DOKUMENTACJA I SPECYFIKACJA WYMAGAŃ (SRS)

Projekt: Analiza pliku CSV, analiza częstości słów, Topic Modelling
Data: 07.06.2025
Autorzy: Dominik Sztychmiler, Alicja Rączka, Liliana Buchner

1.Wprowadzenie

Niniejszy dokument opisuje specyfikację wymagań dla skryptu R, który realizuje analizę Topic Modelling oraz analizę częstości słów na podstawie zawartości pliku csv, który zawiera podział na pozytywne oraz negatywne opinie na temat hotelu. System używa techniki czyszczenia tekstu. Dodatkowo generowane są wizualizacje częstości słów w postaci chmury słów oraz wykresów częstości słów w czterech tematach oddzielnie dla opinii pozytywnych i negatywnych.

2.Cele systemu

- Przeprowadzenie analizy częstości słów
- Przeprowadzenie analizy Topic Modelling dla pozytywnych i negatywnych opinii
- Wizualizacja częstości słów
- Podział na obszary tematyczne oddzielnie dla pozytywnych i negatywnych opinii

3. Wymagania funkcjonalne:

- **Wczytanie danych**
    - Skrypt powinien umożliwiać wczytanie danych z lokalnego pliku .csv
- **Analiza Topic Modelling**
    - Skrypt powinien przeprowadzić czyszczenie tekstu
    - Skrypt powinien zapisać dane w formie oczyszczonych korpusów
    - Skrypt powinien poprawnie zaimplementować metodę LDA dla 4 obszarów tematycznych
    - Skrypt powinien wyodrębnić najbardziej charakterystyczne słowa dla zidentyfikowanych tematów
- **Wizualizacja wyników**

- Skrypt powinien wyświetlić wykresy z podziałem na grupy tematyczne oddzielnie dla pozytywnych i negatywnych opinii
- Skrypt powinien wyświetlić chmury najczęstszych słów oddzielnie dla pozytywnych i negatywnych słów

4.Wymagania niefunkcjonalne

- **Wydajność**
  - System powinien być wydajny czasowo
- **Użyteczność**
  - System powinien mieć przejrzysty i intuicyjny interfejs użytkownika
  - Wyniki analizy powinny być przedstawione czytelnie
- **Niezawodność**
  - System powinien wykrywać i odpowiednio obsługiwać błędy
- **Bezpieczeństwo**
  - System powinien zapewniać bezpieczeństwo i poprawność danych wyjściowych
- **Kompatybilność**
  - System powinien być kompatybilny z wersją R 4 lub nowszą
  - System powinien obsługiwać biblioteki tidyverse, tidytext, topicmodels, wordcloud, RColorBrewer, tm

5. Interfejsy użytkownika:

- **Wejście:**
  - Plik z danymi w formacie .csv.
- **Wyjście:**

- Chmury słów i wykresy:
  - Chmura słów dla recenzji pozytywnych.
  - Chmura słów dla recenzji negatywnych.
  - Wykresy Top Terms w tematach LDA dla recenzji pozytywnych i negatywnych.
- Lista dziesięciu najczęściej występujących słów w recenzjach pozytywnych i negatywnych.

6. Wymagania dotyczące danych

- System przyjmuje dane w formacie .csv, kodowane w UTF-8**.**
- Skrypt zakłada, że dane tekstowe są w języku angielskim.
- System wymaga kolumn:
  - Hotel_Name.
  - Postive_Review.
  - Negative_Review.
- System nie obsługuje plików większych niż 100 MB.
- System usuwa puste recenzje oraz recenzje zawierające wyłącznie stopwords lub znaki specjalne.

Słownictwo dokumentacji:

- **Recenzja**: Opinia użytkowników na temat hotelu, zawarta w pliku CSV w kolumnach Negative_Review i Positive_Review.

- **Czyszczenie tekstu:** Proces usuwania z tekstu znaków specjalnych, linków, cyfr, stopwords, wielokrotnych spacji oraz konwersji tekstu do małych liter.
- **LDA (Latent Dirichlet Allocation):** Algorytm statystyczny służący do modelowania tematów w zbiorze dokumentów, który przypisuje tematy na podstawie współwystępowania słów w tekstach.

Przypadki użycia (use cases)

- **Użytkownik:**
    - Wczytuje plik .csv z recenzjami hotelowymi.
    - Uruchamia analizę tekstu i modelowanie tematów.
    - Wyświetla wykresy i chmury słów.
    - Przegląda listę najczęściej używanych słów w recenzjach pozytywnych i negatywnych.

- **Skrypt/system:**
    - Wczytuje dane z pliku .csv.
    - Czyści dane tekstowe z niepotrzebnych znaków, linków, stopwords itd.
    - Dzieli dane na recenzje pozytywne i negatywne.
    - Tworzy korpusy i przekształca je do formatu akceptowanego przez algorytm LDA.

- Wykonuje topic modeling osobno dla pozytywnych i negatywnych recenzji.
- Wyświetla w konsoli dziesięć najczęściej używanych słów w każdej grupie recenzji.
- Generuje:
  - Wykresy Top Terms dla każdego z tematów.
  - Chmurę słów dla recenzji pozytywnych.
  - Chmurę słów dla recenzji negatywnych.

Testowe przypadki użycia:

- Test z plikiem .csv zawierającym tylko pozytywne recenzje.
- Test z plikiem .csv zawierającym tylko negatywne recenzje.
- Test z plikiem .csv zawierającym zarówno pozytywne, jak i negatywne recenzje.
- Test z plikiem .csv zawierającym puste lub niepełne recenzje.

Scenariusze użytkownika (user stories)

Scenariusz 1: Przeprowadzenie analizy różnicy opinii między hotelami

- **Jako**: Analityk sieci hoteli w agencji turystycznej
- **Chcę**: Przeprowadzić analizę, jakie tematy i słowa pojawiają się w recenzjach dla poszczególnych hoteli
- **Aby**: Przedstawić raport, które hotele powinny być rozważane przy tworzeniu ofert wyjazdów.

Kryteria akceptacji:

- System umożliwia filtrowanie danych po kolumnie Hotel_Name.

- Użytkownik może wyświetlić wyniki dla wskazanego hotelu.
- Wygenerowane chmury słów i wykresy przedstawiają wybrany zakres danych.
- Użytkownik ma dostęp do Top Terms i chmur słów z podziałem na hotele.

Scenariusz 2: Wyszukiwanie najlepszego hotelu na wakacje

- **Jako**: Turysta planujący wakacje bez udziału agencji turystycznych
- **Chcę**: Przeanalizować dostępne na portalach hoteli ze względu na ich recenzje
- **Aby**: Wybrać hotel, który najlepiej odpowiada moim oczekiwaniom i unikać miejsc z powtarzającymi się problemami.

Kryteria akceptacji:

- Użytkownik może wczytać plik .csv z recenzjami hoteli.
- System przeprowadza analizę tekstów oraz analizę tematów dla pozytywnych i negatywnych recenzji.
- System generuje chmury słów i wykresy tematów, dzięki którym użytkownik może zorientować się, co jest najczęściej chwalone i krytykowane w danym hotelu.
- Użytkownik może porównać kilka hoteli z uwagi na pozytywne i negatywne opinie.