



AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

Przetwarzanie Języka Naturalnego

Lab 6

Zbigniew Kaleta
zkaleta@agh.edu.pl

Wydział IEiT
Katedra Informatyki

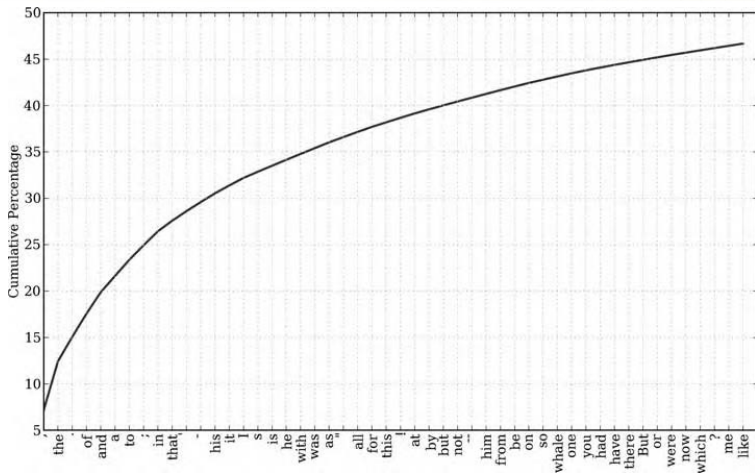
22.04.2015



Częstotliwość występowania wyrazu w tekście jest odwrotnie proporcjonalna do numeru rankingu powstałego przez uporządkowanie wyrazów względem ich częstości występowania.

„Zasada Pareto” w lingwistyce.

Prawo Zipfa



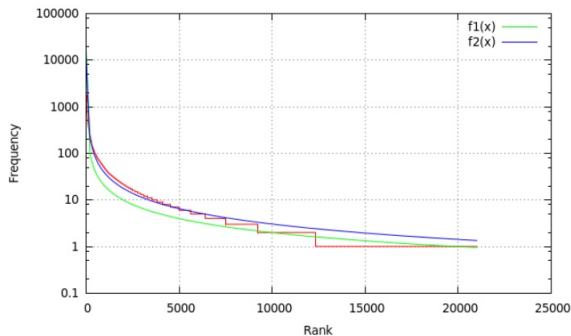
- ✚ najliczniejsze wyrazy są wspólne dla większości tekstów
- ✚ znaczenie tekstu zawarte jest w wyrazach najrzadszych
- ✚ wiele wyrazów występuje w tekście tylko raz - *hapax legomena*
- ✚ w przybliżeniu: wyraz na 50. pozycji w rankingu będzie występował 3-krotnie częściej niż wyraz na pozycji 150. A więc, dla f - częstotliwości, r - pozycji w rankingu, powinna istnieć taka stała k , że:

$$f \cong \frac{k}{r}$$

- ✦ prawo Zipfa oddaje charakter statystyczny wielu problemów związanych z modelowaniem zachowań ludzkich, lecz nie jest możliwe precyzyjne odwzorowanie na całej dziedzinie problemu
- ✦ prawo Madelbrota - uszczegółowienie prawa Zipfa
- ✦ dla pewnych stałych B, d, P :
$$\log(f) = \log(P) - B \cdot \log(r + d)$$

$$f1 = k/x$$

$$f2 = p / ((x+d) ** B)$$



- 1 Wykorzystując plik *odm.txt* (lista słów z odmianami z SJP), sprowadzić wszystkie wyrazy z pliku *potop.txt* do formy podstawowej, a następnie stworzyć posortowaną listę rankingową częstości wystąpień poszczególnych wyrazów (1 pkt.)
- 2 Dla powstałej listy narysować wykres ilustrujący Prawa Zipfa i Mandelbrota (1 pkt)
- 3 Zliczyć *hapax legomena* i ilość wyrazów, które obejmują 50% tekstu (0.5 pkt.)
- 4 Zebrać statystyki występowanie di- i trigramów (0.5 pkt.)

Materiały:

<http://home.agh.edu.pl/~zkaleta/pjn/lab6.tar.gz>