

Specyfikacja Wymagań Systemowych (SRS)

Projekt: Analiza sentymentu i słownictwa zagranicznych artykułów prasowych odnośnie wyborów prezydenckich w Polsce z wykorzystaniem słowników w plikach CSV.

Wersja dokumentu: 1.0

Data: 06.06.2025

Autorki: Barbara Chojnacka, Aleksandra Domańska

1. Wprowadzenie

Dokument opisuje specyfikację wymagań dla skryptu R, który realizuje analizę text mining i analizę sentymentu na podstawie zawartości plików tekstowych. System łączy tradycyjne techniki czyszczenia tekstu, tokenizację, stemming z oceną sentymentu przy użyciu słowników w plikach CSV (AFINN, Bing, NRC, Loughran), a także słowników z pakietu SentimentAnalysis (GI, HE, LM, QDAP). Celem projektu jest stworzenie systemu do analizy artykułów prasowych, umożliwiającego ocenę sentymentu, analizę częstości słów oraz porównanie kilku artykułów opisujących to samo wydarzenie (wyniki wyborów prezydenckich w Polsce). System wspiera użytkowników w zrozumieniu, jak różne media przedstawiają dane wydarzenie.

2. Cele systemu

- Przetwarzanie i analiza tekstów artykułów prasowych (plik.txt), z odpowiednim kodowaniem (UTF-8).
- Przetwarzanie i oczyszczanie tekstu (tokenizacja, stemming, usunięcie stopwords).
- Zliczanie częstości słów oraz tworzenie chmury słów.
- Automatyczna ocena sentymentu tekstów (pozytywny, neutralny, negatywny).
- Analiza słownictwa: częstość występowania słów z pominięciem stopwords.
- Porównywanie kilku artykułów opisujących to samo wydarzenie.
- Wizualizacja wyników analizy w czytelnej formie.

3. Wymagania funkcjonalne

1. Umożliwienie wczytania artykułów w formacie tekstowym (.txt).
2. Analiza sentymentu z podziałem na pozytywny, negatywny i neutralny.

3. Wydobywanie i prezentacja najczęściej występujących słów w tekście.
4. Porównywanie dwóch lub więcej artykułów pod względem sentymentu i słownictwa.
5. Wizualizacja wyników (np. wykresy słów, skale sentymentu).
6. Eksport wyników analizy do raportu (.html).

4. Wymagania niefunkcjonalne

1. Wydajność: czas analizy pojedynczego artykułu ma być jak najkrótszy.
2. Interfejs użytkownika musi być responsywny i intuicyjny.
3. System musi działać poprawnie dla tekstów w języku angielskim.
4. Kod powinien być modularny i umożliwiać rozszerzenie o dodatkowe funkcje w przyszłości.

5. Interfejsy użytkownika i wymagania dotyczące danych

- **Interfejs użytkownika:**

- Miejsce do dodawania artykułów.
- Widok analizy sentymentu (np. wykres kołowy lub słupkowy).
- Widok analizy częstości słów (np. chmura słów).
- Możliwość porównania artykułów.

- **Dane wejściowe:**

- Treść artykułu (plik tekstowy, plik.txt).

- **Dane wyjściowe:**

- Ocena sentymentu.
- Lista i wizualizacja słów najczęściej używanych.
- Wykresy słupkowe rodzaju sentymentu według słowników.
- Analiza porównawcza dla wybranych artykułów.

6. Wymagania dotyczące danych

- Skrypt zakłada, że dane tekstowe (.txt) są w języku angielskim i nie obsługuje analizy innych języków.
- Skrypt wykorzystuje słowniki sentymentów dostępne w plikach .CSV oraz w pakiecie SentimentAnalysis.

7. Słownictwo dokumentacji

Termin	Definicja
Artykuł	Tekst dziennikarski opisujący wydarzenie lub zjawisko
Sentyment	Emocjonalna ocena tekstu: pozytywny, neutralny lub negatywny
Stopwords	Słowa często występujące, nie wnoszące istotnej treści (np. "i", "oraz", "the")
Analiza częstości słów	Liczenie, jak często występują poszczególne słowa
Tokenizacja	Proces dzielenia tekstu na pojedyncze słowa lub wyrażenia
Porównanie artykułów	Analiza różnic i podobieństw w sentymencie i słownictwie

7. Przypadki użycia (use cases)

UC1: Dodanie artykułu

Jako: Użytkownik

Opis: Użytkownik dodaje artykuł do systemu.

UC2: Analiza artykułu

Jako: System

Opis: System analizuje treść pod kątem sentymentu i częstości słów.

UC3: Porównanie artykułów

Jako: Użytkownik

Opis: Użytkownik wybiera dwa lub więcej artykułów i porównuje wyniki analizy.

UC4: Eksport wyników

Jako: Użytkownik

Opis: Użytkownik zapisuje wyniki analizy do pliku (CSV lub PDF).

8. Scenariusze użytkownika (user stories)

Scenariusz 1: Analiza przekazu emocjonalnego.

Jako: Dziennikarz

Chcę: Sprawdzić wydźwięk artykułów z różnych źródeł.

Aby: Porównać ich przekaz emocjonalny.

Kryteria akceptacji:

- System analizuje teksty i przypisuje im emocjonalny wydźwięk (np. pozytywny, negatywny, neutralny) na podstawie treści.
- Wynik analizy emocjonalnej dla każdego artykułu jest prezentowany w czytelnej formie (np. wykres, tabela, kolorystyczne oznaczenia).

Scenariusz 2: Badanie wiarytelności mediów.

Jako: Badacz mediów

Chcę: Analizować najczęstsze słowa w artykułach.

Aby: Zobaczyć, jak konstruowane są narracje.

Kryteria akceptacji:

- System może oznaczyć słowa nacechowane emocjonalnie, aby ułatwić dalszą analizę narracji.
- System umożliwia porównanie częstości użycia konkretnych słów pomiędzy artykułami z różnych źródeł.
- System przeprowadza analizę statystyczną słownictwa w treści artykułów i generuje listę najczęściej występujących słów (z wyłączeniem tzw. stop-wordów, np. „i”, „że”, „na”).

Scenariusz 3: Tworzenie materiałów naukowych.

Jako: Student

Chcę: Wygenerować raport z wynikami

Aby: Włączyć go do prezentacji lub pracy naukowej.

Kryteria akceptacji:

- Raport tworzony jest z użyciem stylu akademickiego (obiektywny, klarowny, formalny język).
- System na podstawie danych generuje raport zawierający: streszczenie wyników, wizualizacje w postaci wykresów czy chmury słów.
- Wygenerowany dokument nie zawiera błędów formatowania, a wykresy i dane są poprawnie podpisane.