Narzędzie do analizy statystycznej słów i bigramów.

Michał Bobowski, Marcin Cieślikowski

2014-01-16

1 Wstęp

Niniejszy dokument stanowi podsumowanie projektu z przedmiotu WEDT. Zawiera opis interfejsu użytkownika i logiki programu oraz przedstawienie struktury kodu źródłowego.

2 Koncepcja programu

Program ma za zadanie obliczyć statystyki słów i bigramów z jednego lub wielu plików wejściowych. Obsługiwane powinny być najbardziej popularne formaty plików w tym pliki MS Word. Wyniki obliczeń powinny być przechowywane na dysku w rozsądnym formacie i możliwe do wykorzystania w przyszłości.

Program powinien zawierać prosty graficzny interfejs użytkownika. Jego zadaniem jest pobranie parametrów obliczeń oraz wyświetlenie wyników symulacji. Możliwa powinna być również filtracja wyświetlanych danych danych.

3 Przypadki użycia

3.1 Przeprowadzenie obliczeń

Podstawowym zadaniem programu jest przeprowadzenie obliczeń i wyświetlenie ich na ekranie. Scenariusz tworzą następujące zdarzenia:

1. Użytkownik wybiera parametry.
2. System przeprowadza obliczenia.
3. System zapisuje wyniki do pliku.

4. System wyświetla wyniki na ekranie.

3.2 Wczytanie wyników

1. Użytkownik wybiera ścieżkę do pliku z zapisanymi wynikami.
2. System wyświetla wyniki na ekranie.

3.3 Filtracja

Operacja filtracji staje się dostępna dopiero po wykonaniu któregoś z wcześniejszych przypadków użycia.

1. Użytkownik wybiera jedną z tabel wynikowych.
2. Użytkownik wypełnia filtry.
3. System wyświetla przefiltrowane wyniki na ekranie.

4 Specyfikacja szczegółowa

W tej części doprecyzowane zostały wymagania dotyczące danych wejściowych i wyjściowych.

4.1 Parametry wejściowe

Przed przeprowadzeniem obliczeń użytkownik może zdefiniować następujące parametry:

1. Ścieżka do pliku wejściowego lub katalogu zawierającego wiele plików wejściowych.
2. Typ bigramu: obliczany dla kolejnych słów lub wszystkich słów w tekście.
3. Części mowy dla słów z bigramu.
4. Nazwa pliku wyjściowego.

4.2 Prezentacja danych

Statystyki słów/bigramów są liczone i prezentowane dwa razy - dla słów z odmianą oraz dla formy podstawowej. Statystyki dla słów:

1. Liczba wystąpień w zbiorze.
2. Liczba zdań w których wystąpiło słowo.
3. Liczba dokumentów w których wystąpiło słowo.
4. Procent dokumentów w których wystąpiło słowo.
5. Miara tf-idf.

Statystyki dla bigramów:

1. Liczba wystąpień w zbiorze.
2. Liczba zdań w któtch wystąpił bigram.
3. Liczba dokumentów w których wystąpił bigram.
4. Procent dokumentów w których wystąpił bigram.
5. Miara tf-idf.
6. Prawdopodobieństwo słowa 1.
7. Prawdopodobieństwo słowa 2.
8. Prawdopodobieństwo bigramu złożonego ze słów 1 i 2.

4.3 Format danych wyjściowych

Dane wyjściowe są zapisywane do pliku bazy danych sqlite. Baza danych zawiera 8 nastepujących tabel:

1. Tabela words zawiera statystyki słow. Tabela ta zawiera następujące pola:
   (a) word - Nazwa słowa.
   (b) count - Liczba wystąpień słowa w korpusie.
   (c) numberOfSentences - Liczba zdań zawierających słowo.
   (d) numberOfDocuments - Liczba dokumentów zawierających słowo.

(e) percentOfDocuments - Procent dokumentów zawierających słowo.

2. Tabela Lemmas zawiera statystyki dla słów w formie podstawowej. Pola tej tabeli są identyczne, jak w tabeli words.

3. Tabela WordTFIDF - Zawiera wskaźnik TF-IDF. W tej tabeli znajdują się następujące pola:

   (a) word - Słowo, dla którego obliczany jest wskaźnik TF-IDF.

   (b) document - Nazwa dokumentu, dla którego oblicznay jest wskaźnik TF-IDF.

   (c) TFIDF - Wartość wskaźnika TFIDF

4. Tabela bigrams zawiera statystyki bigramów. Tabela ta zawiera następujące pola:

   (a) word1 - Pierwsze słowo w bigramie.

   (b) word2 - Drugie słowo w bigramie.

   (c) count - Liczba wystąpień bigramu w korpusie.

   (d) numberOfSentences - Liczba zdań zawierające bigram.

   (e) numberOfDocuments - Liczba dokumentów zawierających bigram.

   (f) percentOfDocuments - Procent dokumentów zawierających bigram.

   (g) percentOfSentence - Procent zdań zawierających bigram.

   (h) percentOfSentencew1 - Procent zdań zawierających pierwsze zdanie.

   (i) percentOfSentencew2 - Procent zdań zawierających drugie zdanie.

5. Tabela bigramsLemma zawiera statystyki bgramów w formie podstawowej. Pola tabeli są takie same, jak w tabeli bigrams.

6. Tabela bigramsTFIDF zawiera wskaźnik TF-IDF dla bigramów. Tabela zawiera następujące pola:

   (a) word1 - Pierwsze słowo w bigramie.

   (b) word2 - Drugie słowo w bigramie.

   (c) TFIDF - Wartość wskaźnika TFIDF

7. Tabela bigramsLemmaTFIDF zawiera wskaźnik TF-IDF dla bigramów w formie podstawowej. Pola tabeli pokrywaja się z polami tabel bigramsTFIDF

5 Wybór narzędzi i technologii

Program został zrealizowany w języku Java. Kod tworzyliśmy przy użyciu środowiska Eclipse oraz częściowo Netbeans (edytor interfejsu użytkownika). Do kontroli kodu wykorzystaliśmy system Git.

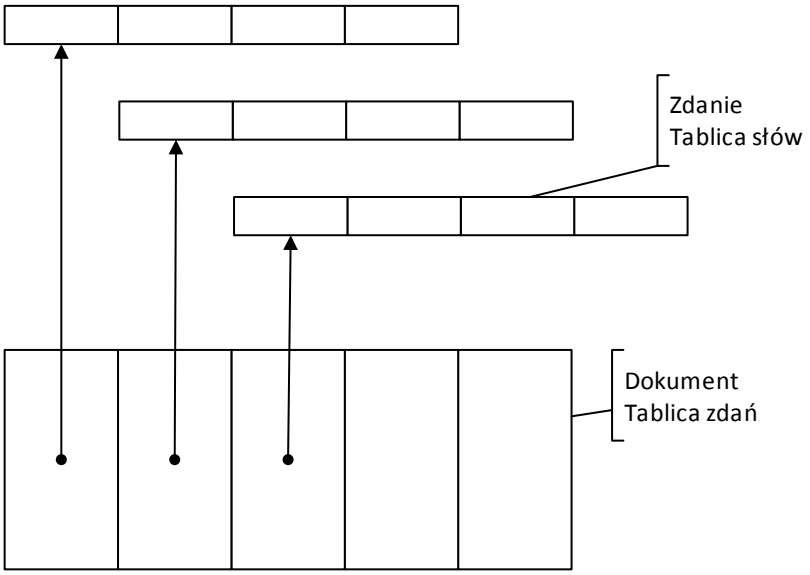
Do oznaczenia części mowy wykorzystaliśmy bibliotekę Gate. Korzysta ona wewnętrznie z biblioteki TIKA, dzięki czemu uzyskaliśmy wsparcie dla wielu formatów tekstowych m. in. txt, html, odt i doc.

6 Opis kodu źródłowego

7 Algorytmy

Program dla podanych dokumentów uruchamia funkcje bibliotęki GATE, które dzielą tekst na słowa, zdania. Dodatkowo dla słów są określanę części mowy i formy podstawowe słów. Dane zwróconę przez GATE są przetwarzane do następującej formy:

danych.pdf



Rysunek 1: Struktura danych