| Wydział Informatyki Politechniki Białostockiej | Data: 15.01.2024<br>Przetwarzanie języka naturalnego |
|---|---|
| Projekt, Tworzenie podsumowań tekstu<br>Piotr Zalewski, Kacper Świderek | Prowadzący:<br>dr inż. Tomasz Łukaszuk |

# 1 Przedstawienie zadania projektowego

Celem projektu było opracowanie prostego narzędzia do tworzenia podsumowań dokumentów tekstowych. Opracowane narzędzie daje możliwość wybrania metod/y użytej do tworzenia podsumowania oraz ew. porównania jakości podsumowania na podstawie metryk BERTScore oraz ROUGE.

# 2 Przedstawienie rozszerzenia zadania do oceny 5.0

Rozszerzeniem zadania jest dodanie możliwości łączenia metod tworzących podsumowania, gdzie przy metodach ekstrakcyjnych metody łączone są przy użyciu sumy ważonej wyników jakie każda metoda przypisuje do zdań i następne wybranie N zdań z najwyższym wynikiem, z kolei w przypadku metod abstrakcyjnych polega na zawężaniu obszerności podsumowania stopniowo każdą kolejną metodą (kolejność podawana jest przez użytkownika). Przykładowo, najpierw z tekstu A tworzone jest podsumowanie metodą X o wielkości 20% tekstu pierwotnego, następnie wykorzystywana jest metoda Y, która na podstawie tekstu podsumowania stworzonego za pomocą metody X, tworzy podsumowanie podsumowania obszerności 25% tego podsumowania.

# 3 Wprowadzenie teoretyczne (naukowe) do zagadnienia

Przetwarzanie języka naturalnego to dziedzina skupiająca wiedzę z zakresu językoznastwa oraz informatyki zajmująca się automatyzacją analizy, rozumienia, tłumaczenia i generowania tekstu/języka naturalnego. Podsumowanie tekstu polega na streszczeniu znaczenia tego tekstu, używając innego tekstu, o mniejszym rozmiarze. Wtórnym do zagadnienia tworzenia podsumowań jest zagadnienie analizy tekstu, czyli sposobu pozyskiwania i kodowania informacji o tekście tj. występujące w nim słowa i semantyka. Wykorzystując metody kodowania informacji o tekście można projektować algorytmy, które na podstawie tych reprezentacji generują podsumowania. Wyróżnia się dwie grupy metod służących do tworzenia podsumowań, metody ekstrakcyjne, polegające na wybraniu najbardziej znaczących zdań z tekstu oraz metody abstrakcyjne, polegające na tworzeniu nowych zdań na podstawie tekstu. Z reguły, z racji na większą złożoność, metody abstrakcyjne tworzą lepsze podsumowania, aczkolwiek ze względu na to, że tworzone podsumowania to całkowicie nowy tekst, mogą zaistnieć w nim przekłamania tekstu pierwotnego. Przy metodach ekstrakcyjnych wybierane są jedynie zdania kluczowe z tekstu, przez co są one w tym kontekście bezpieczniejsze.

## 3.1 Metody ekstrakcyjne

### 3.1.1 Wybór na podstawie długości zdania

Naiwną metodą typowania zdań do podsumowania jest tworzenie rankingu na podstawie ilości słów w zdaniu (nie biorąc pod uwagę *stop words* oraz różnych morfologii tych samych słów). Należy zauważyć, że przy tej metodzie ekstrakcyjnej podsumowanie będzie zawsze możliwie nadłuższe.

### 3.1.2 Wybór pierwszego i ostatniego zdania

Kolejną naiwną metodą jest wybór pierwszego i ostatniego zdania bazując na założeniu, że są one w większości przypadków podsumowywujące.

### 3.1.3 TF-IDF

TF-IDF to metoda reprezentacji tekstu oparta na bag-of-words. Przy tej metodzie tekst reprezentowany jest jako macierz której każdy wiesz odpowiada danemu zdaniu z tekstu, a każda kolumna odpowiada danemu słowu. Wartości macierzy wyliczane są za pomocą wzoru (1).

$$w_{ij} = tf_{ij} \cdot log(\frac{N}{df_i}) \tag{1}$$

Gdzie w przypadku opisywanego rozwiązania $tf_{ij}$ oznacza ilość wystąpień słowa $i$ w zdaniu $j$, N oznacza ilość zdań, a $df_i$ oznacza ilość zdań zawierających słowo $i$. Przy tworzeniu podsumowania tworzony jest ranking w którym dla każdego zdania sumowane są otrzymane wartości $w_{ij}$ zgodnie ze wzorem (2). Następnie wybierane jest $K$ zdań z najwyższym wynikiem.

$$S_j = \sum_{i=0}^{M} w_{ij} \tag{2}$$

#### 3.1.4 TextRank

TextRank to algorytm bazujący na algorytmie PageRank opublikowany w 2004 roku [1]. W algorytmie tworzony jest graf którego wierzchołki to, jak w opisywanym rozwiązaniu, zdania, słowa kluczowe lub inne jednostki tekstu. W przypadku zdań krawędzie pomiędzy wierzchołkami reprezentują podobieństwo pomiędzy zdaniami. Podobieństwo może być określane chociażby za pomocą ilości wspólnych słów/tokenów bądź przy użyciu podobieństwa wektorów wziętych np. z macierzy stworzonej za pomocą algorytmu TF-IDF. Następnie przechodząc po grafie wylicza się wynik każdego wierzchołka według wzoru (3) [1].

$$WS(V_i) = 0.15 + 0.85 \cdot \sum_{V_j \in In(V_i)} \frac{w_{ij}}{\sum_{V_k \in Out(V_j)} w_{jk}} \cdot WS(V_j) \tag{3}$$

Jak w algorytmie PageRank we wzorze (3) dany wierzchołek otrzymuje tym wyższy wynik im więcej wierzchołków nań wskazuje. Co więcej, wynik wierzchołków wskazujących na ten wierzchołek również ma wpływ na wynik tego wierzchołka. We wzorze (3) dodatkowo dochodzą wspomniane wagi krawędzi, im wyższa waga tym większe podobieństwo pomiędzy tymi wierzchołkami tym większa kontrybucja do końcowego wyniku. Sortując wartości wierzchołków otrzymuje się listę zdań uporządkowaną malejąco według stopnia ich reprezentatywności całego tekstu.

### 3.2 Metody abstrakcyjne

#### 3.2.1 T5

#### 3.2.2 BART

#### 3.2.3 Pegasus

## 4 Opis realizacji technicznej (wykorzystane biblioteki, dane treningowe)

### 4.1 Biblioteki

Do wykonania zadania skorzystano z następujących bibliotek:

- spacy - segmentacja tekstu, TextRank
- sklearn - TF-IDF
- transformers - abstrakcyjne podsumowania, BART, Pegasus, T5
- rouge_score - wyliczanie rouge_score podsumowania.
- bert_score - wyliczanie bert_score podsumowania.

### 4.2 Dane treningowe

Dane treningowe wytworzono korzystając z biblioteki gutenberg [2] zawierającej w pełni darmowe cyfrowe kopie książek tj. *War and peace*, którym posłużono się do tworzenia podsumowań. Podsumowania wzorcowe tworzono w większości przy pomocy *ChatGPT* ale również ręcznie.

# 5 Instrukcja korzystania z programu

# 6 Przykłady użycia programu (dane wejściowe i otrzymane wyniki)

# Literatura


[1] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 404–411, Barcelona, Spain, 2004. University of North Texas, Association for Computational Linguistics.

[2] Project Gutenberg. Project gutenberg: Free ebooks. https://www.gutenberg.org, 2025. Accessed: 2025-01-04.