Bartosz Hanc, Paweł Jarosz, Bartłomiej Walczak

High-Dimensional Text Data Visualization

Zbiór danych

Planujemy wykorzystać zbiór ***News Category Dataset***, który zawiera około 200 tysięcy artykułów z portalu HuffPost z lat 2018-2022. Każdy wpis składa się z daty publikacji, nagłówka oraz krótkiego abstraktu. Dane obejmują różnorodne kategorie tematyczne, takie jak polityka, biznes, technologia, zdrowie, sport oraz rozrywka, co pozwala na szeroką analizę tematyczną. Ponieważ istotnym elementem projektu jest analiza zmian tematycznych artykułów w czasie, informacja o dacie publikacji odgrywa kluczową rolę. Umożliwia to śledzenie trendów i dynamiki zmian w obrębie poszczególnych kategorii tematycznych, a także identyfikację okresów wzmożonego zainteresowania danymi zagadnieniami.

Alternatywnie rozważamy użycie zbioru ***MIND*** (Microsoft News Dataset), który jednak obejmuje krótszy okres (kilka tygodni) i nie zawiera jawnie podanej daty publikacji, co wymagałoby dodatkowej obróbki danych. MIND zawiera jednak dodatkowe informacje na temat interakcji użytkowników z artykułami, co mogłoby dostarczyć interesujących insightów na temat popularności i odbioru poszczególnych tematów. Jeśli zdecydujemy się na jego wykorzystanie, konieczne będzie wzbogacenie zbioru o dane czasowe poprzez zewnętrzne źródła lub rekonstrukcję przybliżonych dat na podstawie dostępnych metadanych.

Narzędzia

W celu uzyskania reprezentacji wektorowych tekstu prawdopodobnie wykorzystamy funkcje z modułu *feature_extraction* z biblioteki ***Scikit-learn***, która oferuje szeroki wachlarz metod przetwarzania danych tekstowych, w tym TF-IDF oraz CountVectorizer. Do bardziej zaawansowanych osadzeń tekstów zamierzamy wykorzystać bibliotekę ***Gensim***, która specjalizuje się w modelowaniu tematycznym oraz analizie semantycznej tekstu. Umożliwia ona efektywne trenowanie modeli takich jak word2vec, doc2vec oraz LDA, co pozwali na ekstrakcję i grupowanie tematów w dużych zbiorach danych tekstowych. Gensim cechuje się optymalizacją pod kątem pracy z dużymi korpusami tekstowymi, dzięki czemu nadaje się do przetwarzania zbiorów takich jak News Category Dataset.

W przypadku algorytmów wizualizacji UMAP, t-SNE zapewne wykorzystamy ich implementacje z biblioteki Scikit-learn, natomiast w przypadku PaCMAP i TriMAP – ich oficjalne implementacje udostępnione przez autorów algorytmów.

Interaktywna Wizualizacja

Zgodnie z zaleceniami w treści zadania, chcemy zbudować prosty i interaktywny *front-end* w frameworku **Streamlit**, który będzie pozwalał na eksploracje wyników w czasie rzeczywistym. Możemy wykorzystać interaktywność streamlita do zaangażowania użytkownika w wybór algorytmu do wizualizacji lub wyboru jego hiperparametrów. Dodatkowo w klastrach tekstu możemy wykorzystać funkcjonalność *hover* oraz wyświetlać snippety artykułów, czy *zoom-in/out,* aby lepiej eksplorować klastry. Dodatkowo możemy

wyświetlać ewolucje popularnych tematów zależnie od czasu publikacji (np. sortowane miesiącami).