**Dokumentacja specyfikacji wymagań**

**Wprowadzenie**

Celem systemu jest analiza recenzji o filmie „Mamma Mia”, wyreżyserowanego w 2008 roku przez Phyllidę Lloyd. Głównymi elementami badania są: klasyfikacja opinii jako pozytywne lub negatywne, identyfikacja najczęściej występujących słów i tematów, ogólna analiza sentymentu oraz analiza zmian sentymentu w czasie. Skrypt napisany jest w języku R i do przeprowadzenia analizy wykorzystuje bibliotekę SentimentAnalysis (słowniki GI, HE, LM, QDAP) oraz słowniki w plikach CSV – Afinn, Bing oraz NRC. Wyniki przeprowadzonej przez system analizy mogą zostać wykorzystane do badania rynku oraz preferencji społecznych, tworzenia systemów rekomendacyjnych oraz odpowiedniej strategii marketingowej.

**Cele systemu**

* Podział recenzji na pozytywne i negatywne – pozwala określić, jaki jest ogólny odbiór filmu
* Identyfikacja najczęściej pojawiających się w recenzjach słów i utworzenie dla nich chmury słów – umożliwia zidentyfikowanie głównych aspektów filmu, na które zwracają uwagę konsumenci
* Identyfikacja najczęściej pojawiających się słów w recenzjach z uwzględnieniem podziału na opinie negatywne i pozytywne i utworzenie dla nich chmur słów – umożliwia zidentyfikowanie powielających się schematów w negatywnych i pozytywnych recenzjach filmu
* Modelowanie tematów LDA – pozwala na określenie, jakie są najczęściej poruszane w recenzjach filmu tematy
* Wizualizacja wyników w postaci wykresów i chmur słów – tworzy przystępną dla użytkownika formę przekazu informacji

**Wymagania funkcjonalne**

* Odczyt pliku CSV
* Czyszczenie zbioru dokumentów tekstowych (korpusu) – konwersja znaków do kodowania UTF-8, normalizacja tekstu (zamiana na małe litery), usunięcie liczb, stop words, znaków interpunkcyjnych
* Przeprowadzenie analizy danych tekstowych z wykorzystaniem słowników CSV: Bing,Affin, NRC
* Utworzenie chmury słów dla analizowanych recenzji z podziałem na recenzje pozytywne i negatywne, jak i bez podziału
* Utworzenie raportu html na podstawie rezultatów przeprowadzonej analizy w postaci czytelnej dla użytkownika
* Modelowanie tematów występujących w danych tekstowych na podstawie modelu LDA
* Sporządzenie wykresu asocjacji słów
* Wizualizacja uzyskanych wyników poprzez utworzenie wykresów dla słowników Bing, Affin, NRC oraz słowników biblioteki SentimentAnalysis – GI, HE, LM, QDAP

**Wymagania niefunkcjonalne**

* Wygenerowanie przystępnych, czytelnych dla użytkownika wykresów oraz chmur słów, ilustrujących uzyskane rezultaty analizy
* Estetyka: Wykresy oraz chmury słów powinny być podpisane, zawierać odpowiednie etykiety
* Wydajność systemu: analiza danych tekstowych (100 opinii) nie powinna przekroczyć czasu 2 minut
* Skrypt powinien wykorzystywać bibliotekę SentimentAnalysis
* Kompatybilność: Skrypt systemu powinien być kompatybilny z system R w wersji 4.0 lub nowszej

**Interfejsy użytkownika**

Wejście:

* Plik tekstowy .CSV
* Parametry do przeprowadzenia analizy

Wyjście:

* Chmura słów
* Wykresy dla analizy sentymentu NRC,Bing,Affin, SentimentAnalysis

**Wymagania dotyczące danych**

* Dane tekstowe w języku angielskim
* Dane tekstowe w postaci .CSV

**Słownictwo dokumentacji**

* Analiza sentymentu – wykorzystanie przetwarzania języka naturalnego i analizy tekstu do systematyzowania, określania ilościowego i wyodrębniania informacji subiektywnych
* Chmura słów – sposób wizualizacji danych tekstowych, który prezentuje słowa o największej częstotliwości jako największe w chmurze
* Korpus – zbiór plików tekstowych, które analizuje system
* Stop words – często występujące słowa, które nie wnoszą nic do analizy (np. spójniki)
* Macierz DTM (Document-term matrix) – macierz opisująca częstość występowania słowa w każdym z dokumentów tekstowych; kolumny macierzy DTM odpowiadają terminom, a wiersze odpowiadają poszczególnym dokumentom; wartość w komórce jest częstością występowania danego terminu w danym dokumencie
* Macierz TDM (Term-document matrix) – transpozycja macierzy DTM
* Modelowanie tematów LDA (utajona alokacja Dirichleta) – algorytm topic modeling, który pozwala na wykrycie tematów poruszanych w danych tekstowych

**Przypadki użycia – use cases**

Użytkownik:

* Wybiera plik CSV
* Uruchamia analizę danych
* Definiuje liczbę tematów i wielkość chmur
* Wizualizuje wyniki
* Generuje wykresy oraz raport html

System:

* Wczytuje pliki tekstowe
* Czyści dane tekstowe
* Tworzy korpus
* Analizuje częstość występowania słów
* Tworzy chmurę słów wyrażeń, które występują najczęściej
* Tworzy chmurę słów oddzielnie dla recenzji pozytywnych i negatywnych
* Buduje model tematów na podstawie modelu LDA

**Scenariusze użytkownika – user scenarios**

Scenariusz 1: Analiza odbioru filmu

Jako: Analityk marketingowy

Chcę: Określić, jaki jest ogólny przeważający odbiór filmu (pozytywny czy negatywny)

W celu: Ustalenie preferencji społecznych i dopasowania do niego odpowiedniej strategii marketingowej.

Kryteria klasyfikacji:

* Użytkownik ma możliwość wprowadzenia tekstu w postaci CSV
* System klasyfikuje recenzje jako pozytywne i negatywne z wysoką dokładnością
* System identyfikuje najczęściej używane słowa oddzielnie dla obu typów recenzji, generuje dla nich chmury słów

Scenariusz 2: Tworzenie systemu rekomendacyjnego

Jako: Programista tworzący system rekomendujący do platformy streamingowej

Chcę: Ustalić schematy w analizowanych recenzjach poprzez wykrycie najczęściej używanych słów i najczęściej poruszanych tematów w obu typach recenzji (pozytywnych i negatywnych)

W celu: Dopasowaniu rekomendacji innych filmów dla klienta na podstawie wystawionych przez niego recenzji oraz porównania do recenzji innych odbiorców

Kryteria klasyfikacji:

* Użytkownik ma możliwość wprowadzenia tekstu w postaci CSV
* System klasyfikuje recenzje jako pozytywne i negatywne z dużą dokładnością
* System identyfikuje najczęściej używane słowa w analizowanych recenzjach i tworzy dla nich chmurę słów
* System modeluje tematy, które są najczęściej poruszane w analizowanych recenzjach
* System sporządza wykres asocjacji danego słowa dla obu typów recenzji