Dialog Obywatelski w dyskursie parlamentarnym – przykład integralnego zastosowania infrastruktury CLARIN

Maciej Ogrodniczuk

Zespół Inżynierii Lingwistycznej

Instytut Podstaw Informatyki PAN

maciej.ogrodniczuk@ipipan.waw.pl

Tomasz Walkowiak

Wydział Elektroniki

Politechnika Wrocławska

tomasz.walkowiak@pwr.edu.pl

Agnieszka Hess

Instytut Dziennikarstwa, Mediów i Komunikacji Społecznej

Uniwersytet Jagielloński

agnieszka.hess@uj.edu.pl

CLARIN

CLARIN-PL

Common Language Resources and Technology Infrastructure

24 czerwiec 2021

Elementy infrastruktury CLARIN użyte w badaniach

- Narzędzia dostępne na ws.clarin-pl.eu
- Modelowanie tematyczne
  - wykrywanie ukrytych tematów
- Wykrywanie wyrażeń wielowyrazowych/terminologii
  - Mewex
  - TermoPL

Tomasz Walkowiak

tomasz.walkowiak@pwr.edu.pl

Fundusze Europejskie Inteligentny Rozwój
Rzeczpospolita Polska
Unia Europejska Europejski Fundusz Rozwoju Regionalnego

Modelowanie tematyczne (1)

- Probabilistyczne modele statystyczne służące odkrywaniu tematów występujących w zbiorze dokumentów

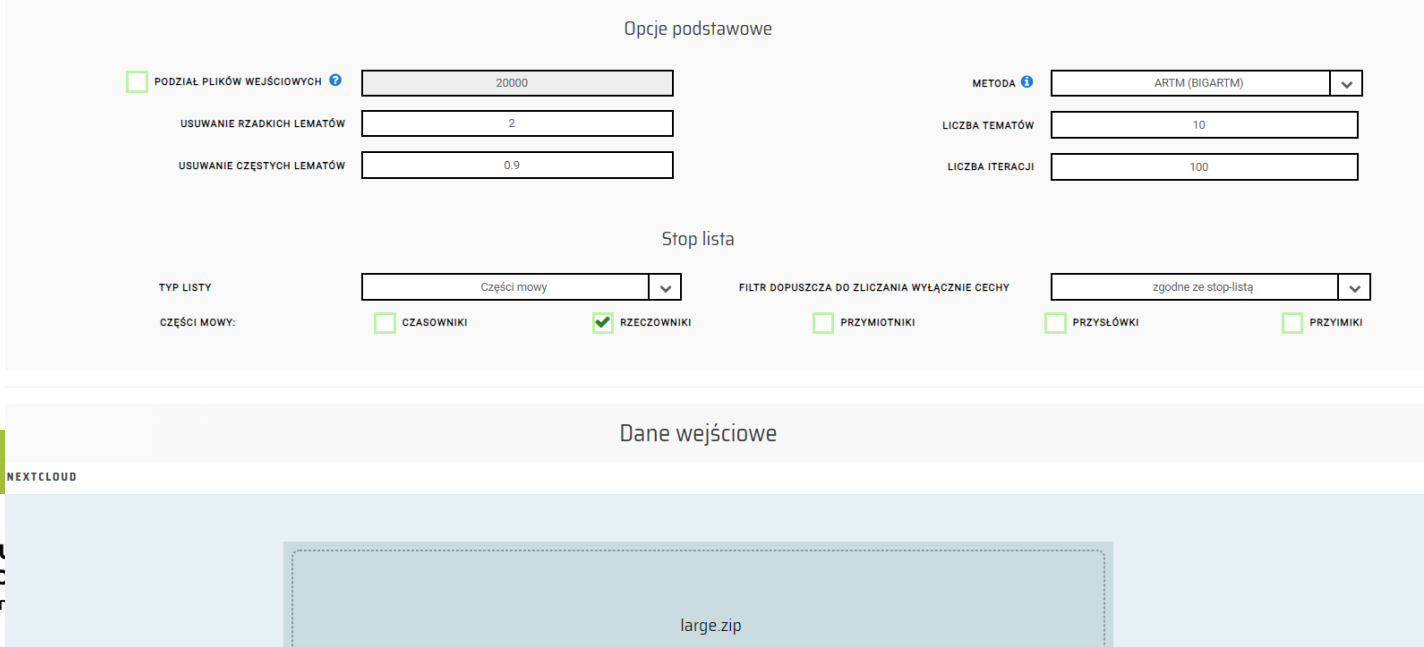
- Dokument
  - modelowany jako zbiór par: słowo i jego liczności
  - **nie istotna jest kolejność wyrazów**
  - **niezbyt długi i nie za krótki**
- LDA – Latent Dirchlet Allocation (Ukryta Alokacja Dirichleta)
  - każdy tekst stanowi mieszaninę pewnych tematów
  - temat jest rozkładem prawdopodobieństwa słów
  - zadania LDA:
    - identyfikacja (ukrytych) tematów (lista wyrazów)
    - oszacowanie rozkładu prawdopodobieństwa występowania wyrazów dla każdego tematu
    - oszacowanie rozkładu prawdopodobieństwa występowania tematów w rozpatrywanych dokumentach
- ARTM – biblioteka BigARTM

Fundusze Europejskie Inteligentny Rozwój
Rzeczpospolita Polska
Unia Europejska Europejski Fundusz Rozwoju Regionalnego

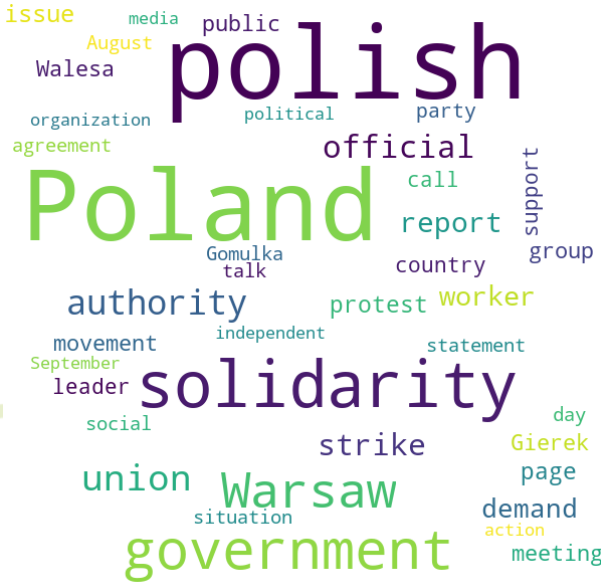
Modelowanie tematyczne (2)
ws.clarin-pl.eu/topic

- podział dokumentów na mniejsze (rozmiar ok. 10-20 kB)
- lematy
- filtrowanie
  - po części mowy (np. rzeczowniki), słowniku
  - max_df - odfiltrowanie leksemów, które pojawiły się w ponad (np. 80%) dokumentów

Wynik



ws.clarin-pl.eu/topicml
Background Reports - Radio Wolna Europa



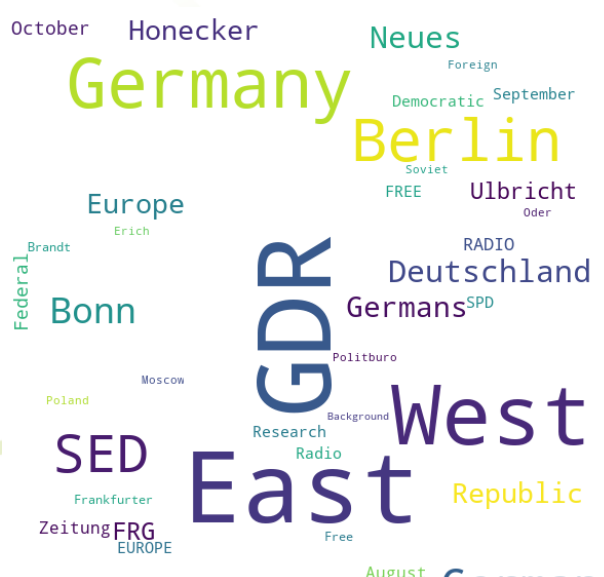
Tematy z współwystępowania słów w dokumentach

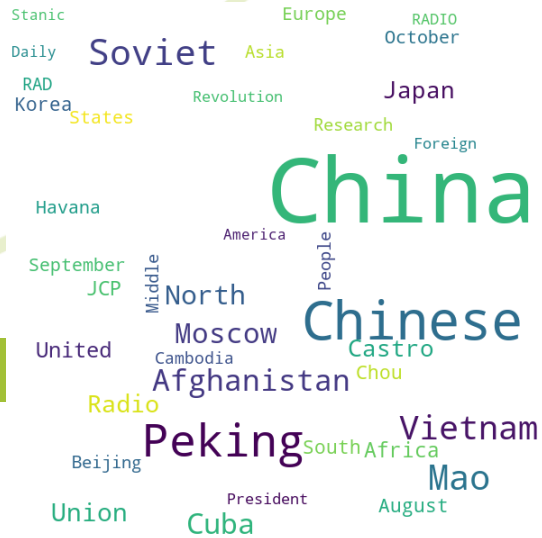
Inteligentny Rozwój

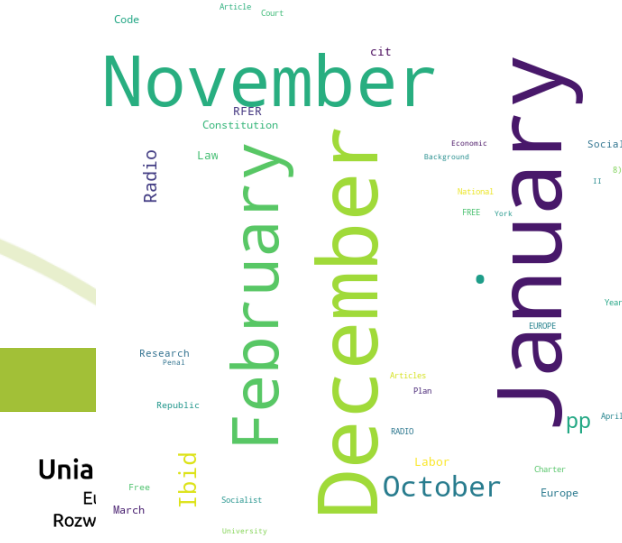
Unia Europejska
Europejski Fundusz Rozwoju Regionalnego

Proces iteracyjny, filtrowanie, podział dokumentów

Tylko nazwy własne







Inteligentny Rozwój

Nie tylko pojedyncze słowa

- Cel: Słownik wielowyrazowych jednostek leksykalnych
- Metoda:
  - kolokacje
  - kryteria
- Wychodzimy od kolokacji, które wydobywamy w sposób automatyczny, ale oceniamy za pomocą kryteriów lingwistycznych
- Kolokacje: często zaobserwowane i nieprzypadkowe połączenie słów występujących w bliskim sąsiedztwie:
  - *białe wino*
  - *czerwona kartka*

Fundusze Europejskie Inteligentny Rozwój
Rzeczpospolita Polska
Unia Europejska Europejski Fundusz Rozwoju Regionalnego

Usługi CLARIN-PL do wydobywania MWE

- MEWEX - wielowyrazowe jednostki leksykalne
  - ws.clarin-pl.eu/mewex
- TermoPL – wydobywanie terminologii
  - Aplikacja do uruchomienia lokalnie
    - zil.ipipan.waw.pl/TermoPL
  - Usługa online
    - ograniczona funkcjonalność
    - ws.clarin-pl.eu/termopl
- Porównywanie korpusów
  - ws.clarin-pl.eu/compcorp
  - przykładowy wynik

Wynik

Tabela | Pobierz

#	Ranking	Pojęcie	Pojęcie zlematyzowane	C-value
1	3	akcjonariusz większościowy	[akcjonariusz większościowy]	7.0
2	3	sąd wysoki	[sąd najwyższy]	7.0
3	3	użytkownik wieczysty	[użytkownik wieczysty]	7.0
4	3	podział nieruchomość	[podział nieruchomości]	7.0
5	3	wartość nieruchomość	[wartość nieruchomości]	7.0
6	4	opłata adiacencki	[opłata adiacencka]	6.8
7	5	użytkownik wieczysty nieruchomość	[użytkownik wieczysty nieruchomości]	6.3
8	5	umowa kredyt indeksować	[umowa kredytu indeksowanego]	6.3
9	5	obniżyć wartość nieruchomość	[obniżenie wartości nieruchomości]	6.3
10	7	średnica obiektyw	[średnica obiektywu]	6.0
11	7	nowy płyta	[nowa płyta]	6.0

Fundusze Europejskie Inteligentny Rozwój
Rzeczpospolita Polska
Unia Europejska Europejski Fundusz Rozwoju Regionalnego