OPI

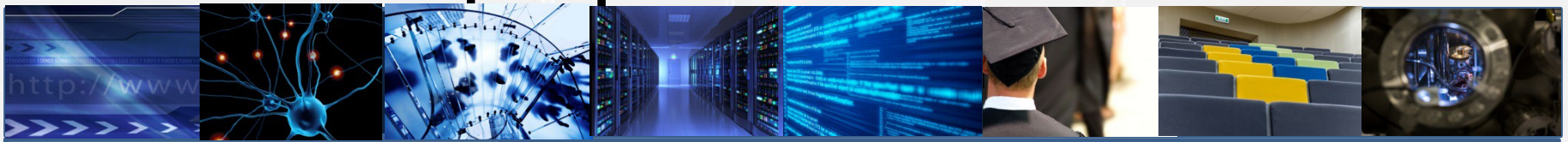
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

OPI – PIB

Warszawa 13.03.17

Interdyscyplinarny instytut

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY



Badania naukowe i prace rozwojowe

Systemy informatyczne i bazy danych

Fundusze strukturalne

- Zarządzanie projektami badawczymi
- Transfer technologii
- Interakcje: człowiek-technologie informacyjne
- Postawy i wartości naukowców (etos nauki)
- Młodzi w nauce
- Detekcja plagiatów
- Algorytmy uczenia maszynowego
- Semantyczne metody grupowania i porównywania tekstów

POL-ON ZINTEGROWANY SYSTEM INFORMACJI O NAUCE I SZKOLNICTWIE WYŻSZYM

inventorum

OSF

Z S U N Zintegrowany System Usług dla Nauki

NAUKA POLSKA
WWW.NAUKA-POLSKA.PL

JSA

SWISS CONTRIBUTION

Fundusze Europejskie
Inteligentny Rozwój

Unia Europejska
Europejski Fundusz Rozwoju Regionalnego

Przykładowe tematy prowadzonych badań

OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

- Analiza nienawiści w Internecie
- Konkurs AAIA 2016
- Konkurs ISMIS 2017
- Systemy rekomendacji ekspertów
- Narzędzia do budowania dedykowanych modeli języka na podstawie CommonCrawl
- Organizowanie unikalnego oraz pierwszego w historii konkursu na inteligentne metody detekcji plagiaryzmy
- Wirtualni konsultanci

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Eko-system BabelNet

Warszawa 13.03.17

O produkcie

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 jezyków, i powstal w ramach automatycznej integracji: WordNeta, Wikipedii, OmegaWIki, Wiktionary, WikiData itd.

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

Zakres danych

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

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

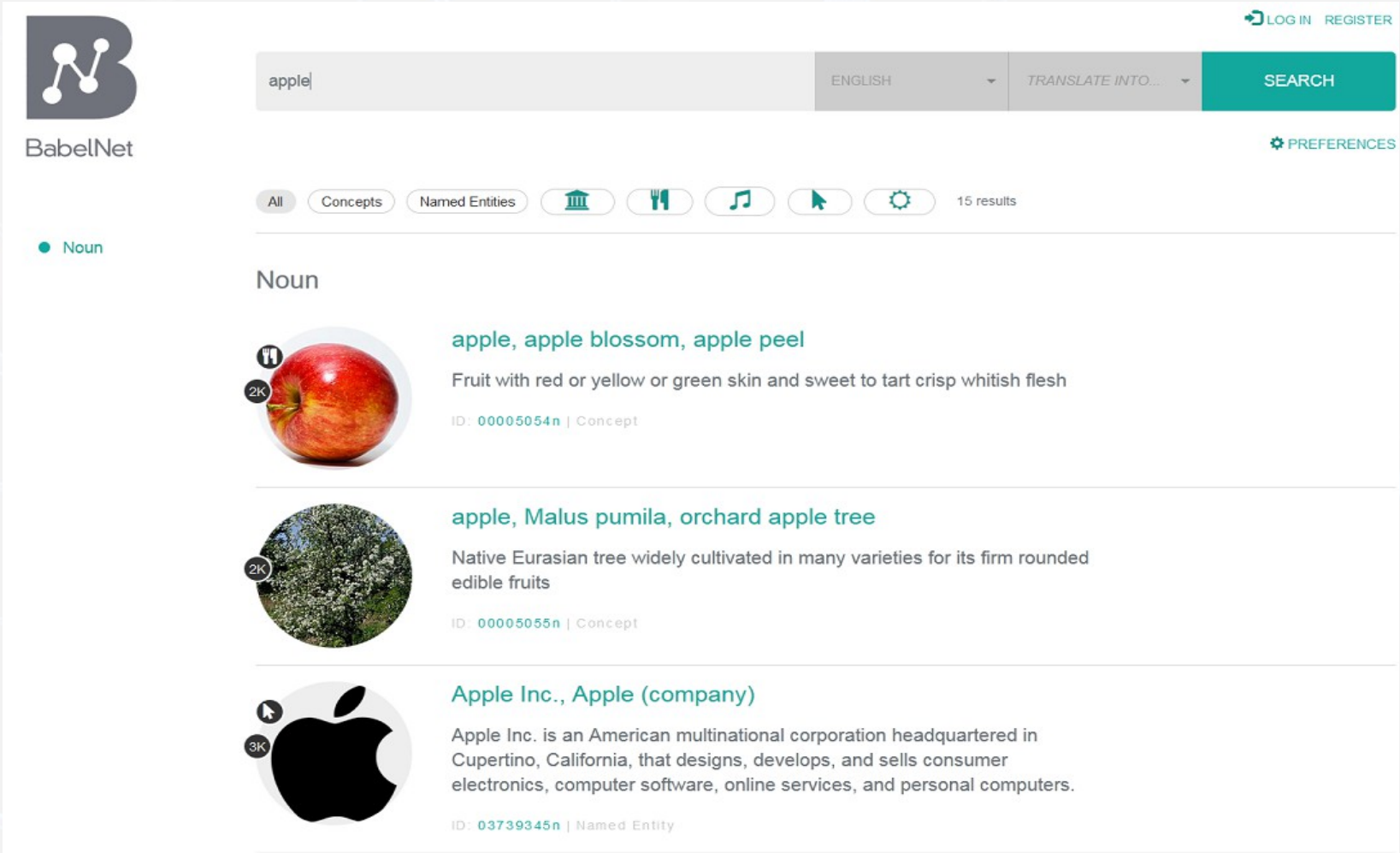
- WordNet, a popular computational lexicon of English (version 3.0).
- Open Multilingual WordNet, a collection of wordnets available in different languages (downloaded in August 2015).
- Wikipedia, the largest collaborative multilingual Web encyclopedia (November 2014 dump).
- OmegaWiki, a large collaborative 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 Dicember 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

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY



BabelNet

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

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...

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 تفاح, Malus domestica, أنواع التفاح, التفاح, تفاح ثنائي, تفاح مستأنس, تفاحة, شجرة تفاح, فوائد التفاح, التفاح الأخضر, domestica malus

ZH 苹果, 蘋果, Malus domestica, 大蘋果市, 柔, 蘋果樹, 苹果, , 青苹果,

EN apple, apple blossom, apple peel, Apples, pomiculture, *Green Apples, American Delicious, An apple a day, Aplle, 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, Pommeraie, pomme verte*

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

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

HE תפוח, תַּפּוּחַ, Malus, גרנד אלכסנדר, גרני סמית, דליישס, ענה, תפוח עץ, malus domestica, תפוח ירוק, תפוחים

HI सेब, *सेव, malus domestica, सीन एप्पल*

BabelNet - JSON

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Response example


```
{
  "senses": [
    {
      "lemma": "Apple_System_on_Chips",
      "simpleLemma": "Apple_System_on_Chips",
      "source": "WIKIRED",
      "sensekey": "",
      "sensenumber": 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": "",
      "sensenumber": 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

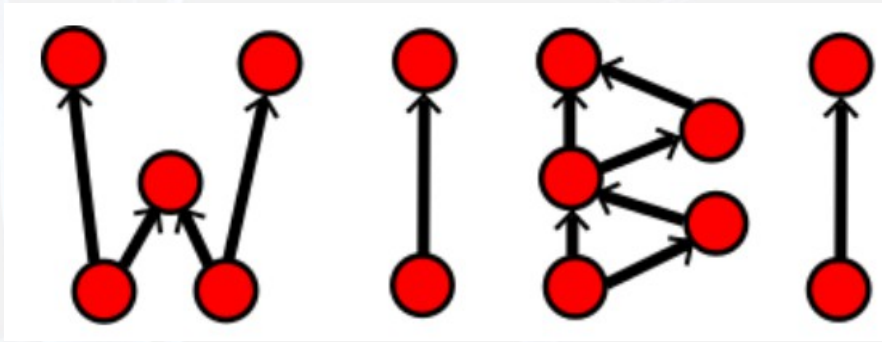
A **BabelSense** is a particular, language-specific lexicalization occurring in a given **BabelSynset**. Each **BabelSense** is tied to a particular source (WordNet, Wikipedia, Wiktionary, automatic translations, etc.).

Pokłosie BabelNetu

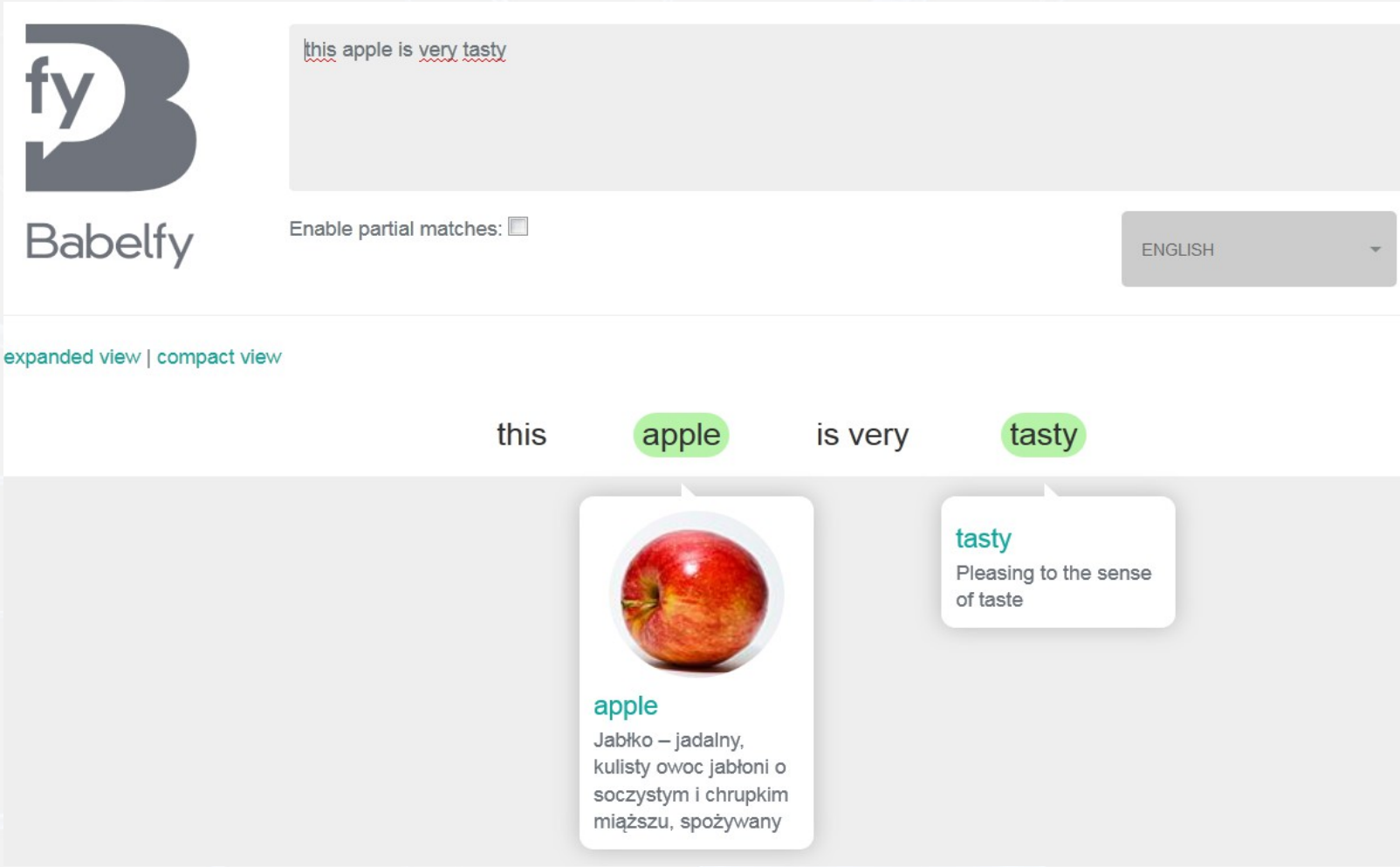
W oparciu o Babelnet (http://babelnet.org/) powstały dwa narzędzia:

1) Babelfy – wielojęzykowy system ujednoznaczniania

2) Wikipedia Bi taksonomia – taksonomia stron Wikipedii dopasowana do taksonomii kategorii.



Babelfy


this apple is very tasty
Babelfy
Enable partial matches:
ENGLISH
expanded view | compact view
this
apple
is very
tasty
apple
Jabłko – jadalny, kulisty owoc jabłoni o soczystym i chrupkim miąższu, spożywany
tasty
Pleasing to the sense of taste

Babelfy

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Babelfy

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

Enable partial matches:

POLISH

xpanded 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

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

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 poziomowe 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 wyszukiwarkę 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

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Algorithm	RI	ARI	JI	F1
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

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

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

Trzeci eksperyment

Approach	RI	ARI	JI	F1
Lingo	62.52	18.09	30.76	49.01
babelCl1	50.60	1.67	26.87	41.53
babelCl2	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?).
- Klasteryzja 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

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Whose line is it anyway?

Czyli charakterystyka emocjonalna filmu na podstawie napisów

Warszawa 13.03.17

Problem

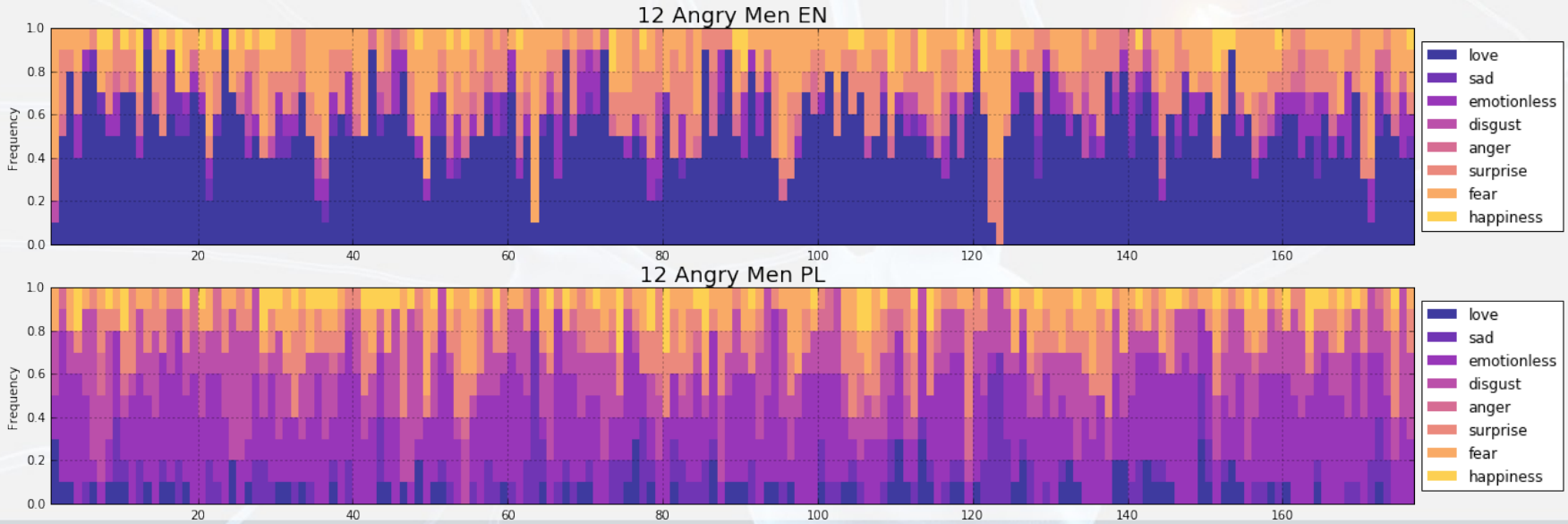
- Dla każdej linii dialogu w danym filmie przypisać jeden z ośmu stanów emocjonalnych: *'love', 'happiness', 'surprise', 'emotionless', 'sad', 'disgust', 'anger'* oraz *'fear'.*
- Znaleźć "profil emocjonalny" kultowych filmów
- Aby dodać pikanterii zadanie wykonać dla dwóch języków
- Sprawdzić, czy wyniki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



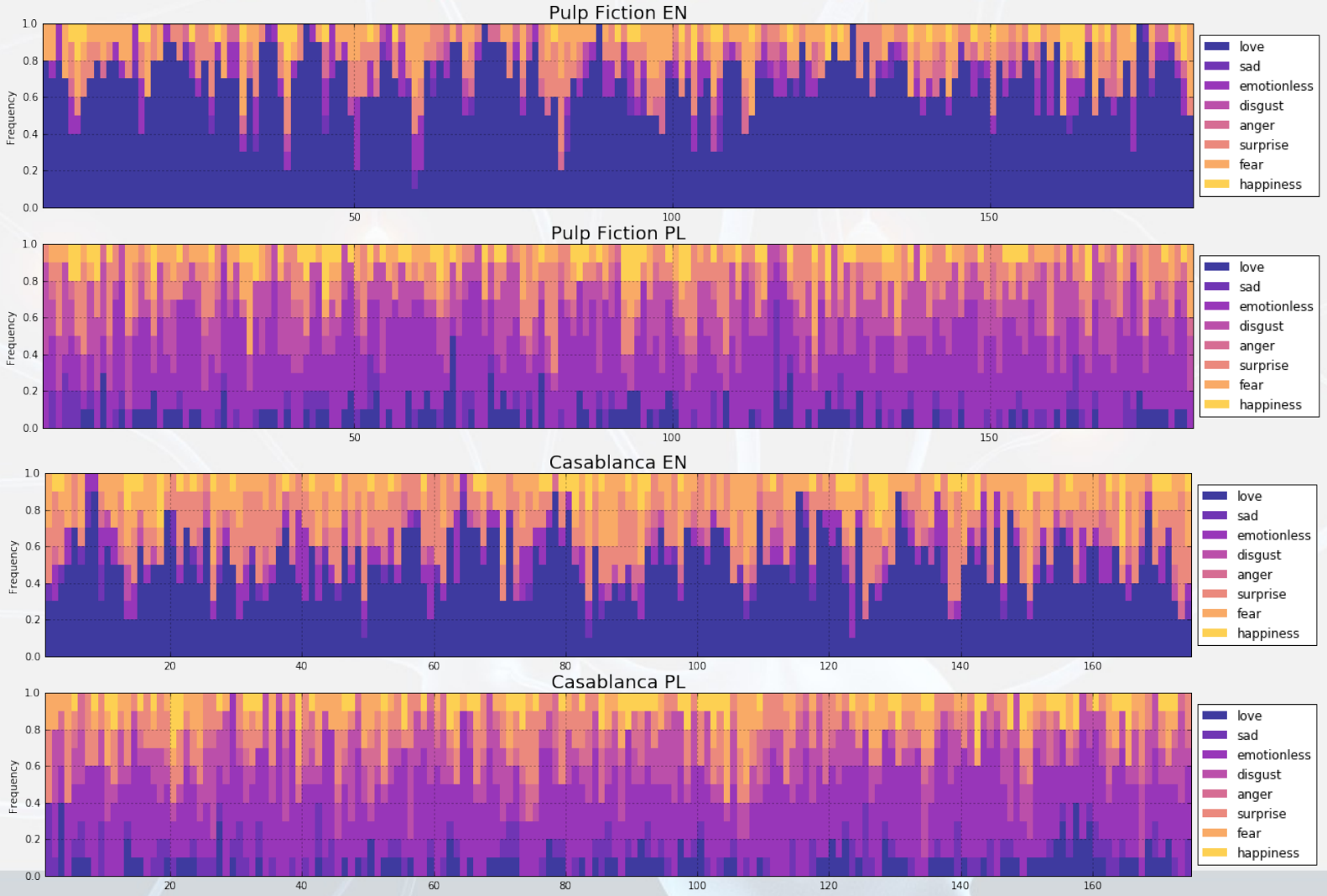
Przykładowe wyniki

Table 1: Comparison of emotions

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

Przykładowe wyniki

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY



Oczywiste rozszerzenia

- Stosowanie lasów zamiast drzew
- Zastąpienie nienadzorowanego, naiwnego algorytmu bazującego na mierze kosinusowej algorytmami nadzorowanymi (to właśnie testujemy)
- Odejście w kierunku klasyfikacji wieloetykietowej (problem wąskiego gardła czyli braku dobrej jakości zbiorów uczących)

LanguageCrawl: A Generic Tool for Building Language Models Upon Common-Crawl

Szymon Roziewski, Wojciech Stokowiec

National Information Processing Institute, Natural Language Processing Laboratory, Warsaw

Introduction

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Overview

- Motivation
- Building scalable framework for textual Big Data processing of The Common Crawl corpus for a given language
- Making Polish N-gram collection
- Characteristics of Polish N-gram collection
- Trained word2vec model on gathered and pre-processed corpus

Motivation

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Why have we implemented LanguageCrawl?

- To help NLP researchers to construct large-scale corpora based on Internet Data, with handy toolkit
- Processing Common Crawl Archive
- Filtering documents for a given language
- Building N-gram corpora
- Training continuous Skip-gram language model

Motivation

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Why N-grams?

- Statistical natural language processing
- Machine translation application
- Machine learning
- Pattern matching
- Information retrieval
- Text generation

Motivation

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Why word2vec model?

- Machine translation
- Topic modelling, tagging
- Word sense disambiguation
- Bioinformatics: BioVectors, ProtVec, GeneVec

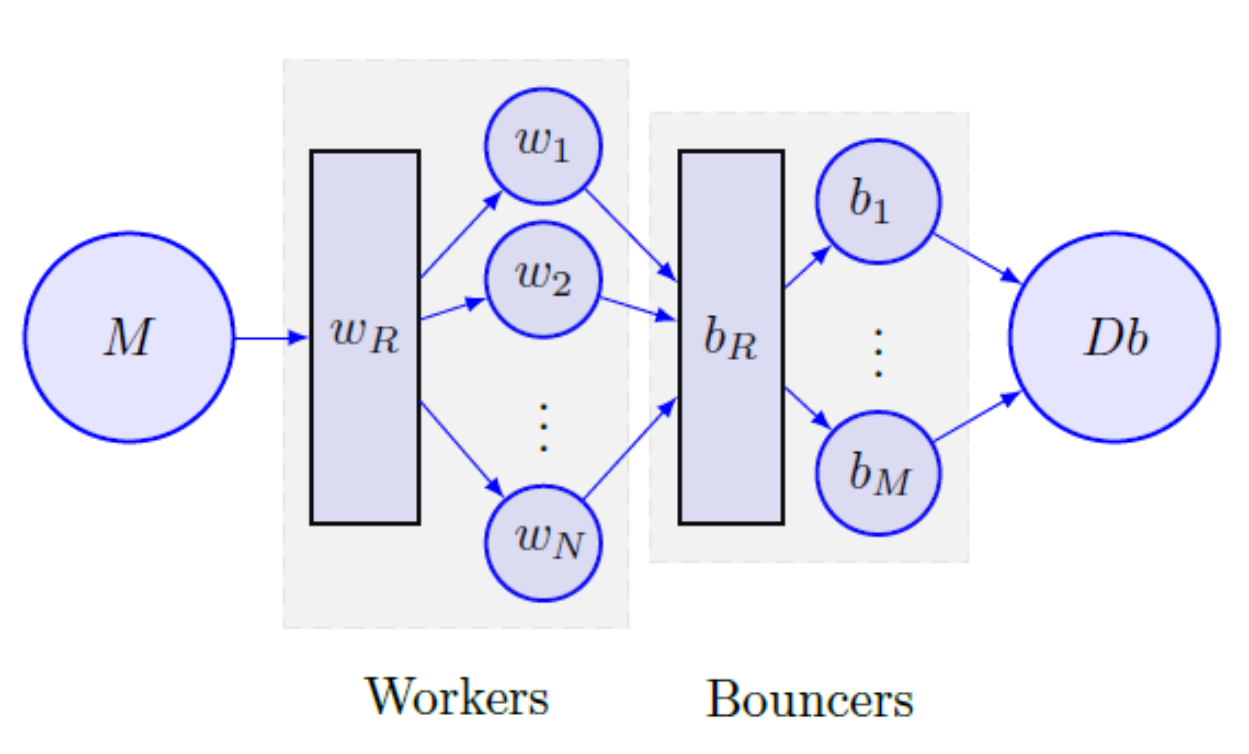
Scalable Framework

Application of effective technologies

- akka for Scala – actor based framework for building highly concurrent, distributed applications
- Apache cassandra – efficient, scalable, distributed database with proven fault-tolerance

Framework Infrastructure

Logic for the first stage processing



Data Preparation

- January 2015 Crawl Archive has been processed
  consisting of 140 TB in size, 1.82 bln web pages
  25000 WET files, 300 MB each,
  ≈ 7.5 TB of data has been analyzed,
  using actor based akka framework
- CommonCrawl Datastore contains
  petabytes of data collected over
  7 years

Data Processing

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

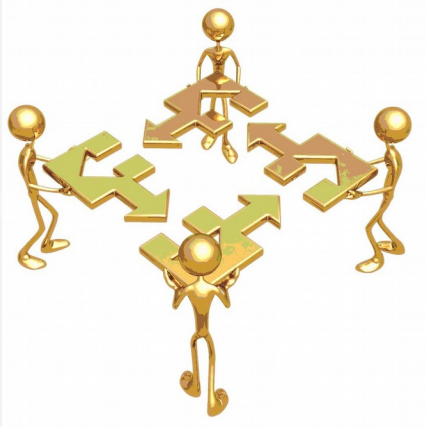
Concurrent Distributed Work

- Actor System for starting the app and managing

  1 File Master to create and manage File Worker actors
- 24 File Workers for downloading and sending data for

  language identification
- 36 Bouncer actors

  filtering Polish web sites using CLD2 library



Data Processing 2

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Concurrent Distributed Work

- Converting to lowercase letters
- Splitting into sentences by using NLTK tokenizer
- Removing non-alphanumerical characters
- Truncating fused and misprinted words corrected
- Polish diacrtitical marks and stop words has been preserved
- Writing to Cassandra Data Storage

N-gram unique number occurrences

N-gram Type	Total #	Size
unigram	3 mln	50 MB
bigram	2.6 mln	60 MB
trigram	3.8 mln	113 MB
four-gram	4.2 mln	159 MB
five-gram	3.6 mln	163 MB
total	17.3 mln	545 MB

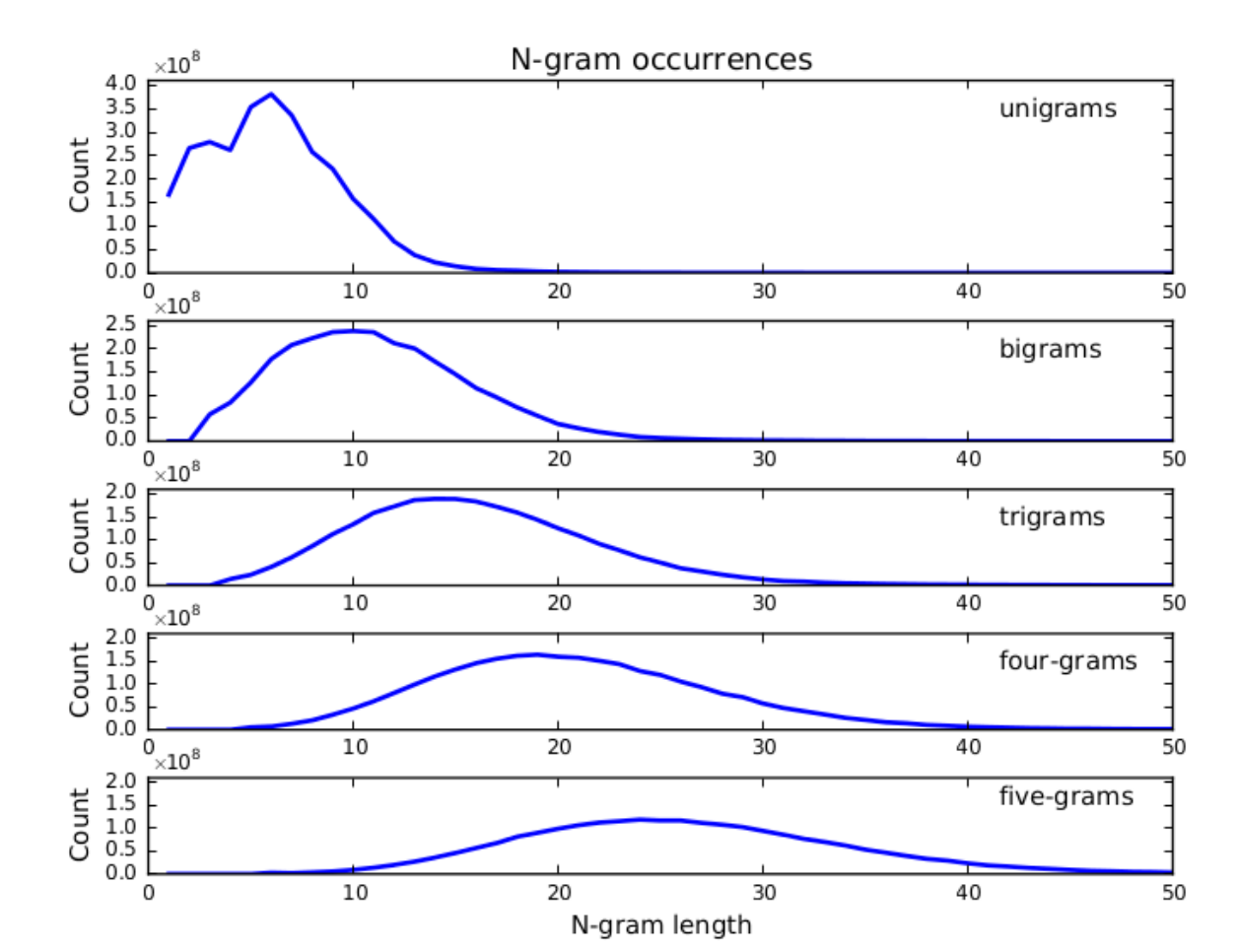
- After N-gram aggregation, model size highly diminished, by 99.4 %
- After cleaning from misprints and fused words, data size decreased on average by 63 %

N-gram Statistics

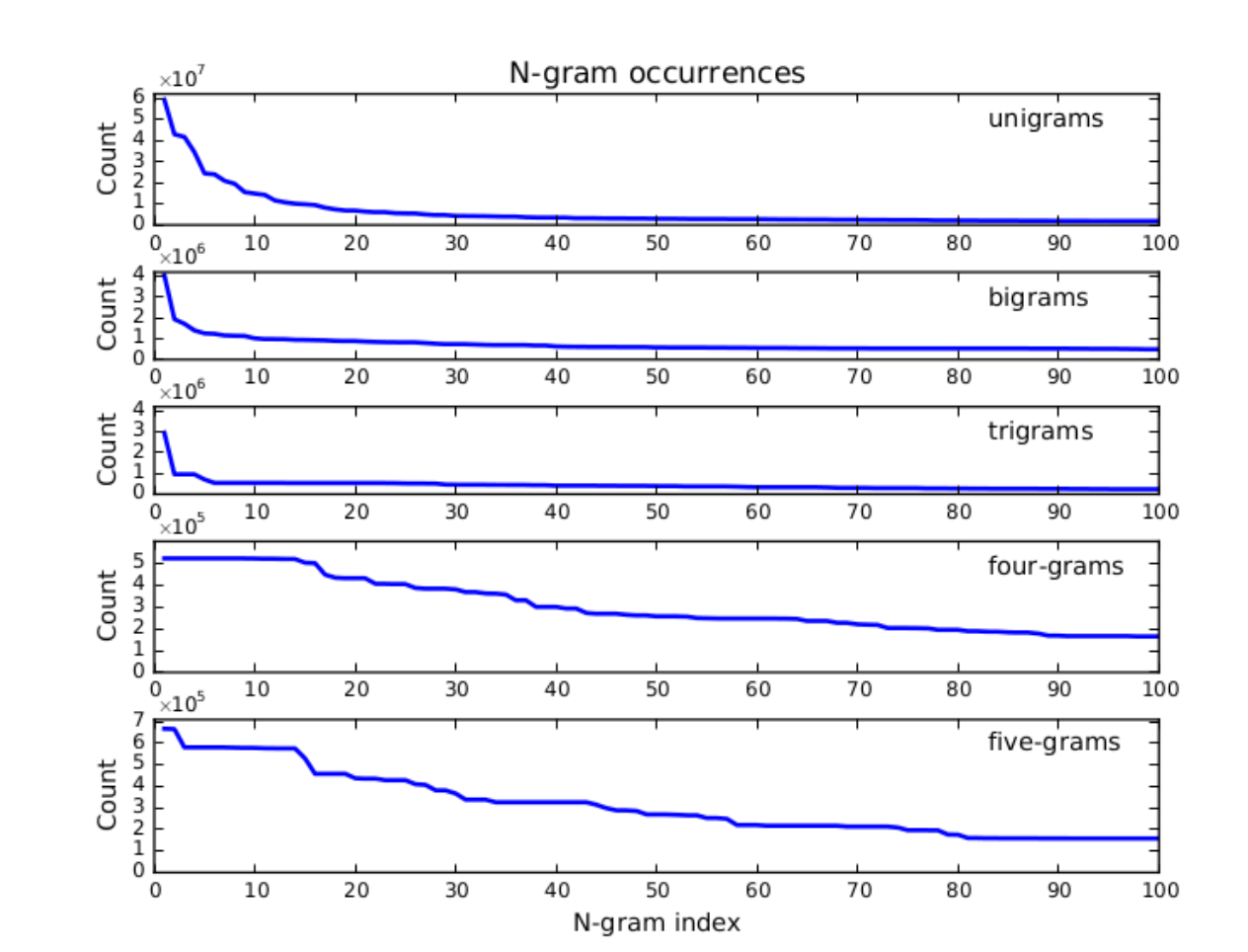
N-gram Type	Mean	SE	Median	10th PCTL	90th PCTL
unigram	8	3	9	5	13
bigram	15	4	15	10	21
trigram	22	6	22	15	30
four-gram	30	7	29	21	39
five-gram	38	8	38	28	49

- Statistics of unique N-gram lengths

N-gram occurrences



N-gram occurrences, position in index table



Word2vec Skip-gram

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Model training

- Efficient Gensim implementation of word2vec
- Skip-gram with hierarchical softmax
- Training with usage of iterators, online training
- Model trained on Polish Internet Corpus
- The resulting model is 5 GB large

Word2vec Skip-gram

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Results

Word	Most Similar	Distance
król (*king*)	cesarz (*emperor*)	0.73
Tusk (*former PM*)	Donald (*his name*)	0.74
kobieta (*woman*)	dziewczyna (*girl*)	0.80
mężczyzna (*man*)	chłopak (*boy*)	0.85
dziewczyna (*girl*)	rozochocna (*horny*)	0.82
apple	tablety (*tablets*)	0.79
Dublin	Blessington	0.85
Sushi	Pizza	0.76

Word2vec Skip-gram

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Results cd.

Expression	**Nearest Token**
król − mężczyzna + kobieta (*king − man + woman*)	Edyp (*Oedipus*)
wiekszy − duży + mały (*bigger − big + small*)	mniejszy (*smaller*)
Włochy − Rzym + Francja (*Italy − Rome + France*)	Paryż (*Paris*)
dżungla + król (*jungle + king*)	Tarzan | król lew (*Tarzan*) | (*lionking*)

Summary

- Polish N-gram Corpus of Polish Internet has been established
- A lot of duplicated data on the Internet
- Most frequent unigrams are stop words
- The long tails of the collection unworthy
- Trained word2vec model obtained
- LanguageCrawl tool will be available soon :)

Thank you!

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Ośrodek Przetwarzania Informacji
Państwowy Instytut Badawczy

al. Niepodległości 188 b
00-608 Warszawa

tel.: +48 22 570 14 00
e-mail: opi@opi.org.pl