Dokumentacja Specyfikacji Wymagań (SRS)

Projekt: Analiza wypowiedzi (jednego z 4 plików - wybór interaktywny) Jerome Powell'a oraz ich wpływ na giełdę (przetwarzanie tekstu oraz analiza sentymentu za pomocą słowników w plikach CSV: Loughran, NRC)
Wersja dokumentu: 1.0
Data: 06.06.2025
Autorzy: Adam Bagiński, Dominika Pawluczuk, Julia Ruta

1. Wprowadzenie:

Niniejszy dokument opisuje wymagania funkcjonalne i niefunkcjonalne dotyczące skryptu w języku R, którego celem jest przeprowadzenie analizy text mining oraz analizy sentymentu na podstawie zawartości pliku tekstowego .txt. System wykorzystuje takie techniki jak text mining, tokenizacja czy stemming, a także ocenę sentymentu przy użyciu słowników w plikach CSV (NRC, Loughran), i słowników z pakietu SentimentAnalysis (GI, HE, LM). Ukazane zostały wizualizacje częstości występowania słów w postaci chmury słów, wykresów rodzaju sentymentu oraz wykresów zmiany sentymentu w czasie. Celem systemu jest ukazanie potencjalnej korelacji pomiędzy tonem analizowanych wypowiedzi a sytuacją na rynkach finansowych.

2. Cele systemu:

• Wczytanie tekstu wejściowego (plik .txt) z odpowiednim kodowaniem (UTF-8).
• Przetwarzanie i oczyszczanie tekstu (normalizacja wielkości liter, rozbieżności kodowań, form skróconych, akcentów, popularnych skrótów, interpunkcji, tokenizacja, stemming).
• Usunięcie zbędnych ciągów znaków, znaków specjalnych, białych znaków, stopwords, cyfr i liczb.
• Zliczanie najczęściej występujących słów oraz ich wizualizacja w formie chmury i wykresów słupkowych.
• Przeprowadzenie analizy sentymentu z użyciem słowników: o plikach CSV (NRC i Loughran), wbudowanych w pakiet SentimentAnalysis (GI, HE, LM).
• Wizualizacja wyników sentymentu za pomocą wykresów słupkowych.
• Porównanie wyników sentymentu między używanymi słownikami.
• Umożliwienie analizy zmian sentymentu w czasie za pomocą wykresów (modelowanie LDA)
• Analiza wpływu wypowiedzi Powell'a na sytuację na giełdzie

3. Wymagania funkcjonalne systemu:

Wczytywanie danych:

• Skrypt powinien umożliwiać wczytanie danych tekstowych z lokalnych plików .txt, obsługuje kodowanie UTF-8.

Przetwarzanie i oczyszczanie tesktu:

• Skrypt powinien umożliwiać wykonanie stemmingu, uzupełnienia rdzeni słów.
• Skrypt powinien umożliwiać normalizację wielkości liter, rozbieżnych kodowań znaków, form skróconych, akcentów, popularnych skrótów (przez rozwinięcie).

• Skrypt powinien umożliwiać usunięcie zbędnych ciągów znaków, znaków specjalnych, białych znaków, cyfr i liczb, interpunkcji oraz stopwords.
• Skrypt powinien umożliwić wykonanie tokenizacji i stemmingu.

Analiza częstości:

• Skrypt powinien umożliwiać zliczenie wystąpienia słów oraz sortuje według ich częstości, co przedstawione jest w postaci chmury słów.

Analiza sentymentu i prezentacja danych

• Skrypt powinien umożliwiać wczytanie słowników: Loughran.csv i NRC.csv.
• Skrypt powinien przeprowadzać analizę sentymentu tekstu z wykorzystaniem biblioteki SentimentAnalysis.
• Skrypt powinien przeprowadzać analizę i dopasowanie sentymentu do konkretnych słów dla słowników Loughran i NRC.
• Skrypt powinien generować wykresy słupkowe pokazujące rozkład sentymentu.
• Skrypt powinien umożliwiać wizualizację wyników.
• Skrypt powinien przeprowadzać analizę sentymentu dla kawałków tekstu tzn. np. podział na negatywny i pozytywny i zmieniającego się sentymentu w tekście w zależności od momentu tekstu dla słowników GI, HE, LM.
• Skrypt powinien generować wykres słupkowy przedstawiający podział sentymentu w całym tekście dla trzech wspomnianych słowników na raz.
• Skrypt powinien dołączać zdjęcia zachowania rynku podczas wystąpienia.
• Skrypt powinien generować wykresy ważności słów w temacie (LDA).

4. Wymagania niefunkcjonalne systemu:

• Szybkość oraz wydajność.

Niezawodność:

• Skrypt zapewnia poprawność danych wyjściowych, poprawnie obsługuje brakujące wartości.

Użyteczność:

• Skrypt powinien przedstawiać wykresy w sposób czytelny oraz zawierając odpowiednie etykiety.
• Skrypt powinien umożliwiać wykonywanie wykonywanie wizualizacji z użyciem ggplot2

Kompatybilność:

• Skrypt powinien być kompatybilny z R w wersji 4.0 lub nowszej, korzysta z takich bibliotek jak: tm, tidytext, stringr, worldcloud, ggplot2, RColorBrewer, ggthemes, SentimentAnalysis, SnowballC, tidyverse, topicmodels, dplyr.

5. Interfejsy użytkownika:

• Wejście:
- ➔ Plik tekstowy .txt.
- ➔ Plik zdjęciowy .png
- ➔ Pliki słowników w formacie .csv

• Wyjście:
- ➔ Chmura słów.

- ➔ Wykresy słupkowe rodzaju sentymentu (słowniki: NRC, Loughran, GI, HE, LM).
- ➔ Wykres liniowy ukazujący zmianę sentymentu w czasie.
- ➔ Wykresy słupkowe pokazujące ważność słów.
- ➔ Zdjęcie analizowanej sytuacji na rynkach finansowych.

6. Wymagania dotyczące danych:

• System analizuje dane tekstowe w języku angielskim, nie obsługując analizy sentymentu dla innych języków.

• System wykorzystuje słowniki sentymentów będących w plikach .CSV i w pakiecie SentimentAnalysis.

• System nie obsługuje analizy sentymentu dla danych tekstowych pochodzących z innych źródeł niż pliki .txt i o rozmiarze powyżej 100MB.

Słownictwo dokumentacji:

• *Chmura słów*: wizualizacja przedstawiająca najczęściej występujące słowa.

• *Sentyment*: emocjonalny ton wyrażony w tekście.

• *Słownik sentymentów*: lista słów oraz ich ocen według posiadanego sentymentu.

• *Stem*: forma słowa po sprowadzeniu go do rdzenia.

• *Ważność słowa w temacie*: waga przydzielana słowom na podstawie liczby ich wystąpień w analizowanym tekście.

• *Stopwords*: słowa niewnoszące wartości semantycznej do analizy.

Przypadki użycia (use cases)

• Użytkownik:
- ➔ wybiera plik .txt.
- ➔ uruchamia analizę
- ➔ wyświetla wyniki

• Skrypt:
- ➔ przetwarza tekst,
- ➔ oczyszcza tekst,
- ➔ generuje chmurę słów (zgodnie z częstotliwością występowania),
- ➔ analizuje sentyment tekstu przy użyciu słowników NRC i Loughan oraz przy użyciu pakietu SentimentalAnalysis (słownik GI, HE, LM),
- ➔ generuje wykres porównujący rodzaj sentymentu wg słowników (GI, HE, LM),
- ➔ generuje wykresy zmiany sentymentu w czasie,
- ➔ generuje obraz odpowiadający analizowanej sytuacji na rynku finansowym,
- ➔ generuje wykresy ważności słów (LDA) dla 4 tematów,

Scenariusze użytkownika (user stories)

Scenariusz 1:

• **Jako:** Inwestor (analityk rynku)

• **Chcę:** Automatycznie analizować wypowiedzi Jerome’a Powella, aby ocenić ich sentyment

• **Aby:** Móc reagować i dostosowywać swoje strategie inwestycyjne

Kryteria akceptacji:
• Użytkownik może wczytać plik tekstowy z wypowiedziami.
• Skrypt przeprowadza analizę sentymentów za pomocą różnych słowników i pakietów (bibliotek).
• Skrypt generuje wykresy sentymentów dla poszczególnych słowników oraz wykresy porównujące.
• Skrypt generuje wykresy ewolucji sentymentu w czasie.
• Skrypt generuje wykresy ważności słów w temacie (model LDA).
• Użytkownik może ulepszyć swoją strategię inwestycyjną.