Dokumentacja Specyfikacji Wymagań (SRS)

Projekt: Analiza przemówień politycznych z okresu II wojny światowej z zastosowaniem metod klastrownia

Wersja dokumentu: 1.0

Data: 08.06.2025

Autor: Mateusz Sosnowski

1. Wprowadzenie:

Niniejszy dokument zawiera specyfikację wymagań dla skryptu w języki R, który realizuje analizę tekstów przemówień politycznych z okresu II wojny światowej trzech liderów głównych państw alianckich: Winstona Churchilla, Franklina Delano Roosevelta oraz Józefa Stalina. Celem projektu jest pobranie, czyszczenie, eksploracja oraz klastrowanie przemówień na podstawie ich treści. System realizuje przetwarzanie tekstu z wykorzystaniem bibliotek tm, SnowballC, wordcloud, cluster, factoextra i ggplot2, a także oferuje zaawansowane wizualizacje oraz interaktywne tabele z wykorzystaniem DT.

2. Cele systemu:

- Pobranie przemówień z określonych źródeł internetowych i zapisanie ich jako pliki .txt
- Przetworzenie i oczyszczenie treści (usunięcie znaków specjalnych, stemming, uzupełnienie rdzeni)
- Tokenizacja tekstu i budowa macierzy częstości
- Generowanie chmur słów na podstawie częstości
- Identyfikacja optymalnej liczby klastrów
- Klastrowanie dokumentów z zastosowaniem algorytmu kmeans
- Prezentacja wyników w formie wykresów i interaktywnych tabel

3. Wymagania funkcjonalne:

- Wczytywanie danych:
    - System powinien umożliwiać pobranie tekstów z podanych adresów URL
    - System zapisuje pliki przemówień w formacie .txt i zakłada ręczne oczyszczenie z fragmentów niebędących przemówieniami
- Przetwarzanie tekstu:
    - Oczyszczanie tekstu z niepożądanych znaków: \, |, www, ~, â€“ itd.
    - Transformacja tekstu do małych liter
    - Usunięcie liczb, interpunkcji oraz stopwords (z języka angielskiego)
    - Wykonanie stemmingu oraz jego uzupełnienie poprzez stemCompletion
    - Utworzenie korpusu dokumentów i jego inspekcja
- Analiza częstości:
    - Zliczenie częstości słów i zapisanie ich w ramkach danych
    - Generowanie globalnej chmury słów z filtrowaniem przez min.freq
- Klastrowanie dokumentów:

    - Zastosowanie metody kmeans
    - Automatyczny dobór liczby klastrów z użyciem metody silhouette / sylwetki
    - Generowanie wizualizacji klastrów i słów charakterystycznych
    - Prezentacja wyników klastrowania w tabeli DT

4. Wymagania niefunkcjonalne:

- Wydajność:
    - Skrypt powinien analizować zestaw do 6 dokumentów w czasie poniżej 10 sekund na etapie analizy i wizualizacji
- Niezawodność:
    - Skrypt powinien wykrywać i pomijać puste fragmenty tekstu
- Użyteczność:
    - Prezentacja wyników powinna być możliwa zarówno w formacie HTML, jak i interaktywnym (np. DataTable)
    - Wykresy powinny być czytelne, opatrzone tytułami, z odpowiednią kolorystyką
- Kompatybilność:
    - Wersja R: 4.0 lub nowsza
    - Wymagane biblioteki: tm, SnowballC, rvest, stringr, readr, wordcloud, factoextra, cluster, ggplot2, DT, ggrepel, RColorBrewer, dplyr

5. Interfejsy użytkownika:

- Wejście:
    - Adresy URL przemówień
    - Folder z plikami .txt po scrapowaniu
- Wyjście:
    - Macierze częstości (TDM, DTM)
    - Tabela słów z częstością
    - Chmura słów
    - Tabela z przypisaniem dokumentów do klastrów
    - Wizualizacje klastrów
    - Słupkowy wykres przypisania dokumentów

6. Wymagania dotyczące danych:

- Teksty muszą być w języku angielskim
- Skrypt nie analizuje sentymentu
- Pliki powinny zawierać jedynie tekst właściwy przemówień
- Rozmiar tekstów nie powinien przekraczać 1 MB na dokument

Słownictwo dokumentacji:

- **Token**: słowo będące podstawową jednostką analizy
- **Stopwords**: często występujące słowa nieposiadające istotnego znaczenia
- **Stem**: skrócona wersja słowa, służąca ujednoliceniu formy
- **Klaster**: grupa dokumentów o podobnej strukturze słów
- **TermDocumentMatrix**: macierz częstości słów (wiersze: słowa, kolumny: dokumenty)
- **Korpus:** zbiór dokumentów tekstowych używanych jako dane wejściowe do analizy