HY463 - Συστήματα Ανάκτησης Πληροφοριών
Information Retrieval (IR) Systems

Στατιστικά Κειμένου
Text Statistics

Γιάννης Τζίτζικας

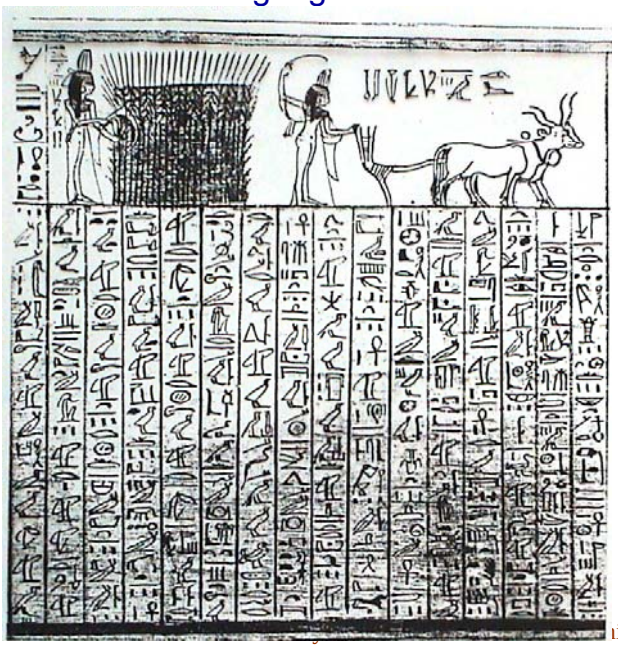
Διάλεξη : 14α
Ημερομηνία :

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

- *Η συχνότητα εμφάνισης των λέξεων, τι κατανομή ακολουθεί;*
- *Πόσο γρήγορα μεγαλώνει το λεξιλόγιο σε σχέση με το μέγεθος της συλλογής κειμένων;*
- *Ποιο είναι το μέσο μήκος των λέξεων;*

Η γνώση των παραπάνω μπορεί να αξιοποιηθεί στη σχεδίαση συστημάτων ανάκτησης πληροφοριών.

.

Συχνότητα Λέξεων

- **Λίγες** λέξεις εμφανίζονται πολύ **συχνά**
  - στις δύο πιο συχνά εμφανιζόμενες λέξεις της αγγλικής (που είναι οι λέξεις “the” και “of”) αντιστοιχεί το 10% των εμφανίσεων λέξεων
- Οι **περισσότερες** λέξεις εμφανίζονται **σπάνια**
  - Οι μισές περίπου λέξεις εμφανίζονται μόνο μία φορά! (*hapax legomena)*

Αυτή η κατανομή συχνά ονομάζεται “**heavy tailed**” (δηλαδή με ... βαριά ουρά)

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

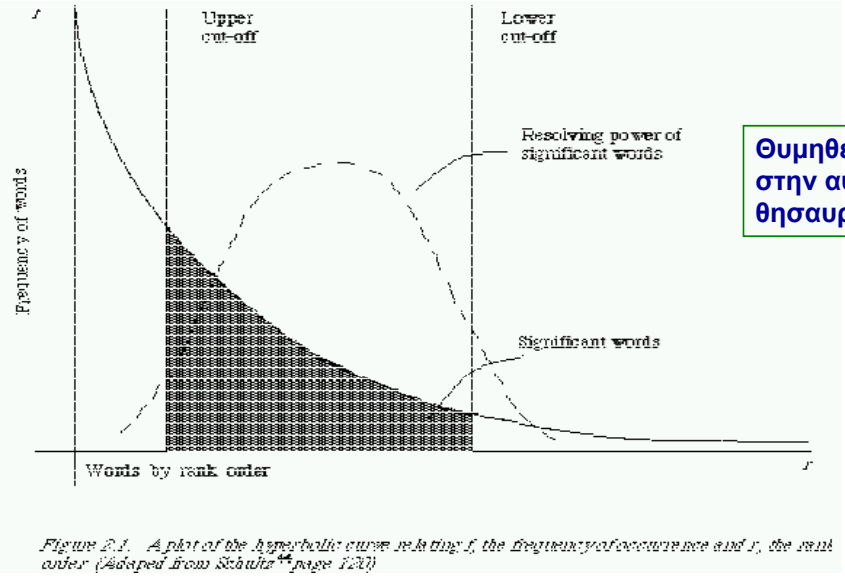
Νόμος του Zipf: Επιπτώσεις στην Ανάκτηση Πληροφοριών

- **Καλά Νέα**: Οι λέξεις αποκλεισμού (stopwords) αντιστοιχούν σε πολλές εμφανίσεις, άρα η απαλοιφή τους μειώνει δραστικά το μέγεθος του ευρετηρίου.

- Άσχημα Νέα: Για τις περισσότερες λέξεις, η συλλογή επαρκών δεδομένων για ασφαλή στατιστική ανάλυση (π.χ. εντοπισμός συνώνυμων λέξων μέσω συνεμφάνισης) είναι δύσκολη διότι οι λέξεις αυτές εμφανίζονται πολύ λίγες φορές.

Ο νόμος του Zipf και η Βάρυνση Όρων Zipf and Term Weighting

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



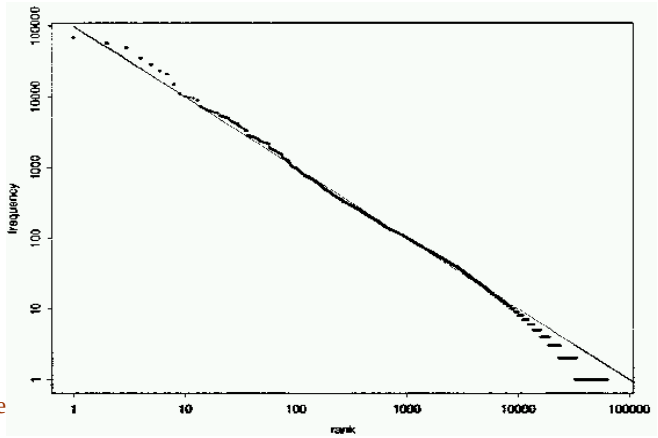
Θυμηθείτε την επιλογή όρων στην αυτόματη κατασκευή θησαυρών

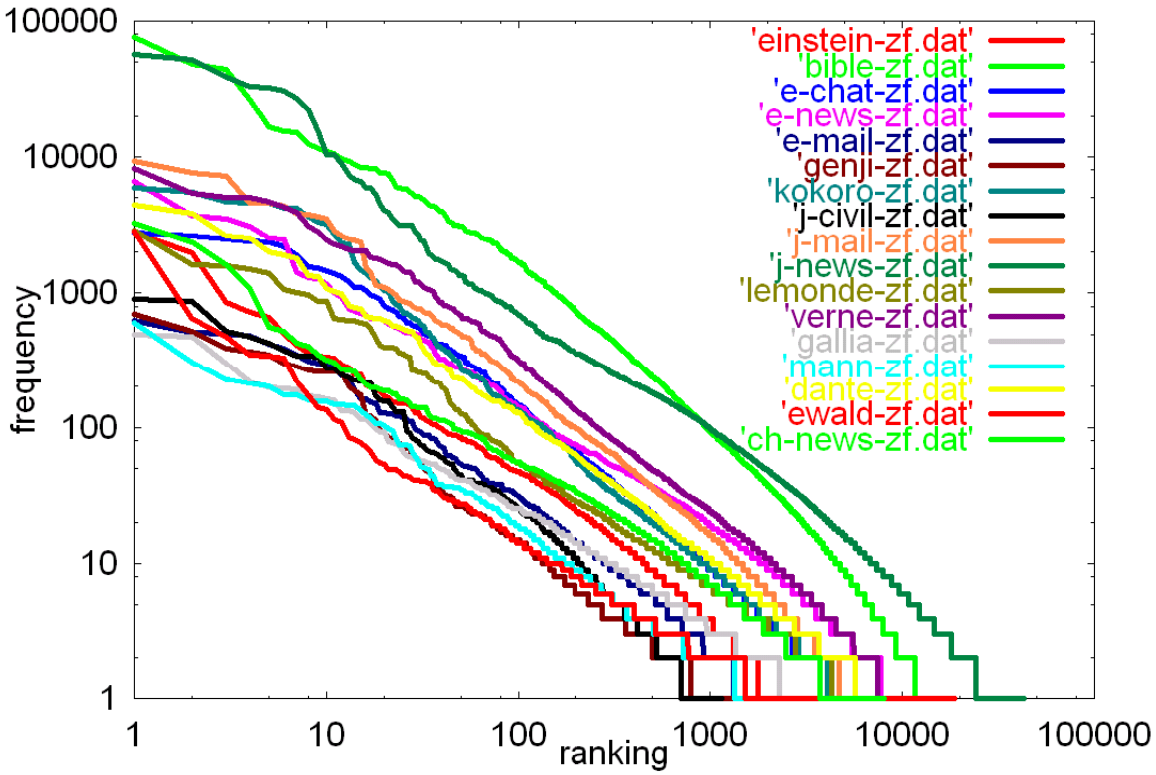
Does Real Data Fit Zipf's Law?

- 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.

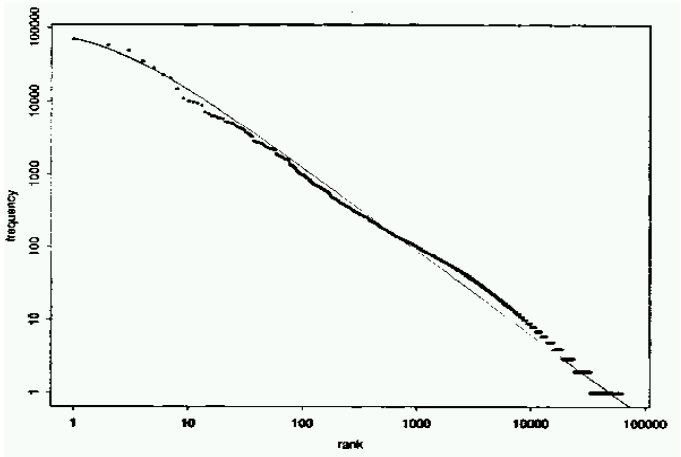




Σημείωση: Ο Χ και Υ έχουν λογαριθμική κλίμακα

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/)

Ανάπτυξη Λεξιλογίου και ο Νόμος του **Heaps** (Vocabulary Growth and **Heaps' Law**)

- Παρατήρηση: Το **μέγεθος του λεξιλογίου** στην πράξη **δεν είναι φραγμένο** (λόγω των κύριων ονομάτων και των ορθογραφικών λαθών).
- Ερώτημα: *Πως αναπτύσσεται το **μέγεθος του λεξιλογίου** (δηλαδή το πλήθος των διαφορετικών λέξεων) σε σχέση με το **μέγεθος της συλλογής** των κειμένων;*
- Μια απάντηση μας έχει δώσει ο Heaps
  - Πηγή: Harold Stanley Heaps. *Information Retrieval: Computational and Theoretical Aspects*. Academic Press, 1978.
    - *Heaps' law is proposed in Section 7.5 (pages 206–208).*

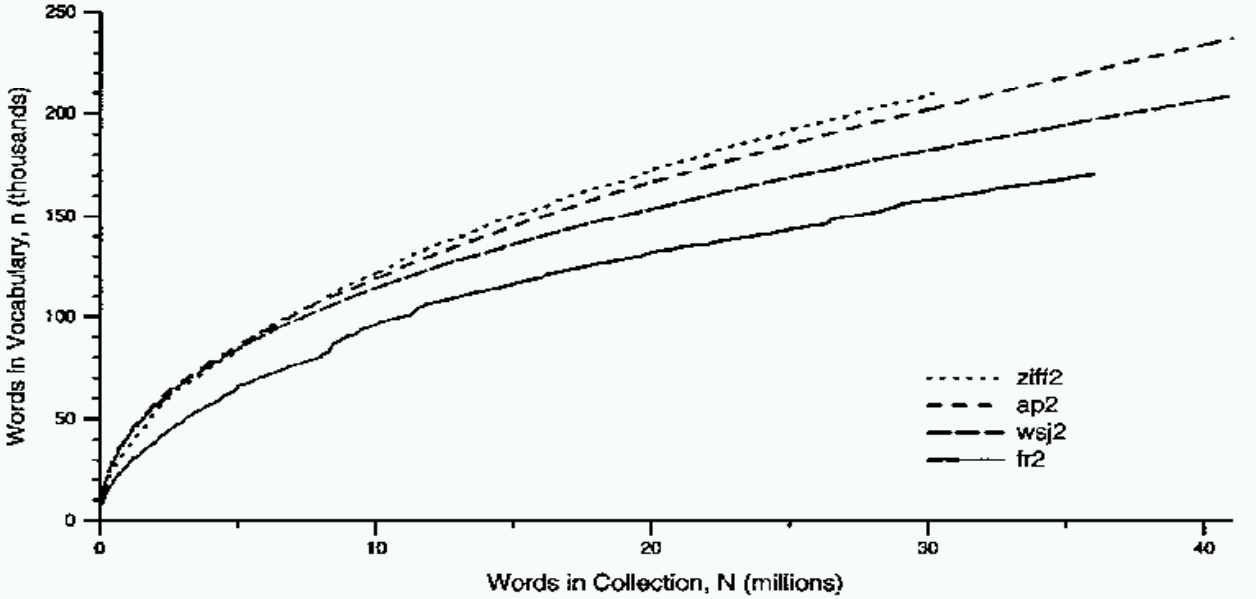
Ο Νόμος του **Heaps**

- Εάν ***V*** είναι το μέγεθος του λεξιλογίου (δηλαδή ο αριθμός των διαφορετικών λέξεων) και ***n*** το μέγεθος της συλλογής κειμένου σε λέξεις, τότε:

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

- Σχετικά με τις σταθερές K και β:
  - $K \approx 10–100$
  - $\beta \approx 0.4–0.6$ (δηλαδή τετραγωνική ρίζα)

Ο Νόμος του **Heaps**



Παρατηρήσεις

- 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