Analiza Sentymentu Wypowiedzi Polityków

Dokumentacja SRS

1. Wprowadzenie

Cel dokumentu

Dokument określa wymagania dotyczące systemu służącego do analizy sentymentu wypowiedzi polityków zaczerpniętych z transkryptów przemówień. System wykorzystuje metody przetwarzania języka naturalnego oraz wybrane słowniki sentymentu.

Zakres systemu

System analizuje pliki tekstowe zawierające transkrypcje wypowiedzi polityków oraz przyznaje każdemu słowu wskaźnik sentymentu (pozytywny, neutralny, negatywny) za pomocą czterech różnych słowników:

- GI (General Inquirer)
- HE (Harvard-IV)
- LM (Loughran-McDonald)
- QDAP (Quantitative Data Analysis Package)

Wyniki są agregowane dla poszczególnych partii politycznych i prezentowane w formie wykresów.

2. Cele systemu

Celem systemu jest dostarczenie narzędzia do automatycznej i obiektywnej analizy sentymentu wypowiedzi polityków na podstawie transkryptów. System ma na celu wsparcie badań politologicznych oraz analizy dyskursu publicznego i nastrojów wśród przedstawicieli różnych partii politycznych.

Cele szczegółowe

1. Analiza emocjonalnego wydźwięku wypowiedzi
2. Porównania międzygrupowe
    - Agregacja wyników na poziomie partii politycznych w celu porównania ogólnego sentymentu różnych ugrupowań.
    - Identyfikacja trendów i zmian sentymentu wśród różnych ugrupowań politycznych.
3. Wsparcie dla badań i dziennikarstwa
    - Przygotowanie obiektywnych materiałów służących do analiz politologicznych i socjologicznych.
    - Wsparcie dla mediów w monitorowaniu trendów sentymentu w debacie publicznej.

3. Wymagania funkcjonalne

3.1 Przetwarzanie danych wejściowych

- System powinien obsługiwać pliki .txt zawierające transkrypty wypowiedzi.
- System powinien usuwać fragmenty plików tekstowych, które nie stanowią bezpośredniej części wypowiedzi polityków - adnotacje w nawiasach.
- System powinien konwertować znaki specjalne (np. \n) na spacje.

3.2 Analiza sentymentu

- System powinien przeprowadzać analizę sentymentu za pomocą słowników: GI, HE, LM, QDAP.

- System powinien klasyfikować każde zdanie, jako pozytywne, neutralne albo negatywne.

3.3 Agregacja i wizualizacja wyników

- System powinien generować przejrzyste wykresy słupkowe przedstawiające rozkład sentymentu.
- System powinien zapewniać możliwość porównania wyników uzyskanych przy pomocy różnych słowników.

4. Wymagania niefunkcjonalne

4.1 Wydajność

- System powinien przetwarzać teksty do 1000 zdań w czasie poniżej 1 minuty.
- Wyniki powinny być dostępne w czasie rzeczywistym po przetworzeniu danych.

4.2 Użyteczność

- System powinien być dostosowany do obsługi przez osoby posiadające jedynie podstawową znajomość języka R.
- Użytkownik powinen móc korzystać z systemu, poprzez uruchamianie skryptu programu R w środowisku systemu zawierającym folder z plikami .txt, w których przechowywane są transkrypty wypowiedzi.
- Wykresy powinny być czytelne, należycie opisane, opatrzone zawierać legendą.

4.3 Kompatybilność i bezpieczeństwo

- System powinien działać w środowisku R (wersja 4.4.3 lub nowsza).
- System powinen działać lokalnie, na urządzeniu użytkownika. Nie powinna być wymagana łączność internetowa.

5. Interfejs użytkownika i wymagania dotyczące danych

Wejście:

- Pliki tekstowe (.txt) z danymi do analizy.
- Pliki słowników w formacie CSV

Wyjście:

- Zliczenie liczby słów i określenie ich nacechowania emocjonalnego
- Wykresy słupkowe przedstawiające natężenie sentymentów arytmetyczne dla każdego z słowników
- Wykresy słupkowe przedstawiające natężenie sentymentów sumaryczne dla każdego z słowników
- Raport HTML

Wymagania dotyczące danych:

- Dane tekstowe muszą być w języku angielskim.
- Skrypt nie obsługuje analizy sentymentu dla innych języków.
- Skrypt nie obsługuje plików większych niż 100 MB.
- Nazwy plików tekstowych zawierających wypowiedzi polityków powinny być w formacie: "ImięNazwisko_nazwaFrakcji.txt". Nazwy frakcji powinne być jedolite dla wszystkich wypowiedzi wykorzystywanych przy każdorazowej analizie.
- Skrypt wykorzystuje słowniki sentymentów dostępne w plikach .CSV oraz w pakiecie SentimentAnalysis.
- Skrypt nie obsługuje analizy sentymentu dla danych tekstowych z innych źródeł niż pliki .txt.

6. Słownictwo dokumentacji

Analiza sentymentu - proces automatycznej klasyfikacji emocjonalnego wydźwięku tekstu (pozytywny/neutralny/negatywny) przy użyciu słowników lub modeli NLP.

Sentyment - Nacechowanie emocjonalne wypowiedzi (np. pozytywne, negatywne, neutralne).

Słownik sentymentu - zbiór słów przypisanych do kategorii emocjonalnych.

Modele przetrwarzania języka naturalnego (NLP) - modele służące do analizy i przetwarzania tekst lub mowy mające na celu odwzorowanie wzorców postrzegania i myślenia u ludzi.

Tokenizacja - podział tekstu na pojedyncze słowa lub zdania.

Agregacja wyników - sumowanie lub uśrednianie sentymentu dla grupy (np. dla całej partii politycznej).

Data cleaning - proces czyszczenia tekstu (usuwanie nawiasów, znaków specjalnych itp.).

Środowisko systemu - w tym przypadku folder zawierający skrypt R odpowiadający za wykonanie analizy, pliki csv z danymi wykorzystywanymi przez słowniki analizy sentymentu, oraz folder z plikami .txt zawierającymi transkrypty wypowiedzi polityków.

Raport - HTML - dokument prezentujący i podsumowujący wyniki uruchomionego skryptu.

7. Przypadki użycia (use cases)

Użytkownik:

-Wczytuje folder plików tekstowych .txt.

-Uruchamia analizę.

-Wyświetla wyniki analizy.

-Generuje wykresy sentymentu oraz raport html

Skrypt:

- przetwarza teksty
- oczyszcza teksty
- analizuje sentyment tekstów przy użyciu słowników

- wylicza sumę i średnią arytmetyczną dla każdego z tekstów
- generuje wykresy do każdego z słowników, zarówno do sumy, jak i do średniej arytmetycznej

Testowe przypadki użycia:

1.Przeprowadzenie testu z plikiem .txt zawierającym tekst o wydźwięku pozytywnym.
2.Przeprowadzenie testu z plikiem .txt zawierającym tekst o wydźwięku negatywnym.
3.Przeprowadzenie testu z plikiem .txt zawierającym tekst o neutralnym wydźwięku.
4.Przeprowadzenie testu z plikiem .txt zawierającym tekst o mieszanym sentymencie.
5.Przeprowadzenie testu z plikiem .txt zawierającym brakujące dane.
6.Przeprowadzenie testu z plikiem .txt zawierającym znaki specjalne.

8. Scenariusze użytkownika (user stories)

Scenariusz 1: analiza sceny politycznej Wielkiej Brytanii przez dziennikarzy.

kto: dziennikarz analizujący sytuację polityczną w Wielkiej Brytanii
co chce osiągnąć: poznanie nacechowania emocjonalnego słów użytych w wystąpieniach poszczególnych polityków
w jakim celu: przekazanie informacji obywatelom dotyczących tego zakresu

Scenariusz 2: analiza reakcji obywateli na przemówienia przez socjologów

kto: socjolog skupiający się na analizie reakcji społeczeństwa na dane wystąpienie
co chce osiągnąć: przełożenie danych z analizy sentymentu poszczególnych przemówień na reakcję obywateli na nie, przekładając je następnie na analizy socjologiczne
w jakim celu: poszerzenie badań nad zachowaniem społeczeństwa Wielkiej Brytanii

Scenariusz 3: analiza retorycznej strony przemówień

kto: osoby zajmujące się retoryczną stroną przemówień: specjaliści, inni politycy, członkowie danej partii
co chce osiągnąć: poprzez analizę sentymentu można wskazać nastawienie emocjonalne w

danej mowie, co można następnie dalej wykorzystać jako punkt w dyskusji

w jakim celu: pozyskanie źródła argumentów w dyskusji publicznej