**Dokumentacja Specyfikacji Wymagań (SRS)**

**Projekt: Pełna analiza Text mining jednego pliku tekstowego**

**Wersja: 1.0**

**Data: 07.06.2025**

**Autorzy: Julia Szmidt, Bartosz Kozarzewski**

1. **Wprowadzenie**

Niniejszy dokument zawiera specyfikację wymagań dla projektu w R wykonującego pełną analizę Text mining jednego pliku tekstowego. Program umożliwia wczytanie pliku tekstowego, przetwarzanie danego tekstu, w tym tokenizację oraz stemming. Dalej analizę częstości słów, analizę sentymentu z użyciem słowników AFINN, Bing, NRC, Loughran, klastrowanie z podziałem na fragmenty. System domyślnie generuje również wykresy i raporty wizualne.

1. **Cele systemu**

System umożliwia:

* Wczytanie i przetworzenie pliku tekstowego w kodowaniu UTF-8
* Oczyszczanie tekstu (usunięcie interpunkcji, liczb, znaków specjalnych, stopwords)
* Tokenizację i stemming tekstu
* Analizę częstości występowania słów oraz ich wizualizację (chmury słów, wykresy)
* Analizę sentymentu z użyciem słowników: AFINN, Bing, NRC, Loughran
* Klastrowanie fragmentów tekstu
* Generowanie wykresów opartych na wynikach (ggplot2, theme\_gdocs, RColorBrewer)

**3. Wymagania funkcjonalne**

**Wczytywanie danych**

* Skrypt powinien umożliwiać wczytywanie danych tekstowych z lokalnego pliku z rozszerzeniem .txt.
* Skrypt powinien obsługiwać kodowanie UTF-8

**Przetwarzanie tekstu**

* Zamiana wszystkich znaków na małe litery
* Usunięcie interpunkcji, liczb i zbędnych znaków specjalnych
* Usunięcie słów bezwartościowych (stopwords)
* Tokenizacja tekstu (unnest\_tokens i strsplit)
* Stemming słów z użyciem biblioteki SnowballC

**Analiza częstości słów**

* Zliczanie i sortowanie wystąpień słów
* Generowanie wykresów
* Możliwość filtrowania najczęściej występujących słów

**Analiza sentymentu**

* Analiza sentymentu przy użyciu słowników: AFINN, Bing, NRC, Loughran
* Przypisanie słowom pozytywnych i negatywnych emocji
* Filtrowanie słów o silnym sentymencie (AFINN)
* Tworzenie zestawień najczęstszych słów według sentymentu
* Obsługa wielu słowników przez jedną funkcję

**Klastrowanie**

* Dzielenie tekstu na fragmenty (co 100 słów)
* Dobór liczby klastrów
* Korzystanie z metody K-średnich

**Wizualizacja**

* Wykresy słupkowe i liniowe w ggplot2
* Wykresy dla kluczowych słów według sentymentu
* Wykresy z wykorzystaniem theme\_gdocs i palet RColorBrewer
* Wizualizacja klasteryzacji
* Interaktywna tabela

1. **Wymagania niefunkcjonalne**

**Wydajność**

* Analiza fragmentów do 10 000 słów powinna zakończyć się w czasie poniżej 15 sekund

**Bezpieczeństwo**

* System powinien zapewnić poprawność danych wyjściowych

**Niezawodność**

* Skrypt powinien obsługiwać każde pliki tekstowe, niezależnie od znaków specjalnych

**Użyteczność**

* Wszystkie wykresy powinny zawierać etykiety i tytuły
* Skrypt powinien generować chmurę słów z wykorzystaniem kolorystyki RColorBrewer

**Kompatybilność**

* Kompatybilność z R ≥ 4.3.3
* Wymagane biblioteki: tm, SnowballC, cluster, ggplot2, wordcloud, factoextra, RColorBrewer, dplyr, ggrepel, DT, tidytext, stringr, SentimentAnalysis, ggthemes, tidyverse, textdata, forcats

1. **Interfejs użytkownika**

**Wejście:**

* Plik formatu .txt zawierający tekst
* Plik słowników w formacie .csv

**Wyjście:**

* Wykresy słupkowe (dla sentymentu i częstości)
* Chmura słów
* Wizualizacja klasteryzacji
* Interaktywna tabela
* Raport z wykresami w formacie HTML

1. **Wymagania dotyczące danych**

* Skrypt wymaga danych tekstowych w języku angielskim
* Skrypt obsługuje wyłącznie pliki tekstowe .txt z kodowaniem UTF-8
* Skrypt wykonuje analizę sentymentu tylko dla słowników AFINN, Bing, NRC i Loughran

1. **Słownictwo dokumentacji**

| Termin | Znaczenie |
| --- | --- |
| Token | Pojedyncze słowo lub element tekstu |
| Stopwords | Słowo bez wartości semantycznej (np. "and", "the") |
| Sentyment | Pozytywny lub negatywny ładunek emocjonalny słów |
| Stemming | Sprowadzanie słów do ich rdzenia |
| Chmura słów | Graficzna reprezentacja częstotliwości słów |
| Klastrowanie | Pojedyncza grupa podobnych danych |

1. **Przypadki użycia**

**Użytkownik:**

* Wczytuje dane z pliku
* Uruchamia analizę
* Wyświetla wyniki
* Generuje wykresy i inne wizualizacje

**Skrypt/system:**

* Przetwarza i oczyszcza tekst
* Analizuje częstość słów
* Analizuje sentyment tekstu przy użyciu słowników
* Generuje chmurę słów
* Generuje wykres częstości słów
* Generuje wykresy sentymentu według słowników
* Generuje wizualizację klasteryzacji
* Generuje interaktywną tabelę

**Testowe przypadki użycia:**

* Przeprowadzenie testu z plikiem .txt zawierającym znaki specjalne
* Przeprowadzenie testu z plikiem .txt o pozytywnym sentymencie
* Przeprowadzenie testu z plikiem .txt o negatywnym sentymencie
* Przeprowadzenie testu z plikiem .txt o neutralnym sentymencie
* Przeprowadzenie testu z plikiem .txt o mieszanym sentymencie

**8. Scenariusze użytkownika**

**Scenariusz 1:** Analiza książki do pracy badawczej

* **Jako:** student literatury
* **Chcę:** przeanalizować dominujące słowa i emocje w powieści
* **Aby:** napisać analizę nastroju bohaterów i stylu autora

**Kryteria akceptacji:**

* Użytkownik może wczytać plik z powieścią
* Skrypt przeprowadza analizę sentymentu za pomocą różnych słowników
* Skrypt porównuje wyniki z różnych słowników
* Skrypt przeprowadza klastrowanie fragmentami tekstu
* Użytkownik może zidentyfikować ogólny sentyment powieści i przeanalizować nastroje bohaterów

**Scenariusz 2:** Krytyka twórczości G. Orwella

* **Jako:** krytyk literacki
* **Chcę:** przeanalizować powieść pod względem nacechowania emocjonalnego
* **Aby:** zwiększyć skuteczność swojej działalności

**Kryteria akceptacji:**

* Użytkownik może wczytać plik z powieścią
* Skrypt przeprowadza analizę częstości słów
* Skrypt przeprowadza analizę sentymentu za pomocą różnych słowników
* Skrypt porównuje wyniki z różnych słowników
* Użytkownik może zidentyfikować ogólny charakter powieści i dokonać odpowiedniej oceny