Συστήματα Ανάκτησης Πληροφορίας
Zipf's Law in Information Retrieval
Ημ.: 2010 – 11 – 16

Φοιτητής:
Ιωάννης Κανακαράκης p3060190

Ο Νόμος του Zipf

Ο υπολογισμός της συχνότητας των λέξεων και των βαθμών τους έγινε με βάση ενός POSIX shell script σε bash. Τα βασικά εργαλεία που χρησιμοποιήθηκαν είναι:

- elinks : text-browser utf-8 compliant για την ανάκτηση των Ελληνικών κειμένων.
- sed : stream editor για την επεξεργασία των κειμένων και τη παραγωγή λιστών από λέξεις
- sort : για τη ταξινόμηση των λέξεων που παρήχθησαν
- uniq : για τη μέτρηση της συχνότητα εμφάνισης των λέξεων
- pr : για να ανάκτηση του βαθμού κάθε λέξης
- bc : για τον υπολογισμό μαθηματικών πράξεων
- awk : για την εξαγωγή επιθυμητών δεδομένων από τις μετρήσεις
- gnuplot : για την αναπαράσταση των δεδομένων σε γράφημα

Ο κώδικας βρίσκεται online εδώ και διανέμεται κάτω από την άδεια GNU GPLv3.

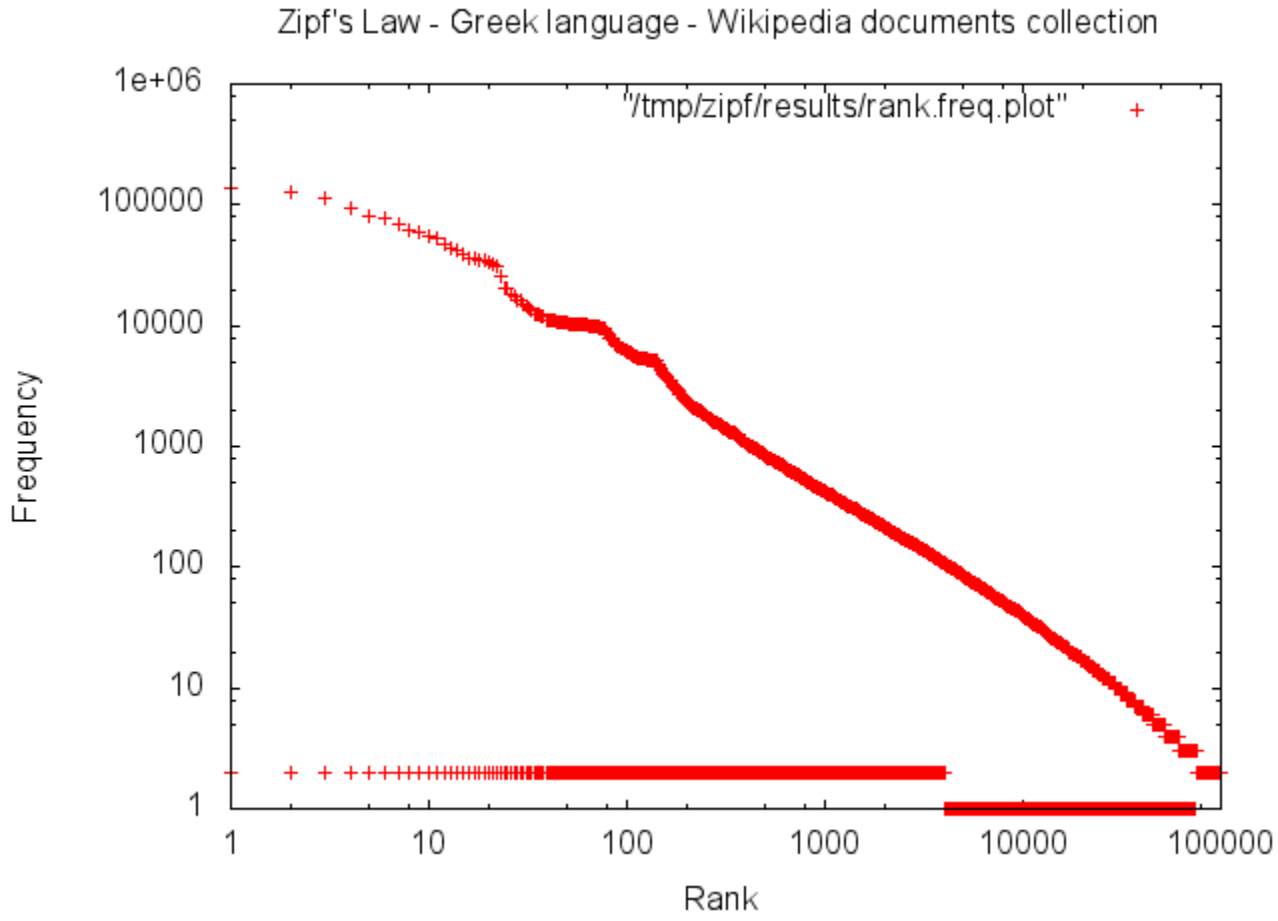
Η συλλογή κειμένων έγινε από την Ελληνική Wikipedia. Συλλέχθηκαν 10.000 τυχαία κείμενα και μετά από την κατάλληλη επεξεργασία εξάχθηκαν τα αποτελέσματα.

Για τον υπολογισμό της τιμής του b μετατράπηκε ο τύπος από το νόμο του Zipf όπως φαίνεται παρακάτω:

$$Freq_{word(i)} \sim rank_{word(i)}^{-b} \Leftrightarrow \log\left(Freq_{word(i)}\right) \sim -b * \log\left(rank_{word(i)}\right)$$

$$\Rightarrow b = -\frac{\log\left(Freq_{word(i)}\right)}{\log\left(rank_{word(i)}\right)} \Leftrightarrow b = -\log_{rank_{word(i)}}\left(Freq_{word(i)}\right)$$

→ Παρακάτω φαίνεται το log-log γράφημα για τα αποτελέσματα επεξεργασίας των 10.000 κειμένων.



Αποτελέσματα:

- Total documents: `10.000`
- Total words: `4.984.085`
- Vocabulary size (unique words) : `174.258`
- Words occurring more than 10 times: `31.133`
- Words occurring once: `70.247`
- Final `b` for all words is `-1.06015791025300522471`

→ Παρακάτω φαίνονται οι 30 πιο συχνά εμφανιζόμενες λέξεις:

```
  1  28959 και
  2  26559 του
  3  23144 το
  4  19388 της
  5  16626 από
  6  16024 την
  7  14115 η
  8  12409 ο
  9  11937 να
 10  11557 με
 11  10503 ν
 12   9224 σε
 13   8975 που
 14   8809 για
 15   7898 τη
 16   7516 των
 17   7406 τον
 18   7211 στην
 19   7088 τους
 20   6811 τα
 21   6614 στο
 22   6533 είναι
 23   5674 οι
 24   4035 αναζήτηση
 25   4001 πλοήγηση
 26   3518 ήταν
 27   3410 στη
 28   3208 τις
 29   2948 ως
 30   2813 α
 31   2803 ή
 32   2798 στις
 33   2752 μπορεί
 34   2578 δεν
 35   2525 ένα
 36   2489 χ
 37   2476 βικιπαίδεια
 38   2347 ανακτήθηκε
 39   2327 ς
 40   2302 στα
 41   2288 π
 42   2274 μια
 43   2270 στον
 44   2248 περιοχές
 45   2199 ότι
 46   2152 δείτε
 47   2124 πρ
 48   2111 αυτό
 49   2093 κατά
 50   2089 συμμετοχή
```

→ Παρακάτω φαίνονται οι 50 πιο συχνά εμφανιζόμενες λέξεις μαζί με τη τιμή της σχετικής συχνότητας Pn|Freq και του b για κάθε λέξη:

Rank	Frequency	Pn-Frequency	b-value	Word
1	136421	.78286343890415989991	0	και
2	125937	.72270011878869957936	.46853096345705138927	του
3	112792	.64726642526354449411	.39595159107439964709	το
4	92962	.53347029421722837844	.45326007834683161732	της
5	79905	.45854159612989859921	.48445750107799765697	την
6	78922	.45290056754600910139	.44206975818268147680	από
7	69614	.39948582282694150660	.47154130016882652934	η
8	60834	.34910105073482574788	.50609446524110133410	ν
9	58942	.33824364882158166866	.49334455834386001667	ο
10	53997	.30986634836651191617	.50882558578809645381	με
11	52499	.30126994875444022977	.50033401793975906835	να
12	46844	.26881825328964357651	.52867972572220156849	για
13	43208	.24795275997222525091	.54368209293793334676	τη
14	42527	.24404478391359987145	.53443459301680545771	σε
15	39624	.22738567304988551523	.54692771291548429533	που
16	36214	.20781709983415490735	.56665343158701701790	τα
17	35526	.20386895368388433309	.56129831478932833151	είναι
18	35257	.20232527444780470449	.55282805845511429229	των
19	34765	.19950189086359958452	.54744947206901744586	τον
20	34111	.19574885658703424213	.54441540044015874998	στην
21	31763	.18227466013233175904	.55911580311809134140	τους
22	31047	.17816583361548040560	.55807726322449574837	στο
23	25714	.14756196236636271297	.61027288707850276300	οι
24	20137	.11555787649418394458	.67902681783793225250	αναζήτηση
25	20007	.11481186050648746980	.67242745718699234437	πλοήγηση
26	18498	.10615233646468762015	.68840197149745384455	ως
27	17625	.10114255217807975484	.69518742665030428980	βικιπαίδεια
28	16244	.09321756695493489575	.71208687256901703949	στη
29	16055	.09213297448051463625	.70814163384723660772	ήταν
30	14924	.08564263538755530560	.72256084169839206376	τις
31	14617	.08388088993968747668	.72171423938066398617	α
32	13818	.07929576090761452780	.73132248943751873948	στις
33	13218	.07585261019516925955	.73758260660107774782	μπορεί
34	12496	.07170935217119345342	.74726735588800454506	χ
35	12485	.07164622774146529017	.74142242304092848658	επεξεργασία
36	12381	.07104941495130811034	.73792819410515272415	ή
37	12175	.06986726654003523490	.73697549865311214457	άρθρο
38	11846	.06797927223271107948	.73910343901862146652	ανακτήθηκε
39	11175	.06412868201929312115	.74977953779960684423	περιοχές
40	11104	.06372124251832043108	.74636141644589486999	δεν
41	11103	.06371550393379968896	.74142289598985988548	ένα
42	10968	.06294079502349950361	.73991578668347772800	δημιουργία
43	10936	.06275716031883575597	.73606362694236255141	δείτε
44	10857	.06230381214169712898	.73350781369571465739	στα
45	10694	.06136842286481616444	.73315138108735671000	δ
46	10631	.06100689204000941127	.73048586882367306166	αλλαγές
47	10625	.06097246053288495859	.72655214106133447512	συμμετοχή
48	10622	.06095524477932273225	.72267375637483863838	π
49	10619	.06093802902576050591	.71891753331254292289	πύλη