Dokumentacja Specyfikacji Wymagań (SRS)

Wprowadzenie

Celem projektu jest analiza programów polityków kandydujących na prezydenta RP w 2025 roku (przetłumaczonych na język angielski). Chcemy zbadać, jak pokrywają się postulaty pomiędzy poszczególnymi kandydatami, z jakim nacechowaniem mówią o konkretnych postulatach oraz które kwestie są dla nich najważniejsze. W skrypcie R generujemy chmury słów, przedstawiające częstości słów, dla kolejnych wczytanych z folderu dokumentów txt., dokonujemy analizy ich sentymentu, przy użyciu słowników w plikach CSV (Bing, NRC, Afinn), oraz (dla całego folderu - wszystkich postulatów) dokonujemy podziału na tematy i generuje wykresy słów o najwyższej informatywności przy użyciu LDA (ukrytej alokacji dirichleta) dla określonej liczby tematów.

Cele systemu

System ma za zadanie umożliwić głębsze poznanie poglądów kandydatów i ułatwienie podjęcia decyzji wyborczych. Program można też wykorzystywać w szerszych analizach socjologicznych i politologicznych w celu obserwowania zależności w świecie polityki, porównania programów wyborczych partii i ich konkretnych kandydatów.

System:

- wczytuje pliki txt. z wybranego folderu
- przetwarza i oczyszcza tekst (normalizacja, tokenizacja, stemming)
- usuwa nieistotne słowa “stop words”
- dla każdego pliku txt. zlicza słowa i generuje chmurę słów, które wystąpiły w tekście co najmniej 4 razy
- przeprowadza analizę sentymentu z użyciem słowników w plikach CSV.: Bing, NRC, Afinn
- analiza sentymentu przeprowadzana jest w ujęciu proporcjonalnym - co pozwala na ograniczenie różnic wynikających z różnych długości tekstów
- wykresy dla słowników Bing i Afinn prezentują zsumowaną wartość sentymentów wyrazów dla dokumentów (dla słownika Bing wartości positive przypisujemy wartość 1, a negative -1)
- dla słownika NRC generowane są dwa rodzaje wykresów

1. zestawienie oddzielnych wykresów dla różnych emocji z procentowym udziałem tych emocji w różnych dokumentach
2. zbiorczy wykres przedstawiający procentową zawartość emocji w kolejnych dokumentach

- następnie dokonuje podziału na tematy (dla wskazanej liczby tematów) i generuje wykresy wizualizujące słowa o największej informatywności, przy użyciu ukrytej alokacji Dirichleta (LDA)

Wymagania funkcjonalne:

- System umożliwia dodanie programu wyborczego kandydata w formacie tekstowym, poprzez wybór folderu zawierającego pliki txt. (w języku angielskim)
- Aplikacja działa jako skrypt uruchamiany w środowisku RStudio
- System wykonuje wstępne czyszczenie danych: usunięcie znaków specjalnych, adresów stron internetowych (ciągów wyrazowych zakończonych ".com", ".pl",) liczb, interpunkcji oraz normalizacja tekstu do małych liter, dokonuje stemming i uzupełnia rdzenie
- Dla każdego dokumentu generowana jest osobna chmura słów, pokazująca najczęściej występujące wyrazy
- system pozwala na wczytanie słowników Bing, NRC i Afinn z plików CSV
- System oblicza łączny sentyment wypowiedzi kolejnych dokumentów dla słowników Bing, Afinn, oraz procentowy udział konkretnych emocji w kolejnych dokumentach dla słownika NRC i obrazuje wyniki na wykresach
- System wskazuje tematy występujące we wszystkich wypowiedziach (analiza dla całego folderu) i przedstawia na wykresach słowa o najwyższej informatywności z poszczególnych tematów
- System generuje raport z analizy w formie czytelnego dokumentu HTML

Wymagania niefunkcjonalne

- System powinien przeanalizować do 10 plików o długości maks. 3 stron w ciągu maks. 3 minut
- System nie powinien generować błędów związanych ze słownictwem
- Obsługiwanie języka angielskiego
- Możliwość podglądu tekstu po przetworzeniu

- generowane wykresy powinny być czytelne i dobrze opisane

Interfejsy użytkownika

Wejście:

- Plik tekstowy .txt
- Folder plików tekstowych
- Pliki słowników sentymentów w formacie csv.
- Parametry analizy tekstowej (liczba słów w chmurze, tematów)

Wyjście:

- Podgląd oczyszczonego dokumentu
- Tabela najczęstszych słów w dokumencie
- Chmura słów
- Wykresy słupkowe pokazujące rodzaju sentymentu (AFINN ,BING ,NRC)

Wymagania dotyczące danych

- Dane tekstowe w postaci .txt
- Kodowanie UTF-8, do obsługi pojedynczych polskich znaków
- Dane tekstowe w języku angielskim

Słownictwo dokumentacji

- BING / NRC / AFINN - słowniki sentymentów. Klasyfikujące kolejno słowa jako pozytywne lub negatywne, przypisujące emocje i ocenę w skali od -5 do 5.
- Stopwords - słowa nie wnoszące wartości semantycznej do analizy
- Skumulowany sentyment - suma (lub różnica) ocen sentymentu dla wszystkich tokenów w dokumencie
- Korpus - zbiór dokumentów używany do analizy.

Przypadki użycia

Użytkownik:

- Wybiera folder zawierający pliki .txt (np. postulaty wyborcze)
- Definiuje liczbę tematów oraz wybiera wielkość chmur słów
- Uruchamia analizę
- Wizualizuje wyniki
- Generuje wykresy i raport html

System:

- Wczytuje pliki tekstowe
- Czyści dane tekstowe
- Tokenizuje teksty i przekształca je w strukturę analityczną
- Tworzy chmury słów
- Przeprowadza analizę sentymentu
- Buduje model tematów
- Tworzy wykresy najważniejszych słów, sentymentu

Scenariusze użytkownika

Scenariusz 1: Analiza zmian poglądów kandydata

- Jako: Dziennikarz
- Chcę: szybko i wygodnie przeanalizować, jak zmieniały się poglądy danego kandydata na przestrzeni czasu
- Aby móc: zadawać mu trafne, oparte na faktach pytania podczas wywiadów i reportaży

Scenariusz 2: Monitorowanie działań kontrkandydatów

- Jako: Kierownik kampanii wyborczej
- Chcę: śledzić ruchy kontrkandydatów
- Aby móc: elastycznie dostosowywać strategię kampanii i reagować na ich taktyki w czasie rzeczywistym

Kryteria akceptacji:

- System usuwa szum informacyjny (np. linki, znaki specjalne, stop-słowa) i przygotowuje dane do analizy.
- Użytkownik może analizować wypowiedzi wielu polityków (każdy dokument traktowany jako osobny przypadek)

- System umożliwia automatyczne wykrywanie tematów pojawiających się w wypowiedziach — można porównać, które tematy są dominujące u którego kandydata
- Można wizualizować emocjonalne różnice pomiędzy kontrkandydatami za pomocą wykresów procentowego udziału emocji