Dokumentacja Specyfikacji wymagań (SRS)

Projekt: Analiza Asocjacji i Sentymentu

Wersja dokumentu: 1.0

Data: 31.05.2025

Autor: Maria Żukowska, Emma Panasiuk, Lena Skrzypiec

1.Wprowadzenie:

Niniejszy dokument opisuje specyfikację wymagań dla skryptu R, który realizuje analizę asocjacji i sentymentu na podstawie pliku CSV zawierającego dane tekstowe. System umożliwia użytkownikowi wybór słów, których asocjacje chciałby zbadać,
a następnie dla najsilniej skorelowanych z nimi wyrazów przeprowadza analizę sentymentu. Skrypt wykorzystuje współczynnik korelacji Pearsona do analizy asocjacji oraz słowniki sentymentu (NRC, Bing, AFFIN) do oceny sentymentu. Dodatkowo generowane są wizualizacje wykresów asocjacji oraz rodzaju sentymentu według słowników.

2.Cele systemu:

- Wczytanie danych wejściowych (plik CSV) z odpowiednim kodowaniem (UTF-8).
- Przetwarzanie i oczyszczanie tekstu (normalizacja, tokenizacja).
- Usunięcie nieistotnych słów (stopwords).
- Przeprowadzenie analizy i wizualizacja asocjacji z wykorzystaniem współczynnika korelacji Pearsona dla wybranych przez użytkownika słów.
- Przeprowadzenie analizy i wizualizacja sentymentu z wykorzystaniem różnych słowników dla najsilniej skorelowanych wyrazów z każdym z wybranych słów.
- Przeprowadzenie analizy i wizualizacja skumulowanego sentymentu dla całego pliku tekstowego z wykorzystaniem różnych słowników.

3. Wymagania funkcjonalne

- **Wczytywanie danych:**
    - Skrypt powinien umożliwiać wczytanie danych tekstowych z lokalnego pliku CSV.
    - Skrypt powinien obsługiwać kodowanie UTF-8.
- **Przetwarzanie i czyszczenie tekstu**
    - Skrypt powinien umożliwiać utworzenie korpusu dokumentów.
    - Skrypt powinien umożliwiać konwersję tekstu na małe litery.
    - Skrypt powinien umożliwiać usunięcie pustych elementów, cyfr, znaków interpunkcyjnych oraz znaków specjalnych.
    - Skrypt powinien umożliwiać usunięcie stopwords z pakietów tidytext i tm.
    - Skrypt powinien umożliwiać generowanie term-document matrix (TDM).
    - Skrypt powinien umożliwiać wykonanie tokenizacji.
- **Analiza asocjacji:**
    - Skrypt powinien przeprowadzać analizę asocjacji z wykorzystaniem współczynnika korelacji Pearsona.
- **Analiza sentymentu:**
    - Skrypt powinien wykorzystywać słowniki sentymentów: NRC, Bing, AFFIN.
    - Skrypt powinien umożliwiać dopasowanie słów do słowników i zliczanie sentymentów.

- **Wizualizacja danych:**
    - Skrypt powinien umożliwiać wizualizację wyników (wykresy ggplot2).
    - Skrypt powinien generować wykresy asocjacji.
    - Skrypt powinien generować wykresy sentymentu oraz skumulowanego sentymentu dla każdego słownika.
- **Agregacja danych:**
    - Skrypt powinien usuwać brakujące wartości (NA).

4. Wymagania niefunkcjonalne:

- **Wydajność:**
    - Skrypt powinien być w stanie przetwarzać pliki CSV o wielkości do 500 KB w czasie poniżej 1 minuty.
- **Bezpieczeństwo:**
    - System powinien zapewnić poprawność danych wyjściowych.
- **Niezawodność:**
    - Skrypt powinien poprawnie obsługiwać brakujące wartości.
- **Użyteczność:**
    - Wykresy powinny być czytelne i zawierać odpowiednie etykiety.
    - Skrypt powinien umożliwiać wykonanie wizualizacji z użyciem ggplot2 i motywu theme_gdocs dla lepszej czytelności.
- **Kompatybilność:**
    - Skrypt powinien być kompatybilny z R w wersji 4.0 lub nowszej.
    - Skrypt powinien korzystać z bibliotek tm, tidyverse, tidytext, worldcloud, ggplot2, ggthemes, textdata.

5. Interfejsy użytkownika:

- **Wejście:**
    - Plik CSV zawierający dane tekstowe
- **Wyjście:**
    - Wykresy asocjacji (ggplot2)
    - Wykresy słupkowe rodzaju sentymentu wg słowników (AFFIN, Bing, NRC)

6. Wymagania dotyczące danych:

- Skrypt zakłada, że dane są w języku angielskim
- Skrypt nie obsługuje analizy sentymentu oraz analizy asocjacji dla innych języków.
- Skrypt nie obsługuje analizy sentymentu i asocjacji dla danych tekstowych z innych źródeł niż pliki CSV, które zawierają jedną kolumnę danych tekstowych bez nagłówka.

Słownictwo dokumentacji:

- **Token:** pojedynczy element tekstu (słowo)
- **Stopwords:** słowa niewnoszące wartości semantycznej do analizy.
- **Asocjacje:** mierzone za pomocą współczynnika korelacji Pearsona powiązania między słowami.
- **Sentyment:** emocjonalne nacechowanie słów.
- **Słownik sentymentu:** lista słów i ich ocen wg sentymentu.
- **Skumulowany sentyment:** suma ocen sentymentu dla całego pliku tekstowego.

Przypadki użycia (use cases)

- Użytkownik:
    - Wczytuje plik CSV zawierający dane tekstowe
    - Wybiera słowa do analizy
    - Uruchamia analizę
    - Wyświetla wyniki
- Skrypt/system:
    - Przetwarza tekst
    - Analizuje asocjacje dla wybranych słów
    - Generuje wykresy asocjacji dla wybranych słów
    - Analizuje sentyment dla najsilniej skorelowanych z każdym słowem wyrazów
    - Generuje wykresy sentymentu dla najsilniej skorelowanych z każdym słowem wyrazów
    - Analizuje sentyment dla całego pliku tekstowego
    - Generuje wykresy sentymentu dla całego pliku tekstowego

Testowe przypadki użycia:

- Test z plikiem CSV zawierającym tekst o pozytywnym sentymencie.
- Test z plikiem CSV zawierającym tekst o pozytywnym sentymencie.
- Test z plikiem CSV zawierającym tekst o mieszanym sentymencie.
- Test z plikiem CSV zawierającym brakujące wartości.
- Test z plikiem CSV zawierającym znaki specjalne.

Scenariusze użytkownika (user stories)

Scenariusz 1: Analiza recenzji gości dotyczących usług hotelowych

- **Jako:** Analityk danych w branży hotelarskiej
- **Chcę:** Przeanalizować recenzje gości dotyczące ich pobytu w hotelu zamieszczane w serwisie internetowym.
- **Aby:** Zidentyfikować poziom zadowolenia gości ze świadczonych przez hotel usług oraz znaleźć obszary wymagające poprawy.

Kryteria akceptacji:

- Użytkownik może wczytać dane tekstowe z serwisu internetowego do pliku CSV.
- Skrypt przeprowadza analizę asocjacji i sentymentu oraz generuje wykresy sentymentu.
- Użytkownik może monitorować ogólny sentyment.
- Użytkownik może badać sentyment związany z konkretnymi obszarami, którymi jest zainteresowany.
- Użytkownik może generować raporty z analizy sentymentu.