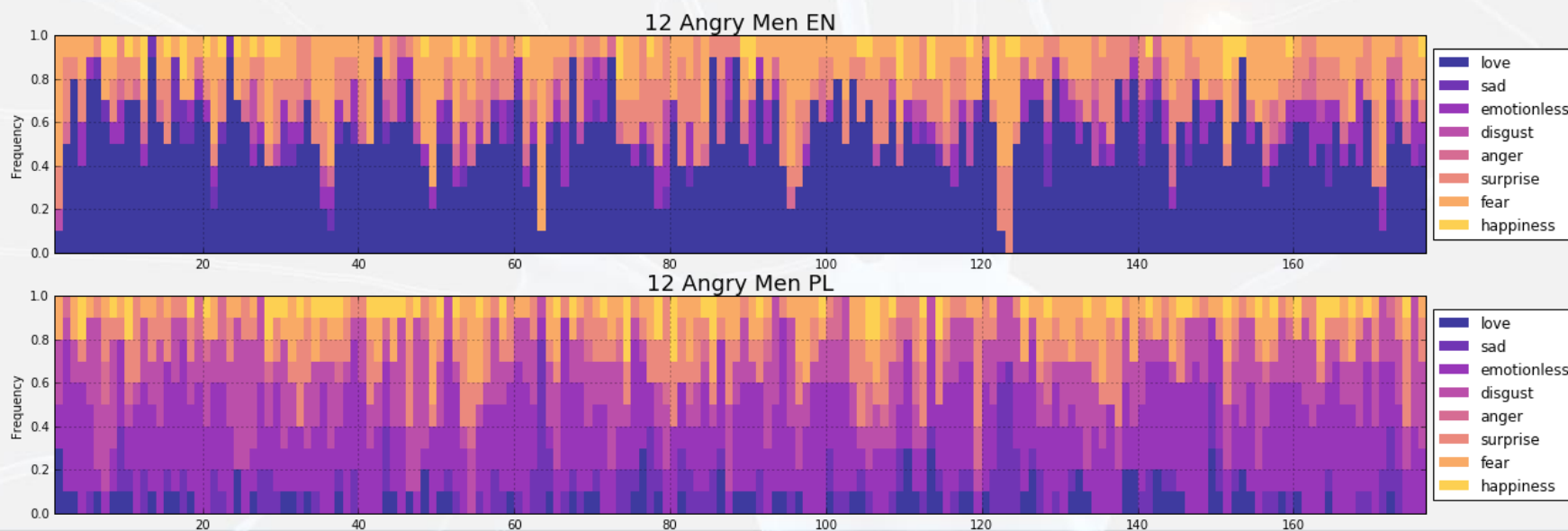


# Metoda – Emocje

- Wytrenować model Skip-Gram z próbkowaniem negatywnym na polskiej i angielskiej wikipedii w celu uzyskania wektorowej reprezentacji słów
- Za pomocą BabelNet'a stworzyć *drzewo emocji*, gdzie korzeniem jest stan emocjonalny a potomkami są sensy będące jego hyponimami (np. Dla sensu 'love' jego hyponimem jest 'admiracja').
- Każdemu wierzchołkowi przypisujemy wektorową reprezentację słowa opisującego dany sens
- Policzyc średnią ważoną wierzchołków drzewa z wagami tym mniejszymi, im dalej wierzchołek jest od korzenia

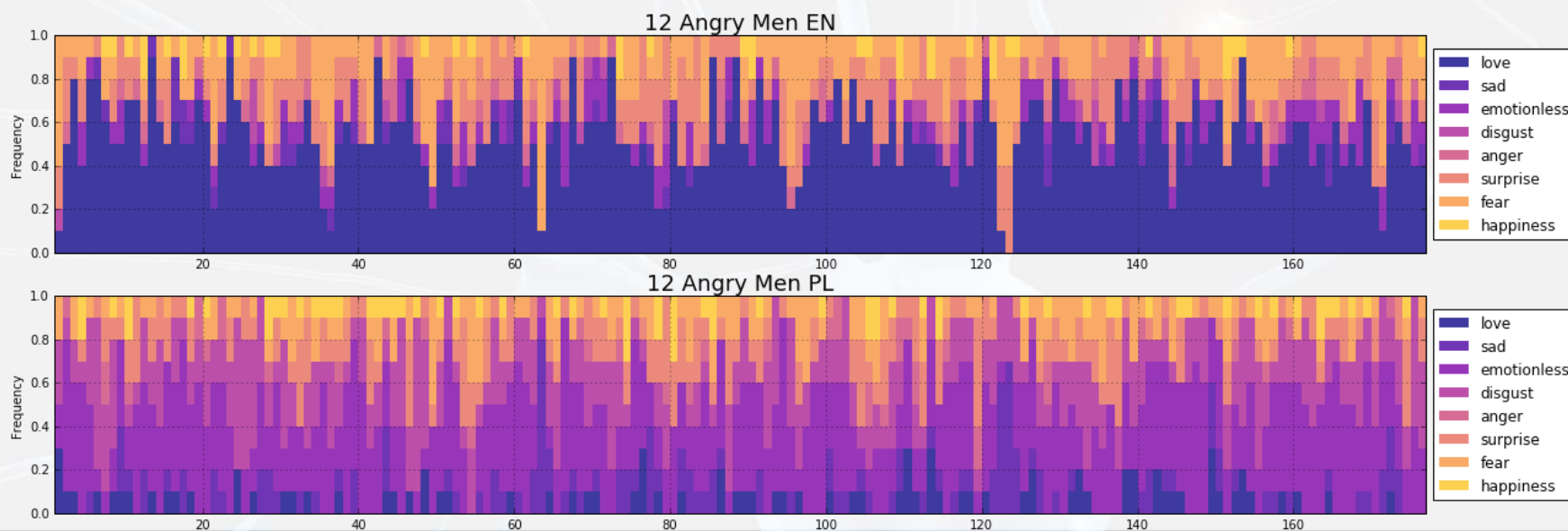
# Metoda – Zdania

- Z każdej liniki usunąć wyrazy znajdujące się na stop liście.
- Uśrednić wektorowe reprezentacje słów pozostałych po filtrowaniu
- ... trzymać kciuki :)



# Metoda – Zdania

- Z każdej liniki usunąć wyrazy znajdujące się na stop liście.
- Uśrednić wektorowe reprezentacje słów pozostałych po filtrowaniu
- ... trzymać kciuki :)

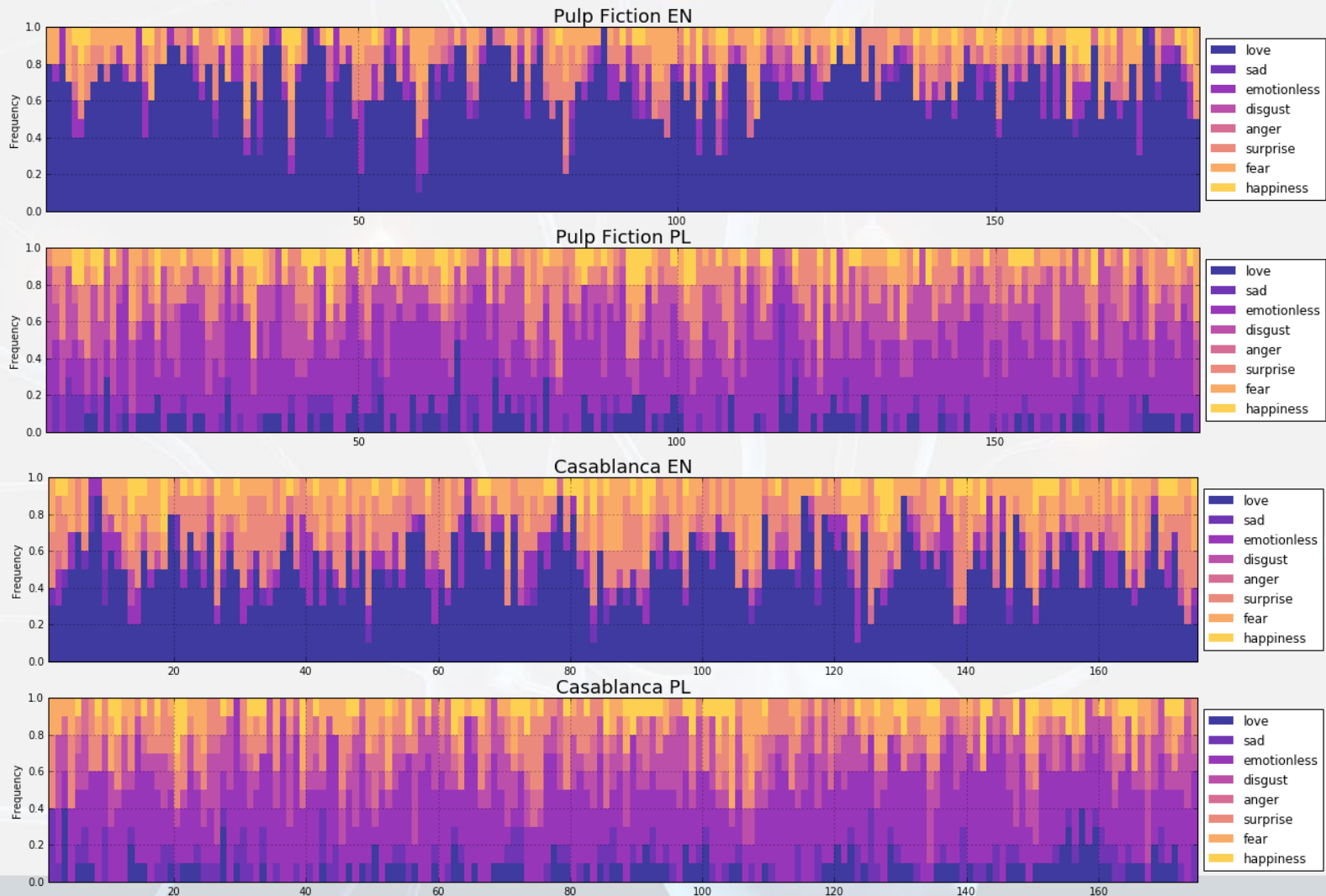


# Przykładowe wyniki

Table 1: Comparison of emotions

	Casablanca		Pulp Fiction	
	EN	PL	EN	PL
<i>love</i>	851	116	1210	112
<i>sad</i>	42	113	33	78
<i>emotionless</i>	133	694	121	722
<i>disgust</i>	33	302	17	407
<i>anger</i>	56	94	61	58
<i>surprise</i>	320	94	157	234
<i>fear</i>	241	128	168	123
<i>happiness</i>	68	87	50	83
<b>sum</b>	<b>1744</b>	<b>1744</b>	<b>1817</b>	<b>1817</b>

# Przykładowe wyniki





# Oczywiste rozszerzenia

- Stosowanie lasów zamiast drzew
- Zastąpienie nienadzorowanego, naiwnego algorytmu bazującego na mierze kosinusowej algorytmami nadzorowanymi (to właśnie testujemy)
- Odejście od zagadnienia klasyfikacji w kierunku klasyfikacji wieloetykietowej (ktoś chętny do tagowania zbioru?)

## Koniec części 2

Ośrodek Przetwarzania Informacji – Państwowy Instytut Badawczy  
al. Niepodległości 188B, 00-608 Warszawa  
tel. 22 570 14 00  
[www.opi.org.pl](http://www.opi.org.pl)

# Who was it anyway?

Czyli próba oceny postaci historycznych

Warszawa 05.04.16



- Oprócz tego, że BabelFy i BabelNet zawierają w sobie koncepty, nazwy własne w postaci grafów, mogą również posłużyć do wyszukiwania osób, o których informacje możemy znaleźć na przykład w Wikipedii.

★ • bn:00002460n • NOUN • Named Entity • Categories: Nobliści • fizyka, Urodzeni w 1879, Niemieccy Żydzi, Teoria względności...

**Albert Einstein** • **Einstein**

Albert Einstein – niemiecki **fizyk** żydowskiego pochodzenia, jeden z największych fizyków-teoretyków XX wieku, twórca ogólnej i szczególnej **teorii względności**, współtwórca korpuskularno-falowej **teorii światła**, odkrywca **emisji wymuszonej**.

IS A: człowiek • fizyk • Fizyka teoretyczna

ACADEMIC ADVISOR: Heinrich Friedrich Weber

ARCHIVES AT: EN Swiss Literary Archives

AUTHOR: Michio Kaku

AWARD RECEIVED: EN Franklin Medal • Prix Jules-Janssen • Copley Medal

BIRTH PLACE: Cesarstwo Niemieckie • Ulm • EN University of Louisiana at Monroe

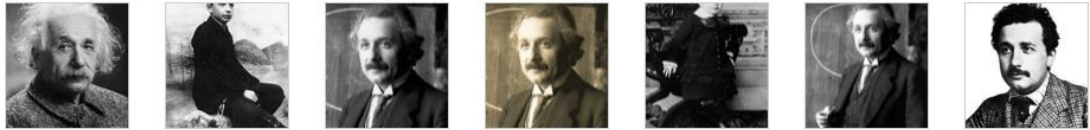
CHILD: Hans Albert Einstein • Eduard Einstein • Lieserl Einstein

CHILDREN: Hans Albert Einstein • EN Einstein family • Eduard Einstein

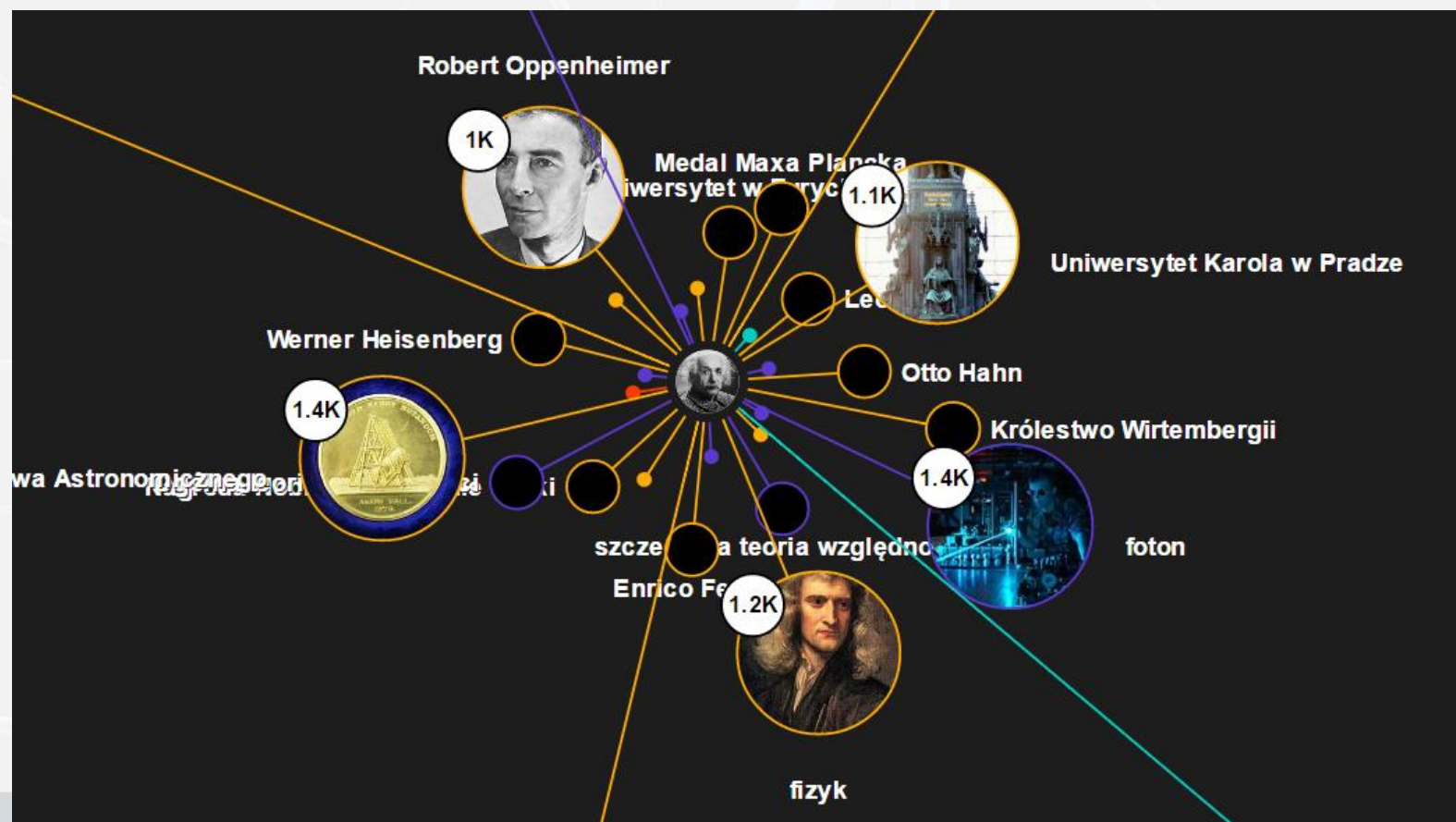
COUNTRY OF CITIZENSHIP: Stany Zjednoczone • Szwajcaria • Republika Weimarska

+ More relations

EXPLORE NETWORK



- Możemy również eksplorować sieć semantyczną, powiązaną z główną postacią (szukaną frazą).





# Więc co możemy zrobić?

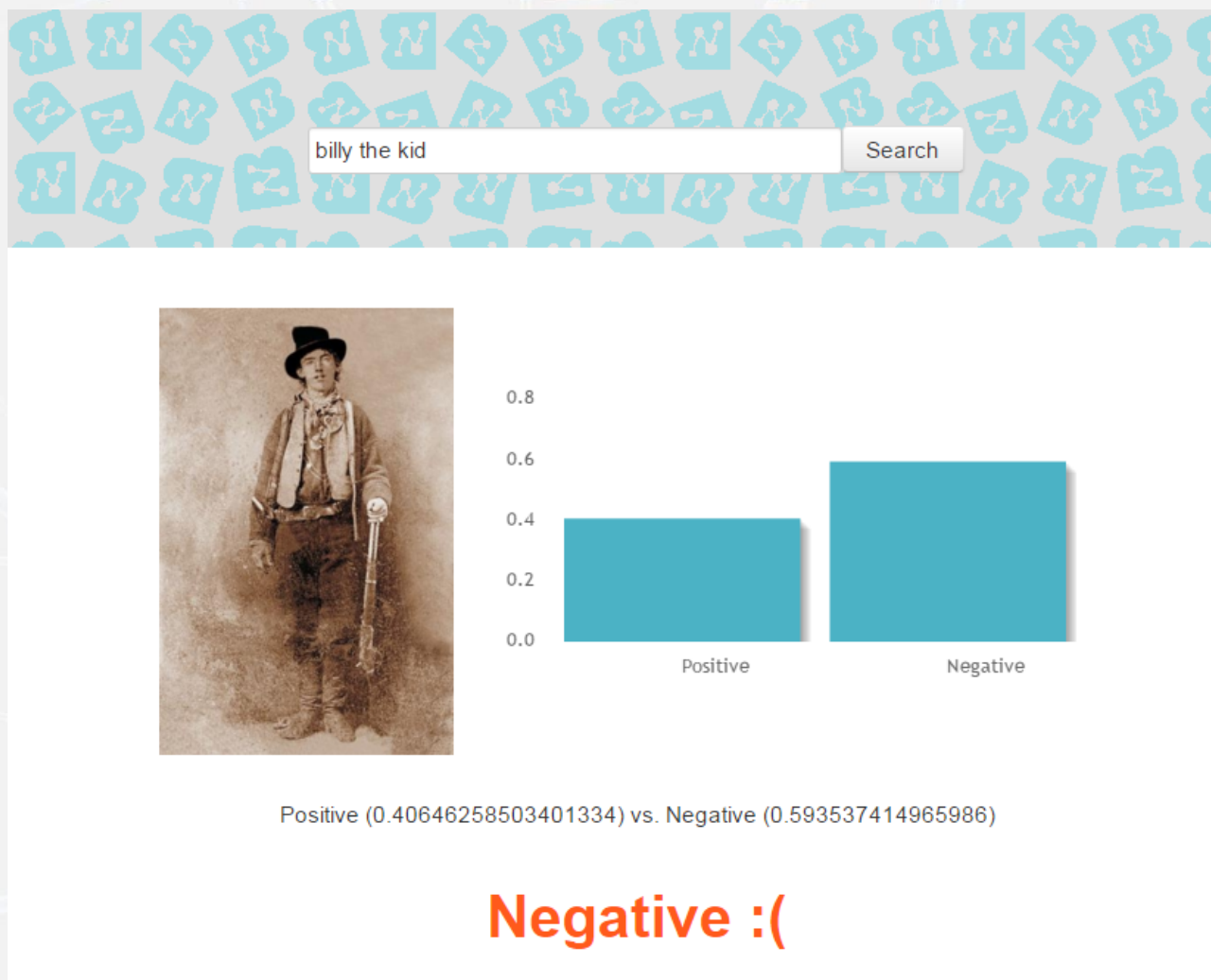
- Odpytać sieć semantyczną o informacje, które zawiera o zadanej postaci historycznej,
- Zebrać teksty pojawiające się w znalezionych węzłach sieci.
- Za pomocą słowników semantycznych ocenić zbadać sentyment zebranych definicji.
  - *lexicon* (<http://www.cs.uic.edu/liub/>)
  - *sentiwordnet* (<http://sentiwordnet.princeton.edu/>)
- Opakować w zgrabną aplikację :)



# <https://bhs.opi.org.pl/>



# <https://bhs.opi.org.pl/>



# Jak to działa?

albert einstein



albert einstein

Albert Einstein –  
niemiecki fizyk  
żydowskiego  
pochodzenia, jeden...



albert

Prince consort of  
Queen Victoria of  
England (1819-1861)

- Dla szukanej frazy, za pomocą BabelFy wykonujemy dysambiguację, odnajdujemy najbardziej pasujący węzeł sieci,
- Dla znalezionej gałęzi, za pomocą BabelNet odwiedzamy gałęzie: hypernoms, hyponyms, meronym,
- Dla odwiedzonych gałęzi zbieramy definicje (gloss),
- Tokenizujemy znaleziony korpus i oceniamy go za pomocą słowników. Wynik jest normalizowany do 1,

# Oceniamy!

- Postaci pozytywne

Postać	Wynik pozytywny	Wynik negatywny	Ocena
Albert Einstein	0.5519	0.4480	Pozytywna
Marilyn Monroe	0.5719	0.4280	Pozytywna
Isaac Newton	0.6286	0.3713	Pozytywna

- Postaci negatywne

Postać	Wynik pozytywny	Wynik negatywny	Ocena
Al Capone	0.4768	0.5231	Negatywna
Stalin	0.4573	0.5426	Negatywna
Billy the Kid	0.4064	0.5935	Negatywna

# A jak to jest wśród zwolenników / przeciwników kotów?

- Osoby znane z miłości do kotów

Postać	Wynik pozytywny	Wynik negatywny	Ocena
Abraham Lincoln	0.4263	0.5736	Negatywna
Ernest Hemingway	0.6171	0.3828	Pozytywna
Theodore Roosevelt	0.5969	0.4030	Pozytywna

- Osoby znane z nienawiści do kotów

Postać	Wynik pozytywny	Wynik negatywny	Ocena
Dwight Eisenhower	0.5147	0.4852	Pozytywna
Napoleon Bonaparte	0.5428	0.4571	Pozytywna
Henry III of England	0.3653	0.6346	Negatywna

# Problemy

- Dysambiguacja nie zawsze pomaga: Stalin vs. Joseph Stalin. W przypadku postaci historycznych lepsze wyniki otrzymywane były w przypadku przeszukiwania sieci z pominięciem BabelFy,
- Iterowanie dużej ilości węzłów,
- Brak zbiorów referencyjnych,



# Koniec części 3 ... i ostatniej :)

Ośrodek Przetwarzania Informacji – Państwowy Instytut Badawczy  
al. Niepodległości 188B, 00-608 Warszawa  
tel. 22 570 14 00  
[www.opi.org.pl](http://www.opi.org.pl)

Rekrutujemy do pracy przy projektach typu:

- Konkurs na najlepszy algorytm antyplagiatowy
  - Centralny system antyplagiatowy
    - Wirtualni konsultanci
    - Semantyczne wyszukiwarki
- Centralna Polska Bibliografia Naukowa

**Kontakt:** [dko@opi.org.pl](mailto:dko@opi.org.pl)

**Blog:** <http://opi-lil.github.io/>