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.

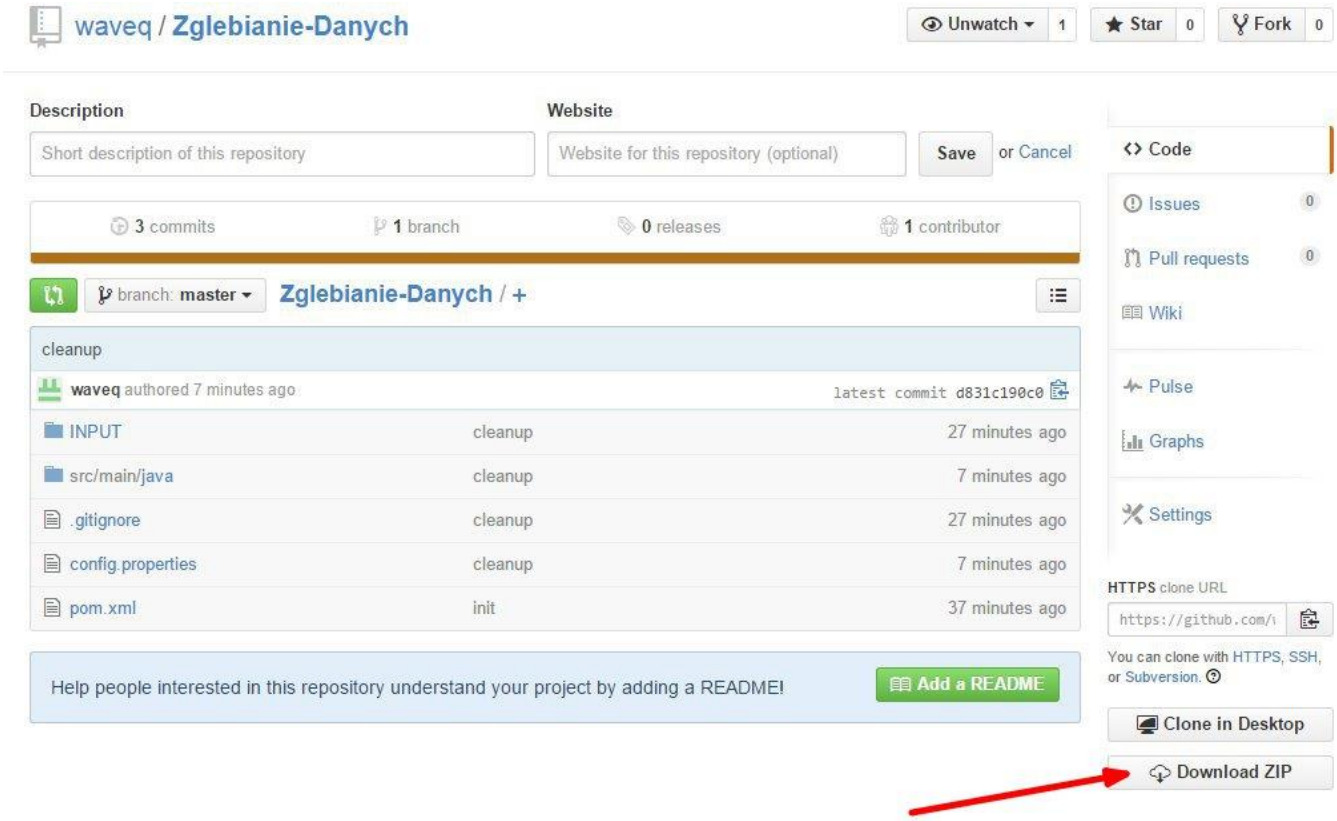
TEORIA

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

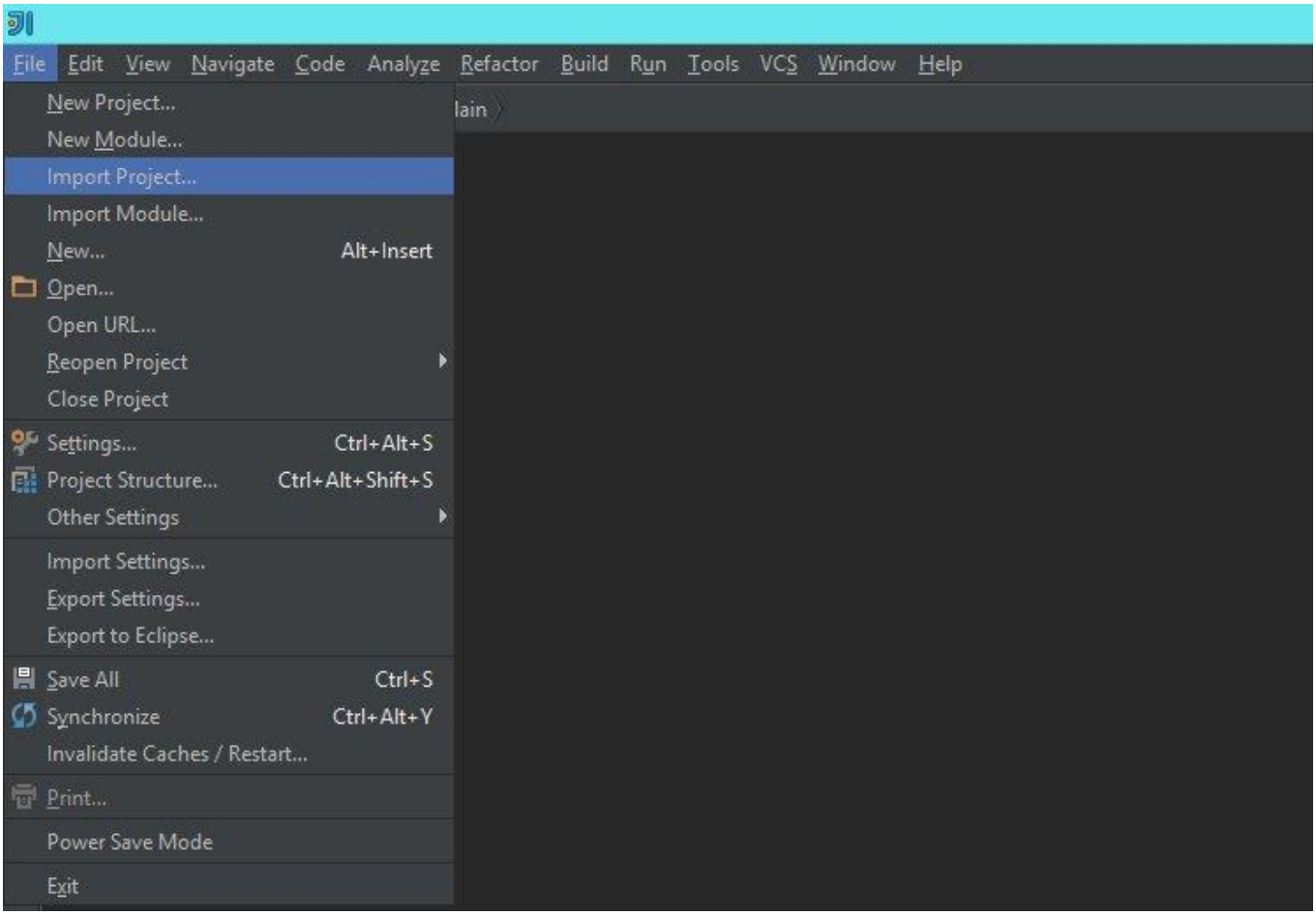
0) Zainstaluj javę 8 http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html
1) Udaj się do https://github.com/waveq/Zglebianie-Danych

2) Pobierz repozytorium

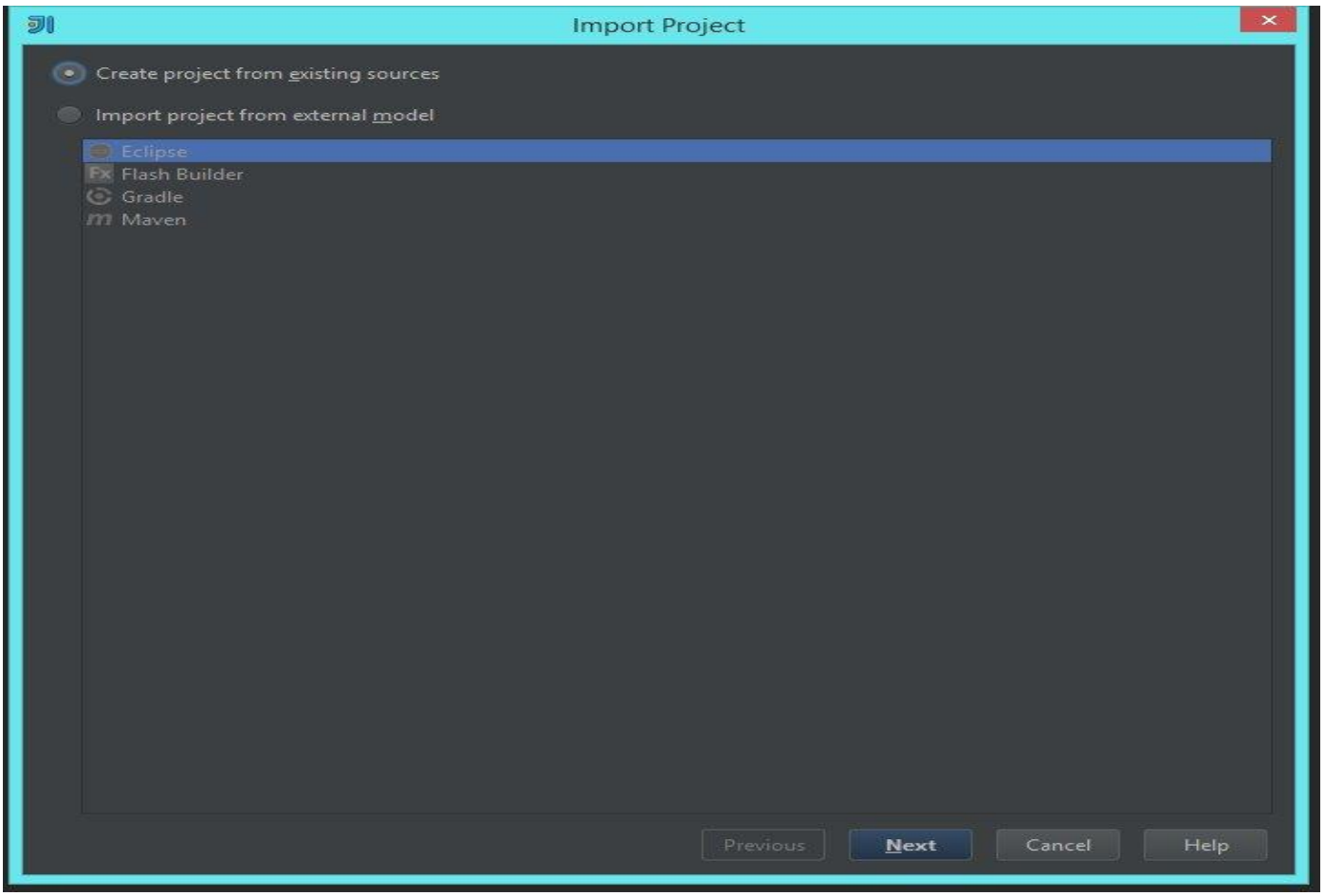


3) Wypakuj repozytorium i uruchom IntelliJ Idea Community

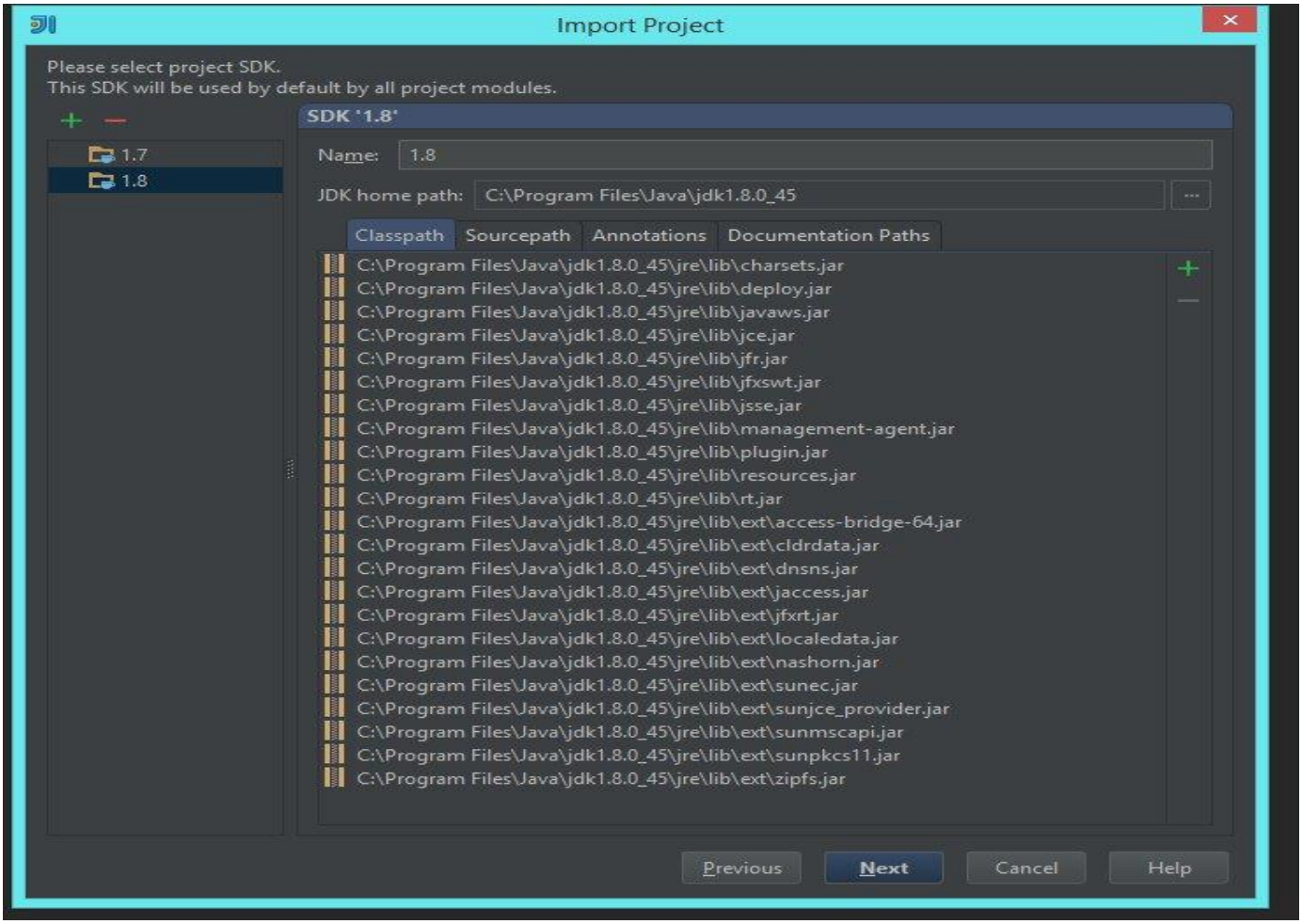
4) W intelliJ wybierz File > Import Project...



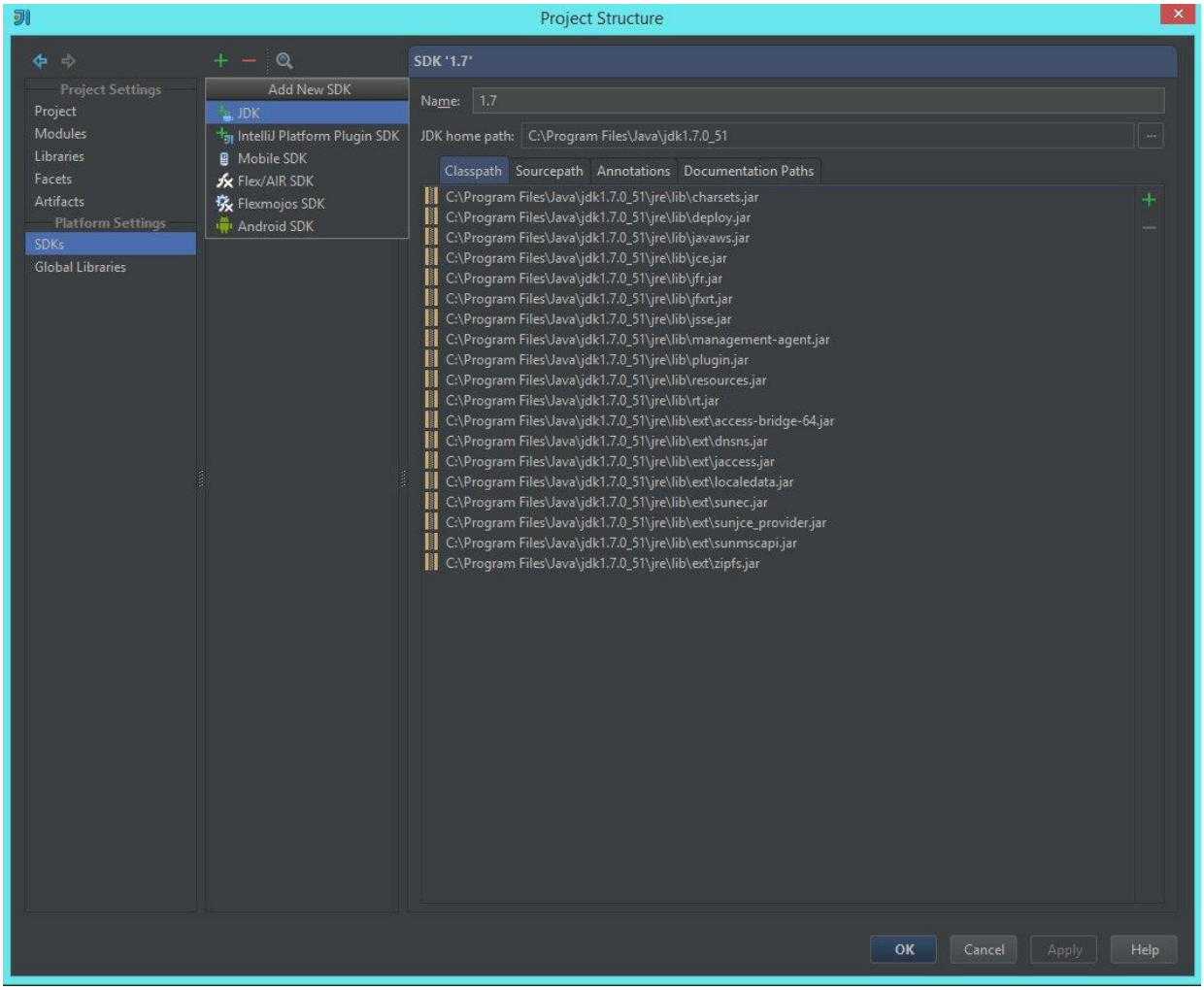
5) Wybierz swój rozpakowany projekt i w następnym oknie zaznacz przycisk "Create project from existing sources" i przechodź dalej aż natrafisz na wybór SDK



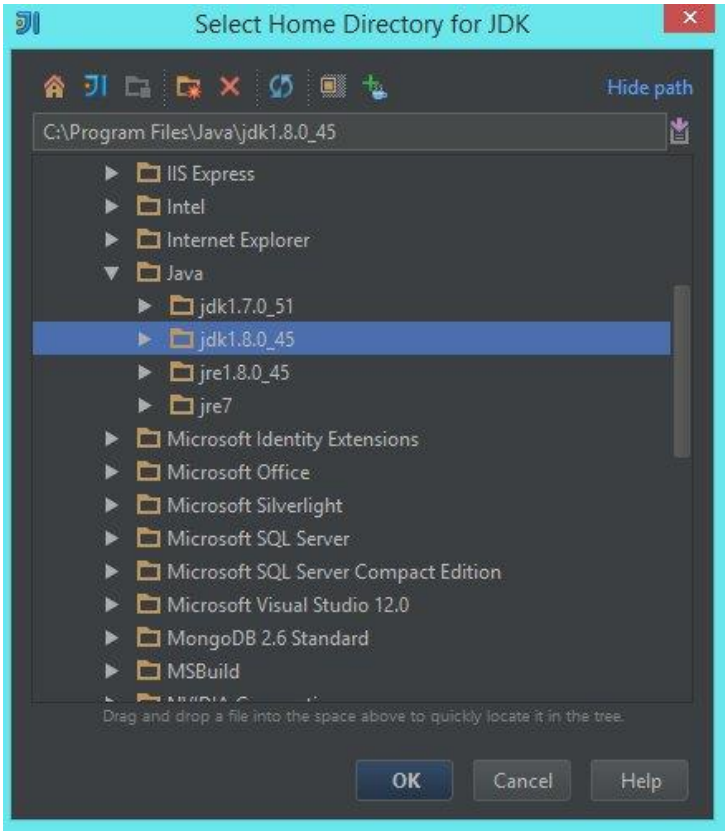
6) Spośród dostępnych SDK wybierz wcześniej zainstalowaną javę 8



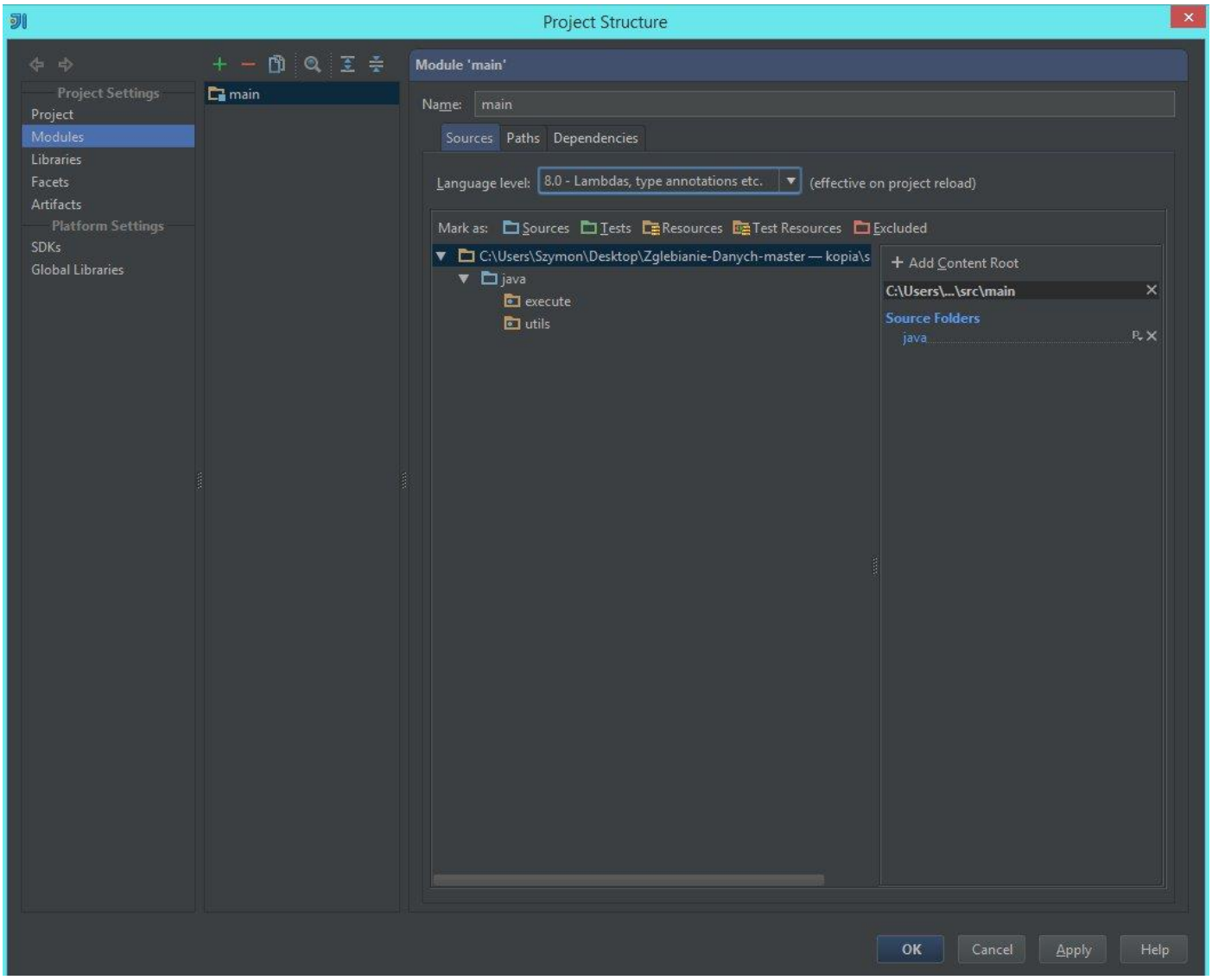
6.1) Jeśli zainstalowałeś javę 8 a nie masz jej do wyboru naciśnij zielony plus a następnie JDK



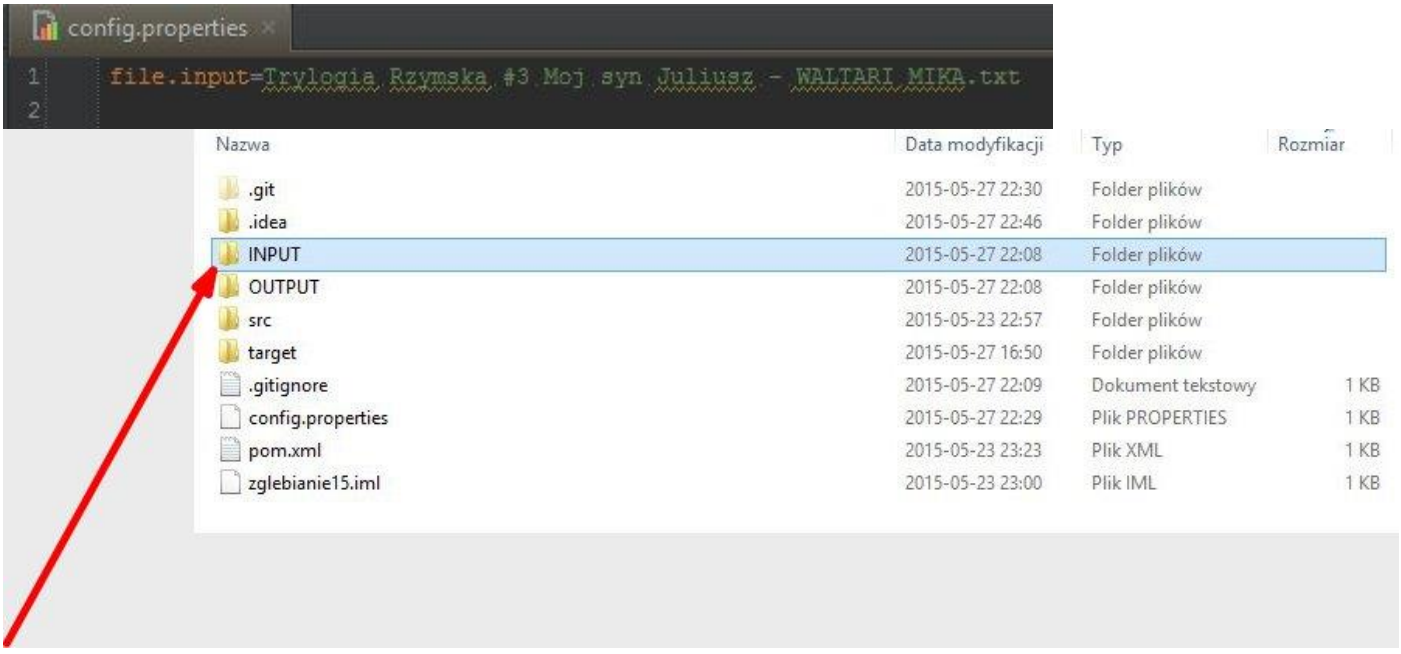
6.2) Podaj ścieżkę do zainstalowanej javy 8



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

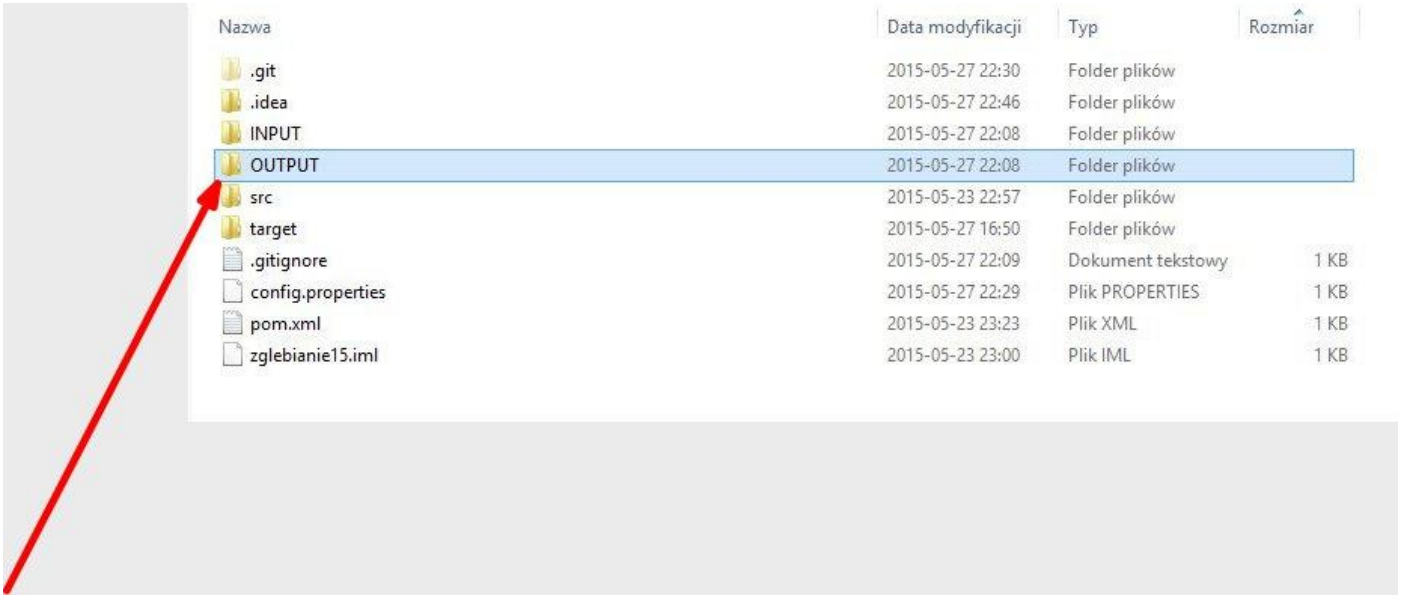


8) Otwórz plik config.properties i jako wartość zmiennej file.input podaj nazwę pliku w formacie .txt znajdującego się w katalogu INPUT, na którym chcesz zacząć badania.



9) Otwórz klasę Main.java (ctrl + shift + n) a następnie ją uruchom (ctrl + shift + f10)

10) Wynik badań znajduje się w katalogu OUTPUT.



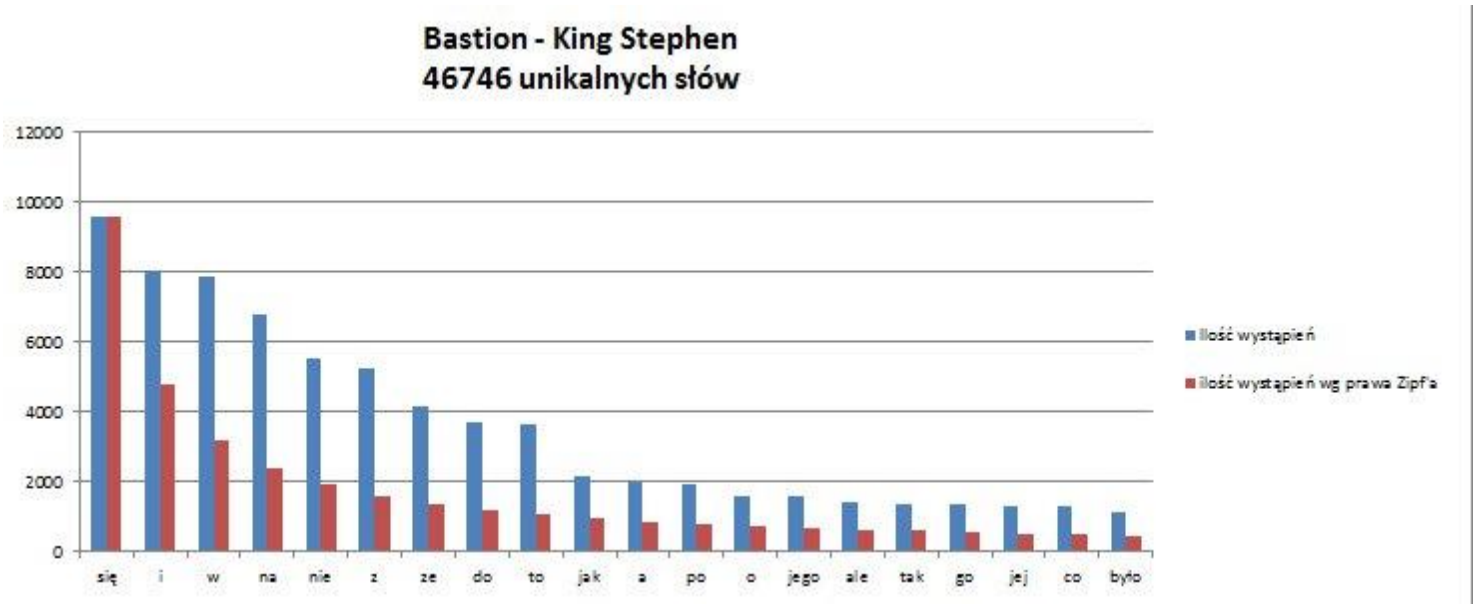
Nazwa	Data modyfikacji	Typ	Rozmiar
.git	2015-05-27 22:30	Folder plików	
.idea	2015-05-27 22:46	Folder plików	
INPUT	2015-05-27 22:08	Folder plików	
OUTPUT	2015-05-27 22:08	Folder plików	
src	2015-05-23 22:57	Folder plików	
target	2015-05-27 16:50	Folder plików	
.gitignore	2015-05-27 22:09	Dokument tekstowy	1 KB
config.properties	2015-05-27 22:29	Plik PROPERTIES	1 KB
pom.xml	2015-05-23 23:23	Plik XML	1 KB
zglebianie15.iml	2015-05-23 23:00	Plik IML	1 KB

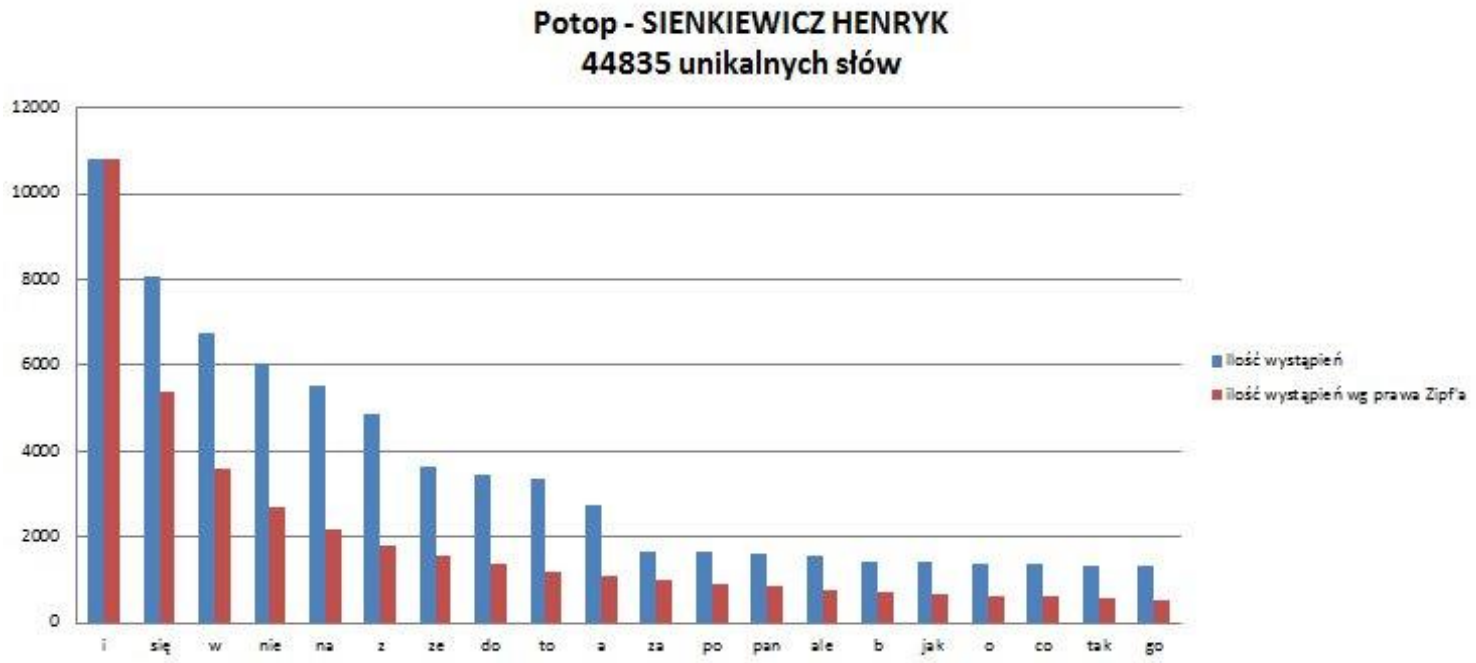
11) W pierwszej linii pliku z wynikami jest liczba wszystkich słow, a w kolejnych posortowane malejąco słowa i ilość ich występowań. Po znaku '|' wyliczona została wartość, ile razy słowo powinno występować wg. prawa Zipf'a.

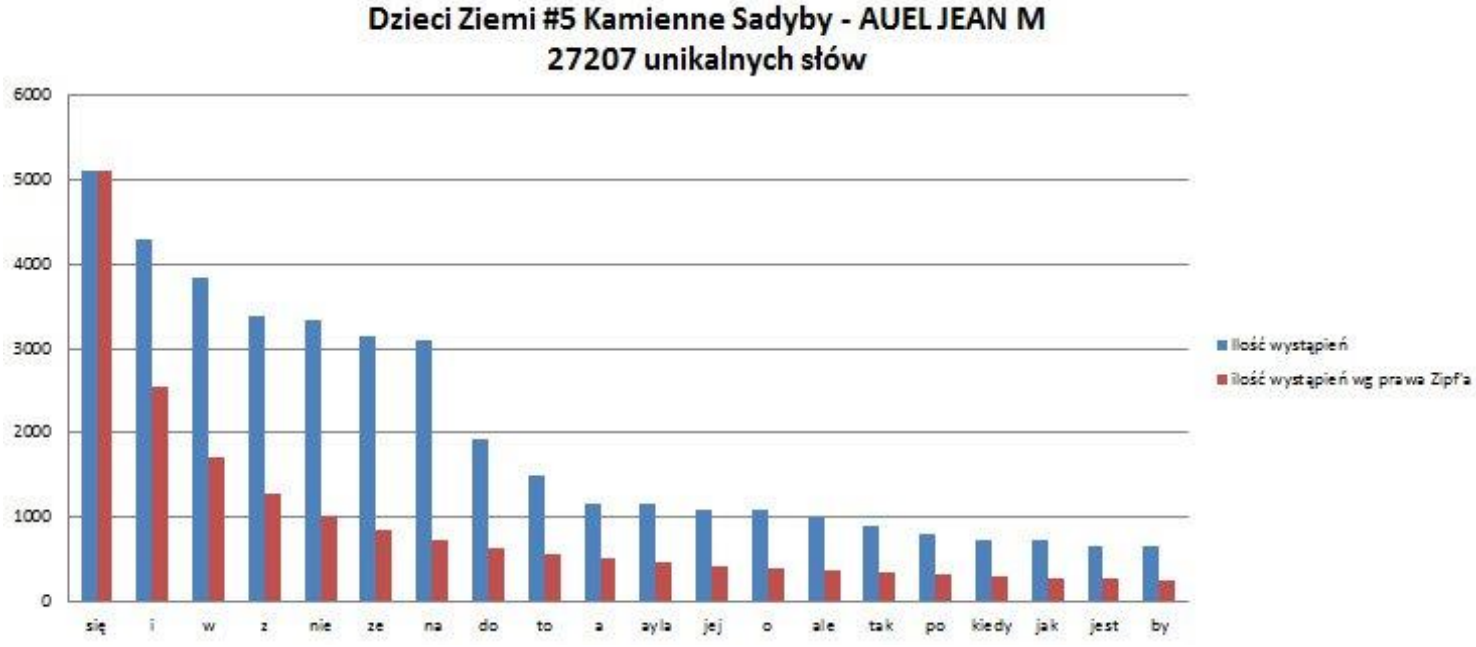
```
NUMBER OF WORDS: 19453

i - 2104 | 2104/1 = 2104
w - 1856 | 2104/2 = 1052
sie - 1760 | 2104/3 = 701
nie - 1645 | 2104/4 = 526
ze - 1602 | 2104/5 = 420
na - 1398 | 2104/6 = 350
z - 1156 | 2104/7 = 300
do - 972 | 2104/8 = 263
o - 644 | 2104/9 = 233
to - 545 | 2104/10 = 210
a - 463 | 2104/11 = 191
za - 407 | 2104/12 = 175
po - 388 | 2104/13 = 161
mnie - 386 | 2104/14 = 150
aby - 356 | 2104/15 = 140
jego - 350 | 2104/16 = 131
neron - 330 | 2104/17 = 123
ale - 298 | 2104/18 = 116
jak - 291 | 2104/19 = 110
jest - 287 | 2104/20 = 105
go - 284 | 2104/21 = 100
tak - 279 | 2104/22 = 95
byl - 245 | 2104/23 = 91
gdy - 243 | 2104/24 = 87
przez - 237 | 2104/25 = 84
mi - 236 | 2104/26 = 80
od - 232 | 2104/27 = 77
nerona - 227 | 2104/28 = 75
tylko - 224 | 2104/29 = 72
co - 206 | 2104/30 = 70
nawet - 204 | 2104/31 = 67
jej - 204 | 2104/32 = 65
tym - 201 | 2104/33 = 63
dla - 201 | 2104/34 = 61
ich - 198 | 2104/35 = 60
bo - 190 | 2104/36 = 58
juz - 186 | 2104/37 = 56
tego - 176 | 2104/38 = 55
bylo - 174 | 2104/39 = 53
mu - 169 | 2104/40 = 52
ja - 151 | 2104/41 = 51
ktory - 142 | 2104/42 = 50
choc - 141 | 2104/43 = 48
przeciez - 138 | 2104/44 = 47
wiec - 138 | 2104/45 = 46
jeszcze - 137 | 2104/46 = 45
sobie - 130 | 2104/47 = 44
```

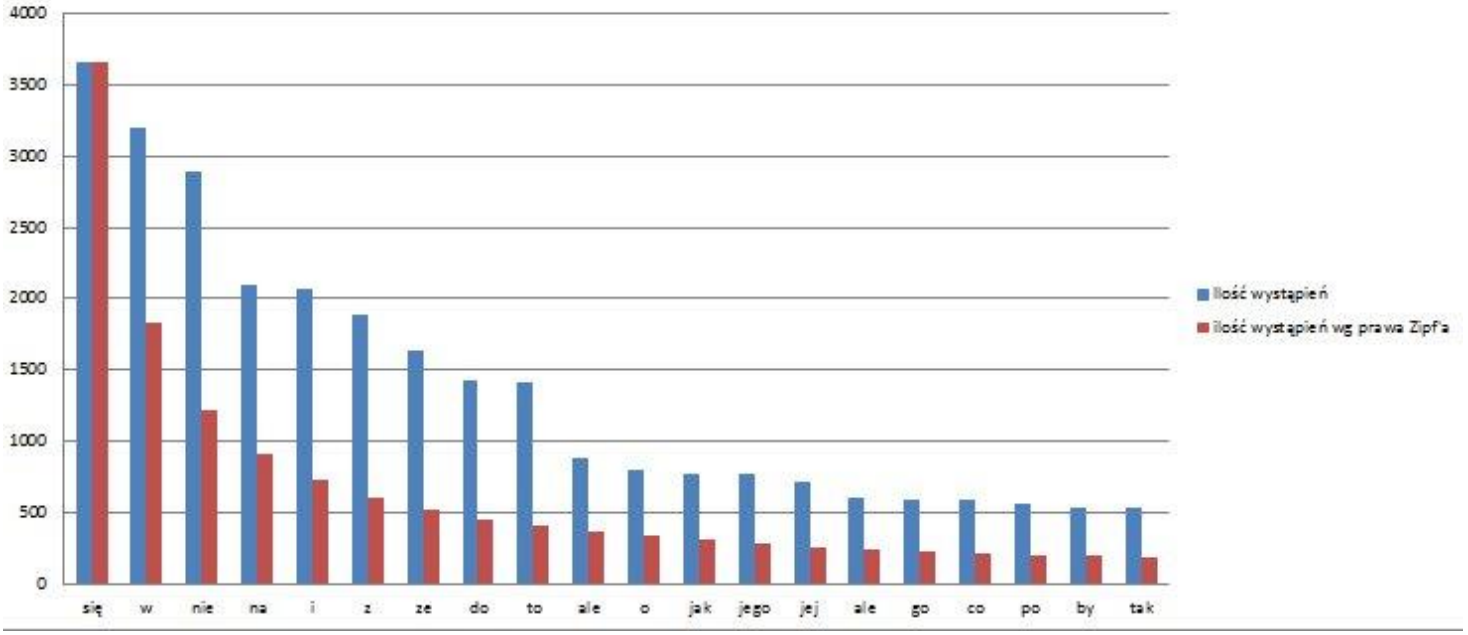
Wyniki


Bastion - King Stephen
46746 unikalnych słów
12000
10000
8000
6000
4000
2000
0
się i w na nie z ze do to jak a po o jego ale tak go jej co było
Ilość wystąpień
Ilość wystąpień wg prawa Zipf'a

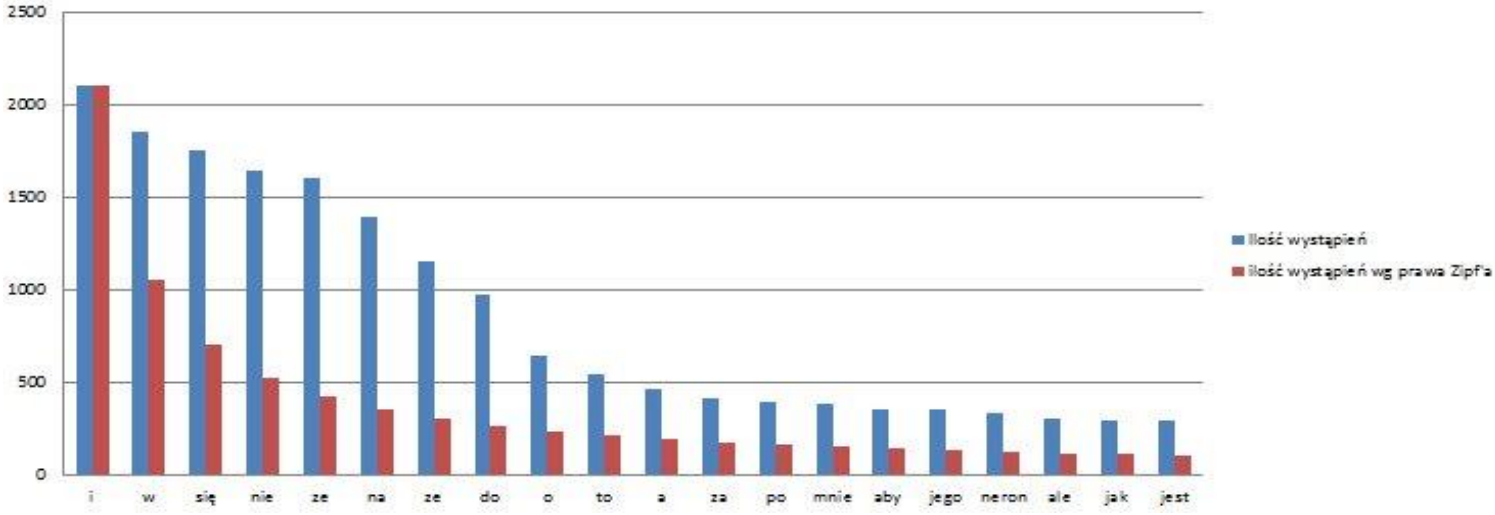

Potop - SIENKIEWICZ HENRYK
44835 unikalnych słów
12000
10000
8000
6000
4000
2000
0
i się w nie na z ze do to a za po pan ale b jak o co tak go
Ilość wystąpień
Ilość wystąpień wg prawa Zipf'a


Dzieci Ziemi #5 Kamienne Sadyby - AUEL JEAN M
27207 unikalnych słów
6000
5000
4000
3000
2000
1000
0
się i w z nie ze na do to a ayla jej o ale tak po kiedy jak jest by
Ilość wystąpień
Ilość wystąpień wg prawa Zipf'a

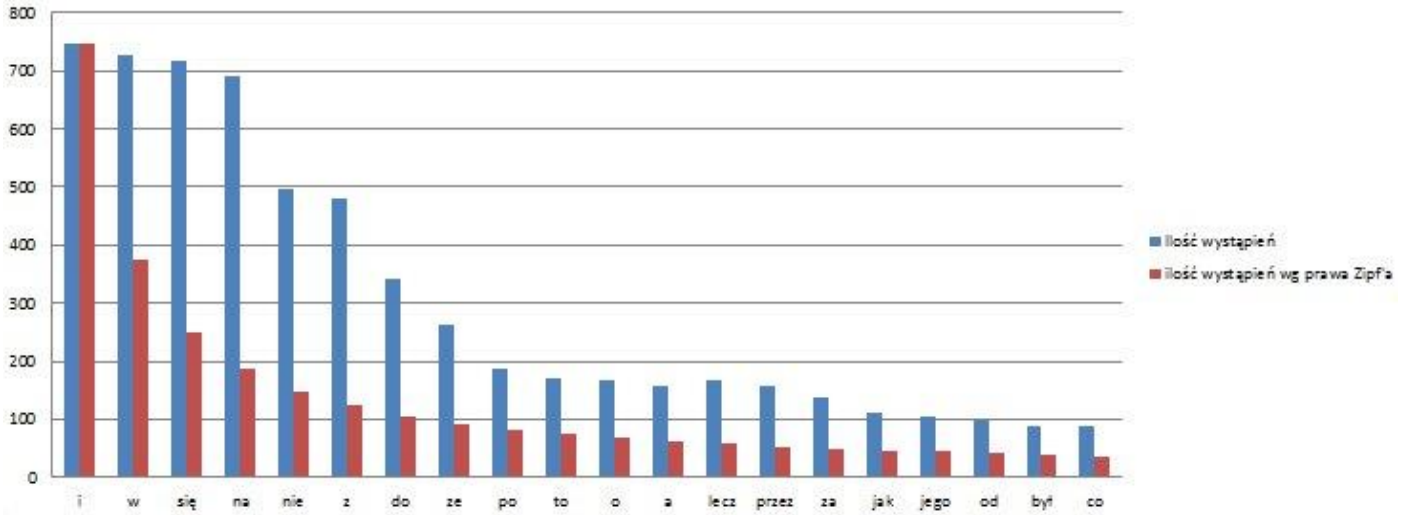
Wielkie Sekretne Widowisko - BARKER CLIVE
24127 unikalnych słów

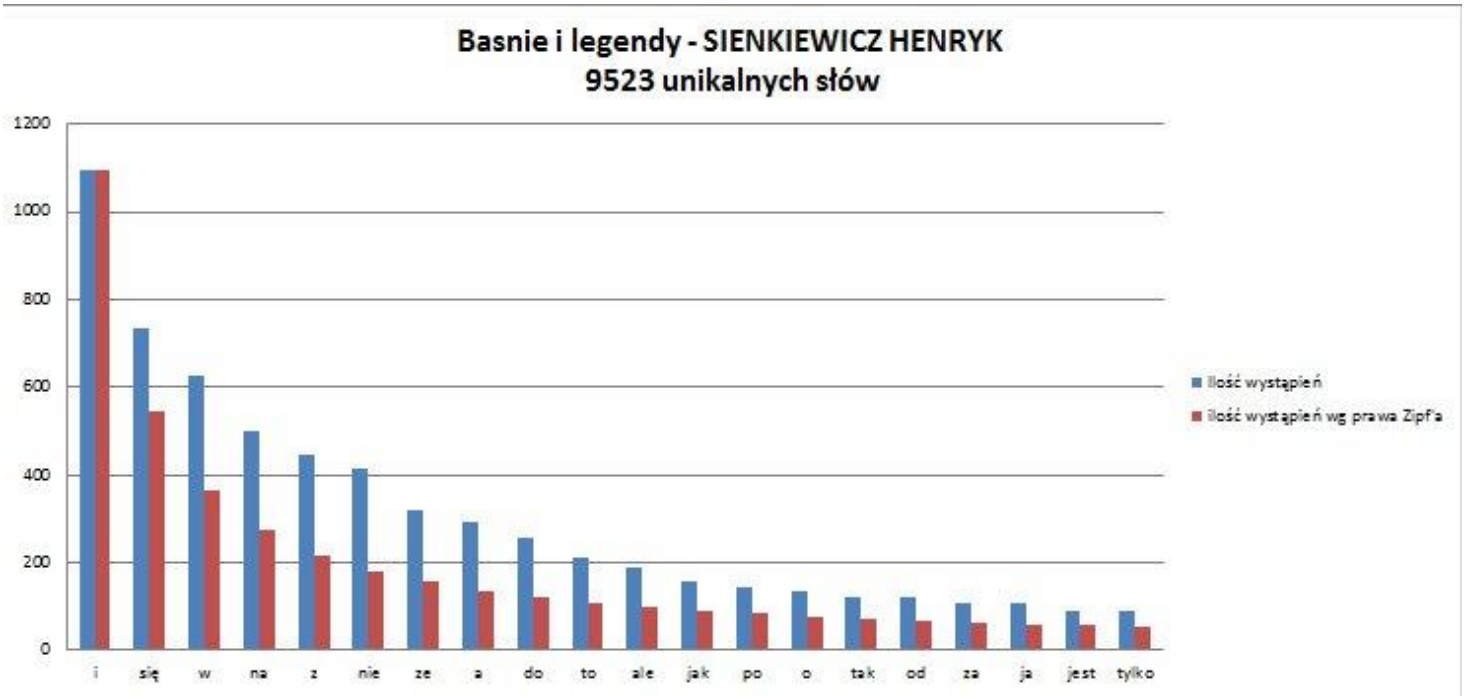


Trylogia Rzymska #3 Moj syn Juliusz - WALTARI MIKA
19453 unikalnych słów



Bezkresne morze - MACLEAN ALISTAIR
10656 unikalnych słów





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.