Przetwarzanie Języka Naturalnego		
Temat:	**Ekstrakcyjne podsumowanie artykułów przy użyciu grafu podobieństwa zdań**	**Data:**
Autorzy:	**Krzysztof Gonet** **Tomasz Hołda**	**29.05.2022**

Streszczanie (podsumowanie) tekstów przy pomocy uczenia maszynowego lub innych metod statystycznych można zasadniczo podzielić na dwa rodzaje:

1. Podsumowanie ekstrakcyjne – podsumowanie nie jest zupełnie nowym, przeredagowanym tekstem, a składa się ze zdań lub fragmentów zdań zawierających się z tekście źródłowym.
2. Podsumowanie abstrakcyjne – podsumowanie jest tworzone na podstawie tekstu źródłowego, ale składa się z nowych zdań o podobnym znaczeniu jak oryginał.

W naszym projekcie zaprezentowaliśmy podejście do podsumowania ekstrakcyjnego polegajcie na wstępnym przetworzeniu tekstów (głównie przy pomocy biblioteki *nltk*) oraz wykorzystaniu grafu podobieństwa zdań w danym tekście.

Zbiór danych stanowi zestaw 300 tys. artykułów z magazyny *CNN Daily Mail*. W celu uproszczenia i przyspieszenia operacji wykorzystaliśmy jedynie 10 pierwszych tekstów, ale nasza metoda oraz dostępny w notatniku kod źródłowy jest dostosowany do przetwarzania dowolnie dużego zbioru danych (należy jednak pamiętać, że taka operacja zajmie odpowiednio więcej czasu).

Wstępne przetworzenie tekstów polega na usunięciu zbędnych, mało znaczących słów tzw. *stopwords*, zamianie wszystkich liter na małe, poprawieniu słów, które są uznawane za slang, np. „you’re” na „you ‘are” oraz rozbiciu tekstu na pojedyncze zdania. Dodatkowo każde słowo zostało poddane *stemmingowi* czyli procesowi sprowadzenia słów do ogólnego, podstawowego formatu.

Następnie wykorzystaliśmy wytrenowany model *word2vec* pobrany ze zbiorów biblioteki *gensim* do osadzenia słów we wspólnej, 300 wymiarowej przestrzeni, w której słowa o podobnym znaczeniu znajdują się bliżej siebie. Proces ten nosi nazwę *word embedding* i jest powszechnie używany do przejścia do przestrzeni, w której słowa są zakodowane jako wektory o współrzędnych rzeczywistych, określających znaczenie słów.

Następnie każde zdanie z każdego artykułu zostało zakodowane w tej przestrzeni przy użyciu wspomnianego modelu *word2vec*. Szczegóły techniczne zostały szerzej opisane w dołączonym notatniku Jupytera.

Końcowym etapem naszego projektu było stworzenie macierzy podobieństw zdań w obrębie każdego artykułu. Wykorzystaliśmy miarę podobieństwa cosinusowego. Następnie stworzyliśmy graf podobieństwa, którego węzłami są zdania artykułu, a krawędzie reprezentują podobieństwo miedzy zdaniami. Wybraliśmy kilka najbardziej podobnych zdań i z nich stworzyliśmy końcowe podsumowanie.

Wszystkie etapy niniejszego projektu zostały dodatkowo opisane w dołączonym notatniku, który jest gotowy do uruchomienia np. w środowisku Google Colab.