ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ
COMPUTER SCIENCE DEPARTMENT
UNIVERSITY OF CRETE

Συστήματα Ανάκτησης Πληροφοριών
ΗΥ-463

4η Σειρά Ασκήσεων

Ψαράκη Μαρία-Γεωργία ΜΕΤ 556
psaraki@csd.uoc.gr

Εαρινό Εξάμηνο 2008-2009

Άσκηση 1η

Η απόσταση ενημέρωσης (Edit Distance) μεταξύ των λέξεων «αυτοκινητο» και «αυτοματο» είναι 4 (2 αντικαταστάσεις και 2 διαγραφές από την πρώτη στην δεύτερη).

Εφαρμόζοντας τον αλγόριθμο δυναμικού προγραμματισμού έχουμε: Έστω Τ = «αυτοκινητο» και Ρ = «αυτοματο». Ο πίνακας που προκύπτει είναι ο παρακάτω:

		α	υ	τ	ο	κ	ι	ν	η	τ	ο
	0	0	0	0	0	0	0	0	0	0	0
α	1	0	1	1	1	1	1	1	1	1	1
υ	2	1	0	1	2	2	2	2	2	2	2
τ	3	2	1	0	1	2	3	3	3	2	3
ο	4	3	2	1	0	1	2	3	4	3	2
μ	5	4	3	2	1	1	2	3	4	4	3
α	6	5	4	3	2	2	2	3	4	5	4
τ	7	6	5	4	3	3	3	3	4	4	5
ο	8	7	6	5	4	4	4	4	4	5	**4**

Άσκηση 2η

(α) Σύμφωνα με τον νόμο του Heaps το μέγεθος του λεξιλογίου εκτιμάται ως: V = Kn^β με n τον συνολικό αριθμό λέξεων και σταθερές K, β, όπου K ≈ 10−100 και β ≈ 0.4−0.6 (δηλαδή τετραγωνική ρίζα).
Οπότε έχουμε:

- Για K=10: V = Kn^β = 10*2,000,000^0.5 = 14,142 λέξεις
- Για K=100: V = Kn^β = 100*2,000,000^0.5 = 141,421 λέξεις

Άρα, το πλήθος των διαφορετικών λέξεων θα είναι μεταξύ 14,142 και 141,421 λέξεων.

(β) Σύμφωνα με τον νόμο του Zipf θα έχουμε ότι η 20η πιο συχνά εμφανιζόμενη λέξη θα εμφανίζεται: 1/20^θ φορές την πιο συχνή, όπου θ μεταξύ 1.5 και 2.
Οπότε έχουμε: 550,000*1/20^1.5 = 6,149 και 550,000*1/20^2 = 1,375
Άρα θα εμφανίζεται μεταξύ 1,375 και 6,149 φορές.

Άσκηση 3η

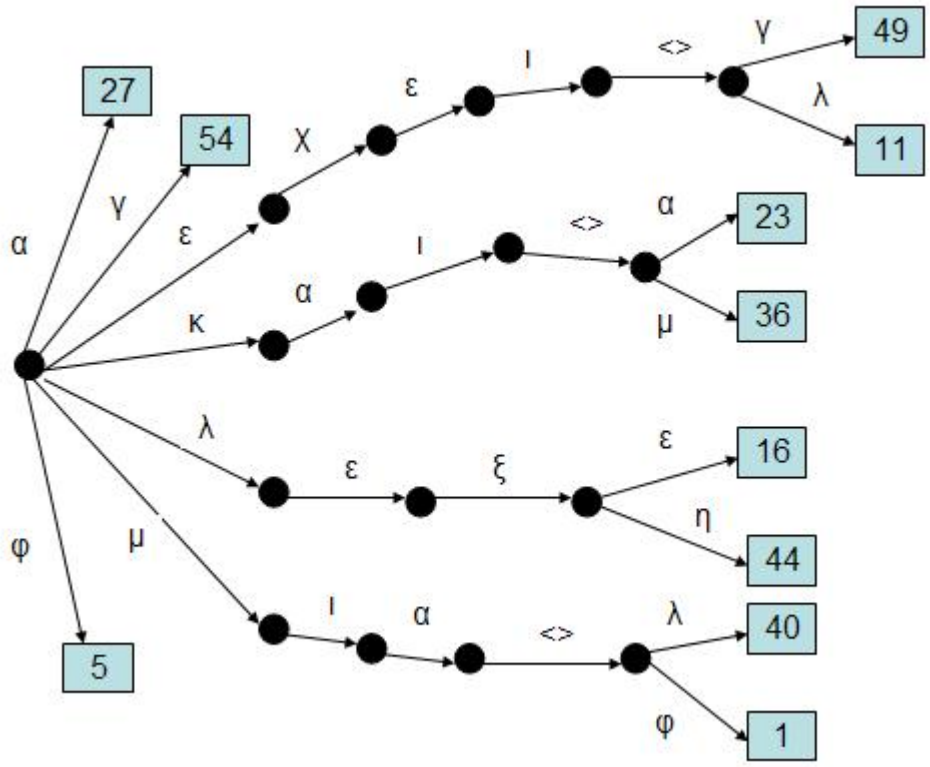
(α) Οι καταλήξεις του κειμένου είναι:

μια φράση έχει λέξεις και αριθμούς και μια λέξη έχει γράμματα
φράση έχει λέξεις και αριθμούς και μια λέξη έχει γράμματα
έχει λέξεις και αριθμούς και μια λέξη έχει γράμματα
λέξεις και αριθμούς και μια λέξη έχει γράμματα
και αριθμούς και μια λέξη έχει γράμματα
αριθμούς και μια λέξη έχει γράμματα
και μια λέξη έχει γράμματα
μια λέξη έχει γράμματα
λέξη έχει γράμματα
έχει γράμματα
γράμματα

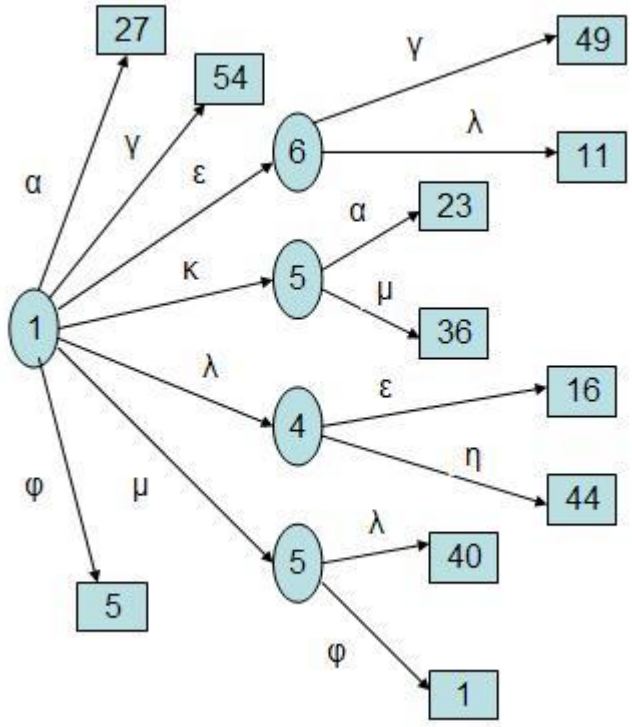
Ταξινομώντας τις καταλήξεις λεξικογραφικά έχουμε:

αριθμούς και μια λέξη έχει γράμματα
γράμματα
έχει γράμματα
έχει λέξεις και αριθμούς και μια λέξη έχει γράμματα
και αριθμούς και μια λέξη έχει γράμματα
και μια λέξη έχει γράμματα
λέξεις και αριθμούς και μια λέξη έχει γράμματα
λέξη έχει γράμματα
μια λέξη έχει γράμματα
μια φράση έχει λέξεις και αριθμούς και μια λέξη έχει γράμματα
φράση έχει λέξεις και αριθμούς και μια λέξη έχει γράμματα

Οπότε, το δένδρο καταλήξεων του κειμένου θεωρώντας ως σημεία ευρετηρίου τις αρχές των λέξεων είναι:



Το Patricia Tree είναι:

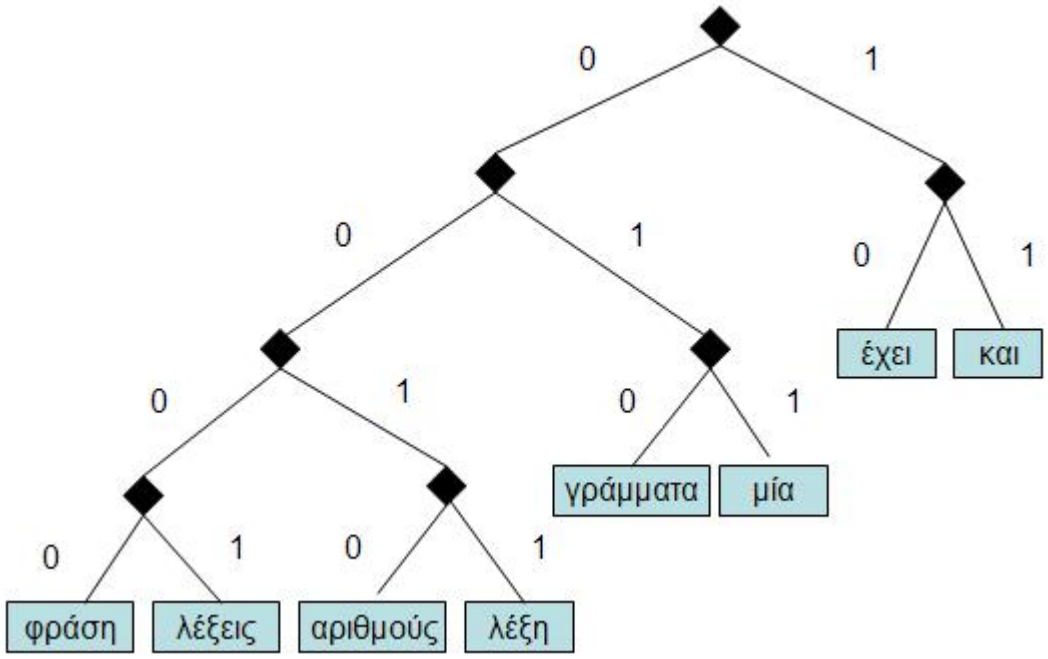


(β) Κωδικοποίηση κατά Huffman

Κείμενο: «μια φράση έχει λέξεις και αριθμούς και μια λέξη έχει γράμματα»

Συχνότητες:
«μια»(2), «έχει»(2), «και»(2), «φράση»(1), «λέξεις»(1), «αριθμούς»(1), «λέξη»(1), «γράμματα»(1)

Canonical Tree:



Το ύψος του αριστερού υποδένδρου για κάθε κόμβο δεν είναι ποτέ μικρότερο από αυτό του δεξιού υποδένδρου και όλα τα φύλλα είναι σε αύξουσα σειρά πιθανοτήτων από αριστερά προς δεξιά.

Άσκηση 4η

Έστω το εξής τμήμα ανεστραμμένου ευρετηρίου:

Gates	1; 2; 3; 4;
IBM	4; 7;
Microsof	1; 2; 3; 5;

Αν θεωρήσουμε την λίστα των ids των εγγράφων σαν μία ακολουθία από κενά μεταξύ των αριθμών των εγγράφων, έχουμε:

[1; 2; 3; 4;] → [1; 1; 1; 1;]
[4; 7] → [4; 3;]
[1; 2; 3; 5;] → [1; 1; 1; 2;]

Με αυτόν τον τρόπο έχουμε μικρές τιμές στην λίστα του ευρετηρίου μας. Μπορούμε να επιτύχουμε καλύτερη συμπίεση κωδικοποιώντας μικρές τιμές με συντομότερους κωδικούς. Έτσι, χρησιμοποιώντας την τεχνική Elias-γ έχουμε:

1 = 2^0 + 0 = 1
2 = 2^1 + 0 = 010
3 = 2^1 + 1 = 011
4 = 2^2 + 0 = 00100

Οπότε η συμπιεσμένη μορφή του ευρετηρίου έχει ως εξής:

Gates	1; 1; 1; 1;
IBM	00100; 011;
Microsof	1; 1; 1; 010;

Άσκηση 5η

Έστω τα έγγραφα d1,…,d9, και sim(di, dj) = (i+j)/20

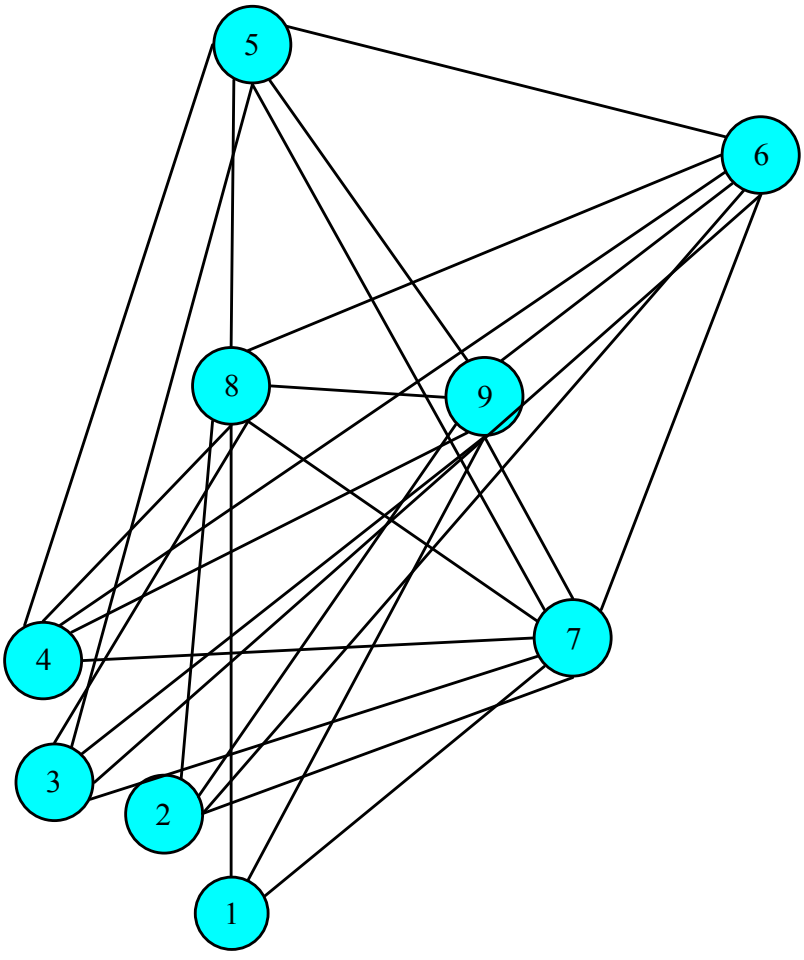
Έχουμε:
sim(d1, d9) = 10/20 = 0.5
sim(d1, d8) = 9/20 = 0.45
sim(d1, d7) = 8/20 = 0.4
sim(d1, d6) = 7/20 = 0.35
sim(d1, d5) = 6/20 = 0.3
sim(d1, d4) = 5/20 = 0.25

sim(d1, d3) = 4/20 = 0.2
sim(d1, d2) = 3/20 = 0.15
sim(d2, d9) = 11/20 = 0.55
sim(d2, d8) = 10/20 = 0.5
sim(d2, d7) = 9/20 = 0.45
sim(d2, d6) = 8/20 = 0.4
sim(d2, d5) = 7/20 = 0.35
sim(d2, d4) = 6/20 = 0.3
sim(d2, d3) = 5/20 = 0.25
sim(d3, d9) = 12/20 = 0.6
sim(d3, d8) = 11/20 = 0.55
sim(d3, d7) = 10/20 = 0.5
sim(d3, d6) = 9/20 = 0.45
sim(d3, d5) = 8/20 = 0.4
sim(d3, d4) = 7/20 = 0.35
sim(d4, d9) = 13/20 = 0.65
sim(d4, d8) = 12/20 = 0.6
sim(d4, d7) = 11/20 = 0.55
sim(d4, d6) = 10/20 = 0.5
sim(d4, d5) = 9/20 = 0.45
sim(d5, d9) = 14/20 = 0.7
sim(d5, d8) = 13/20 = 0.65
sim(d5, d7) = 12/20 = 0.6
sim(d5, d6) = 11/20 = 0.55
sim(d6, d9) = 15/20 = 0.75
sim(d6, d8) = 14/20 = 0.7
sim(d6, d7) = 13/20 = 0.65
sim(d7, d9) = 16/20 = 0.8
sim(d7, d8) = 15/20 = 0.75
sim(d8, d9) = 17/20 = 0.85

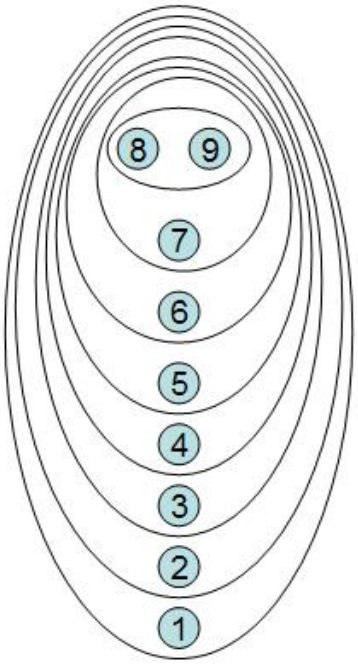
Έχοντας ως κατώφλι ομοιότητας την τιμή 0.4 κρατάμε τις εξής συσχιτίσεις:
sim(d1, d9) = 10/20 = 0.5
sim(d1, d8) = 9/20 = 0.45
sim(d1, d7) = 8/20 = 0.4
sim(d2, d9) = 11/20 = 0.55
sim(d2, d8) = 10/20 = 0.5
sim(d2, d7) = 9/20 = 0.45
sim(d2, d6) = 8/20 = 0.4
sim(d3, d9) = 12/20 = 0.6
sim(d3, d8) = 11/20 = 0.55
sim(d3, d7) = 10/20 = 0.5

sim(d3, d6) = 9/20 = 0.45
sim(d3, d5) = 8/20 = 0.4
sim(d4, d9) = 13/20 = 0.65
sim(d4, d8) = 12/20 = 0.6
sim(d4, d7) = 11/20 = 0.55
sim(d4, d6) = 10/20 = 0.5
sim(d4, d5) = 9/20 = 0.45
sim(d5, d9) = 14/20 = 0.7
sim(d5, d8) = 13/20 = 0.65
sim(d5, d7) = 12/20 = 0.6
sim(d5, d6) = 11/20 = 0.55
sim(d6, d9) = 15/20 = 0.75
sim(d6, d8) = 14/20 = 0.7
sim(d6, d7) = 13/20 = 0.65
sim(d7, d9) = 16/20 = 0.8
sim(d7, d8) = 15/20 = 0.75
sim(d8, d9) = 17/20 = 0.85

Οπότε ο γράφος που προκύπτει είναι:

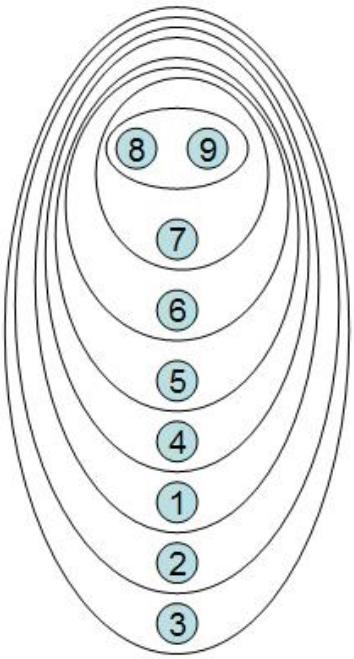


Με ομαδοποίηση κατά single link έχουμε:



sim(d1, d2-3-4-5-6-7-8-9) = 10/20 = 0.5
sim(d2, d3-4-5-6-7-8-9) = 11/20 = 0.55
sim(d3, d4-5-6-7-8-9) = 12/20 = 0.6
sim(d4, d5-6-7-8-9) = 13/20 = 0.65
sim(d5, d6-7-8-9) = 14/20 = 0.7
sim(d6, d7-8-9) = 15/20 = 0.75
sim(d7, d8-9) = 16/20 = 0.8
sim(d8, d9) = 17/20 = 0.85

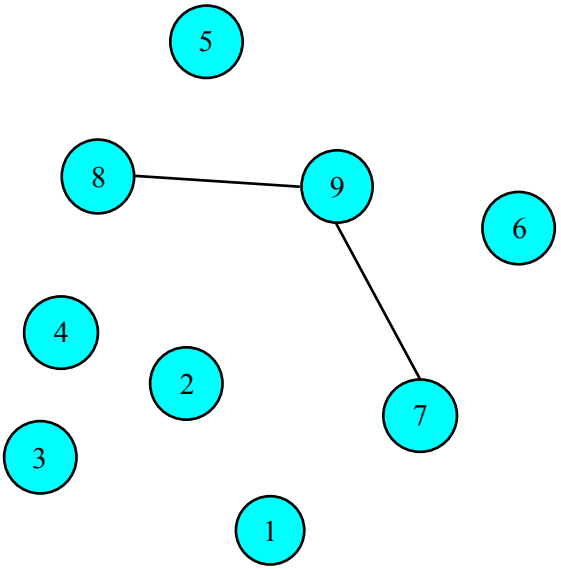
Με ομαδοποίηση κατά complete link έχουμε:

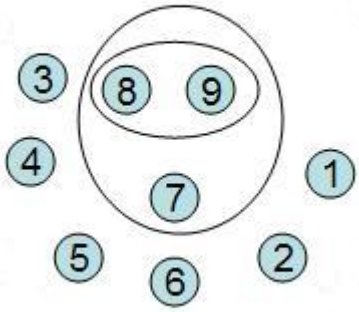


sim(d1, d7-8-9) = 8/20 = 0.4
sim(d2, d6-7-8-9) = 8/20 = 0.4
sim(d3, d5-6-7-8-9) = 8/20 = 0.4
sim(d4, d5-6-7-8-9) = 9/20 = 0.45
sim(d5, d6-7-8-9) = 11/20 = 0.55
sim(d6, d7-8-9) = 13/20 = 0.65
sim(d7, d8-9) = 15/20 = 0.75
sim(d8, d9) = 17/20 = 0.85

Έχοντας ως κατώφλι ομοιότητας την τιμή 0.8 κρατάμε τις εξής συσχιτίσεις:
sim(d7, d9) = 16/20 = 0.8
sim(d8, d9) = 17/20 = 0.85

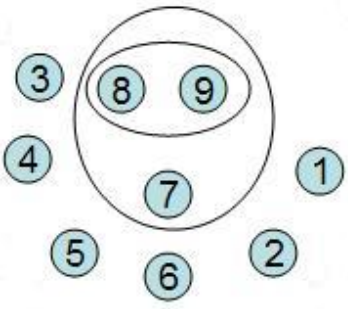
Οπότε ο γράφος που προκύπτει είναι:





Με ομαδοποίηση κατά single link έχουμε:

sim(d7, d9) = 16/20 = 0.8
sim(d8, d9) = 17/20 = 0.85



Με ομαδοποίηση κατά complete link έχουμε:

sim(d7, d9) = 16/20 = 0.8
sim(d8, d9) = 17/20 = 0.85

Άσκηση 6η

Μέγεθος ελληνικού παγκόσμιου ιστού

Αριθμός σελίδων: 3,7 εκ σελίδες
Μέσο μέγεθος σελίδων: 22 Kb
Αριθμός σελίδων ανά site: 150
Πλήθος εισερχόμενων συνδέσμων: 10.3
Πλήθος εξερχόμενων συνδέσμων: 17.2
Μέσος αριθμός των διαφορετικών ελληνικών Web sites που δείχνουν σε κάποιο δοσμένο (ελληνικό) Web site (average indegree per-Web site): 1.6
Μέσος αριθμός των διαφορετικών ελληνικών Web sites που δείχνονται από κάποιο δοσμένο (ελληνικό) Web site (average outdegree per-Web site): 4.8
Πληθυσμός (ελληνικός): 11.212.468
Χρήστες διαδικτύου: 3.800.000
Αύξηση χρηστών(‘00-‘05): 280%

Πηγή: [Characterization of National Web Domains] -2007

Site summary
Number of sites ok: 29,191
Number of sites with valid page age: 22,090
Average internal links: 1,093.67
Average pages per site: 146.90
Average static pages per site: 85.42
Average dynamic pages per site: 61.48
Average of age of oldest page in months: 16.86
Average of age of average page in months: 12.23
Average of age of newest page in months: 9.74
Average in-degree: 5.37
Average out-degree: 5.37
Average site max depth: 3.65
Average site size in MB: 2.61

Webpages summary
Total Web pages: 4,051,326
Unique: 3,781,912 93.35%
Duplicates: 269,414 6.65%
Static: 2,524,270 62.31%
Dynamic: 1,527,056 37.69%

Πηγή: [Characterization of National Web Domains] – 2004

Για την σχεδίαση του ευρετηρίου μιας μηχανής αναζήτησης για τον ελληνικό ιστό, και την εκτίμηση του μεγέθους του λεξιλογίου θα πρέπει πρώτα να εκτιμήσουμε το μέγεθος της συλλογής των κειμένων. Το μέγεθος του ελληνικού λεξιλογίου αντιστοιχεί σε 135.000 περίπου λέξεις([http://www.greeknewsonline.com/modules.php?name=News&file=print&sid=794]), συν άλλες 43.000 περίπου αγγλικές λέξεις, δηλαδή συνολικά θα αποτελείται από 178.000 διακριτές λέξεις. Θεωρώντας ότι κάθε λέξη θα έχει κατά μέσο όρο 10 χαρακτήρες (10 bytes) έχουμε:
Μέγεθος λεξιλογίου = 178.000*10 bytes = 1.780.000 bytes = 1,78 Mb

Όσον αφορά το συνολικό πλήθος λέξεων, έχουμε ότι ο ελληνικός ιστός αποτελείται από 4.000.000 περίπου σελίδες με μέσο μέγεθος σελίδας 22Kb. Κάθε λέξη έχει μέγεθος περίπου 10bytes, οπότε έχουμε:
Μέγεθος συλλογής κειμένων = 4.000.000*22.000/10 = 8.800.000.000 λέξεις.
Θεωρώντας ότι για την αποθήκευση κάθε εμφάνισης μιας λέξεις απαιτούνται 4 bytes έχουμε:
Μέγεθος λιστών εμφάνισης = 8.800.000.000*4 bytes = 35.200.000.000 bytes = 35,2 Gb

Επομένως, ο χώρος που απαιτείται για το ευρετήριο αντιστοιχεί σε 35,2 Gb + 1,78 Mb = 35,203 Gb.
Το ίδιο το λεξιλόγιο θα μπορούσε να βρίσκεται μόνιμα στην κύρια μνήμη για καλύτερες επιδόσεις.