AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

Przetwarzanie Języka Naturalnego
Lab 6

Wojciech Korczyński
`wojciech.korczynski@agh.edu.pl`

Wydział IEiT
Katedra Informatyki

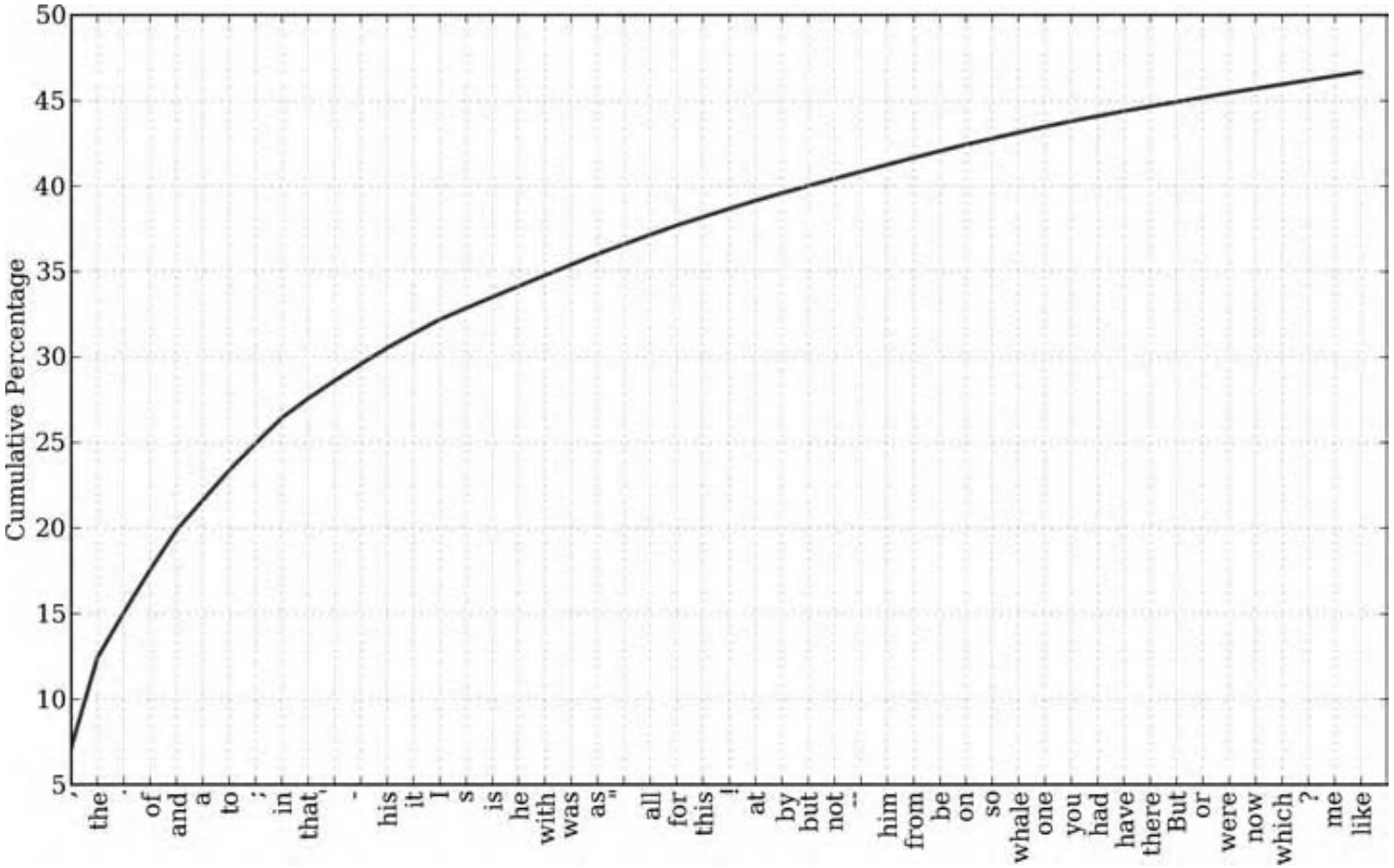
22.04.2015

Prawo Zipfa

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


Cumulative Percentage
50
45
40
35
30
25
20
15
10
5
' the . of and a to ; in that , - his it I s is he with was as " all for this ! at by but not -- him from be on so whale one you had have there But or were now which ? me like

Konsekwencje

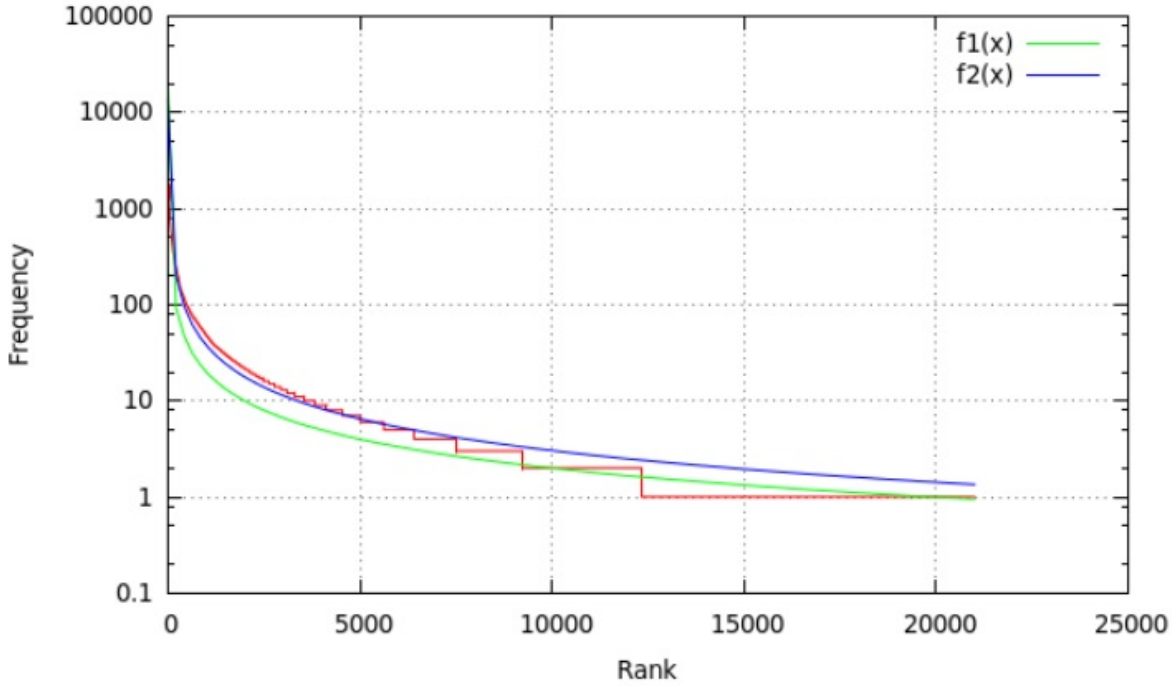
- najliczniejsze wyrazy są wspólne dla większości tekstów
- znaczenie tekstu zawarte jest w wyrazach najrzadszych
- wiele wyrazów występuje w tekście tylko raz - *hapax legomena*
- w przybliżeniu: wyraz na 50. pozycji w rankingu będzie występował 3-krotnie częściej niż wyraz na pozycji 150. A więc, dla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

Prawo Mandelbrota

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

Prawo Mandelbrota

```
f1 = k/x
f2 = P/((x+d)**B)
```



Prawo Mandelbrota

Jak znaleźć wartości stałych P, d, B?

$$log(f) = log(P) - B \cdot log(r + d)$$
$$log(f) = log(P) - log(r + d)^B$$
$$log(f) = log(\frac{P}{(r+d)^B})$$
$$f = \frac{P}{(r+d)^B}$$

Dopasowanie zdefiniowanej przez nas funkcji do zbioru danych umożliwia np. funkcja `fit` w Gnuplocie.

- `http://gnuplot.sourceforge.net/docs_4.2/node82.html`
- `http://people.duke.edu/~hpgavin/gnuplot.html` (rozdział 7.)

Zadania

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 (1.5 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.)

Materiały:
`http://home.agh.edu.pl/~wojtek/pjn2015/lab6.tar.gz`