Dokumentacja Specyfikacji Wymagań (SRS)

Projekt analizy tekstu z filmów z serii „Shrek" w języku R

Maciej Kawelczyk, Jakub Maszkowski, Szymon Pawłowski

8 czerwca 2025

1 Wprowadzenie

Celem niniejszej dokumentacji jest przedstawienie szczegółowej specyfikacji wymagań dla systemu analizującego scenariusze z filmów z serii „Shrek". System realizowany jest w języku R i koncentruje się na analizie sentymentu, porównaniach językowych oraz eksploracji zależności między tekstami.

2 Cele systemu

System ma na celu:

- Przeprowadzenie analizy sentymentu scenariuszy filmów, przy użyciu odpowiednich słowników.
- Zliczanie popularności słów , w tym generowanie chmury słów.
- Obliczanie korelacji pomiędzy poszczególnymi filmami.
- Analiza asocjacji słów .
- Wizualizacja wyników.

3 Wymagania funkcjonalne

1. System powinien umożliwiać wczytywanie plików tekstowych z dialogami z poszczególnych części „Shreka", dokumenty w formacie .txt.
2. System powinien wczytywać folder z plikami scenariusza wszystkich części.
3. System powinien oczyszczać tekst z najczęściej pojawiających się fraz oraz znaków specjalnych.
4. System powinien wykonać analizę sentymentu dla każdego z filmów.
5. System powinien pokazywać najczęściej występujące słowa w filmach i wizualizować wyniki w formie chmur słów.
6. System powinien wizualizować ewolucję natężenia sentymentu dla wszystkich słowników.
7. System powinien obliczać korelacje występowania słów między wszystkimi częściami.
8. System powinien umożliwiać analizę asocjacji słów z możliwością edytowania parametrów algorytmu.

9. System powinien wizualizować wyniki skumulowanego sentymentu dla każdego słownika na wykresie.

4 Wymagania niefunkcjonalne

- System powinien być zaimplementowany w języku R (wersja 4.0 lub wyższa).
- Analizy i wizualizacje generowane w formatach graficznych.
- Wyniki powinny być prezentowane w sposób jasny oraz czytelny. Powinny umożliwiać interpretację przez użytkownika, wykonanie wizualizacji z użyciem ggplot2.
- Czas analizy wszystkich części nie powinien przekraczać 240 sekund.

5 Interfejsy użytkownika

- System działa jako skrypt R uruchamiany lokalnie.
- Dane wejściowe: pliki tekstowe z dialogami w formacie .txt (jeden plik na film) w jednym folderze.
- Dane wyjściowe: wykresy, chmury słów, tabele, macierze korelacji.

6 Wymagania dotyczące danych

- Dane tekstowe w języku angielskim w plikach .txt.
- System wczytuje teksty tylko w języku angielskim.
- Pliki znajdują sie w jednym folderze.

6.1 Słownictwo dokumentacji

- **Analiza sentymentu** – proces przypisywania emocji (np. pozytywnych/negatywnych) do tekstu.
- **Stop Words** – słowa niewnoszące nic do fabuły.
- **Chmura słów** – graficzna reprezentacja częstości słów.
- **Korelacja słów** – miara współwystępowania słów między zbiorami danych.
- **Wartości kierunkowe** – konwersja ciągłych wartości sentymentu na kategorie.
- **Asocjacje** – związki pomiędzy słowami wyznaczone metodą asocjacyjną.
- **Tokenizacja** – podział tekstu na słowa lub frazy.

7 Przypadki użycia (Use Cases)

UC1: Wczytanie plików z danymi

- **Aktor:** Użytkownik
- **Opis:** Użytkownik wskazuje lokalizację plików tekstowych z dialogami.
- **Rezultat:** System wczytuje dane i przygotowuje do analizy.

UC2: Analiza sentymentu

- **Aktor:** System
- **Opis:** System analizuje sentyment wypowiedzi w każdym filmie.
- **Rezultat:** Generowane są wyniki w formie wykresów, tabel oraz chmury słów.

UC3: Porównanie filmów

- **Aktor:** System
- **Opis:** System porównuje korelacje między poszczególnymi filmami.
- **Rezultat:** Powstają macierze korelacji.

8 Scenariusze użytkownika (User Stories)

- **Jako badacz języka,** chcę przeanalizować emocjonalny wydźwięk dialogów w poszczególnych częściach „Shreka", aby porównać tonację i wydźwięk filmów. Analizuję różnice pomiędzy tematyką i sposobem ekspresji w kolejnych częściach.
  Kryteria akceptacji: Skrypt generuje wykresy porównujące sentyment w poszczególnych częściach.
- **Jako analityk danych,** chcę zobaczyć najczęściej używane słowa w każdym filmie, aby zidentyfikować kluczowe tematy. Sprawdzam, jak różnią się pomiędzy kolejnymi częściami. Dokonuję porównania tematów pomiędzy filmami.
  Kryteria akceptacji: Skrypt generuje tabele z najczęściej pojawiającymi się słowami. System generuje chmurę słów. Przeprowadza asocjacje najczęściej współwystępujących słów.
- **Jako użytkownik,** chcę zobaczyć graficzne porównanie słów pomiędzy filmami, żeby szybko zorientować się w różnicach pomiędzy kolejnymi częściami.
  Kryteria akceptacji: Skrypt generuje mapę ciepła podobieństw dokumentów.