**Część II. Dokumentacja Specyfikacji Wymagań**

**1. Wprowadzenie**

Projekt dotyczy systemu do analizy sentymentu dialogów postaci z plików tekstowych. Kod w R wykorzystuje biblioteki do przetwarzania tekstu i analizy sentymentu, takie jak tidyverse, SentimentAnalysis, tidytext czy ggplot2. System umożliwia wczytanie dialogów, ich oczyszczenie ze zbędnych słów i znaków, przeprowadzenie analizy sentymentu według różnych słowników (QDAP, Bing) oraz wizualizację wyników.

Celem projektu jest zapewnienie reproducible research, czyli powtarzalności i transparentności analizy, poprzez dobrze udokumentowany kod i klarowne dane wejściowe.

**2. Cele systemu**

* Umożliwienie wczytania i oczyszczenia dialogów z plików tekstowych dla wybranych postaci.
* Przeprowadzenie analizy sentymentu wypowiedzi według różnych słowników sentymentu.
* Wizualizacja wyników analizy sentymentu w postaci wykresów porównawczych dla postaci.
* Zapewnienie pełnej reprodukowalności analizy poprzez modularny i czytelny kod.
* Ułatwienie dalszej rozbudowy systemu o kolejne metody analizy i nowe postaci.

**3. Wymagania funkcjonalne**

* System powinien umożliwiać wybór plików z dialogami dla dowolnej liczby postaci (w przykładzie: Chandler, Rachel, Monica).
* System musi czyścić tekst z niepotrzebnych stop-słów, znaków interpunkcyjnych oraz normalizować tekst (małe litery, ujednolicenie apostrofów).
* System powinien realizować analizę sentymentu z użyciem co najmniej dwóch słowników: QDAP i Bing.
* System musi generować wykresy słupkowe rozkładu sentymentu dla każdej postaci oddzielnie oraz porównawcze dla sentymentu pozytywnego i negatywnego.
* System musi raportować ewentualne ostrzeżenia i błędy podczas wczytywania plików (np. niekompletna linia).

**4. Wymagania niefunkcjonalne**

* Kod musi być modularny i czytelny, aby ułatwić modyfikację i rozszerzanie.
* Środowisko pracy musi być jasno określone (R w wersji 4.4.3, określone wersje pakietów tidyverse, SentimentAnalysis, tidytext, ggplot2 itp.).
* Analiza musi być powtarzalna - ten sam kod na tych samych plikach powinien dać identyczne wyniki i wykresy.
* Wykresy muszą być czytelne, estetyczne i odpowiednio opisane (tytuły, osie).
* System powinien działać efektywnie, obsługując teksty o rozmiarze typowym dla dialogów z seriali lub podobnych źródeł.

**5. Interfejsy użytkownika i wymagania dotyczące danych**

* Interfejs użytkownika to interaktywne okno wyboru plików (funkcja file.choose()), umożliwiające wskazanie plików tekstowych z dialogami postaci.
* Dane wejściowe to pliki tekstowe w formacie UTF-8, zawierające dialogi, gdzie każda linia ma format z tabulatorem oddzielającym metadane (np. kto mówi, kiedy itp.) od treści.
* System oczekuje, że pliki zawierają poprawne linie dialogowe; linie puste lub uszkodzone są ignorowane.
* Wynikiem są wykresy generowane w sesji R, które mogą być zapisane lub eksportowane do dalszej analizy.

**6. Słownictwo dokumentacji**

* **Postać** - bohater (np. Chandler) reprezentowany przez plik tekstowy z dialogami.
* **Stop-słowa** - zbiór często występujących słów, które są usuwane przed analizą (np. „i”, „yeah”, „ok”).
* **Analiza sentymentu** - ocena emocjonalnego wydźwięku tekstu (pozytywny, negatywny, neutralny).
* **Słownik QDAP/Bing** - zestawy słów z przypisanymi wartościami sentymentu.
* **Wykres słupkowy** - graficzne przedstawienie ilości słów o danym sentymencie.
* **Reproducible Research** - podejście zapewniające powtarzalność i transparentność wyników.
* **Modularność** - podział kodu na funkcje i bloki ułatwiające zarządzanie i rozwój.

**7. Przypadki użycia (use cases)**

|  |  | **Cel** | **Opis działania** |
| --- | --- | --- | --- |
| 1 | Analityk | Wczytanie dialogów postaci | Użytkownik wybiera pliki tekstowe z dialogami, system je wczytuje i oczyszcza z niepotrzebnych elementów. |
| 2 | Analityk | Przeprowadzenie analizy sentymentu | System analizuje sentyment każdego zdania według wybranych słowników (QDAP, Bing). |
| 3 | Analityk | Wizualizacja wyników | System generuje wykresy sentymentu z podziałem na postaci i kategorie sentymentu. |
| 4 | Analityk | Eksport wyników | Użytkownik może zapisać wykresy lub wyniki do dalszej analizy. |

### **8. Scenariusze użytkownika (user stories)**

* Jako użytkownik np. platformy streamingowej, rodzic chcę uwzględnić ograniczenia wiekowe w analizie, aby filtrować dialogi i treści nieodpowiednie dla młodszych użytkowników.
* Jako pracownik działu marketingowego, chcę, aby system wykorzystywał algorytm rekomendacji (algorytm proponujący podobne treści), żeby na podstawie analizy sentymentu i wzorców sugerować użytkownikowi seriale, które mogą mu się spodobać.
* Jako analityk, chcę, żeby system automatycznie kategoryzował seriale do odpowiednich grup tematycznych lub gatunków, bazując na analizie tekstu dialogów i sentymentu, aby ułatwić wyszukiwanie.
* Jako analityk danych, chcę, aby analiza była powtarzalna i dawała takie same wyniki przy ponownym uruchomieniu, aby móc zaufać otrzymanym rezultatom.