**Analiza sentymentu w czasie dla dwóch komedii romantycznych: „How to Lose a Guy in 10 Days” oraz „The Proposal” z użyciem słowników AFINN, NRC i Bing**

*Aniela Wróblewska, Natalia Woźniak, Adam Ćwierzyński*

**Wprowadzenie**

Celem niniejszego dokumentu jest przedstawienie specyfikacji wymagań dla systemu informatycznego analizującego zmiany sentymentu w czasie na podstawie scenariuszy komedii romantycznych. System przetwarza dane tekstowe i identyfikuje emocjonalne zmiany w czasie trwania filmu z wykorzystaniem słowników: AFINN, NRC oraz Bing. Dokument zawiera opis celów systemu, wymagań funkcjonalnych i niefunkcjonalnych, interfejsu użytkownika, przypadków użycia oraz scenariuszy użytkownika.

**Cele systemu**

System ma na celu:

* Analizę zmian sentymentu w czasie w wybranych komediach romantycznych
* Wizualizację dynamiki emocji w trakcie trwania filmów
* Zastosowanie technik analizy danych tekstowych w celu interpretacji istotnych informacji zawartych w dialogach

**Wymagania funkcjonalne** (co system ma umożliwić, jakie zadania wykonywać)

* Wczytywanie danych wejściowych (użytkownik może załadować dane tekstowe z plików w formacie .txt)
* Przetwarzanie tekstu (system automatycznie dzieli tekst na linijki, umożliwiając dalszą analizę)
* Analiza sentymentu (każda linijka analizowana jest pod kątem emocjonalnym przy użyciu wybranych słowników)
  1. System powinien przetwarzać tekst z wykorzystaniem słownika AFINN w celu przypisania wartości sentymentu każdej linii.
  2. System powinien przetwarzać tekst z wykorzystaniem słownika NRC, przypisując emocje do słów i grupując je w czasie.
  3. System powinien przetwarzać tekst z wykorzystaniem słownika Bing, klasyfikując słowa jako pozytywne lub negatywne.
* Wizualizacja wyników (wyniki prezentowane są w postaci wykresów oraz chmury słów dla każdego filmu)
* Eksport danych (system powinien umożliwiać eksport wykresów do plików graficznych oraz danych liczbowych do CSV)

**Wymagania niefunkcjonalne** (jak system ma działać, jakie ma mieć własności)

* Powtarzalność (projekt umożliwia pełne odtworzenie wyników analizy na podstawie dostarczonych danych i kodu źródłowego)
* Przejrzystość i czytelność kodu (kod zawiera komentarze, podział na sekcje oraz logiczną strukturę)
* Wydajność (system analizuje pliki tekstowe różnej wielkości w rozsądnym czasie, zapewniając płynne działanie nawet przy większych zbiorach danych)

**Interfejsy użytkownika i wymagania dotyczące danych**

**Interfejs użytkownika:**

* **Główny interfejs to program w RStudio:**
  + Wprowadzenie scenariusza.
  + Wizualizację analizy sentymentu w formie wykresu (np. linia sentymentu w czasie).
  + Możliwość porównania kilku filmów obok siebie.

**Wymagania dotyczące danych:**

* **Źródła danych:**
  + Scenariusze filmów (format .txt lub .srt)
  + Informacje o czasie (np. timestampy w dialogach)
* **Format danych wejściowych:**
  + Tekst podzielony na segmenty
* **Przetwarzanie:**
  + Tokenizacja tekstu - proces dzielenia tekstu na mniejsze jednostki, zwane tokenami, takie jak słowa, zdania lub znaki, które są wykorzystywane w dalszej analizie językowej
  + Analiza sentymentu przy użyciu wybranego modelu NLP
* **Dane wyjściowe:**
  + Wykresy wartości sentymentu przypisanych do punktów w czasie
  + Wykresy ilości słów nacechowanych emocjonalnie
  + Wykresy emocji w czasie
  + Chmury słów emocjonalnych

**Słownictwo dokumentacji**

* Techniczne: tokenizacja, sentyment, analiza tekstu, wykres
* Użytkowe: film, scena, dialog, czas, słowa, emocje
* Interfejsowe: wykres, legenda, filtr
* Procesowe: ładowanie danych, przetwarzanie, wizualizacja, porównanie, zapis wyników

**Słowniki:**

* AFINN - słownik zawierający słowa ocenione na skali od -5 do +5, gdzie liczby oznaczają siłę sentymentu (ujemne – negatywne, dodatnie – pozytywne).
* Bing - zawiera listę słów podzielonych na dwie kategorie: pozytywne i negatywne.
* NRC - przypisuje słowom jedną lub więcej z 8 emocji oraz oznaczenia pozytywne/negatywne, w tej analizie są to wybrane cztery z nich: anger, joy, sadness, trust.

**Przypadki użycia (use cases)**

**Przypadek 1:** Analiza pojedynczego filmu

* Aktorzy: Użytkownik, System
* Opis: Użytkownik wprowadza tytuł filmu lub przesyła plik scenariusza. System przetwarza dane i prezentuje wykres sentymentu
* Warunek początkowy: Użytkownik ma dostęp do interfejsu i pliki w odpowiednim formacie
* Rezultat końcowy: Wizualizacja sentymentu filmu

**Przypadek 2:** Porównanie dwóch filmów

* Aktorzy: Użytkownik, System
* Opis: Użytkownik wybiera dwa filmy i system przedstawia zestawienie wykresów oraz kluczowe różnice w dynamice sentymentu

**Scenariusze użytkownika (user stories)**

* Student chce porównać wykresy sentymentu dwóch filmów, aby zauważyć różnice w strukturze emocjonalnej i wykorzystać je w analizie
* Użytkownik chce zobaczyć średni sentyment filmu i szybko ocenić jego nastrój, aby podjąć decyzję o jego obejrzeniu
* Nauczyciel chce móc pobrać dane z analizy, aby wykorzystać je w prezentacji o narracji w komediach romantycznych