Dokumentacja Specyfikacji Wymagań (SRS)

Projekt: Pełna Analiza Text Mining z wykorzystaniem modelu Bag of Words
Wersja dokumentu: 1.0
Data: 30.05.2025
Autor: [Jakub Hejduk, Antoni Radecki]

1. Wprowadzenie

Niniejszy dokument opisuje specyfikację wymagań dla skryptu R realizującego analizę text mining z wykorzystaniem modelu Bag of Words (BoW). Skrypt umożliwia przetwarzanie tekstu, analizę częstości słów, generowanie chmur słów oraz przeprowadzenie analizy sentymentu z wykorzystaniem różnych słowników (AFINN, Bing, NRC, Loughran, GI, HE, LM, QDAP). Dokumentacja starannie opisuje funkcjonalność skryptu, wymagania systemu oraz scenariusze użycia w praktyce.

2. Cele systemu:

- Wczytanie tekstu z pliku .txt w kodowaniu UTF-8.
- Przetwarzanie i oczyszczanie tekstu (usunięcie znaków specjalnych, interpunkcji, liczb, stopwords).
- Tokenizacja, stemming i uzupełnienie rdzeni słów.
- Zliczenie częstości występowania słów.
- Generowanie chmury słów i wykresów częstości.
- Analiza sentymentu:
  ++Wykorzystanie słowników z plików CSV (AFINN, Bing, NRC, Loughran).
  ++Wykorzystanie słowników z pakietu SentimentAnalysis (GI, HE, LM, QDAP).
- Wizualizacja wyników za pomocą wykresów słupkowych i czasowych.
- Generowanie wykresów porównujących sentyment z różnych słowników.
- Analiza zmian sentymentu w czasie.

3. Wymagania funkcjonalne

- Wczytywanie danych
  ++Skrypt powinien wczytywać pliki tekstowe (.txt) w kodowaniu UTF-8.
  ++Skrypt powinien obsługiwać kodowanie UTF-8.
- Przetwarzanie tekstu
  ++Normalizacja tekstu (małe litery, usunięcie apostrofów, liczb, interpunkcji).

++Tokenizacja i usunięcie stopwords.

++Stemming i uzupełnienie rdzeni słów.

- Analiza częstości słów

++Zliczenie częstości występowania słów.

++Generowanie chmury słów z wykorzystaniem pakietu wordcloud.

++Wyświetlanie tabeli z najczęściej występującymi słowami.

- Analiza sentymentu

++Wczytanie słowników z plików CSV (AFINN, Bing, NRC, Loughran).

++Przeprowadzenie analizy sentymentu z wykorzystaniem słowników.

++Filtrowanie słów o określonym sentymencie (pozytywnym/negatywnym).

++Wykorzystanie słowników z pakietu SentimentAnalysis (GI, HE, LM, QDAP).

++Konwersja wartości sentymentu na kierunkowe (pozytywny/negatywny/neutralny).

- Wizualizacja danych

++Generowanie wykresów słupkowych dla sentymentu.

++Tworzenie wykresów zmian sentymentu w czasie.

++Porównanie wyników z różnych słowników na jednym wykresie.

- Agregacja danych

++Agregacja wyników z różnych słowników w jedną ramkę danych.

++Usunięcie brakujących wartości (NA).

4. Wymagania niefunkcjonalne

- Wydajność

++Analiza pliku tekstowego o długości 1000 zdań powinna trwać nie dłużej niż 20sekund.

- Bezpieczeństwo

++Skrypt powinien zapewniać poprawność danych wyjściowych.

- Niezawodność

++Obsługa różnych formatów danych tekstowych.

++Poprawna obsługa brakujących wartości.

- Użyteczność

++Wykresy powinny być czytelne i zawierać etykiety.

++Wykonanie wizualizacji z użyciem ggplot2

++Możliwość dostosowania motywów wizualizacji z użyciem theme_gdocs.

- Kompatybilność

++Skrypt powinien działać w R w wersji 4.0 lub nowszej.

++Wymagane pakiety: tm, tidytext, stringr, ggplot2, ggthemes, SentimentAnalysis, SnowballC, tidyverse.

5. Interfejsy użytkownika

- Wejście:
  ++Plik tekstowy (.txt) z danymi do analizy.
  ++Pliki słowników w formacie CSV (AFINN, Bing, NRC, Loughran).
- Wyjście:
  ++Tabela z częstością występowania słów.
  ++Chmura słów.
  ++Wykresy słupkowe i czasowe sentymentu.

6. Wymagania dotyczące danych

- Dane tekstowe muszą być w języku angielskim.
- Skrypt nie obsługuje analizy sentymentu dla innych języków.
- Skrypt wykorzystuje słowniki sentymentów dostępne w plikach .CSV oraz w pakiecie SentimentAnalysis.
- Skrypt nie obsługuje analizy sentymentu dla danych tekstowych z innych źródeł niż pliki .txt.
- Skrypt nie obsługuje plików większych niż 100 MB.

7. Słownictwo dokumentacji

- **Token:** Pojedynczy element tekstu (słowo).
- **Stopwords:** Słowa o małej wartości semantycznej.
- **Sentyment:** Emocjonalne nastawienie w tekście.
- **Stem:** forma słowa po sprowadzeniu go do rdzenia.
- **Stem Completion:** Rdzeń słowa po procesie stemmingu.
- **Wartości kierunkowe (Directional Sentiment):** Konwersja ciągłej wartości sentymentu na kategorie: "pozytywny", "negatywny", "neutralny".
- **Segmentacja tekstu:** Podział tekstu na fragmenty o stałej długości do analizy temporalnej.

8. Przypadki użycia (use cases)

- Użytkownik:
    ++Wczytuje plik tekstowy .txt.
    ++Uruchamia analizę.
    ++Wyświetla wyniki analizy.
    ++Generuje wykresy sentymentu, raport html i chmurę słów.

- Skrypt/system:
    ++Przetwarza i oczyszcza tekst.
    ++Analizuje sentyment tekstu przy użyciu słowników.
    ++Tworzy chmurę słów.
    ++Tworzy wykresy przedstawiające skumulowany sentyment .
    ++Tworzy wykres porównujący typy sentymentu według różnych słowników.
    ++Tworzy wykresy obrazujące zmiany sentymentu w czasie.

Testowe przypadki użycia:

• Przeprowadzenie testu z plikiem .txt zawierającym tekst o wydźwięku pozytywnym.
• Przeprowadzenie testu z plikiem .txt zawierającym tekst o wydźwięku negatywnym.
• Przeprowadzenie testu z plikiem .txt zawierającym tekst o neutralnym wydźwięku.
• Przeprowadzenie testu z plikiem .txt zawierającym tekst o mieszanym sentymencie.
• Przeprowadzenie testu z plikiem .txt zawierającym brakujące dane.
• Przeprowadzenie testu z plikiem .txt zawierającym znaki specjalne.

Scenariusze użytkownika (user stories)

Scenariusz 1: Analiza opinii klientów
Cel: Zrozumienie ogólnego sentymentu klientów
Kroki:
++Wczytanie pliku z opiniami.
++Przeprowadzenie analizy sentymentu.
++Generowanie wykresów i raportów.

++Identyfikacja ogólnego sentymentu klientów i obszarów, które wymagają poprawy

Scenariusz 2: Monitorowanie mediów społecznościowych
Cel: Reagowanie na zmiany sentymentu
Kroki:

++Wczytanie danych z mediów społecznościowych.
++Analiza zmian sentymentu w czasie.
++Wygenerowanie chmury słow.
++Identyfikacja nagłych zmian.

Scenariusz 3: Analiza przemówień
Cel: Identyfikacja dominujących emocji
Kroki:

++Wczytanie tekstu przemówienia.
++Generowanie chmury słów i wykresów sentymentu.
++Generowanie wykresów ewolucji sentymentu w czasie.
++Analiza wyników i użycie danych do dalszych analiz naukowych

Scenariusz 4: Badanie nastrojów wśród pracowników
Cel: Wykryć problemy moralne w zespołach
Kroki:

++Wczytanie pliku .txt z odpowiedziami na pytanie otwarte.
++Usunięcie neutralnych słów (stopwords) i przeprowadzenie stemming'u.
++Generowanie wykresu liniowego zmian emocji w kolejnych miesiącach.
++Analiza i przekazanie wyników do zarządu firmy.