Dokumentacja Specyfikacji Wymagań

Projekt: Analiza text mining albumu Kendricka Lamara “To Pimp a Butterfly”
Autorzy: Aleksandra Górska, Szymon Skawiński, Maciej Lewandowski Projekt
Data: semestr letni 2024/2025

Wprowadzenie

Niniejszy dokument przedstawia specyfikację wymagań funkcjonalnych i niefunkcjonalnych dla systemu analizującego dane tekstowe metodą Bag of Words. Projekt bazuje na tekstach utworów z albumu *To Pimp a Butterfly* autorstwa Kendricka Lamara. Obejmuje analizę częstości słów, podstawowe wizualizacje danych oraz interfejs umożliwiający użytkownikowi analizę własnych plików tekstowych.

Cele systemu

System ma na celu:

- ➔ Przeprowadzenie analizy tekstowej danych z wykorzystaniem technik eksploracji tekstu (text mining),
- ➔ Umożliwienie użytkownikowi załadowania własnych danych w formacie tekstowym,
- ➔ Przedstawienie wyników analizy w formie czytelnych wizualizacji (np. chmury słów, wykresy słupkowe),
- ➔ Zapewnienie edukacyjnej wartości projektu w zakresie przetwarzania języka naturalnego (NLP).

Wymagania funkcjonalne

1. Umożliwienie użytkownikowi wczytania pliku tekstowego (.txt lub .csv z kolumną tekstową).
2. Automatyczne przetworzenie tekstu (tokenizacja, usunięcie stop words).
3. Analiza częstości słów przy użyciu modelu Bag of Words.
4. Generowanie wykresów słupkowych przedstawiających najczęstsze słowa.
5. Generowanie chmur słów na podstawie danych.
6. Wyświetlanie podstawowych statystyk (liczba słów, unikalnych tokenów).

Wymagania niefunkcjonalne

1. Aplikacja zrealizowana w języku R (w formacie R Markdown).

2. Czytelny kod źródłowy, z komentarzami i podziałem na sekcje.

3. Zastosowanie bibliotek takich jak: tidytext, dplyr, ggplot2, wordcloud.

4. Wizualizacje powinny być przejrzyste i estetyczne.

5. Projekt powinien działać lokalnie na standardowym środowisku RStudio

Interfejsy użytkownika i wymagania dotyczące danych

- ➔ Użytkownik uruchamia analizę poprzez R Markdown. Interfejs umożliwia:
  - Wczytanie danych tekstowych,
  - Uruchomienie przetwarzania tekstu i generowania wykresów.
- ➔ Dane wejściowe: tekst w języku angielskim (plik tekstowy lub kolumna w pliku CSV).
- ➔ Dane wyjściowe: wykresy, statystyki, chmury słów.

Słownictwo dokumentacji

- ➔ Bag of Words (BoW): model reprezentacji tekstu jako zbioru słów i ich częstości.
- ➔ Tokenizacja: proces dzielenia tekstu na pojedyncze słowa.
- ➔ Stop words: słowa bez znaczenia analitycznego (np. "the", "and").
- ➔ Wordcloud: graficzne przedstawienie częstości słów.
- ➔ ggplot2: biblioteka R do tworzenia wykresów.

Przypadki użycia (Use Cases)

1: Wczytanie pliku tekstowego

Aktor: Użytkownik
Opis: Użytkownik wybiera plik do analizy.
Warunki wstępne: Plik ma poprawny format tekstowy.
Efekt końcowy: Plik zostaje wczytany do systemu.

2: Uruchomienie analizy słów

Aktor: Użytkownik
Opis: Użytkownik klika przycisk analizuj.
Efekt końcowy: Wyświetlane są wykresy i statystyki.

Scenariusze użytkownika (User Stories)

1: Jako użytkownik, chcę załadować tekst piosenek, aby przeanalizować ich strukturę słowną.

2: Jako użytkownik, chcę zobaczyć, które słowa występują najczęściej, aby zrozumieć tematykę albumu.

3: Jako użytkownik, chcę zobaczyć graficzną chmurę słów, aby szybko rozpoznać dominujące terminy.

4: Jako użytkownik, chcę mieć możliwość analizy własnych plików, aby zastosować narzędzie w innym kontekście.

Technologe i biblioteki

Projekt zrealizowano w języku R.
Wykorzystano biblioteki: tidytext, dplyr, ggplot2, wordcloud.
Środowisko: RStudio, R Markdown (.Rmd).
Dane: teksty utworów z albumu "To Pimp a Butterfly" Kendricka Lamara.

Wnioski

Dokumentacja obejmuje wszystkie kluczowe wymagania projektowe.
Projekt spełnia cele edukacyjne, umożliwiając analizę tekstów i wizualizację wyników.
Dzięki modularnej strukturze może być łatwo rozszerzany.