ΗΥ463 - Συστήματα Ανάκτησης Πληροφοριών
Information Retrieval (IR) Systems

Στατιστικά Κειμένου
Text Statistics

Γιάννης Τζίτζικας

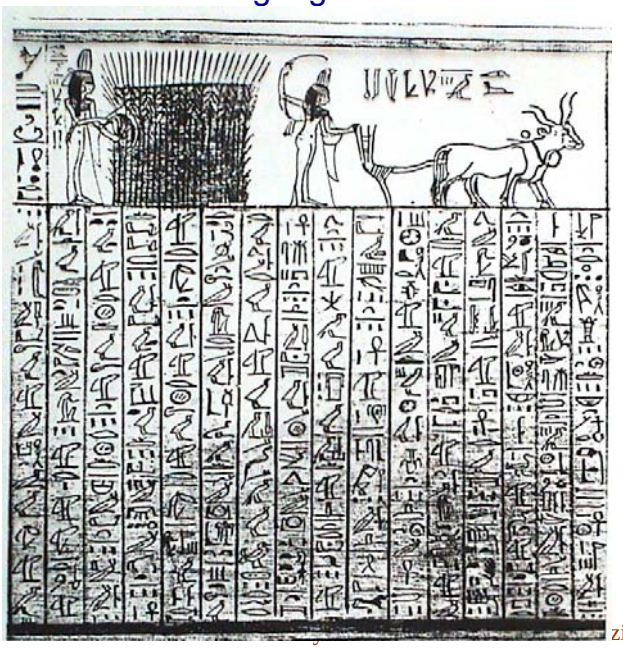
Διάλεξη : 14α
Ημερομηνία :

Διάρθρωση

- Συχνότητα Εμφάνισης Λέξεων
- Ο Νόμος του Zipf
- Ο Νόμος του Heaps

Γραπτός Λόγος - Κείμενο

Starting with hieroglyphs, the first written surfaces (stone, wood, animal skin, papyrus and rice paper), and paper, text has been created everywhere, in many forms and languages.



Στατιστικές Ιδιότητες Κειμένου

- *How is the frequency of different words distributed?*
- *How fast does vocabulary size grow with the size of a corpus?*

Such factors affect the performance of information retrieval and can be used to select appropriate term weights and other aspects of an IR system.

Συχνότητα Λέξεων

- A few words are very common.
  - 2 most frequent words (e.g. “the”, “of”) can account for about 10% of word occurrences.
- Most words are very rare.
  - Half the words in a corpus appear only once, called *hapax legomena* (Greek for “read only once”)
- Called a “heavy tailed” distribution, since most of the probability mass is in the “tail”

Sample Word Frequency Data
(from B. Croft, UMass)

Frequent Word	Number of Occurrences	Percentage of Total
the	7,398,934	5.9
of	3,893,790	3.1
to	3,364,653	2.7
and	3,320,687	2.6
in	2,311,785	1.8
is	1,559,147	1.2
for	1,313,561	1.0
The	1,144,860	0.9
that	1,066,503	0.8
said	1,027,713	0.8

Frequencies from 336,310 documents in the 1GB TREC Volume 3 Corpus
125,720,891 total word occurrences; 508,209 unique words

Ο νόμος του Zipf

- **Rank *r of a word***: The numerical position of the word in a list sorted by decreasing frequency (*f*).
- *Zipf* (1949) "discovered" that: $f \cdot r = k \ \ (\text{for constant} \ \ k)$
- *Πχ:*
  - *f1 * 1 = k*
  - *f2 * 2 = k*
  - *f3 * 3 = k*
  - *...*
  - *fi * i = k*
  - *= f1 * 1 = f1* $\Leftrightarrow$ ***fi = f1 / i***
- *Η συχνότητα της i-th πιο συχνά εμφανιζόμενης λέξης είναι 1/i φορές η συχνότητα της πιο συχνής.*
- *Πιο ακριβές: $1/i^{\theta}$ όπου θ μεταξύ 1.5 και 2*

Sample Word Frequency Data (again)

(from B. Croft, UMass)

Frequent Word	Number of Occurrences	Percentage of Total
the	7,398,934	5.9
of	3,893,790	3.1
to	3,364,653	2.7
and	3,320,687	2.6
in	2,311,785	1.8
is	1,559,147	1.2
for	1,313,561	1.0
The	1,144,860	0.9
that	1,066,503	0.8
said	1,027,713	0.8

- 1 * 5.9 = 5.9
- 2 * 3.1 = 6.2
- 3 * 2.7 = 8.1
- 4 * 2.6 = 10.4
- 5 * 1.8 = 9
- 6 * 1.2 = 7.2
- 7 * 1 =7
- 8 * 0.9 =7.2
- 9 * 0.8 =7.2
- ...

Frequencies from 336,310 documents in the 1GB TREC Volume 3 Corpus
125,720,891 total word occurrences; 508,209 unique words

Zipf's Law Impact on IR

- Good News: Stopwords will account for a large fraction of text so eliminating them greatly reduces inverted-index storage costs.
- Bad News: For most words, gathering sufficient data for meaningful statistical analysis (e.g. for correlation analysis for query expansion) is difficult since they are extremely rare.

Zipf and Term Weighting

- Luhn (1958) suggested that both extremely common and extremely uncommon words were not very useful for indexing.

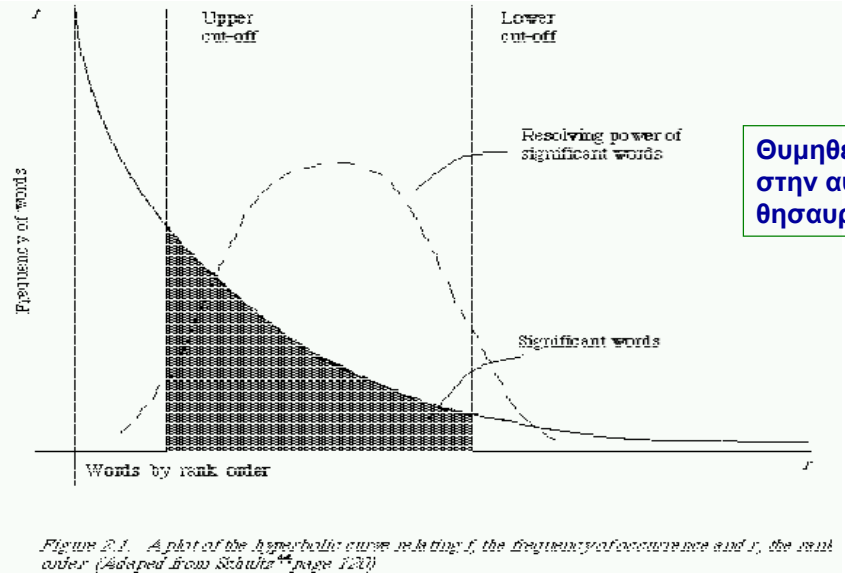


Figure 2.1. A plot of the hyperbolic curve relating f, the frequency of occurrence and r, the rank order (Adapted from Schultz[44] page 120)

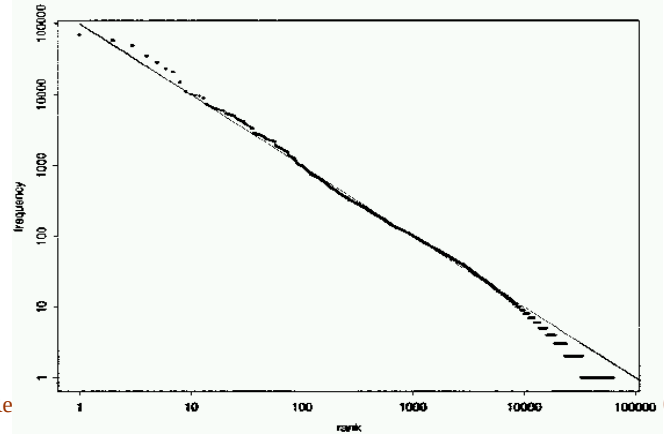
Θυμηθείτε την επιλογή όρων στην αυτόματη κατασκευή θησαυρών

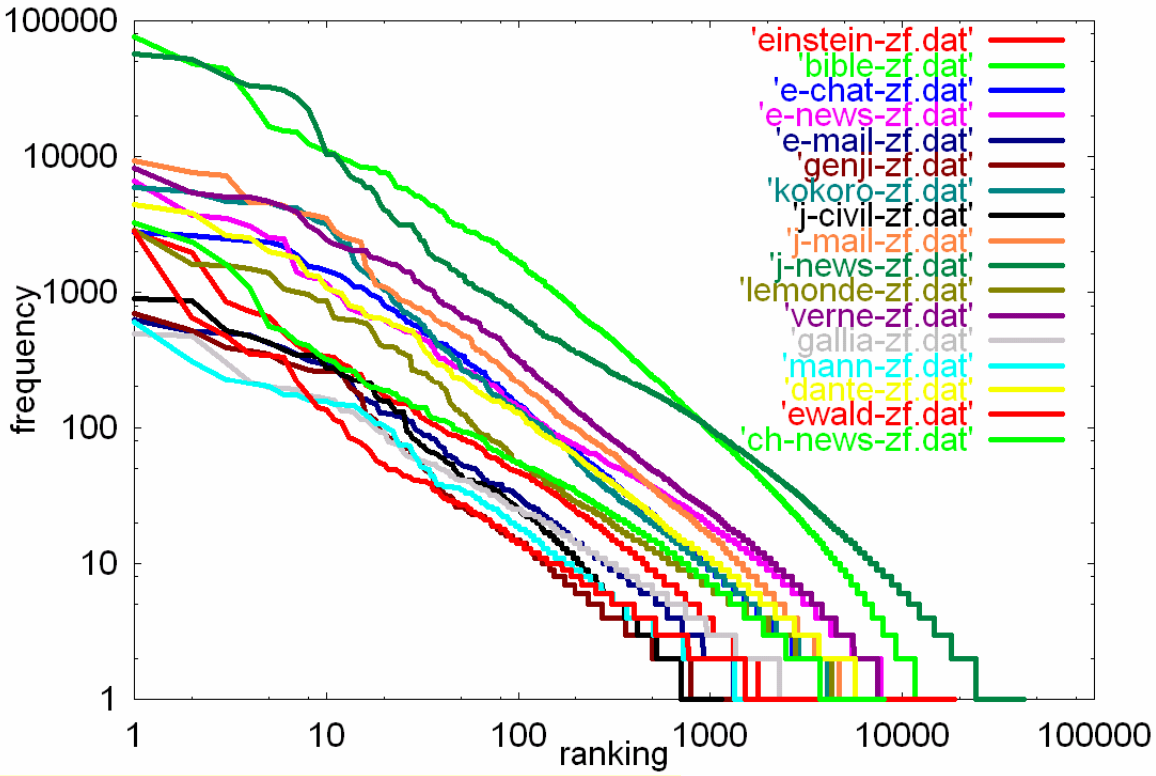
Does Real Data Fit Zipf's Law?

- A law of the form $y = kx^c$ is called a power law.
- Zipf's law ($fi=f1/i$) is a power law with $c = -1$
- On a log-log plot, power laws give a straight line with slope c.

$$\log(y) = \log(kx^c) = \log k + c\log(x) = \log k - \log(x)$$

Zipf is quite accurate except for very high and low rank.

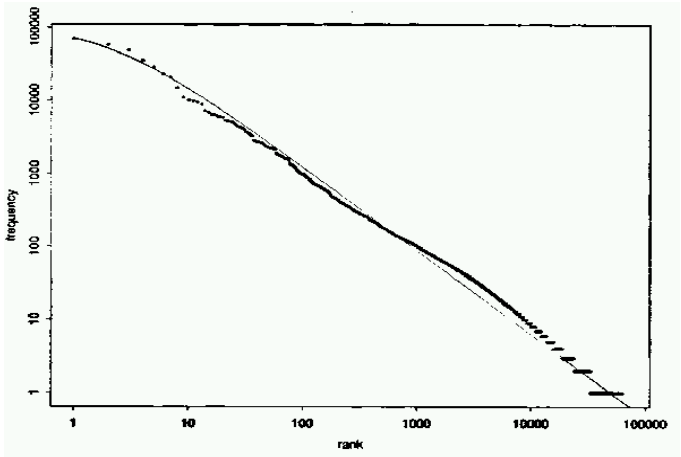




Σημείωση: Ο Χ και Υ έχουν λογαριθμική κλίμακα

Mandelbrot (1954) Correction

- *Ziph's Law:* $fi = f1/i^{\theta}$
- Mandelbrot correction: $fi = f1*k/(c+i)^{\theta}$
  - *c: parameter*
  - *k: so that all frequencies add to N*
  - This formula fits better with the read texts



Mandelbrot's function on Brown corpus

Explanations for Zipf's Law

- Zipf's explanation was his "principle of least effort." Balance between speaker's desire for a small vocabulary and hearer's desire for a large one.
  - Η επανάληψη λέξεων είναι ευκολότερη από την επινόηση/χρήση νέων
- Debate (1955-61) between Mandelbrot and H. Simon over explanation.
- Με επιφύλαξη:
  - Li (1992) shows that just random typing of letters including a space will generate "words" with a Zipfian distribution.
  - (http://linkage.rockefeller.edu/wli/zipf/)

Vocabulary Growth

- How does the size of the overall vocabulary (number of unique words) grow with the size of the corpus?
- This determines how the size of the inverted index will scale with the size of the corpus.
- Vocabulary not really upper-bounded due to proper names, typos, etc.

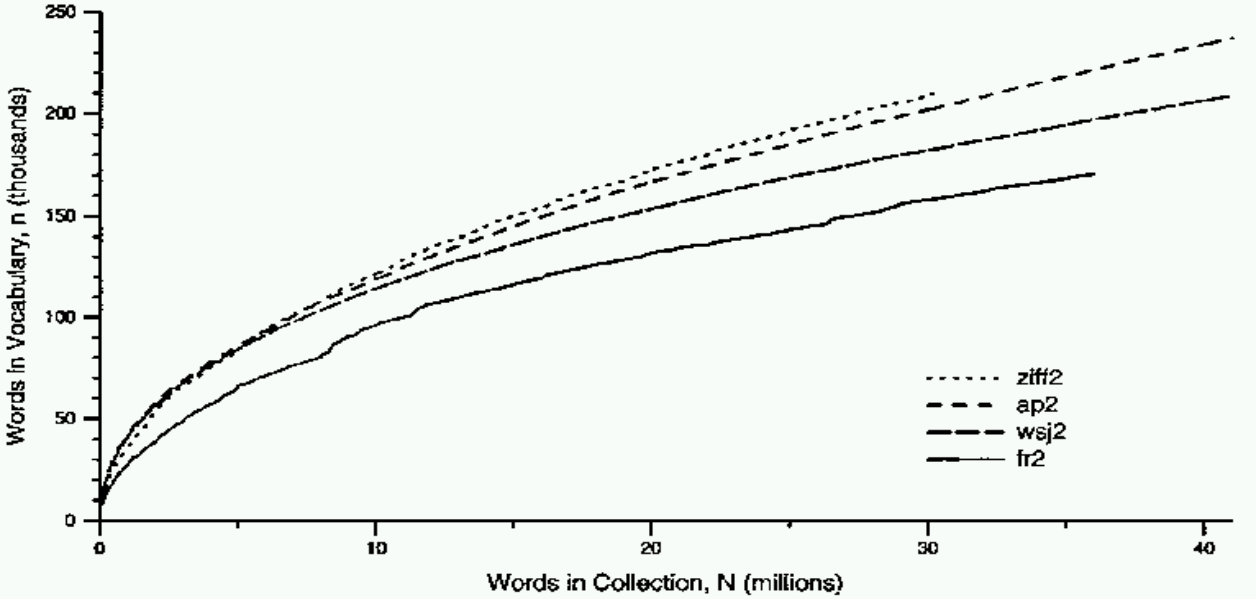
Heaps' Law

- If *V* is the size of the vocabulary (i.e. number of distict words) and the *n* is the length of the corpus in words:

$$V = Kn^{\beta} \quad \text{with constants } K,\ 0 < \beta < 1$$

- Typical constants:
  - $K \approx 10\text{–}100$
  - $\beta \approx 0.4\text{–}0.6$ (approx. square-root)

Heaps' Law Data



- Explanation for Heaps' Law
  - Can be derived from Zipf's law by assuming documents are generated by randomly sampling words from a Zipfian distribution
- Average Length of Words
  - Why? To estimate the storage space needed for the vocabulary.
  - Average word length in TREC-2 = 5 letters
  - If we remove stopwords then average word length: 6-7 letters