**Prawo Zipp'fa**

**Wstęp**

Projekt służy do analizy języka i sprawdzenia czy prawo Zippfa zachodzi w języku polskim.

**Narzędzia**

Program został napisane w javie 8, która jest do pobrania tutaj: http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html.

Używałem środowiska IntelliJ Idea, jednak jest to projekt mavenowy i powinien bez problemu się zaimportować do każdego innego javowego środowiska. Niżej jest przykład jak zrobić to dla IntelliJ

Wykresy zostały stworzone w programie MS Excel.

**Dane**

Książki w formacie .txt pobrałem z jednego z polskich popularnych serwisów do udostępniania plików, są to:

Basnie i legendy - SIENKIEWICZ HENRYK

Bastion - KING STEPHEN

Bezkresne morze - MACLEAN ALISTAIR

Dzieci Ziemi #5 Kamienne Sadyby - AUEL JEAN

Potop - SIENKIEWICZ HENRYK

Trylogia Rzymska #3 Moj syn Juliusz - WALTARI MIKA

Wielkie Sekretne Widowisko - BARKER CLIVE

Książki mają od 46000 - 9500 słów.

**TEORIA**

Prawo Zipf'a mówi, że jeśli w języku najczęstsze słowo występuje x razy, to słowo na

drugim miejscu pod względem częstotliwości wystepuje x/2 razy, na trzecim x/3 razy, itd.

ogólnie: jeśli słowo na i-tym miejscu w

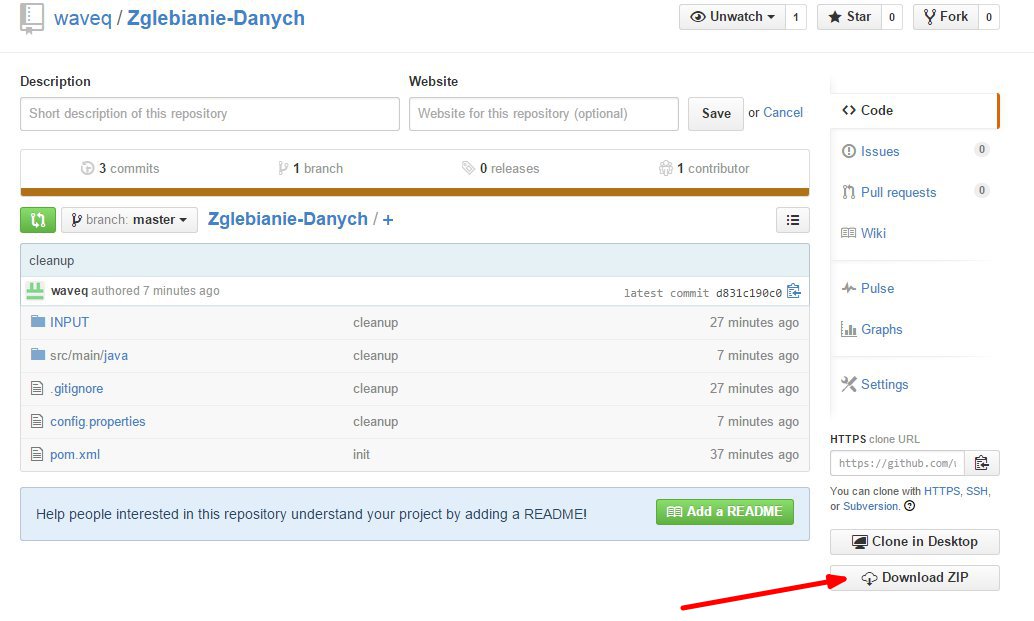
rankingu wszystkich słów w tekście wystepuje P razy, to i\*P=0,1.

**OPIS EKSPERYMENTOW**

0) Zainstaluj javę 8 http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html

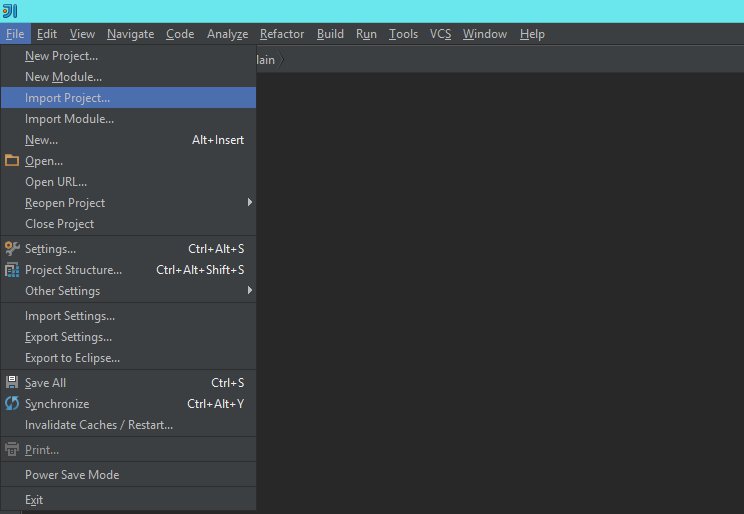
1) Udaj się do https://github.com/waveq/Zglebianie-Danych

2) Pobierz repozytorium

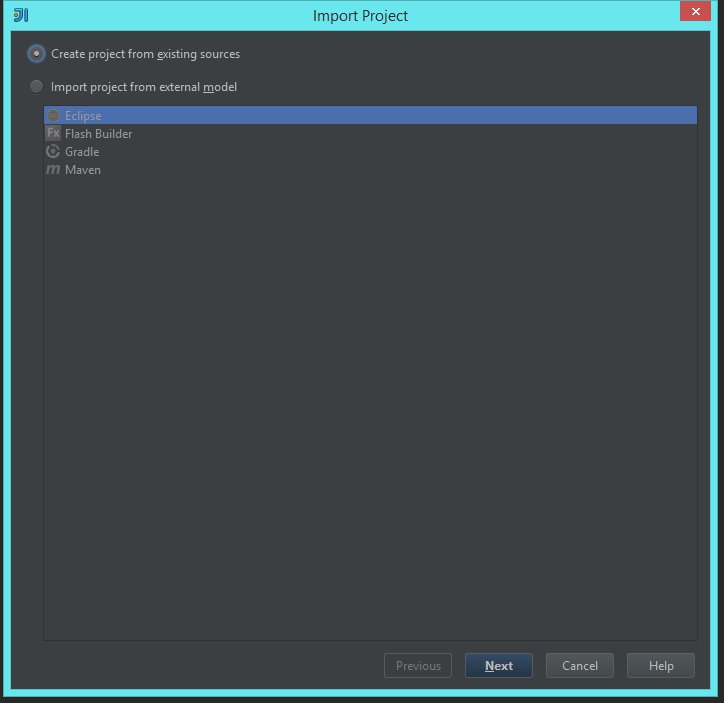


3) Wypakuj repozytorium i uruchom IntelliJ Idea Community

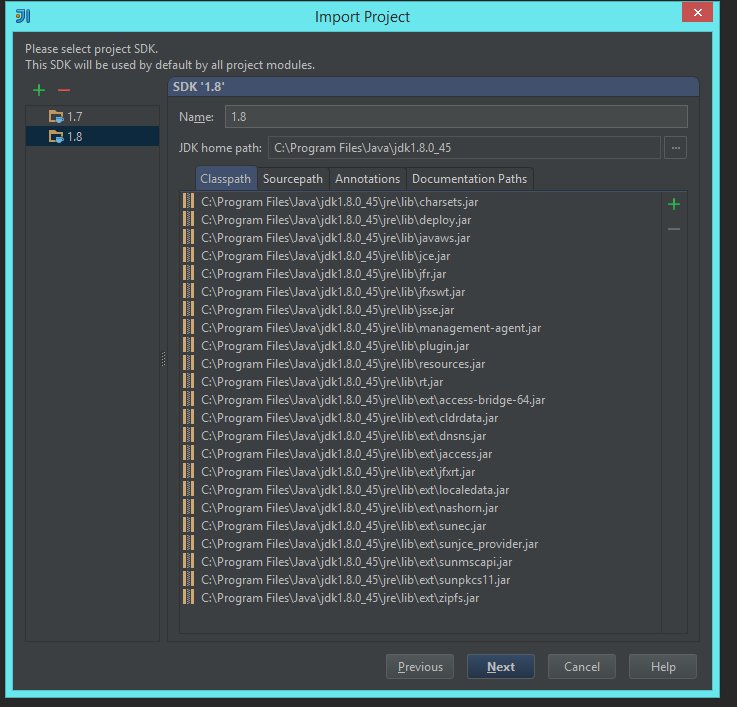
4) W intelliJ wybierz File > Import Project...



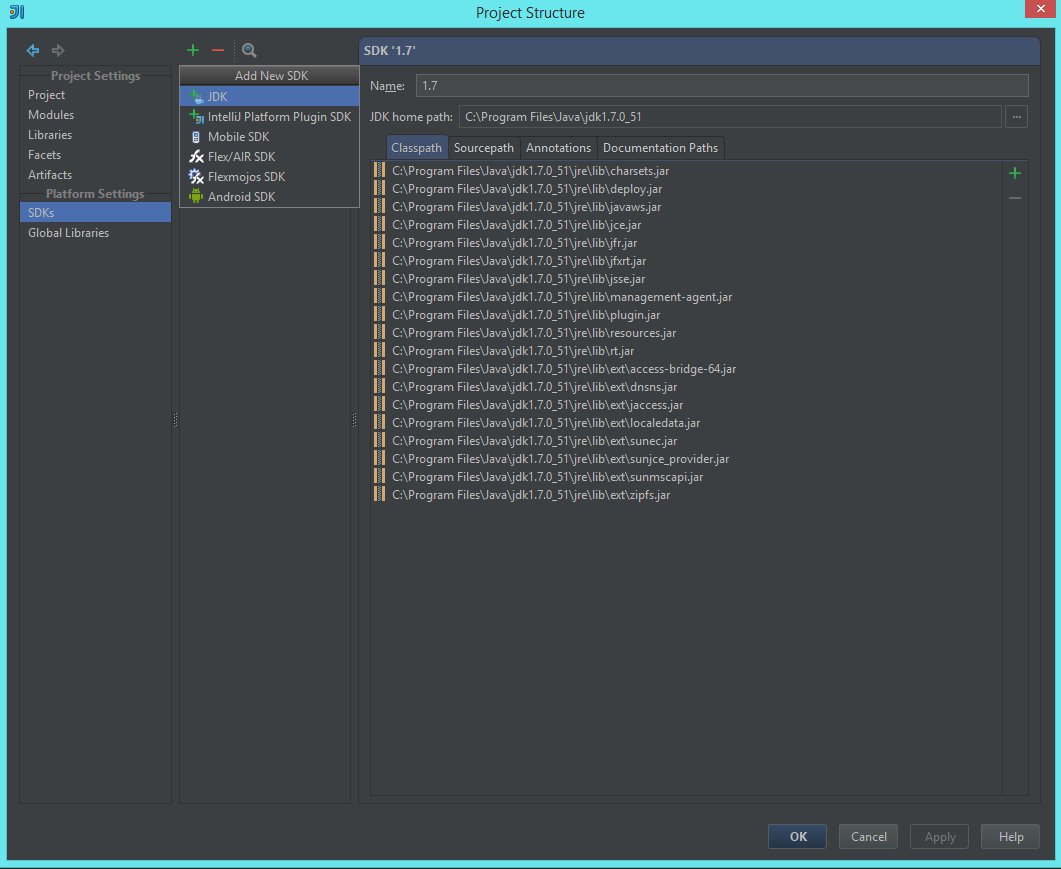
5) Wybierz swój rozpakowany projekt i w następnym oknie zaznacz przycisk "Create project from existing sources" i przechodź dalej aż natrafisz na wybór SDK



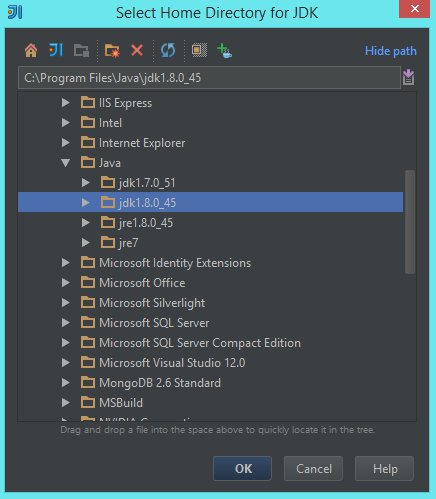
6) Spośród dostępnych SDK wybierz wcześniej zainstalowaną javę 8



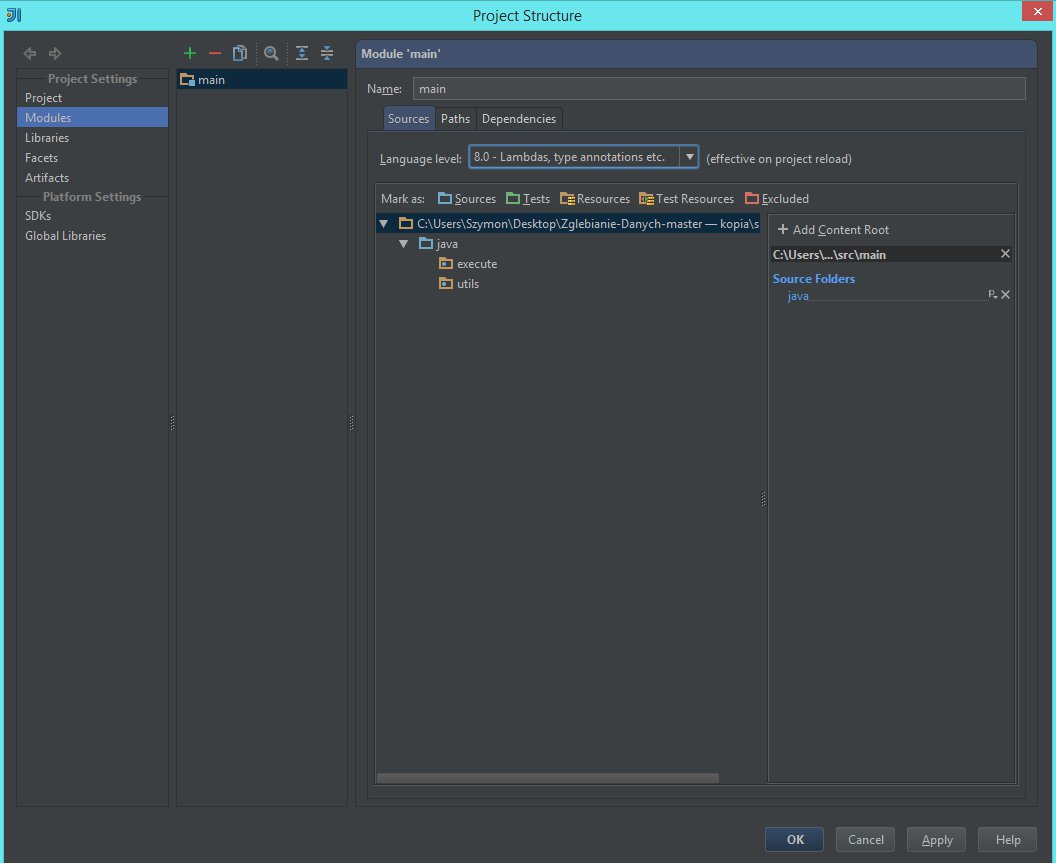
6.1) Jeśli zainstalowałeś javę 8 a nie masz jej do wyboru naciśnij zielony plus a następnie JDK



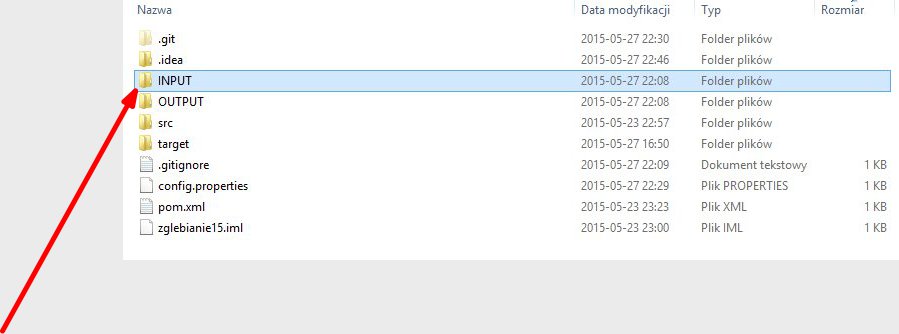
6.2) Podaj ścieżkę do zainstalowanej javy 8



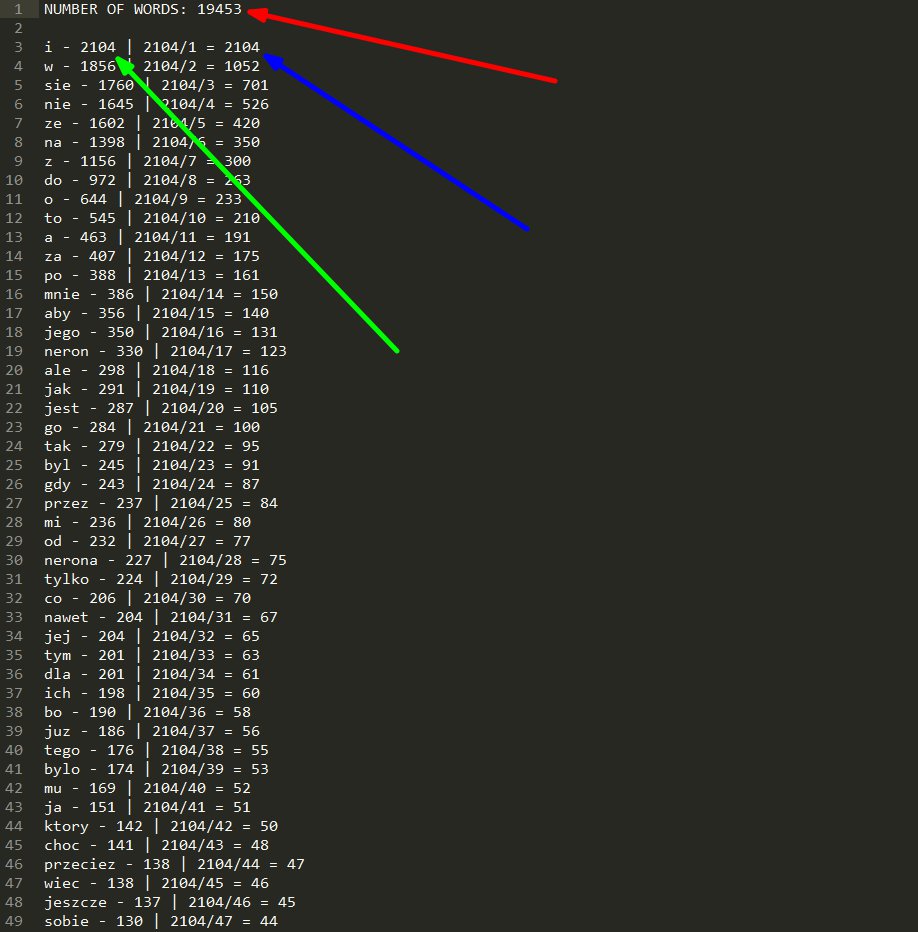
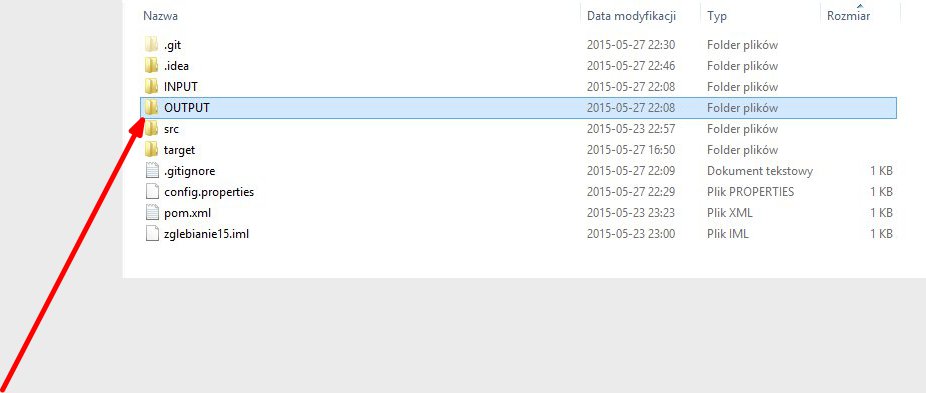
7) Otwórz File > Project Structure i ustaw Language level na 8.0



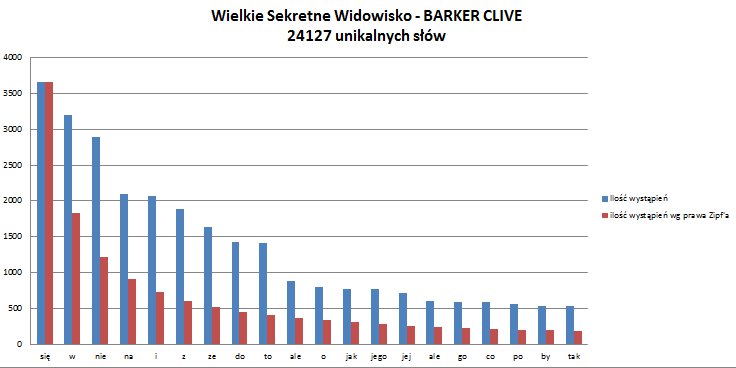
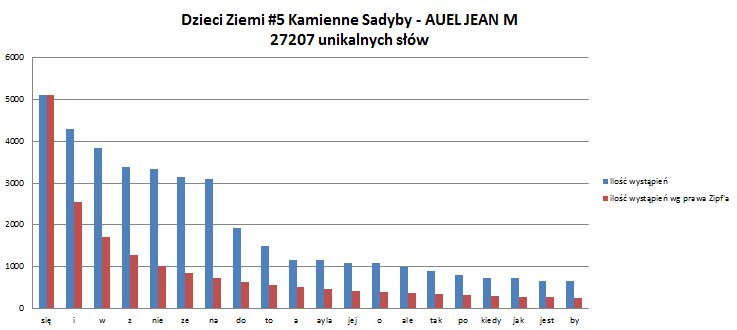
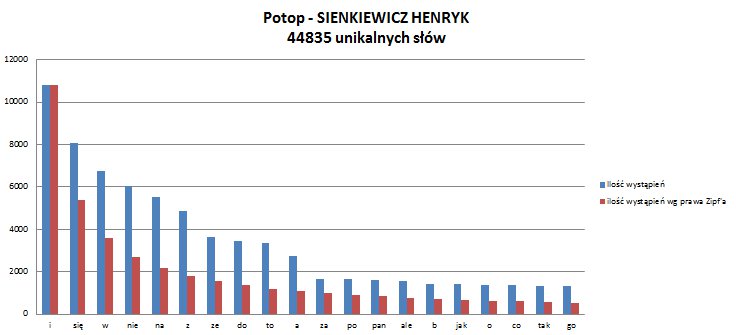
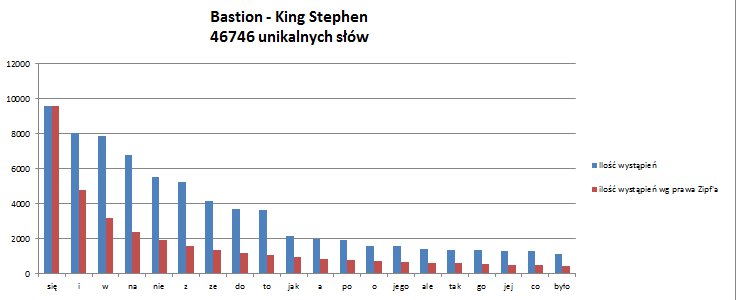
8) Otwórz plik config.properties i jako wartość zmiennej file.input podaj nazwę pliku w formacie .txt znajdującego się w katalogu INPUT, na którym chcesz zacząć badania. 9) Otwórz klasę Main.java (ctrl + shift + n) a następnie ją uruchom (ctrl + shift + f10)

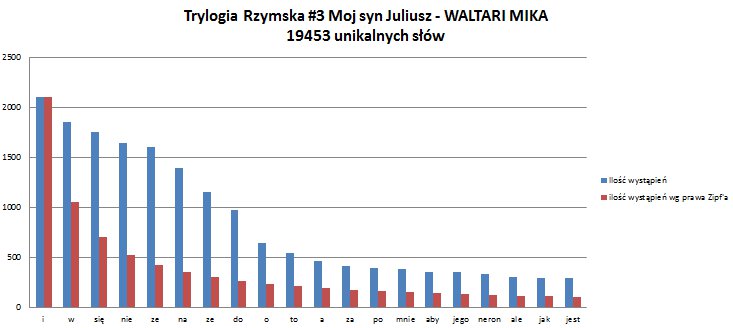


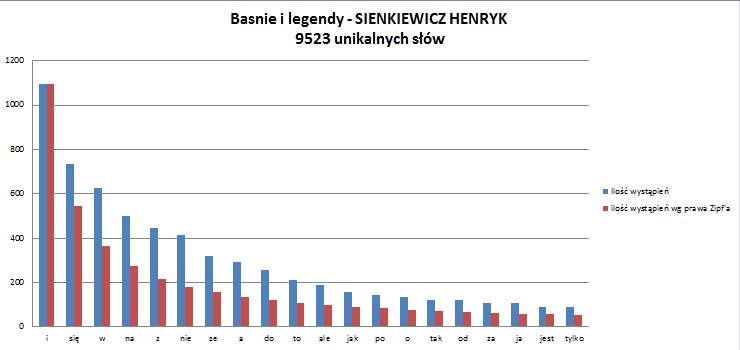
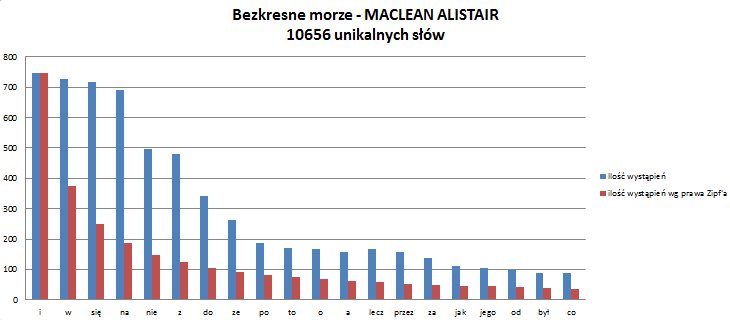
10) Wynik badań znajduje się w katalogu OUTPUT. 11) W pierwszej linii pliku z wynikami jest liczba wszystkich słow, a w kolejnych posortowane malejąco słowa i ilość ich występowań. Po znaku '|' wyliczona została wartość, ile razy słowo powinno występować wg. prawa Zipf'a.



**Wyniki**



****



**Interpretacja wyników**

Na podstawie przeprowadzonych badań można stwierdzić, że prawo Zippf’a w języku polskim nie zachodzi. Występowanie unikalnych słów jest niemal zawsze wyższe niż zakłada prawo. Można jednak dostrzec słowa, które mają znaczną przewagę w niemal każdym dziele zaliczają się do nich:

*i, w, się, na, nie, z, do, ze ,po, to, do.* Zależnie od dzieła w czołówce unikalnych słów pojawia się czasem słowo specyficzne dla danej książki np dla **Trylogii Rzymskiej** jest to *Neron* na 17 pozycji.