1. Wprowadzenie

1.1 Cel dokumentu

Niniejsza dokumentacja określa wymagania funkcjonalne i niefunkcjonalne dla systemu analizującego recenzje filmu *"Szybcy i Wściekli: Tokio Drift"* za pomocą metod przetwarzania języka naturalnego. Głównym celem jest automatyczne przetwarzanie, oczyszczanie i wizualizacja danych tekstowych w celu identyfikacji kluczowych opinii, sentymentu oraz najczęściej występujących tematów wśród 100 recenzji.

1.2 Zakres projektu

Projekt obejmuje pełny proces analizy tekstu - od wczytania danych, przez ich oczyszczenie, analizę statystyczną, aż po wizualizację wyników. System będzie działał w środowisku R i wykorzystywał specjalistyczne pakiety do przetwarzania języka naturalnego.

1.3 Grupa docelowa

Dokumentacja jest skierowana do:

- Analityków danych badających opinie użytkowników
- Zespołów marketingowych monitorujących odbiór produktów
- Badaczy zajmujących się przetwarzaniem języka naturalnego
- Programistów implementujących rozwiązania NLP

2. Cele systemu

- **Analiza częstości słów**
  System ma identyfikować i wyświetlać najczęściej występujące słowa w recenzjach, co pozwoli na szybkie określenie głównych tematów poruszanych przez recenzentów.

- **Wykrywanie sentymentu**
  W przyszłości system ma zostać rozszerzony o funkcjonalność klasyfikacji opinii na pozytywne i negatywne na podstawie analizy słownictwa.

- **Wizualizacja danych**
  System będzie generował intuicyjne wizualizacje w postaci chmur słów i wykresów, ułatwiające interpretację wyników analizy.

- **Automatyzacja przetwarzania**
  Projekt ma zminimalizować potrzebę ręcznej obróbki danych poprzez zautomatyzowanie całego procesu od wczytania danych do prezentacji wyników.

3. Wymagania funkcjonalne

- **Przetwarzanie danych**
  System powinien umożliwiać wczytywanie tekstu oraz jego automatyczne czyszczenie i tokenizację w celu przygotowania danych do analizy.

- **Analiza danych**
  System powinien umożliwiać tworzenie macierzy TDM, obliczanie wartości TF-IDF oraz sortowanie słów według częstotliwości lub znaczenia.

- **Wizualizacja**
  System powinien generować chmurę słów na podstawie przeanalizowanego tekstu.

- **Eksportowanie wyników**
  System powinien umożliwiać eksportowanie wyników analizy w postaci raportu HTML.

4. Wymagania niefunkcjonalne

- **Wydajność**
  System powinien przetwarzać 100 recenzji w czasie nie dłuższym niż 15 sekund na standardowym sprzęcie biurowym.

- **Spójność wyników** Identyczne dane wejściowe muszą generować takie same wyniki niezależnie od systemu operacyjnego (Windows, Linux, macOS), wersji środowiska R, konfiguracji sprzętowej

- **Przejrzystość wyników**
  Intuicyjne wizualizacje taki jak chmury słów z czytelną legendą, wykresy z opisami osi i tytułami, znormalizowane skale dla porównywalności

- **Bezpieczeństwo danych**
  System nie przechowuje ani nie przetwarza żadnych danych osobowych czy poufnych informacji o użytkownikach.

- **Prostota modyfikacji**
  Kod powinien być skonstruowany w taki sposób, aby ułatwiał wprowadzanie zmian, dodawanie nowych funkcji oraz szybkie naprawianie błędów, bez narażania stabilności całego systemu

5. Interfejsy użytkownika i wymagania dotyczące danych

5.1 Dane wejściowe

System oczekuje pliku w formacie CSV z dokładnie jedną kolumną zawierającą recenzje

5.2 Wizualizacje

System generuje dwa główne typy wizualizacji:

1. Chmury słów - graficzna reprezentacja częstości terminów

2. Wykresy słupkowe - precyzyjne przedstawienie 10 najczęstszych słów

5.3 Interfejs

Interfejs systemu składa się z:

- Konsoli R wyświetlającej wyniki pośrednie i końcowe
- Okien graficznych prezentujących wizualizacje
- Możliwości eksportu wyników do plików

6. Słownictwo dokumentacji

- **TF-IDF**
  (Term Frequency-Inverse Document Frequency) - metoda statystyczna służąca do oceny ważności słowa w dokumencie względem całej kolekcji dokumentów. Słowa częste w pojedynczym dokumencie, ale rzadkie w całym korpusie, otrzymują wysokie wagi.

- **Korpus**
  Zbiór dokumentów tekstowych poddawanych analizie. W tym przypadku korpus składa się ze 100 recenzji filmowych.

- **TDM**
  (Term-Document Matrix) - macierz, w której wiersze reprezentują słowa, kolumny reprezentują dokumenty, a wartości w komórkach oznaczają częstość występowania danego słowa w danym dokumencie.

- **Tokenizacja**
  Proces dzielenia ciągów tekstowych na pojedyncze jednostki (tokeny), którymi najczęściej są pojedyncze słowa.

- **Chmura słów**
  wizualizacja najczęściej występujących słów w tekście

7. Przypadki użycia (Use Cases)

1. **Wczytanie pliku CSV** - użytkownik ładuje plik zawierający recenzje do systemu.
2. **Czyszczenie i tokenizacja tekstu** - system automatycznie przetwarza teksty, usuwając zbędne
   elementy.
3. **Tworzenie macierzy TDM** - system tworzy Term-Document Matrix na podstawie
   oczyszczonych danych.
4. **Obliczanie TF-IDF** - system oblicza wagę terminów przy użyciu metody TF-IDF.
5. **Sortowanie słów według częstotliwości** - system umożliwia sortowanie słów wg częstości lub
   znaczenia.
6. **Generowanie wizualizacji** – system tworzy chmurę słów.
7. **Eksport wyników do HTML** - użytkownik eksportuje wyniki analizy do raportu HTML.

8. Scenariusze użytkownika (User Stories)

Jako analityk danych

1. Chcę załadować plik CSV z recenzjami, aby móc je automatycznie przetworzyć i
   przeanalizować.
2. Chcę zobaczyć chmurę słów najczęstszych terminów, aby szybko zrozumieć główne tematy poruszane w recenzjach.
3. Chcę wyeksportować wyniki analizy do pliku HTML, aby podzielić się nimi z zespołem bez konieczności ponownego uruchamiania skryptu.

Jako członek zespołu marketingowego

1. Chcę zobaczyć, jakie słowa najczęściej pojawiają się w recenzjach, aby zrozumieć, jakie
   aspekty filmu są najczęściej komentowane przez widzów.
2. Chcę łatwo porównać wyniki z innymi filmami, aby zobaczyć, czy film budzi pozytywniejsze reakcje niż konkurencja.