Dokumentacja Specyfikacji Wymagań (SRS)

Projekt: Analiza tekstów piosenek Taylor Swift (LDA, asocjacje, bigramy)
Wersja dokumentu: 1.0
Data: 20.05.2025
Autor: Zofia Hasslinger, Antonina Kaniewska

1. Wprowadzenie:

Niniejszy dokument opisuje specyfikację wymagań dla skryptu R, który służy do analizy tektów piosenek Taylor Swift. Skrypt wykorzystuje metodę utajonej alokacji Dirichleta (LDA) do analizy i wizualizacji słów o największej informatywności dla wybranej liczby tematów, analizę asocjacyjną dla wybranych słów oraz znajduje najczęstsze bigramy.

2. Cele systemu:

- Wydobycie dominujących tematów metodą LDA.
- Wizualizacja najbardziej charakterystycznych słów dla każdego tematu.
- Zbadanie współwystępowania słów (asocjacje) i wizualizacja korelacji
- Wykrycie najczęstrzych bigramów

3. Wymagania funkcjonalne:

- **Wczytywanie danych**:
  - Skrypt powinien umożliwiać wczytanie danych z lokalnego pliku .csv.
- **Analiza tekstów**:
  - Skrypt powinien prztworzyć teskt (tokenizacja, konwersja do małych liter, usunięcie znaków interpunkcyjnych i stop words, tworzenie macierzy dokument-term (DTM)).
  - Skrypt powinien przeprowadzać modelowanie tematów z wykorzystaniem metody LDA.
  - Skrypt powiniem umożliwiać uruchomienie modelu LDA z dowolnie zadeklarowaną liczbą tematów.
  - Skrypt powinien znajdować wyrazy często współwystępujące z podanym słowem (asocjacje).
  - Skrypt powinien wykrywać najczęściej występujące bigramy.
- **Wizualizacja danych**:
  - Skrypt powinien powinien umożliwiać wygenerowanie chmury słów (wordcloud) z globalnej częstości słów.
  - Skrypt powinien umożliwiać automatyczne tworzenie wykresu słów najbardziej charakterystycznych dla każdego tematu (z użyciem ggplot).
  - Skrypt powinien umożliwiać tworzenie wykresów słów najbardziej skorelowanych z podanym wyrazem (wykres lizakowy oraz lizakowy z natężeniem z użyciem ggplot).

- **Agregacja danych**:
  - Skrypt powinien tworzyć ramkę danych (data.frame) z wynikami zliczania słów.

4. Wymagania niefunkcjonalne:

- **Wydajność**:
  - System powinien przetwarzać dane (dyskografie) w czasie krótszym niż 60 sekund.
- **Bezpieczeństwo**
  - System powinien zapewnić poprawność danych wyjściowych.
- **Niezawodność**:
  - Skrypt powinien poprawnie obsługiwać różne formaty danych tekstowych.
  - Skrypt powinien poprawnie obsługiwać brakujące wartości.
- **Użyteczność**:
  - Wykresy powinny być czytelne, estetyczne i zawierać odpowiednie etykiety.
- **Kompatybilność**:
  - Skrypt powinien być kompatybilny z R w wersji 4.0 lub nowszej.
  - Skrypt powinien korzystać z bibliotek tm, tidyverse, tidytext, ggplot2, wordcloud, topicmodels.

5. Interfejsy użytkownika:

- **Wejście**:
  - Plik .csv
- **Wyjście**:
  - Chmura najczęstszych słów (workcloud).
  - Wykresy tematów ggplot2.
  - Wykresy asocjacji ggplot2.
  - Lista bigramów

6. Wymagania dotyczące danych:

- Skrypt zakłada, że dane tekstowe są w języku angielskim.
- Skrypt nie obsługuje analizy sentymentu dla innych języków.
- Pliki muszą być poprawnie zakodowane (UTF-8)

Słownictwo dokumentacji:

- **LDA –** Latent Dirichlet Allocation, algorytm do analizy tematów.
- **Bigram** – Para kolejnych słów.
- **Asocjacje** – Skojarzenia między słowami na podstawie ich współwystępowania.
- **Chmura słów** – wizualizacja częstości występowania słów.
- **Stop words** – najczęściej występujące, mało informacyjne słowa (np. "the", "and")

Przypadki użycia (use cases)

• Użytkownik:

- wczytuje plik .csv.
- uruchamia analizę
- wyświetla wyniki

• Skrypt/system:

- przetwarza teksty
- buduje model LDA i przypisuje tematy
- wykrywa asocjacje i bigramy
- generuje wizualizacje (wykresy i chmurę słów)

Testowe przypadki użycia

- Test z plikiem .csv zawierającym brakujące wartości.
- Test ze zmienioną liczbą tematów LDA.
- Test z innymi słowami do zbadania asocjacji.

Scenariusze użytkownika (user stories)

Scenariusz 1: Przygotowanie koncertu

- **Jako:** organizator koncertu Taylor Swift
- **Chcę:** przeanalizować dominujące tematy w jej tekstach
- **Aby:** dopasować scenografię i oprawę koncertu do przekazu piosenek.

Kryteria akceptacji:

- Użytkownik może wczytać plik z tekstami piosenek.
- Skrypt analizuje tematy piosenek za pomocą LDA.
- Wyniki prezentowane są w formie czytelnych list i wykresów.
- Organizator może zidentyfikować często występujące motywy (np. miłość, zemsta, nostalgia).

Scenariusz 2: Dobór utworów do setlisty

- **Jako:** kierownik artystyczny
- **Chcę:** sprawdzić, jak często w piosenkach występują określone tematy lub słowa
- **Aby:** stworzyć setlistę budującą konkretną narrację koncertu

Kryteria akceptacji:

- Użytkownik może wczytać dane tekstowe z interesującymi go piosenkami.
- Użytkownik może wybrać słowo kluczowe.
- Skrypt pokazuje, z jakimi innymi słowami występuje.
- Wyniki umożliwiają stworzenie spójnej i przemyślanej kolejności utworów.