**Tytuł projektu:** ***"Analiza tekstów exposé premierów Wielkiej Brytanii z wykorzystaniem metod eksploracji tekstu i analizy sentymentu"***

**1. Wprowadzenie**

Niniejszy projekt dotyczy analizy tekstów exposé wygłaszanych przez premierów Wielkiej Brytanii (np. Cameron, Johnson czy Starmer). Celem projektu jest eksploracja języka politycznego oraz analiza sentymentu, tonu wypowiedzi oraz występujących tematów w przemówieniach politycznych. Analiza przeprowadzona została w języku R, z użyciem metod text miningu, takich jak analiza częstości słów, model Bag-of-Words, analiza sentymentu, statyczna oraz w czasie, z użyciem różnych słowników (np. Afinn czy QDAP) oraz wizualizacja wyników. Efektem projektu jest skrypt analizujący teksty w formacie .txt oraz raport HTML prezentujący wyniki.

**2. Cele systemu**

System został zaprojektowany w celu:

- porównania treści exposé premierów Wielkiej Brytanii na przestrzeni lat
- identyfikacji emocjonalnego ładunku wypowiedzi (np. strachu, zaufania, gniewu)
- znalezienia dominujących tematów i słów charakterystycznych dla danego polityka
- wsparcia badań z zakresu politologii, socjolingwistyki i komunikacji politycznej
- zapewnienia łatwego w użyciu narzędzia analitycznego, które umożliwia szybkie wnioski na podstawie surowych danych tekstowych

**3. Wymagania funkcjonalne**

System powinien:

- Umożliwić użytkownikowi wczytanie tekstu w formacie .txt
- Przetwarzać tekst: czyszczenie, normalizacja, tokenizacja, usuwanie stopwords, stemming
- Liczyć częstość występowania słów i generować chmurę słów
- Przeprowadzać analizę sentymentu przy pomocy słowników csv (Loughran, NRC, Bing i Afinn)
- Przeprowadzać analizę sentymentu w czasie z pomocą słowników z pakietu SentimentAnalysis (GI, HE, LM, QDAP)
- Wizualizować wyniki za pomocą wykresów (wykresy słupkowe), chmury słów, wykresy czasowe) oraz umożliwić użytkownikowi dostosowanie ich parametrów wizualnych
- Eksportować raport HTML z analizy (domyślnie z użyciem estetycznego layoutu, który użytkownik powinien móc modyfikować)

**4. Wymagania niefunkcjonalne**

System powinien być:

- **Wydajny**: Analiza pojedynczego tekstu nie powinna trwać dłużej niż 15 sekund
- **Niezawodny**: Ten sam tekst powinien dawać spójne wyniki niezależnie od środowiska
- **Przenośny**: Skrypt powinien niezawodnie działać na większości systemów z zainstalowanym językiem programowania R
- **Łatwo modyfikowalny**: Kod powinien być zaprojektowany tak, aby umożliwiał sprawne wprowadzanie zmian, rozwijanie nowych funkcjonalności oraz szybkie usuwanie błędów, bez ryzyka naruszenia stabilności systemu
- **Czytelny**: Wyniki powinny być przedstawione w sposób zrozumiały, również dla osób bez specjalistycznego wykształcenia
- **Użyteczny**: Obsługa systemu powinna być prosta i intuicyjna dla każdego użytkownika

**5. Interfejsy użytkownika i wymagania dotyczące danych**

**Wejście:**

- Plik .txt zawierający pełną treść exposé jednego premiera

**Wyjście:**

- Chmura słów
- Wykresy sentymentu statycznego dla każdego słownika
- Wykresy sentymentu skumulowanego dla każdego słownika
- Wykresy zmiany sentymentu w czasie (surowe i wygładzone)
- Raport HTML

**Wymagania danych:**

- Tekst w języku angielskim,
- Maksymalny rozmiar pliku: 100 MB,
- Pliki w formacie .txt i kodowaniu UTF-8.

**6. Słownictwo dokumentacji**

- **Token** – pojedynczy wyraz uzyskany po tokenizacji tekstu
- **Bag-of-Words (BoW)** – model przekształcający dokumenty w zbiór słów i ich częstości
- **Stopwords** – zbiór słów o niskiej wartości informacyjnej, np. “the”, “and”
- **Stemming** – redukcja słów do rdzeni (np. “running” → “run”)
- **Sentyment** – emocjonalna wartość przypisana słowom lub wypowiedzi,

- **Słownik sentymentu** – zbiór słów o przypisanej wartości emocjonalnej
- **Chmura słów** – wizualizacja najczęściej występujących słów w tekście

**7. Przypadki użycia**

**Użytkownik**:

- Wczytuje plik .txt z exposé
- Uruchamia analizę
- Otrzymuje wizualizacje i raport
- Porównuje przemówienia premierów między sobą

**System**:

- Wczytuje i czyści tekst
- Przetwarza dane (tokenizacja, stemming itd.)
- Oblicza częstość słów
- Generuje chmurę słów i wykresy
- Przeprowadza analizę sentymentu (statyczną i w czasie) z użyciem różnych słowników
- Wizualizuje efekty analizy sentymentu
- Eksportuje raport HTML.

**8. Testowe przypadki użycia**

- Test z exposé Churchilla (język formalny, wojenny ton)
- Test z exposé Borisa Johnsona (język potoczny, ekspresyjny)
- Test z exposé Liz Truss (krótkie exposé)
- Test z tekstem zawierającym neutralne słownictwo
- Test z tekstem pełnym negatywnych emocji (np. z czasów kryzysu)
- Test z nieangielskim tekstem – powinien zgłosić błąd lub ostrzeżenie
- Test z dużym plikiem (> 100 MB) – oczekiwane ostrzeżenie o przekroczeniu limitu
- Test z tekstem pełnym technicznego żargonu

**9. Scenariusze użytkownika**

**Scenariusz 1: Analiza porównawcza tonów exposé polityków**

**Jako**: Analityk polityczny
**Chcę**: Porównać sentyment przemówień trzech premierów
**Aby**: Zobaczyć, czy styl wypowiedzi zmienia się z czasem i sytuacją polityczną

**Kryteria akceptacji**:

- Możliwość uruchomienia kodu dla wielu plików .txt
- Możliwość wygenerowanie raportu html dla każdego pliku
- Możliwość dokonania analizy porównawczej z pomocą uzyskanych rezultatów

**Scenariusz 2: Wykrycie negatywnych tonów**

**Jako**: Dziennikarz polityczny
**Chcę**: Zidentyfikować, które exposé zawierały najwięcej negatywnych emocji
**Aby**: Zwrócić uwagę opinii publicznej na użycie retoryki strachu lub krytyki

**Kryteria akceptacji**:

- Możliwość zidentyfikowania negatywnie nacechowanego sentymentu w przemówieniach
- Możliwość wizualizacji rezultatów analizy sentymentu (statycznej i dynamicznej)
- Łatwość interpretacji wyników nawet przez dziennikarzy nieposiadających specjalistycznej wiedzy technicznej.

**10. Kryteria akceptacji**

- System poprawnie przetwarza pliki .txt z exposé
- Generuje analizę częstości słów i chmurę słów
- Przeprowadza analizę sentymentu dla każdego dokumentu
- Obsługuje wiele słowników sentymentu
- Raport HTML zawiera wykresy, które można łatwo zinterpretować
- Skrypt działa w mniej niż 15 sekund na pliku do 1MB
- Wyniki są powtarzalne