**ΕΡΓΑΣΙΑ 1**

**ΣΤΟ ΜΑΘΗΜΑ: ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΜΗΧΑΝΕΣ ΑΝΑΖΗΤΗΣΗΣ**

1. A) Το αντεστραμμένο ευρετήριο (Inverted index) που προκύπτει είναι το εξής:

|  |  |  |
| --- | --- | --- |
| Term | Freq | docID |
| άνοδο | 1 | 2 |
| αύξηση | 2 | 3,4 |
| διαμερισμάτων | 4 | 1,2,3,4 |
| Ιούλιο | 3 | 2,3,4 |
| νεόκτιστων | 2 | 1,4 |
| ξεπέρασαν | 1 | 1 |
| οι | 2 | 1,2 |
| παρουσιάζουν | 1 | 2 |
| προβλέψεις | 1 | 1 |
| πωλήσεις | 4 | 1,2,3,4 |
| στις | 2 | 3,4 |
| τις | 1 | 1 |
| τον | 3 | 2,3,4 |
| υπάρχει | 1 | 4 |

B) Το ερώτημα <<πωλήσεις AND αύξηση>> επιστρέφει τα έγγραφα Doc3 και Doc4 γιατί {1,2,3,4} AND {3,4} = {3,4}

Το ερώτημα << Ιούλιο AND NOT (άνοδο OR αύξηση)>> ΔΕΝ επιστρέφει κανένα έγγραφο γιατί {2,3,4} AND NOT {2,3,4} = KENO

* Αρχικά θα υπολογίσουμε όλες τις ενώσεις Α = (τραπέζι OR καρέκλα), Β = (λάμπα OR κουρτίνα), Γ = (υπολογιστής OR μολύβι) και έπειτα θα υπολογίσουμε με AND τις μικρότερες τομές των ενώσεων.
* Αν το μήκος της ένωσης δυο postings lists είναι O(x + y) τότε θα υπολογίσουμε με την σειρά (υπολογιστής OR μολύβι) AND (λάμπα OR κουρτίνα) AND (τραπέζι OR καρέκλα). Άρα θα έχουμε (18000+31000) AND (5000+50000) AND (25000+32000) …

1. A) Το ερώτημα θα επεξεργαστεί με την σειρά <<X AND Y AND Z>> γιατί τα μεγέθη των postings lists είναι X<Y<Z και θα έχουμε κόστος 100 + 105 ( που προκύπτει από την πράξη A = X AND Y και θέτουμε το γράμμα Κ ως μήκος) άρα θα έχουμε <<A AND Z>> με κόστος Κ+110. Συνολικά θα έχουμε

ΚΟΣΤΟΣ= Κ+315.

Β) Τώρα αλλάζει η σειρά γιατί μας δίνονται τα μήκη κάποιον πράξεων και θα γίνει <<X AND Z AND Y>> με κόστος 100 + 110 (από την πράξη B = X AND Z) και (από την πράξη B AND Y) θα έχουμε L + 105 αλλά γνωρίζουμε ότι L=0. Συνολικά θα έχουμε ΚΟΣΤΟΣ = 315.

1. A) Το ερώτημα <<μην τρέχεις πολύ>> ταιριάζει στα έγγραφα doc2 (1,2,3), doc4 (8,9,10) και δυο φορές στο doc7(3,4,5 και 13,14,15).

Β) Το ερώτημα << μην τρέχεις πολύ AND υπάρχει κίνδυνος να σκοτωθείς>> ταιριάζει μόνο στο έγγραφο doc4 (8,9,10 και 12,13,14,15).

ΣΗΜΕΙΩΣΗ: Μέσα στις παρενθέσεις είναι οι θέσεις που γίνονται μάτς οι αντίστοιχες λέξεις.

1. Α) Με postings lists θα έχουμε Ο(x + y) συγκρίσεις. Δηλαδή 11 + 1 = 12 συγκρίσεις (σταματάμε στο 11 γιατί το 47 βρίσκεται σε αυτή την θέση).

Β) Με skip pointers και μήκος 4 (ρίζα του 16) θα κάνουμε 4 συγκρίσεις αλλά δεν θα βρούμε τον αριθμό που θέλουμε γιατί θα κάνουμε ‘’μεγαλύτερο’’ άλμα-βήμα. Άρα θα κάνουμε συνολικά 4+1= 5 συγκρίσεις.

1. Θα έχουμε Κ+1 (όπου Κ το άθροισμα των ορών της λέξης) όρους στην postings list. Εστω ότι εχουμε την λεξη <<paok>>. Θα εχουμε 5 ορους στη postings list (paok$, $paok, k$pao, ok$pa, aok$p).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | | | | | **p** | | | | | **a** | | | | **r** | | | | **i** | | | | **s** | |
|  |  | | | |  |  | | | |  |  | | |  |  | | |  |  | | |  |  |  | |
|  | | | | 0 | 1 | | | | 1 | 2 | | | 2 | 3 | | | 3 | 4 | | | 4 | 5 | 5 | |
| **a** |  | | | 1 | | 1 | | | 2 | | 1 | | 3 | | 3 | | 4 | | 4 | | 5 | | 5 | 6 |
|  | | | 1 | | 2 | | | 1 | | 2 | | 1 | | 2 | | 2 | | 3 | | 3 | | 4 | 4 |
| **l** |  | | 2 | | | 2 | | 2 | | | 2 | 2 | | | 2 | 3 | | | 3 | 4 | | | 4 | 5 |
|  | | 2 | | | 3 | | 2 | | | 3 | 2 | | | 3 | 2 | | | 3 | 3 | | | 4 | 4 |
| **i** |  | 3 | | | | 3 | 3 | | | | 3 | 3 | | | 3 | 3 | | | 2 | 4 | | | 4 | 5 | |
|  | 3 | | | | 4 | 3 | | | | 4 | 3 | | | 4 | 3 | | | 4 | 2 | | | 3 | 3 | |
| **c** |  | 4 | | | | 4 | 4 | | | | 4 | 4 | | | 4 | 4 | | | 4 | 3 | | | 3 | 4 | |
|  | 4 | | | | 5 | 4 | | | | 5 | 4 | | | 5 | 4 | | | 5 | 3 | | | 4 | 3 | |
| **e** |  | 5 | | | | 5 | 5 | | | | 5 | 5 | | | 5 | 5 | | | 5 | 4 | | | 4 | 4 | |
|  | 5 | | | | 6 | 5 | | | | 6 | 5 | | | 6 | 5 | | | 6 | 4 | | | 5 | 4 | |

1. Υπολογισμός απόστασης Levenshtein:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | **p** | **a** | **r** | **i** | **s** |
|  | **0** | **1** | **2** | **3** | **4** | **5** |
| **a** | **1** | **1** | **1** | **2** | **3** | **4** |
| **l** | **2** | **2** | **2** | **2** | **3** | **4** |
| **i** | **3** | **3** | **3** | **3** | **2** | **3** |
| **c** | **4** | **4** | **4** | **4** | **3** | **3** |
| **e** | **5** | **5** | **5** | **5** | **4** | **4** |

Οι πράξεις που οδήγησαν από alice σε paris:

|  |  |  |  |
| --- | --- | --- | --- |
| **cost** | **operation** | **input** | **output** |
| 1 | replace | \* | p |
| 0 | (copy) | a | a |
| 1 | replace | l | r |
| 0 | (copy) | i | i |
| 1 | replace | c | s |
| 1 | delete | e | \* |

1. Από τον τύπο του Jaccard coefficient **JC = |X ∩Y| / |X∪ Y|** θα έχουμε:

query: bord -> bo or rd

term: or -> or

JC = return 1/(3+1-1)= 1/3

query: bord -> bo or rd

term: border -> bo or rd de er

JC = return 3/(3+5-3)=3/5

query: bord -> bo or rd

term: lord -> lo or rd

JC = return 2/(3+3-2)= 1/2

query: bord -> bo or rd

term: morbid -> mo or rb bi id

JC = return 1/(3+5-1)= 1/7

query: bord -> bo or rd

term: sordid -> so or rd di id

JC = return 2/(3+5-2)= 1/3

1. Δυο τελείως διαφορετικά κυρία ονόματα που βρήκα είναι το Mark και το March. Έχουν soundex code M620.
2. Δυο κυρία ονόματα που είναι φωνητικά όμοια, είναι το Peter και το Peta. Έχουν soundex code P360 και P300 αντίστοιχα.

**ΑΠΟ ΤΟΝ ΦΟΙΤΗΤΗ: Τσιαούση Σταύρο (dai17173)**