Dokumentacja Specyfikacji Wymagań (SRS)

**Projekt: nazwa**

Wersja: 1.0

Data: 05.06.2025

Autorzy: Michał Potasiński, Kacper Buchalik

Wprowadzenie

Projekt jest w celu stworzenia narzędzia analitycznego służącego do automatyzacji procesu porównywania opinii przedsiębiorstw i analizy wrażeń konsumentów piszących komentarze. Na potrzeby projektu wzięliśmy 60 losowych opinii z google maps, 30 z nich dotyczyło restauracji Mcdonald’s a 30 KFC. Niniejszy dokument zawiera specyfikację wymagań dla projektu analizy tekstu na podstawie dwóch plików wejściowych. System umożliwia wczytywanie tekstu, jego czyszczenie, analizę sentymentu przy użyciu każdej z następujących bibliotek (GI, HE, LM, QDAP), analizę sentymentu w czasie, eksploracyjną analizę danych oraz wyszukiwanie asocjacji. Projekt obejmuje również generowanie wykresów i ich zestawianie

Cel systemu

* Wczytywanie tekstu wyjściowego (plik .csv) z odpowiednim kodowaniem (UTF-8)
* Przetwarzanie i oczyszczanie tekstu (Tokenizacja)
* Usunięcie nieistotnych słów (stopwords)
* Zliczenie częstości występowania słów i ich prezentacja w formie chmury słów
* Przeprowadzanie analizy sentymentów z pomocą słowników (GI, HE, LM, QDAP)
* Wizualizacja wyników sentymentu za pomocą wykresów słupkowych, czasowych
* Porównanie wyników między słownikami zestawienie wykresów obok siebie
* Wyszukiwanie asocjacji wybranych słów na różnych poziomach istotności
* Wizualizacja asocjacji z różnymi terminami za pomocą wykresów słupkowych

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 tokenizację tekstu i konwersję do małych liter
* Skrypt powinien umożliwiać usunięcie stopwords z pakietów tydytext i tm
* Skrypt powinien umożliwiać usunięcie pustych elementów i niepotrzebnych znaków specjalnych
* Skrypt powinien umożliwiać usunięcie liczb, interpunkcji oraz form skróconych zawierających apostrofy
* Analiza częstości:
* Skrypt powinien umożliwiać zliczenie liczby wystąpień słów
* Skrypt powinien umożliwiać dopasowanie słów do słowników i zliczenie częstości
* Skrypt powinien umożliwiać obliczenie wartości TF-IDF dla każdego słowa
* Skrypt powinien umożliwiać wizualizacje częstości słów oraz wartości TF-IDF
* Analiza sentymentu (słowniki w pakiecie SentimentAnalysis)
* Skrypt powinien wykonać analizę sentymentu z wykorzystaniem biblioteki
* Skrypt powinien wykorzystywać słowniki (GI, HE, LM, QDAP)
* Wizualizacja danych
* Skrypt powinien umożliwiać wizualizacje wyników z użyciem biblioteki ggplot2
* Skrypt powinien generować wykresy sentymentu dla każdej biblioteki
* Skrypt powinien generować wykresy zmiany sentymentu w czasie
* Skrypt powinien generować wykresy zestawienia bibliotek
* Skrypt powinien generować wykresy asocjacji
* Asocjacje
* Skrypt powinien znajdować asocjacje wybranych słów na wybranym poziomie istotności

Wymagania niefunkcjonalne

* Wydajność
* System powinien przetwarzać co najmniej 500 dokumentów/minutę przy analizie sentymentu
* Wyszukiwanie asocjacji w zbiorze 10000 dokumentów powinno trwać nie dłużej niż 30 sekund
* Niezawodność
* System powinien działać poprawnie w 99,5% przypadków, niezależnie od formatu tekstu wejściowego
* Błędy przetwarzania muszą być logowane i raportowane w czasie rzeczywistym
* Użyteczność
* System powinien generować czytelne i estetyczne wykresy
* Skrypt powinien umożliwiać wykonanie wykresu z użyciem biblioteki ggplot2
* Kompatybilność
* Skrypt powinien korzystać z bibliotek tm, tidytext, ggplot2, ggthemes, SentimentAnalysis, tidyverse, wordcloud

Interfejsy użytkownika

* Wejście
* Plik .txt lub .csv
* Working directory
* Słowa bazowe do asocjacji
* Poziom istotności asocjacji
* Wyjście
* Chmury słów
* Wykresy słupkowe
* Wykresy liniowe
* Wykresy słupkowe (asocjacje)
* Raport HTML

Wymagania dotyczące danych

* Dane są w języku Angielskim
* Skrypt nie obsługuje plików o rozmiarze powyżej 100 MB

1. Słownictwo dokumentacji

* Token – Pojedynczy element tekstu
* Stopwords – słowa bez własności logicznej
* Sentyment – emocjonalna wartość słowa
* Słownik Sentymentów – lista słów i ich wartość sentymentu
* Skumulowany sentyment – sumaryczna wartość sentymentu
* Asocjacje – słowa o wysokiej korelacji z słowem bazowym

Przypadki użycia:

* Użytkownik:
* Wczytuje plik .txt lub .csv
* Uruchamia analizę
* Wyświetla wyniki
* Generuje wykresy
* Wybiera słowa bazowe do asocjacji
* Skrypt
* Przetwarza tekst
* Czyści tekst i przygotowuje go do analizy
* Analizuje sentyment za pomocą słowników
* Generuje wykresy sentymentu skumulowanych
* Generuje zmianę sentymentu w czasie
* Wyszukuje asocjacje dla wybranych słów bazowych
* Tworzy wykresy asocjacji na wybranym poziomie istotności

Scenariusze użycia

Scenariusz 1: Analiza rynku restauracji porównując opinie dwóch knajp

* Jako: analityk rynkowy
* Chcę: zbadać, która restauracja jest lepiej oceniania
* Aby: wyciągnąć wnioski na temat prawidłowego zarządzania restauracją

Kryteria akceptacji:

* Użytkownik wczytuje opinie w pliku .csv
* System tworzy analizę sentymentu i tworzy wykresy
* System tworzy asocjacje z wybranymi słowami
* System tworzy wykresy analizy sentymentu w czasie
* Użytkownik może porównać wyniki dla różnych restauracji i sprawdzić asocjacje różnych słów co pomoże ocenić konkretne elementy knajp

Scenariusz 2: Analiza opinii o kandydatach prezydenckich

* Jako: dziennikarz
* Chcę: zbadać opinie na temat polityka
* Aby: zrozumieć nastroje wśród wyborców

Kryteria akceptacji:

* Użytkownik wczytuje opinie w pliku .csv
* System tworzy analizę sentymentu i tworzy wykresy
* System tworzy asocjacje z wybranymi słowami
* System tworzy wykresy analizy sentymentu w czasie
* Użytkownik interpretuje wyniki systemu i identyfikuje bardziej popularnego kandydata