# Anotacja korpusów oraz osadzenia słów i tekstów
# Część II: Modele przestrzeni wektorowych

Autorzy: Oliwer Krupa, Adam Bednarski, Jan Masłowski, Łukasz Lenkiewicz

October 22, 2024

# Contents

# 1 Word Embeddings

## 1.1 Osadzenia słów w anotowanym korpusie (Word2Vec i Fasttext)

Zadanie polegało na wykonaniu osadzeń słów z wykorzystaniem co najmniej dwóch różnych modeli. Wybraliśmy następujące modele osadzeń:

- **Word2Vec**:
  - dla języka polskiego: `https://dsmodels.nlp.ipipan.waw.pl`
  - dla języka angielskiego: `https://radimrehurek.com/gensim/index.html`
- **Fasttext**:
  - dla języka polskiego: `https://huggingface.co/clarin-pl/fastText-kgr10`
  - dla języka angielskiego: `https://fasttext.cc`

## 1.2 Wizualizacja danych za pomocą t-SNE

Dane zostały zwizualizowane za pomocą techniki t-SNE, z wykorzystaniem interaktywnych wykresów. Każdy punkt na wykresie odpowiada słowu z anotacją. Po najechaniu na punkt można odczytać zarówno słowo, jak i przypisaną do niego etykietę.

## 1.3 Porównanie k-najbardziej podobnych słów dla dwóch modeli

Dla każdego z modeli (Word2Vec i Fasttext) wygenerowano listy k-najbardziej podobnych słów w oparciu o osadzenia zaanotowanych słów. Następnie porównano listy podobieństw, uwzględniając różnice w reprezentacjach przestrzeni wektorowych.

## 1.4 Dyskusja wyników

Otrzymane wyniki pokazują różnice w sposobie, w jaki modele Word2Vec i Fasttext modelują przestrzeń wektorową. Obserwowane różnice w podobieństwach między słowami mogą wynikać z różnych metod treningu oraz charakterystyki korpusów użytych do trenowania modeli.

# 2 Text Embeddings

## 2.1 Osadzenia zdań / tekstów (Fasttext i TF-IDF)

Dla osadzeń całych zdań i tekstów zastosowano dwa modele:

- **Fasttext** (wykorzystano modele z wcześniejszego ćwiczenia),
- **TF-IDF** z biblioteki sklearn.

## 2.2 Wizualizacja osadzeń tekstów za pomocą t-SNE

Podobnie jak w przypadku osadzeń słów, zastosowano t-SNE do wizualizacji osadzeń tekstów, z odniesieniem do etykiet anotacji. Wykresy mają interaktywny charakter, umożliwiając użytkownikowi odczytanie pełnego zdania i jego anotacji po najechaniu na punkt.

## 2.3 Klasteryzacja osadzeń anotowanych zdań

Osadzenia zdań zostały poddane klasteryzacji z wykorzystaniem algorytmu HDBSCAN, a wyniki klasteryzacji przedstawiono w zredukowanej przestrzeni, z wykorzystaniem t-SNE. Interaktywne wykresy umożliwiają przeglądanie przypisanych klastrów i przypisanych do nich zdań.

## 2.4 Dyskusja wyników

Wyniki klasteryzacji oraz wizualizacji t-SNE pokazują, że różne modele osadzania zdań (Fasttext i TF-IDF) generują różne reprezentacje przestrzeni. TF-IDF jest bardziej czuły na częstość występowania słów, natomiast Fasttext efektywniej modeluje podobieństwa semantyczne, co jest szczególnie widoczne w wynikach klasteryzacji. Różnice te mogą wpływać na wydajność modeli w zadaniach klasyfikacji.