

1. Wprowadzenie

Dokument SRS opisuje wymagania dla systemu analizującego teksty na podstawie pliku "biblia.txt". Celem jest automatyzacja procesu tokenizacji, analizy częstości słów, analizy sentymentu oraz wizualizacji wyników. System ma wspierać badania lingwistyczne i analizy nastrojów w tekstach literackich.

2. Cele systemu

1. **Automatyzacja przetwarzania tekstu** – minimalizacja ręcznej interwencji przy analizie korpusu.
2. **Ekstrakcja częstości słów** – identyfikacja najpopularniejszych tokenów w tekście.
3. **Analiza sentymentu** – klasyfikacja tokenów na pozytywne i negatywne.
4. **Wizualizacja wyników** – generowanie wykresów słupkowych i chmury słów.
5. **Łatwa rozszerzalność** – możliwość modyfikacji poziomu filtrów, progów i źródeł słów.

3. Wymagania funkcjonalne

- **RF1. Instalacja i ładowanie pakietów:**
 - System wykrywa brakujące biblioteki i instaluje je.
- **RF2. Wczytywanie danych:**
 - Odczyt pliku "biblia.txt"; w przypadku braku pliku system zgłasza błąd i przerywa.
- **RF3. Przygotowanie danych:**
 - Tokenizacja tekstu na słowa (unnest_tokens).
 - Usunięcie stop-words (pakiet tidytext).
 - Normalizacja do małych liter.
- **RF4. Analiza częstości:**
 - Zliczanie unikalnych słów i sortowanie malejąco.
- **RF5. Analiza sentymentu:**
 - Pobranie słownika "bing" z pakietu textdata.
 - Klasyfikacja i zliczenie słów w kategoriach pozytywnych/negatywnych.
- **RF6. Wizualizacje:**
 - Wykres słupkowy top 20 słów.
 - Wykres rozkładu sentymentu.
 - Chmura słów z parametrem min.freq = 2, max.words = 100.

4. Wymagania нефunkcjonalne

- **WN1. Wydajność:**
 - Analiza pliku do 1 MB w czasie < 10 s.
- **WN2. Użyteczność:**

- Uruchomienie jednym poleceniem `source("analysis_script.R")`.
- **WN3. Niezawodność:**
 - Obsługa błędów (brak pliku, brak pakietu).
- **WN4. Przenośność:**
 - Działa na Windows, macOS, Linux w R ≥ 4.0 .
- **WN5. Odtwarzalność:**
 - Kod jednoznaczny, komentarze, wersje pakietów.

5. Interfejsy użytkownika i wymagania dotyczące danych

- **Interfejs konsolowy:**
 - Użytkownik wykonuje `source("analysis_script.R")` w konsoli R.
 - Komunikaty informują o postępie i ewentualnych błędach.
- **Dane wejściowe:**
 - Plik tekstowy UTF-8: "biblia.txt".
- **Dane wyjściowe:**
 - Wykresy w oknie graficznym R + pliki PNG (top20_words.png, sentiment_distribution.png).
 - Chmura słów w panelu Plots.

6. Słownictwo dokumentacji

- **Tokenizacja:** Rozbicie tekstu na pojedyncze słowa (tokeny).
- **Stop-words:** Słowa o niskiej wartości semantycznej (np. "the", "and").
- **Sentiment:** Emocjonalne zabarwienie słowa (pozytywne/negatywne).
- **Chmura słów:** Graficzna prezentacja słów, gdzie wielkość odpowiada częstotliwości.
- **Tibble:** Nowoczesny odpowiednik ramki danych w tidyverse.

7. Przypadki użycia (use cases)

Nazwa	Aktor	Cel	Opis
UC1: Start analizy	Analityk	Uruchomienie skryptu	Analityk wywołuje <code>source("analysis_script.R")</code> .
UC2: Walidacja	System	Sprawdzenie istnienia pliku	System weryfikuje, czy "biblia.txt" istnieje.
UC3: Generowanie	System	Przygotowanie wizualizacji	System generuje wykresy słupkowe i chmurę słów.

8. Scenariusze użytkownika (user stories)

1. **Jako analityk**, chcę, aby skrypt automatycznie pobierał i instalował potrzebne pakiety, żeby od razu móc uruchomić analizę.
2. **Jako badacz**, chcę zobaczyć listę najczęściej występujących słów w "biblia.txt", żeby zrozumieć główne tematy tekstu.
3. **Jako lingwista**, chcę poznać ogólny sentyment korpusu, aby ocenić, czy ton tekstu jest pozytywny czy negatywny.

4. **Jako użytkownik**, chcę zobaczyć wyniki graficznie (wykresy, chmurę), aby łatwiej wyciągać wnioski.