Prawo Zipf'a

Wstęp

Prawo Zipf'a – jest to prawo, które mówi, iż częstotliwość występowania słów w tekście powinna być odwrotnie proporcjonalna do pozycji w rankigu, gdzie ranking jest tworzony poprzez zliczenie częstotliwości występowania słów oraz posortowania malejąco powstałej listy. Tzn. pierwsze napotkane przez algorytm słowo będzie występowało około dwa razy częściej niż drugie słowo z rankingu.

Narzędzia

Do rozpatrywania tego zagadnienia użyłem napisanego przeze mnie algorytmu w języku Java. Do uruchomienia projektu lokalnie wystarczy zainstalowany pakie jre w wersji 8.

Dane

Teksty książek do analizy zostały zaczerpnięte z pl.wikisource.org, z portalu autocentrum.pl, oraz ze strony ae-lib.org.ua/texts-c Do testów wybrałem następujące książki/artykuły:

- Henryk Sienkiewicz „Krzyżacy” cz.2
- Bolesław Prus „Lalka” tom I
- 3 najnowsze publikacje z portalu autocentrum.pl
- John R.R. Tolkien „The Lord Of The Rings: Return Of The King” book V

W tekstach zostały usunięte numery oraz tytuły rozdziałów, dane są załączone wraz z kodem programu na githubowym repozytorium tutaj. Program należy odpowiednio uruchomić z linii poleceń systemu z podaniem argumentów w postaci ścieżek do plików, które mają zostać przeanalizowane.

Opis eksperymentów

Do sprawdzenia prawa Zipf'a dla polskich tekstów użyłem metody zliczania słów w tekście, taka metoda pomoże w prosty sposób stwierdzić czy prawo Zipfa jest prawdziwe. Jak wspomniałem wcześniej do zbadania tego zagadnienia wykorzystuję własny algorytm, którego działanie najłatwiej będzie zademonstrować na przykładzie. Załóżmy, że mam do przeanalizowania następujące zdanie:

```
„— Waryat! waryat!... Awanturnik!... Józiu, przynieśno jeszcze piwa. A która to butelka?”.
```

Algorytm w pierwszym kroku zamieni wszyskie duże litery na małe, oraz usunie znaki specjalne, których nie opisuje prawo Zipf'a. Tekst będzie wyglądał w następujący sposób:

```
„ waryat waryat awanturnik józiu przynieśno jeszcze piwa a która to butelka”
```

W kolejnych krokach algorytm podzieli tekst na listę słów oraz usunie słowa które są puste, lista przy użyciu przykładowego zdania będzie wyglądała tak:

```
 [waryat, waryat, awanturnik, józiu, przynieśno, jeszcze, piwa, a, która, to
,butelka]
```

Kolejnym krokiem jest zliczenie występowania słów w tablicy, w tym kroku powstanie mapa słowno-numeryczna:

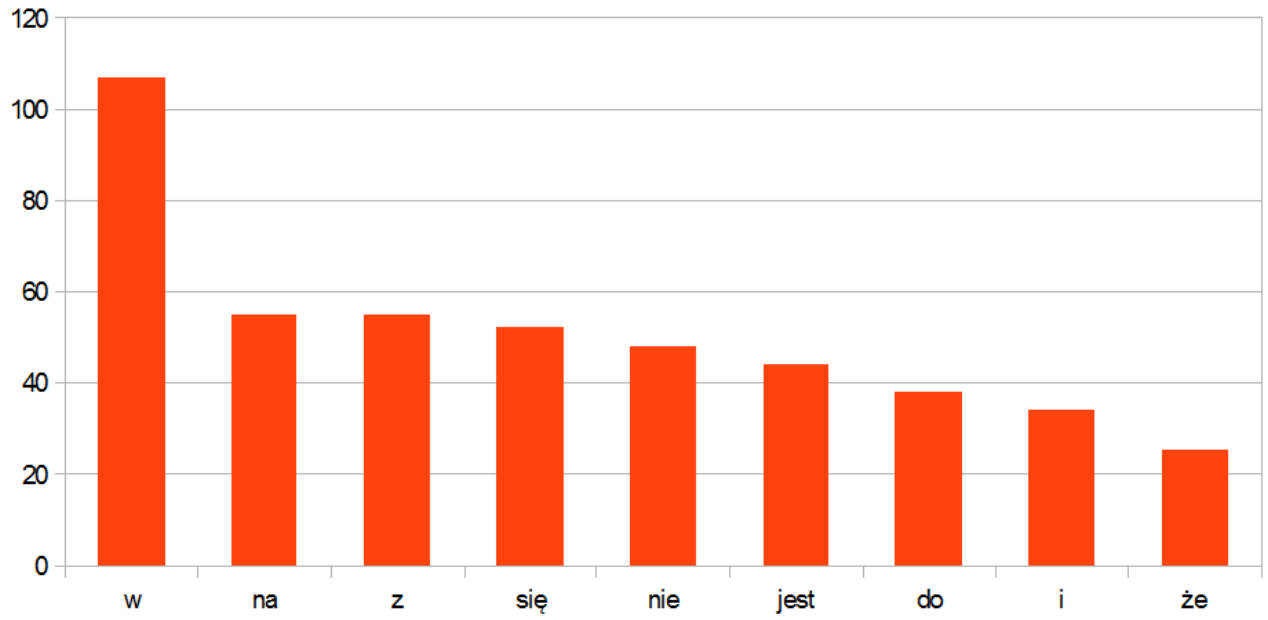
```
{[waryat: 2], [awanturnik: 1], [józiu: 1], [przynieśno: 1], [jeszcze: 1], [piwa:
1], [a: 1], [która: 1], [to: 1], [butelka: 1]}
```

Ostatnim krokiem algorytmu jest posortowanie malejąco powstałej mapy, w tym przykładzie mapa zostanie bez zmian. Przy tym przykładzie ciężko stwierdzić czy prawo Zipf'a działa, ponieważ jest to zbyt krótki przykład.

Wyniki

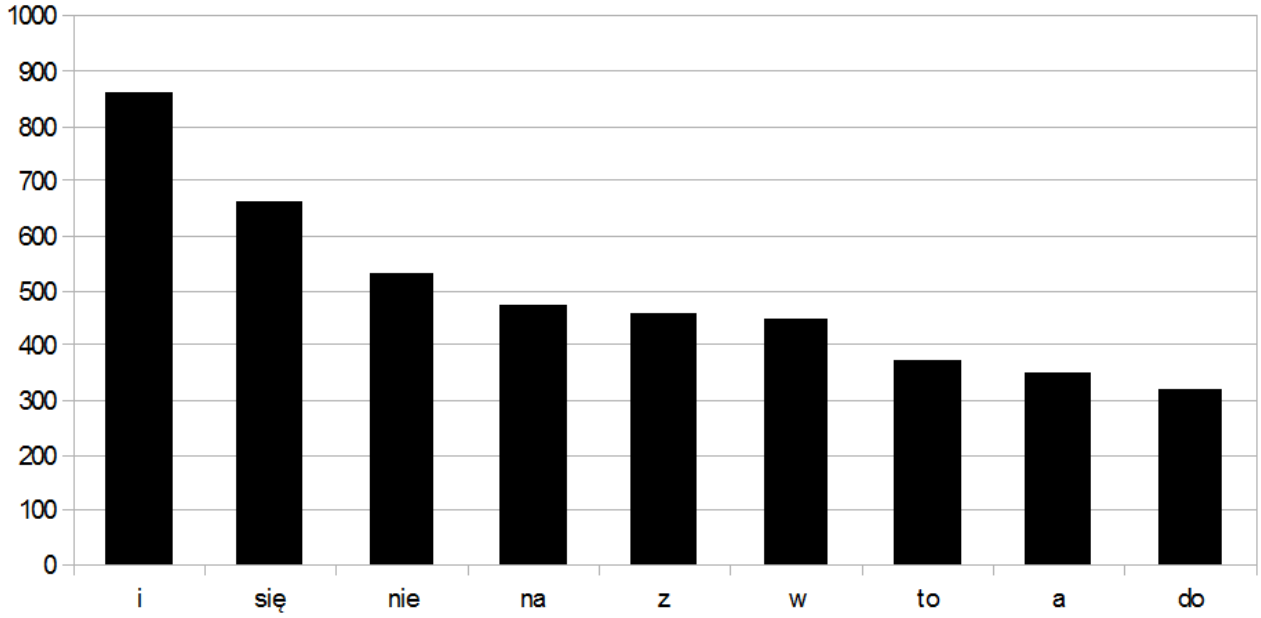
1. *Publikacje z portalu autocentrum.pl*; Liczba słów: 2888

wykres ilości występowań słów w tekście:



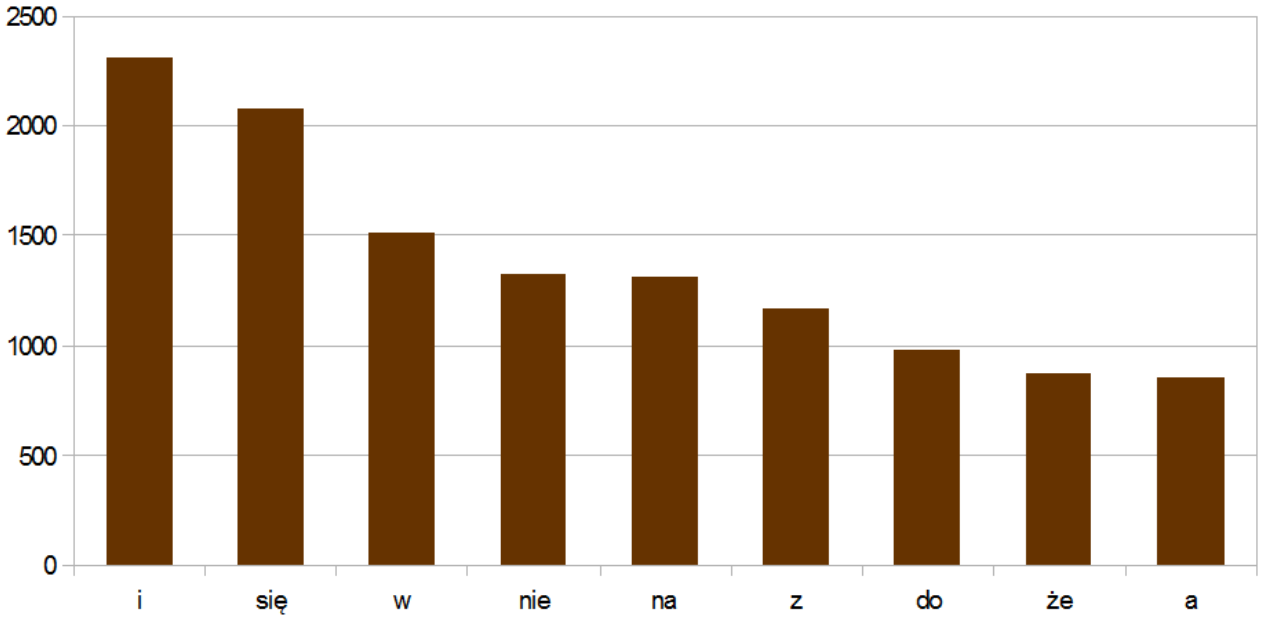
3. *„Krzyżacy”*; Liczba słów: 23379

wykres ilości występowań słów w tekście:



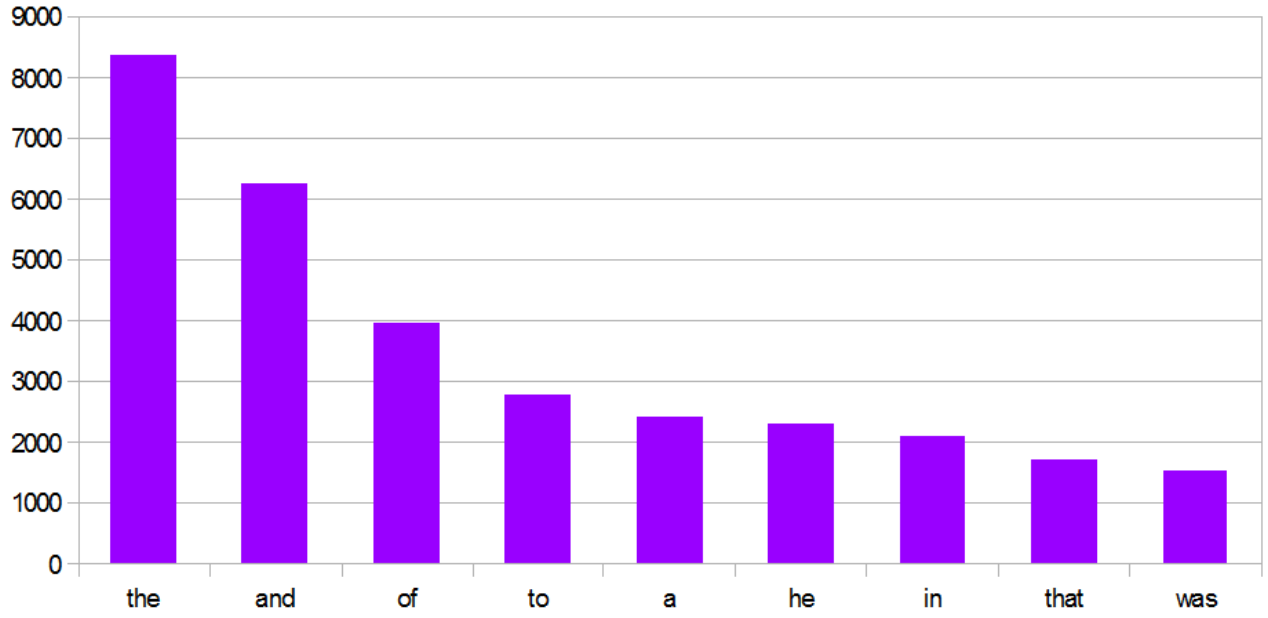
4. *„Lalka”*, Liczba słów: 74923

wykres ilości występowań słów w tekście:



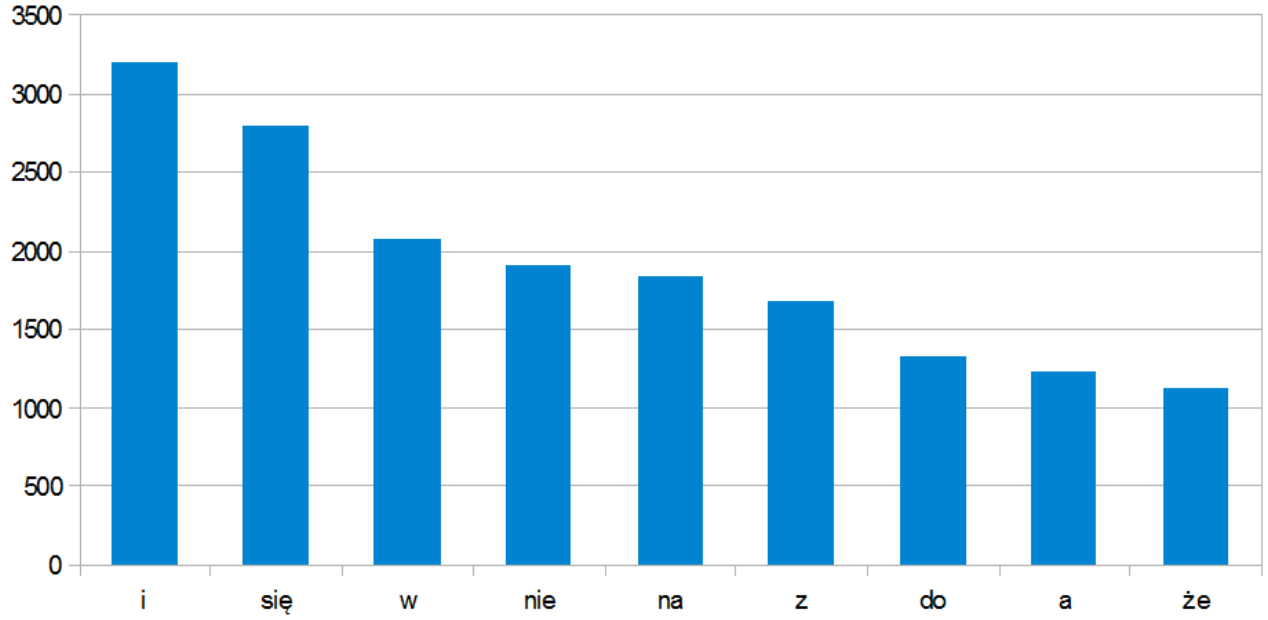
5. *„The Lord Of The Rings: Return Of The King”*, Liczba słów: 132722

wykres ilości występowań słów w tekście:



Interpretacja wyników

Pierwszym moim spostrzeżeniem jest to, że w każdym wykresie widnieją te same słowa (+/-1 słowo), są one tylko lekko pozamieniane kolejnością w poszczególnych wykresach. Słowa te należą do tak zwanych ***stopwords***, czyli słów które same nie posiadają żadnej informacji, jeżeli by spojrzeć na podsumowanie występowań słów, które znajdują się poniżej, można zaobserwować, że są to tylko i wyłącznie takie słówka



Podsumowując, nie udało mi się udowodnić poprawności prawa Zipf'a dla polskich tekstów, przypadek publickacji z autocentrum, wydawał się dość obiecujący, można tam zauważyć spełnienie zależności prawa dla dwóch pierwszych kolumn, niestety im dalej tym mniej różnic, podobnie jest z

resztą polskich źródeł. Można dodać, że mimo większej ilości słów w fragmencie **Lalki**, wykres wydaje się być bardzo zbliżony wizualnie do wykresu **Krzyżaków**, z początku widać tendencję spadkową, niestety reszta wykresu jest płaska. Dla kontrastu spróbowałem tekstu anglojęzycznego - **Lord Of The Rings**, tam prawo Zipf'a jakby bardziej się sprawdza, cztery kolejne słupki są od siebie niższe, niestety też nie jest idealnie. Możliwe, że prawo Zipf'a byłoby lepiej widoczne na jeszcze większych tekstach tzn. na całych książkach.