Dokumentacja Specyfikacji Wymagań (SRS)

**Projekt: Pełna analiza Text Mining jednego pliku**

**Wersja: 1.0**

**Data: 26.05.2025**

**Autorzy: Małgorzata Kowalczyk, Weronika Rudowska, Maja Tęczyńska**

# **1. Wprowadzenie**

Niniejszy dokument zawiera specyfikację wymagań dla projektu analizy tekstu (Text Mining) na podstawie jednego pliku wejściowego. System umożliwia wczytywanie tekstu, jego czyszczenie, analizę częstości słów, ocenę sentymentu z użyciem różnych słowników oraz zastosowanie techniki TF-IDF do oceny ważności słów. Projekt obejmuje także generowanie wykresów i raportów wizualnych.

# **2. Cel systemu**

* Wczytanie tekstu wejściowego (plik .txt) z odpowiednim kodowaniem (UTF-8)
* Przetwarzanie i oczyszczanie tekstu (normalizacja, tokenizacja, stemming)
* Usunięcie nieistotnych słów (stopwords)
* Zliczenie częstości występowania słów oraz ich wizualizacja w formie chmury
* Przeprowadzenie analizy sentymentu z użyciem słowników:
  + w plikach CSV (AFINN, Bing, NRC, Loughran)
  + wbudowanych w pakiet SentimentAnalysis (GI, HE, LM, QDAP)
* Wizualizacja wyników sentymentu za pomocą wykresów słupkowych i czasowych
* Porównanie wyników sentymentu między słownikami
* Umożliwienie wizualizacji danych: wykresy słupkowe, liniowe, chmury słów, analiza w czasie

# **3. Wymagania funkcjonalne**

* Wczytywanie danych
  + Skrypt powinien umożliwiać wczytanie danych tekstowych z lokalnego pliku .txt.
  + Skrypt powinien obsługiwać kodowanie UTF-8
* Przetwarzanie I czyszczenie tekstu:
  + Skrypt powinien umożliwiać normalizację apostrofów na apostrof klasyczny.
  + Skrypt powinien umożliwiać usunięcie liczb, interpunkcji oraz form skróconych zawierających apostrofy.
  + Skrypt powinien umożliwiać usunięcie pustych elementów oraz zbędnych znaków specjalnych.
  + Skrypt powinien umożliwiać usunięcie stopwords z pakietów tidytext i tm.
  + Skrypt powinien umożliwiać wykonanie stemmingu i uzupełnienia rdzeni słów.
  + Skrypt pownien umożliwiać tokenizację tekstu i konwersję do małych liter
* Analiza częstości:
  + Skrypt powinien umożliwiać zliczenie liczby wystąpień słów.
  + Skrypt powinien umożliwiać posortowanie słów według częstości.
  + Skrypt powinien umożliwiać obliczanie wartości TF-IDF dla każdego słowa
  + Skrypt powinien umożliwiać wizualizacę częstości słów oraz wartości TF-IDF (chmury,tabele)
* Analiza sentymentu (słowniki w plikach CSV):
  + Skrypt powinien umożliwiać wczytanie słowników: afinn.csv, bing.csv, nrc.csv, loughran.csv.
  + Skrypt powinien umożliwiać dopasowanie słów do słowników i zliczenie sentymentów. Skrypt powinien umożliwiać filtrowanie słów o sentymencie pozytywnym lub negatywnym.
* Analiza sentymentu (słowniki wbudowane w pakiet SentimentAnalysis):
  + Skrypt powinien przeprowadzać analizę sentymentu tekstu z wykorzystaniem biblioteki SentimentAnalysis.
  + Skrypt powinien wykorzystywać słowniki GI, HE, LM, QDAP.
  + Skrypt powinien konwertować ciągłe wartości sentymentu na wartości kierunkowe. Skrypt powinien umożliwiać podział tekstu na segmenty o ustalonej długości.
* Wizualizacja danych:
  + Skrypt powinien umożliwiać wizualizację wyników (wykresy ggplot2).
  + Skrypt powinien generować wykresy skumulowanego sentymentu dla każdego słownika.
  + Skrypt powinien generować wykres porównujący sentyment na podstawie różnych słowników.
  + Skrypt powinien generować wykresy przedstawiające ewolucję sentymentu w czasie (wykresy liniowe i wygładzone).
* Agregacja danych:
  + Skrypt powinien agregować sentyment z różnych słowników w jednej ramce danych.
  + Skrypt powinien usuwać brakujące wartości (NA).

# **4. Wymagania niefunkcjonalne**

* Wydajność:
  + Analiza pliku o długości 1000 zdań powinna trwać nie dłużej niż 15 sekund.
* Bezpieczeństwo:
  + System powinien zapewnić poprawność danych wyjściowych.
* Niezawodność:
  + Skrypt powinien poprawnie obsługiwać różne formaty danych tekstowych.
  + 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.
  + Skrypt powinien umożliwiać generowanie chmury słów z wykorzystaniem kolorystyki RColorBrewer.
* Kompatybilność:
  + Skrypt powinien być kompatybilny z R w wersji 4.0 lub nowszej.
  + Skrypt powinien korzystać z bibliotek tm, tidytext, stringr, ggplot2, ggthemes, SentimentAnalysis, SnowballC, tidyverse..

**5. Interfejsy użytkownika**

* Wejście:
  + Plik .txt lub .html z treścią (np. recenzją)
  + Pliki słowników .csv
* Wyjście:
  + Tabela z częstością występowania słów
  + Chmury słów
  + Wykresy słupkowe
  + Wykresy liniowe (sentyment, TF-IDF)
  + Raport HTML

# **6. Wymagania dotyczące danych**

* Skrypt zakłada, że dane tekstowe są 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 o rozmiarze powyżej 100 MB.

**Słownictwo dokumentacji:**

* Token: pojedynczy element tekstu (słowo).
* Stopwords: słowa niewnoszące wartości semantycznej do analizy.
* Sentyment: emocjonalne nastawienie w tekście.
* Słownik sentymentów: lista słów i ich ocen wg sentymentu.
* Skumulowany sentyment: suma ocen sentymentu dla całego tekstu.
* Wartości kierunkowe: konwersja ciągłych wartości sentymentu na kategorie (np. pozytywny, negatywny, neutralny).
* Ewolucja sentymentu: zmiana sentymentu w czasie (wzdłuż czasu narracyjnego).
* Stem: forma słowa po sprowadzeniu go do rdzenia.

**Przypadki użycia (use cases)**

* Użytkownik:
  + wczytuje plik .txt lub .csv
  + uruchamia analizę
  + wyświetla wyniki
  + generuje wykresy i raport html

• Skrypt/system:

* + przetwarza tekst
  + oczyszcza tekst
  + analizuje sentyment tekstu przy użyciu słowników
  + analizuje istotność słów w recenzjach za pomocą wskaźnika TF-IDF
  + generuje chmurę słów
  + generuje wykresy skumulowanego sentymentu
  + generuje wykres porównujący rodzaj sentymentu wg słowników
  + generuje wykresy zmiany sentymentu w czasie narracyjnym

**Testowe przypadki użycia:**

• Test z plikiem .txt zawierającym tekst o pozytywnym sentymencie.

• Test z plikiem .txt zawierającym tekst o negatywnym sentymencie.

• Test z plikiem .txt zawierającym tekst o mieszanym sentymencie.

• Test z plikiem .txt zawierającym znaki specjalne.

**Scenariusze użytkownika (user stories)**

**Scenariusz 1: Monitorowanie opinii po wprowadzeniu zmian w menu**

• **Jako:** Menedżer restauracji

• **Chcę:** Porównać sentyment opinii przed i po zmianach w menu

• **Aby:** Sprawdzić, czy nowa oferta wpłynęła pozytywnie na opinie klientów.

**Kryteria akceptacji:**

• Użytkownik może wczytać plik z recenzjami restauracji (.txt lub .csv).

• Skrypt przeprowadza analizę sentymentu za pomocą różnych słowników.

• Skrypt generuje wykresy skumulowanego sentymentu i porównuje wyniki z różnych słowników.

• Skrypt generuje wykresy ewolucji sentymentu w czasie.

• Użytkownik może zidentyfikować ogólny sentyment klientów i obszary, które wymagają poprawy.

**Scenariusz 2: Wykrycie nietypowych i krytycznych opinii klientów**

• **Jako:** Właściciel restauracji

• **Chcę:** Przeanalizować recenzje klientów, aby zidentyfikować rzadko występujące, ale potencjalnie poważne sytuacje opisane w opiniach (np. przypadki naruszenia diety, jak podanie glutenu osobie z celiakią)

• **Aby:** Szybko reagować na incydenty i poprawić jakość obsługi

**Kryteria akceptacji:**

• Użytkownik może wczytać plik z recenzjami restauracji (.txt lub .csv).

• Skrypt wykonuje analizę TF-IDF i identyfikuje słowa lub frazy o wysokim znaczeniu lokalnym.

• Użytkownik może zobaczyć listę recenzji zawierających rzadkie, potencjalnie alarmujące słownictwo (np. „alergia”, „celiakia”, „zatrucie”, „gluten”, „skandaliczne”).

• Skrypt umożliwia jednoczesne przeprowadzenie analizy sentymentu dla zidentyfikowanych recenzji.

• Właściciel restauracji może szybko wychwycić i przeanalizować incydenty, które wymagają pilnej reakcji lub zmiany procedur wewnętrznych.

**Scenariusz 3: Monitorowanie sentymentu w mediach społecznościowych**

• **Jako:** Specjalista ds. mediów społecznościowych

• **Chcę:** Monitorować sentyment w mediach społecznościowych

• **Aby:** Sprawdzić, czy nowa oferta wpłynęła pozytywnie na opinie klientów.

**Kryteria akceptacji:**

• Użytkownik może wczytać plik z recenzjami restauracji z mediów społecznościowych (.txt lub .csv).

• Skrypt przeprowadza analizę sentymentu i generuje wykresy ewolucji sentymentu w czasie.

• Użytkownik może monitorować zmiany sentymentu.

• Użytkownik może identyfikować nagłe zmiany sentymentu i reagować na nie.

• Użytkownik może generować raporty z analizy sentymentu.