Konspekt pracy zaliczeniowej na studiach podyplomowych

„Data Science Program”

Magdalena Benbenek

Spis treści

[1. Cel pracy 1](#_Toc28536624)

[2. Analizowane dane 1](#_Toc28536625)

[2.1. Źródło danych 1](#_Toc28536626)

[2.2. Format danych 1](#_Toc28536627)

[3. Charakterystyka problemu 2](#_Toc28536628)

[4. Etapy projektu 3](#_Toc28536629)

[5. Środowisko pracy 3](#_Toc28536630)

[6. Przegląd publikacji i projektów 3](#_Toc28536631)

[6.1. Projekt Morfeusz 4](#_Toc28536632)

[6.2. Word embeddings 4](#_Toc28536633)

[6.3. spaCy-pl 4](#_Toc28536634)

[6.4. Dodatkowe materiały 4](#_Toc28536635)

# Cel pracy

Celem pracy jest przegląd dostępnych rozwiązań i metod dostępnych w zakresie NLP (*natural language processing*) dla języka polskiego wraz z ich zastosowaniem na przykładowym zbiorze danych tekstowych i analizą rezultatów.

Dodatkowym elementem będzie modelowanie statystyczne wykorzystanych danych np. w celu przewidzenia sentymentu danej wypowiedzi pod warunkiem zadanego tematu, mówcy i punktu w czasie lub wnioskowania o cechach autora wypowiedzi na podstawie jej charakterystyk.

# Analizowane dane

# Źródło danych

Zbiór na którym zostaną przeprowadzone analizy został pobrany z serwisu kaggle.com. Zawiera transkrypcje przemówień polskich polityków z lat 1989 – 2019 oraz profil każdego z mówców.

# Format danych

Dane zostały udostępnione w formie bazy danych zawierającej dwie tabele.

Pierwsza z nich zawiera szczegółowy profil każdego polityka. Do najistotniejszych informacji należą:

|  |  |
| --- | --- |
| Nazwa kolumny w zbiorze | Opis |
| full\_name | Imię i nazwisko |
| elected | Data wybrania na posła |
| graduated\_school | Ukończona szkoła/uczelnia |
| education\_level | Wykształcenie |
| Occupation | Zawód |
| party\_section | Partia |
| number\_of\_votes | Liczba otrzymanych głosów |
| languages | Znane języki |
| last\_party | Ostatnia partia |

W tabeli występuje 2626 posłów, 68 partii z 8 kadencji. Druga tabela zawiera transkrypcje przemówień sejmowych. Do najistotniejszych informacji należą:

|  |  |
| --- | --- |
| Nazwa kolumny w zbiorze | Opis |
| session\_number | Numer sesji |
| date\_ | Data przemówienia |
| Number\_ | Numer porządku obrad |
| speech\_title | Tytuł przemówienia |
| speech\_raw | Tekst przemówienia |

Tabela zawiera 272 321 wierszy. Niektóre wiersze zawierają fragmenty tych samych wystąpień. Unikalnych wystąpień znajduje się z bazie ok. 19 tys.

# Charakterystyka problemu

W ostatnich latach obszar NLP rozwija się bardzo intensywnie. Powstają nowe rozwiązania pozwalające na zaawansowane przetwarzanie i analizę nieustrukturyzowanych danych jakimi są dane tekstowe. Coraz bardziej zaawansowane są modele pozwalające na interpretację jak również generowanie tekstu.

Podstawowe elementy przygotowywania danych do analiz związanych z przetwarzaniem języka naturalnego to:

* tokenizacja, czyli podział tekstu na segmenty, najczęściej pojedyncze słowa,
* stemming ma na celu obcięcie wszystkich przyrostków i przedrostów aby zbliżyć słowo do podstawowej postaci,
* lematyzacja to przypisanie do każdego słowa jego formy podstawowej, która go reprezentuje,
* tworzenie wektorów własnościowych (word embeddings) w uproszczeniu będących wektorową reprezentacją znaczenia danego słowa.

Każdy z tych elementów to istotny element przetwarzania języka naturalnego i każdemu powinna zostać poświęcona odpowiednia uwaga. Po ich przejściu, przetworzone dane można wykorzystać w analizach takich jak:

* modelowanie tematyczne (topic modeling) czyli odkrywanie tematów pojawiających się w dużych zbiorach tekstów a następnie przypisywanie nowym, niezaklasyfikowanym tekstom tematu,
* analiza sentymentu pozwala na określenie jakimi emocjami nacechowany jest dany tekst,
* automatyczne odpowiadanie na pytania na podstawie tekstu,
* generowanie tekstu,
* tłumaczenia,
* budowa chatbotów

i wiele innych.

Większość publikacji w obszarze NLP bazuje na analizach języka angielskiego a ze względu na specyfikę języka polskiego tj. złożoną gramatykę i odmianę fleksyjną, nie wszystkie da się na nim prosto zastosować. Większość dostępnych do pobrania modeli była trenowana na tekstach w języku angielskim.

Dla języka polskiego trudniej jest też o dostępność słowników wspierających analizę sentymentu.

W trakcie analiz szczególny nacisk zostanie położony na zagadnienia:

* modelowanie tematyczne (topic modeling),
* analiza sentymentu (sentiment analysis) oraz
* wizualizacja powyższych zagadnień.

Przygotowanie danych pod analizy również będzie wymagało zastosowania szczególnego podejścia specyficznego dla danych nieustrukturyzowanych.

# Etapy projektu

Projekt będzie realizowany w następujących krokach:

* przygotowanie tekstu do pracy,
  + łączenie danych dotyczących poszczególnych wypowiedzi,
  + dodawanie informacji o autorze,
  + czyszczenie danych.
* przegląd dostępnych technik lematyzacji, wybór optymalnej,
* modelowanie tematyczne – przegląd metod pozwalających na odkrycie segmentów wypowiedzi (tematów),
* analiza metod określenia sentymentu słów,
  + wypracowanie metod na rozszerzenie bazy NAWL (Nencki Affective Word List) zawierającej wymiarowanie ok. 2000 polskich słów w przestrzni opisującej pięć różnych emocji (szcęście, smutek, złość, strach, obrzydzenie)
* budowa modeli przewidujących cechy autora wypowiedzi lub jej senstyment,
  + przygotowanie danych pod modelowanie,
  + wybór metod modelowania dostoswanych do wybranego zagadnienia (co najmniej dwie metody),
  + budowa i analiza otrzymanych modeli.

# Środowisko pracy

Podstawowym środowiskiem przeprowadzania analiz oraz przygotowywania wizualizacji będzie Python.

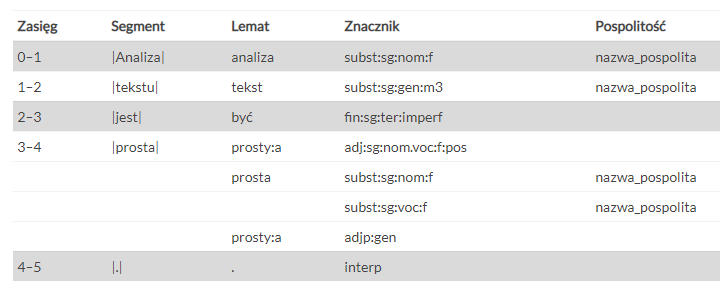
# Przegląd publikacji i projektów

Poniżej przedstawiona jest lista niektórych źródeł, materiałów i projektów, których potencjalne zastosowanie będzie analizowane w niniejszej pracy.

# Projekt Morfeusz

Jednym z szerzej znanych projektów dotyczących rozwoju polskiego NLP jest projekt Instytutu Podstaw Informatyki PAN Morfeusz (<http://morfeusz.sgjp.pl/>), który wykonuje analizę morfologiczną dla języka polskiego.

Poniżej wynik analiza dla przykładowe zdania „Analiza tekstu jest prosta”.



Wyniki można wykorzystać m.in. w procesie lematyzacji tekstu.

# Word embeddings

Metoda reprezentowania słów jako wektory pojawiła się po raz pierwszy w latach sześćdziesiątych ubiegłego stulecia, ale intensywny rozwój tej techniki przypada na ostatnie lata. Niektóre z metod to:

* **word2vec**, Tomas Mikolov et al. (2013) "Efficient Estimation of Word Representations in Vector Space".
* **GloVe**, Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. “GloVe: Global Vectors for Word Representation”
* **fastText**, E. Grave, P. Bojanowsk\*, P. Gupta, A. Joulin, T. Mikolov, “Learning Word Vectors for 157 Languages”.

W szczególności ostatnie podejście pozwalające na pobranie gotowych embeddingów wytrenowanych dla języka polskiego, będzie analizowane w niniejszej pracy.

# spaCy-pl

Projekt IPI PAN dedykowany budowie rozwiązań dla NLP wspierających język polski. W projekcie dostępne są elementy takie jak tokenizacja i lematyzacja (<http://spacypl.sigmoidal.io/#home>).

# Dodatkowe materiały

Inne publikacje, które zostaną przeanalizowane pod kątem wykorzystania w projekcie:

* Adriaan M. J. Schakel, Benjamin J. Wilson “Measuring Word Significance using Distributed Representations of Words”
* Riegel, M., Wierzba, M., Wypych, M. et al. „Nencki Affective Word List (NAWL): the cultural adaptation of the Berlin Affective Word List–Reloaded (BAWL-R) for Polish” Behav Res 47, 1222–1236 (2015) doi:10.3758/s13428-014-0552-1