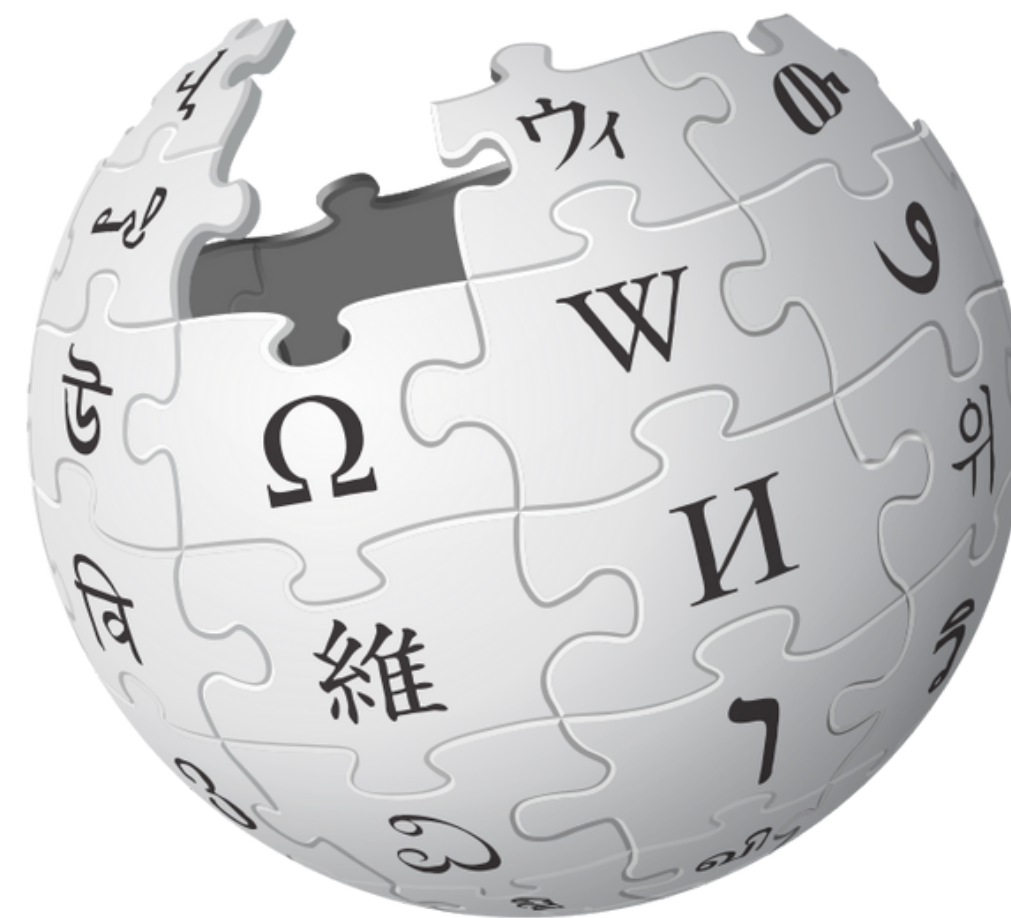


Semantyczna wyszukiwarka

Jakub Baran, Sylwia Zemła, Kacper Ledwoń



WIKIPEDIA

Dziedzina, którą obejmuje wyszukiwarka to **medycyna**.

Baza danych składa się z ok. 40 artykułów z Wikipedii

- w obecnej fazie projektu są to głównie artykuły związane z chorobą astmy.

Artykuły zostały pobrane w formacie **xml** oraz odpowiednio przetworzone za pomocą wyrażeń regularnych **regexp** aby przygotować ich treść do porównania z zapytaniem użytkownika.

Baza danych – MongoDB



- pymongo

Silnik wyszukiwarki – Python



Biblioteki:

- semnatykacja: nltk, scikit-learn, transformers
- dodatkowe: pandas, matplotlib, numpy

GUI – Dear PyGui



Wykorzystane technologie



Transformers

Jak to działa?

Tokenizacja – tekst jest dzielony na mniejsze jednostki, co umożliwia dokładniejsze zrozumienie treści

Wektorowanie – słowa są reprezentowane jako wektory numeryczne. Pozwala to na lepsze zrozumienie relacji między nimi.

Semantyzacja – słowa mają przypisywane znaczenia w kontekście zapytania. Umożliwia to zrozumienie semantyki fraz, uwzględniając kontekst oraz relacje między słowami.

Metryka podobieństw – pomaga ocenić, jak dobrze wyniki wyszukiwania odpowiadają intencjom użytkownika, uwzględniając podobieństwo semantyczne

Testowane parametry

Tokenizacja

- treebank
- casual tokenizer
- biobert (pre-trenowany)
- punkt (pre-trenowany)

Wektoryzacja

- bag of words
- tf-idf

Semantyzacja

- PCA
- LDiA
- SVD

Metryki dystansu/podobieństwa

- Euclidean distance
- Cosine similarity
- Chebyshev distance
- Manhattan distance

Ilość topicków/wektorów semantycznych

- w pierwszych testach ze zbioru: [3, 5, 7, 10]
- w kolejnych z zakresu: [2-35]

Wyniki

Najlepsze połączenia

PCA

- Punt + TF-IDF + manhattan + ok. 23 topicków
- Biobert + TF-IDF + manhattan + ok. 16 topicków

SVD

- Punt + TF-IDF + manhattan + ok. 22 topicków
- Biobert/Treebank/Punkt + TF-IDF + cosine + ok. 8 topicków

LDiA

- Biobert/Punt + TF-IDF + cosine + ok. 15 topicków
- Punkt/Treebank + TF-IDF + cosine/manhattan + ok. 18 topicków

W mniej specjalistycznych
zapytaniach:
Precyzja = 0.7

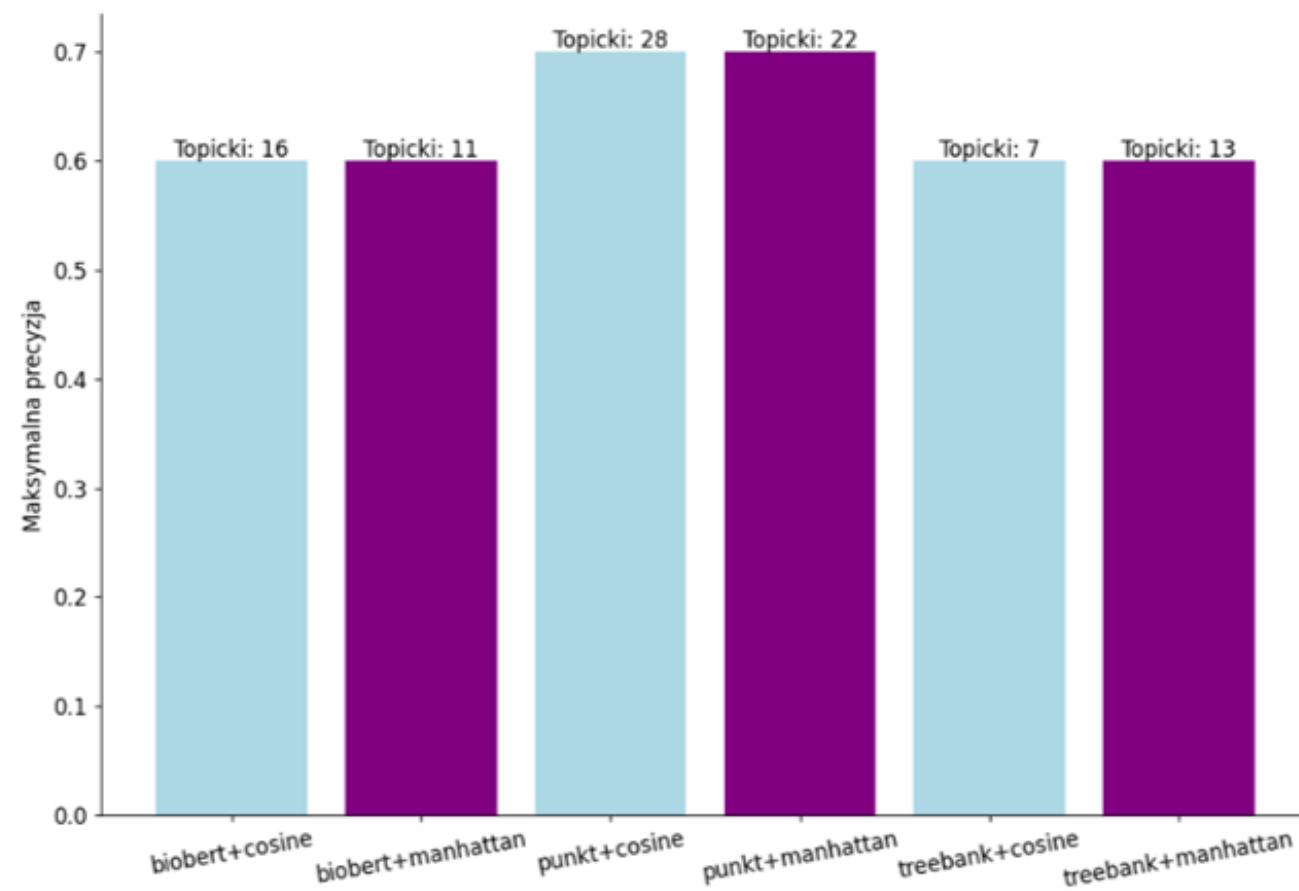
W bardziej specjalistycznych
zapytaniach:
Precyzja = 0.4

W porównaniu z wyszukiwarką
Google

Najlepsze Wyniki

Dla semantyzatora SVD

Najlepsze wyniki zapytania "asthma in children" dla SVD

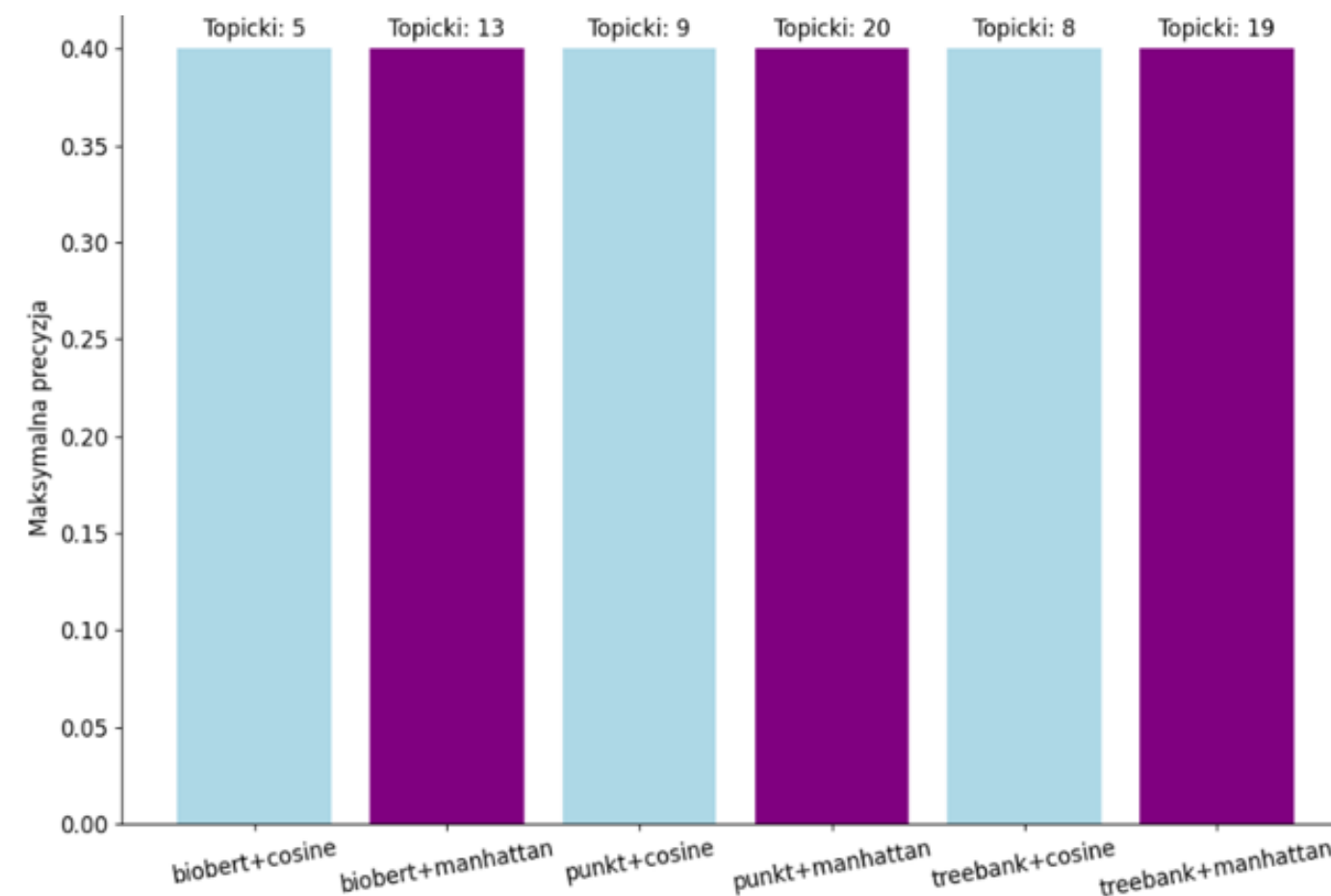


Zapytanie: "asthma in children"

Precyzja: 0.7

7/10 wyników zgadzających się z pierwszymi wynikami Google

Najlepsze wyniki zapytania "asthma inhaler types" dla SVD



Zapytanie: "asthma inhaler types"

Precyzja: 0.4

4/10 wyników zgadzających się z pierwszymi wynikami Google

Menu, wyniki wyszukiwarki

asthma in children	Query
punkt	▼ Tokenizer
tfidf	▼ Vectorizer
svd	▼ Semantizator
manhattan	▼ Similarity Metric
22	▼ Number of topics
Search	
Results:	
Epidemiology of asthma	
Asthma	
Pathophysiology of asthma	
Exhaled nitric oxide	
Asthma trigger	
Asthma-related microbes	
Alcohol-induced respiratory reactions	
Aspirin-exacerbated respiratory disease	
Acute severe asthma	
Reactive airway disease	

Prezentacja działania

Źródła

[1] Semantic search – Wikipedia

https://en.wikipedia.org/wiki/Semantic_search

[2] LDiA – scikit-learn

<https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html>

[3] Truncated SVD – scikit-learn

<https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html>

[4] PCA – scikit-learn

<https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html>

[5] Hobson Lane, Cole Howard, Hannes Hapke, "Natural Language Processing in Action", Manning, 2019, chapter 4

[6] Sebastian Raschka, "Python Machine Learning", Packt Publishing, 2015, chapter 5