# Dokumentacja Specyfikacji Wymagań (SRS)

**Projekt: Text Mining - analiza kultowego serialu Friends**

**Wersja dokumentu: 1.0**
**Data: 07.06.2025**
**Autor: [Maria Jastrzębska, Rozalia Kalisz]**

## 1. Wprowadzenie:

Niniejszy dokument opisuje specyfikację wymagań dla systemu R przeznaczonego do eksploracyjnej analizy danych tekstowych z dialogów serialu *Friends*. System umożliwia analizę częstości słów, analizę asocjacyjną dla wybranych słów, modelowanie tematów metodą LDA oraz wizualizację wyników w postaci chmury słów i wykresów.

## 2. Cele systemu:

- Analiza rozkładu częstości słów w wypowiedziach postaci.
- Wydobycie tematów dominujących w dialogach (LDA).
- Wizualizacja wyników analizy tekstowej w postaci wykresów i chmur słów.
- Wspomaganie badania języka postaci i dynamiki serialu.

## 3. Wymagania funkcjonalne:

- Wczytywanie danych:
  - Skrypt powinien umożliwiać wczytanie danych tekstowych z lokalnego pliku .rda (RData), zawierającego ramkę danych z dialogami postaci.
- Przetwarzanie tekstu:
  - Usuwanie znaków specjalnych, tokenizacja, stemming, konwersja do małych liter.
  - Usunięcie zbędnych słów (stopwords), np. “you”, “the”, "achhh", "ummooh".
  - Budowanie macierzy DTM (Document-Term Matrix).
- Analiza tekstu:
  - Obliczanie częstości słów.
  - Wyszukiwanie asocjacji: analiza współwystępowania słów z imionami głównych bohaterów (*Rachel, Ross, Monica, Chandler, Joey, Phoebe*).
  - Modelowanie tematów metodą LDA.
  - Uruchomienie modelu LDA z dowolnie zadeklarowaną liczbą tematów.
- Wizualizacja danych:
  - Chmura słów - graficzna reprezentacja często występujących słów.
  - Wykresy asocjacji słów – słowa najbardziej skorelowane z podanym wyrazem (wykresy lizakowe).
  - Wykresy słów w tematach (LDA) – wykresy słupkowe (ggplot2) dla każdego tematu, przedstawiające słowa o najwyższym prawdopodobieństwie.

## 4. Wymagania niefunkcjonalne:

- Wydajność:
  - Analiza pełnych danych z sezonu serialu powinna trwać krócej niż obejrzenie serialu🙂

- Niezawodność: system powinien poprawnie działać z różnymi strukturami danych wejściowych.
- Użyteczność:
    - Wykresy i wizualizacje powinny być czytelne, estetyczne i zawierać odpowiednie etykiety.
- Kompatybilność:
    - Wersja R: 4.0 lub nowsza.
    - Biblioteki: *tm, ggplot2, ggthemes, tidyverse, tidytext, topicmodels, wordcloud*.

### 5. Interfejsy użytkownika:

- **Wejście:**
    - Plik .rda zawierający ramkę danych z wypowiedziami postaci (tekst, postać, odcinek).
- **Wyjście:**
    - Chmura najczęstszych słów (workcloud).
    - Wykresy asocjacji ggplot2.
    - Wykresy tematów ggplot2.

### 6. Wymagania dotyczące danych:

- Skrypt zakłada, że dane tekstowe są w języku angielskim.
- Skrypt nie obsługuje analizy dla innych języków.
- Plik wejściowy musi być w formacie .rda, zawierającym data.frame.
- Zalecane kolumny: character (postać), text (wypowiedź), episode (odcinek).

**Słownictwo dokumentacji:**

- LDA (Latent Dirichlet Allocation) - algorytm modelowania tematów.
- DTM (Document-Term Matrix) - macierz częstości słów.
- Chmura słów - graficzna reprezentacja częstości słów.
- Stop words - słowa powszechnie występujące, usuwane w analizie (np. “the”, “and”).

## Przypadki użycia (use cases)

- **Użytkownik:**
    - Wczytuje dane.
    - Uruchamia analizę
    - Wyświetla i interpretuje wyniki.
- **Skrypt/system:**
    - Wczytuje dane.
    - Przetwarza tekst.
    - Przeprowadza analizę.
    - Generuje wizualizacje i zapisuje wyniki.

**Testowe przypadki użycia:**

- Wczytanie prawidłowego pliku .rda.
- Test działania LDA z różną liczbą tematów.
- Test na danych niepełnych (brak jednej kolumny).
- Test działania na małej próbce (np. 1 odcinek).

- Test z innymi słowami do zbadania asocjacji

## Scenariusze użytkownika (user stories)

**Scenariusz 1 - Chmura słów bohatera:**
- **Jako:** fan serialu
- **Chcę:** zobaczyć, jakich słów najczęściej używa Joey
- **Aby:** stworzyć grafikę do publikacji na Instagram.

**Kryteria akceptacji:**
- Wygenerowana zostaje chmura słów tylko dla postaci Joey'a.
- Możliwość eksportu do pliku graficznego lub HTML.

**Scenariusz 2 - Planowanie sequela – analiza asocjacji postaci:**
- **Jako:** scenarzysta pracujący nad kontynuacją serialu
- **Chcę:** zobaczyć, z jakimi słowami najczęściej kojarzy się każda z głównych postaci
- **Aby:** zrozumieć, jaką rolę pełniła w oryginalnym serialu i które tematy można rozwinąć w sequelu.

**Kryteria akceptacji:**
- Użytkownik wybiera imię postaci jako słowo-klucz. Wygenerowany zostaje wykres asocjacyjny dla tego słowa.
- Wykresy są odrębne dla każdej postaci.
- Wyniki są dostępne w raporcie HTML wraz z nazwą postaci i listą skojarzeń.

**Scenariusz 3 - Analiza zmian tematyki w serialu:**
- **Jako:** badacz kultury medialnej
- **Chcę:** przeanalizować, jak zmieniały się dominujące tematy wypowiedzi w serialu *Friends* na przestrzeni sezonów
- **Aby:** zidentyfikować ewolucję narracji, priorytetów postaci i społecznych kontekstów poruszanych w serialu.

**Kryteria akceptacji:**
- Dane można ręcznie filtrować według sezonów. System przeprowadza analizę LDA na wyfiltrowanych danych. Możliwość porównania tematów w osobnych uruchomieniach skryptu.
- Dla każdego sezonu prezentowane są słowa kluczowe dla tematów w formie wykresów.
- Raport zawiera porównanie tematów między sezonami, co umożliwia obserwację zmian w strukturze tematycznej serialu.