ΗΥ-463 Συστήματα Ανάκτησης Πληροφοριών
Εαρινό Εξάμηνο 2008
Τμήμα Επιστήμης Υπολογιστών
Πανεπιστήμιο Κρήτης

Λύσεις 3ης Σειράς Ασκήσεων

Άσκηση 1 *(2 βαθμοί)*

Θεωρείστε ένα έγγραφο με περιεχόμενο:
«στο σημερινό μάθημα μάθαμε περισσότερα σε σχέση με το προηγούμενο μάθημα»
Αγνοώντας τους τόνους, σχεδιάστε

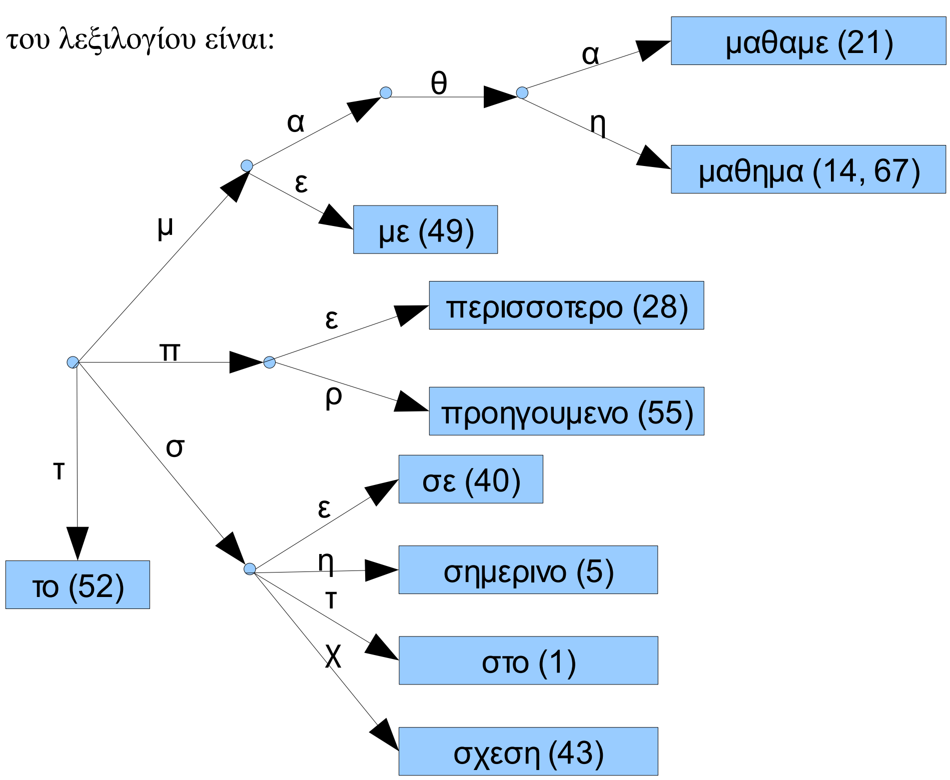
(α) το trie του λεξιλογίου του παραπάνω εγγράφου,
(β) το δένδρο καταλήξεων του παραπάνω εγγράφου θεωρώντας ως σημεία ευρετηρίου (index points) τις αρχές των λέξεων, και
(γ) συμπτύξτε το παραπάνω δένδρο καταλήξεων στην μορφή ενός Patricia tree.

Λύση

(α)
Αγνοώντας τους τόνους έχουμε το εξής έγγραφο:

```
1   5        14     21     28          40 43    49 52 55          67
στο σημερινο μαθημα μαθαμε περισσότερα σε σχεση με το προηγουμενο μαθημα
```

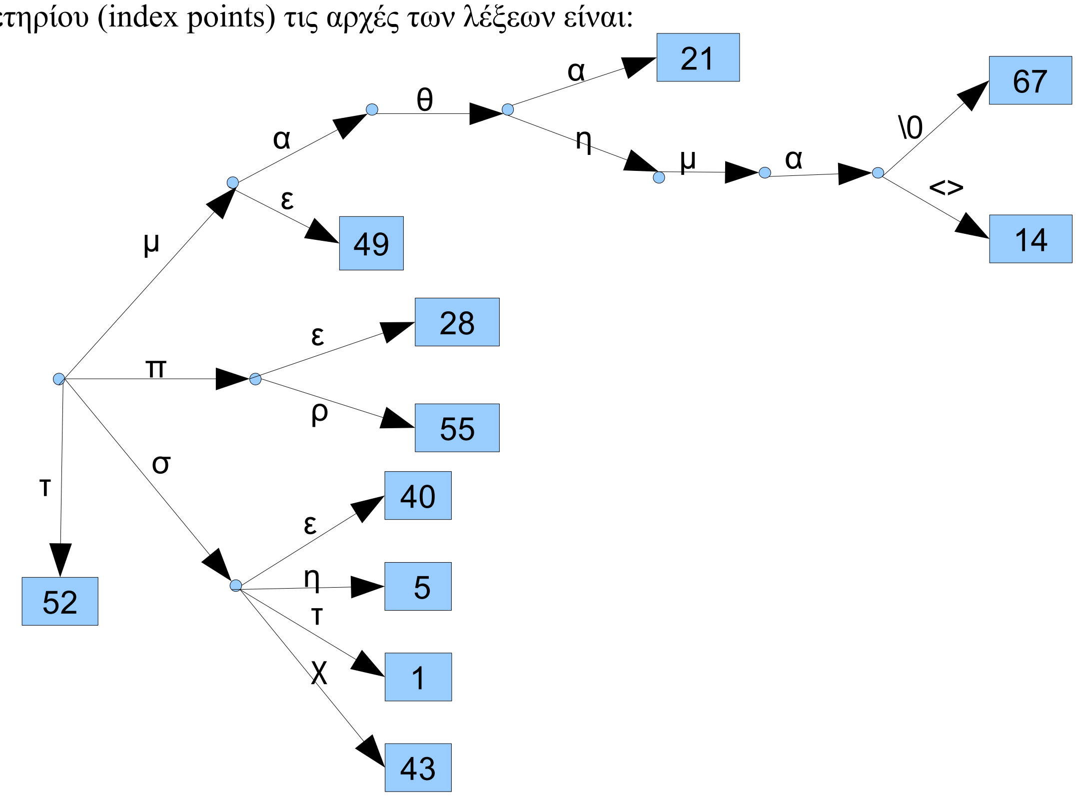
Άρα το trie του λεξιλογίου είναι:



(β)
Οι καταλήξεις του κειμένου βάσει της αρχής των λέξεων είναι:
στο σημερινο μαθημα μαθαμε περισσοτερα σε σχεση με το προηγουμενο μαθημα
σημερινο μαθημα μαθαμε περισσοτερα σε σχεση με το προηγουμενο μαθημα
μαθημα μαθαμε περισσοτερα σε σχεση με το προηγουμενο μαθημα
μαθαμε περισσοτερα σε σχεση με το προηγουμενο μαθημα
περισσοτερα σε σχεση με το προηγουμενο μαθημα
σε σχεση με το προηγουμενο μαθημα
σχεση με το προηγουμενο μαθημα
με το προηγουμενο μαθημα
το προηγουμενο μαθημα
προηγουμενο μαθημα
μαθημα

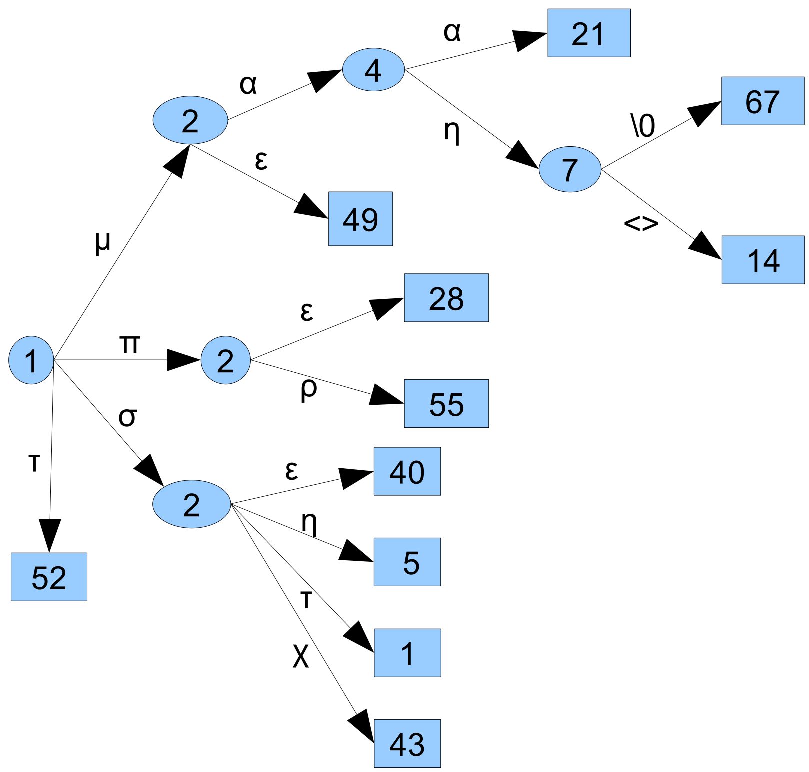
Ταξινομώντας τις παραπάνω καταλήξεις λεξικογραφικά έχουμε:
μαθαμε περισσοτερα σε σχεση με το προηγουμενο μαθημα
μαθημα
μαθημα μαθαμε περισσοτερα σε σχεση με το προηγουμενο μαθημα
με το προηγουμενο μαθημα
περισσοτερα σε σχεση με το προηγουμενο μαθημα
προηγουμενο μαθημα
σε σχεση με το προηγουμενο μαθημα
σημερινο μαθημα μαθαμε περισσοτερα σε σχεση με το προηγουμενο μαθημα
στο σημερινο μαθημα μαθαμε περισσοτερα σε σχεση με το προηγουμενο μαθημα
σχεση με το προηγουμενο μαθημα
το προηγουμενο μαθημα

Το δένδρο καταλήξεων που προκύπτει από τις ταξινομημένες καταλήξεις, θεωρώντας ως σημεία του ευρετηρίου (index points) τις αρχές των λέξεων είναι:



(γ)

Το Patricia δένδρο που προκύπτει από το παραπάνω suffix δένδρο είναι το εξής:



Άσκηση 2 *(2 βαθμοί)*

Θεωρείστε το αλφάβητο {α, β, γ, δ, ε, ζ} και την εξής πρόταση:
"α α β β γ α γ α δ α α α δ β ε β ε ζ"
(α) Βάσει αυτής της φράσης ποιά είναι η εντροπία του αλφαβήτου;
(β) Δώστε τη συμπιεσμένη μορφή της φράσης χρησιμοποιώντας κανονικοποιημένους κώδικες Huffman

Λύση

(α)
Η εντροπία εκφράζει το κατώτερο όριο, μετρημένο σε bits ανά σύμβολο, το οποίο χρησιμοποιείται σε μεθόδους κωδικοποίησης και βασίζεται στην πιθανότητα εμφάνισης κάθε συμβόλου. Υπολογίζεται ως εξής:

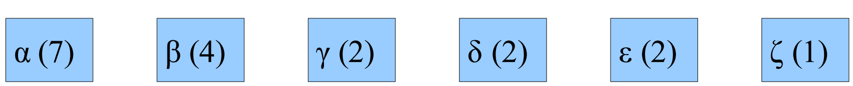
$$E = \sum p_i \log_2 \left(\frac{1}{p_i}\right)$$

Αρά βρίσκουμε τις πιθανότητες να εμφανιστεί κάποιο από τα σύμβολα του αλφαβήτου στην πρόταση που μας δόθηκε και έχουμε:

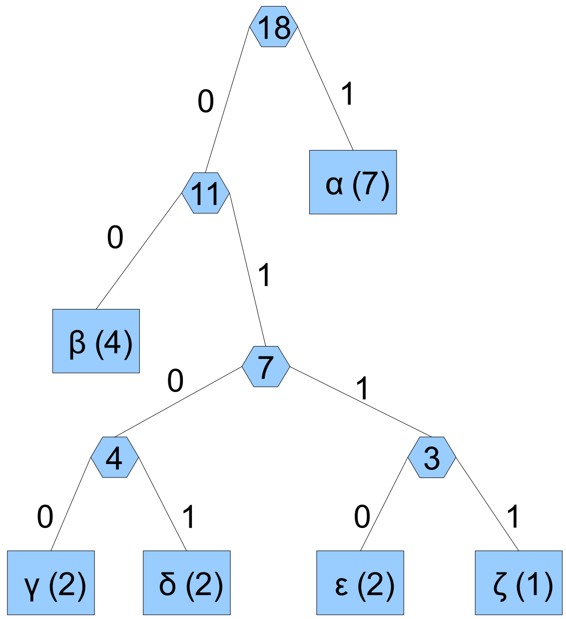
|α| = 7	$P_1 = 7/18 = 0.388$
|β| = 4	$P_2 = 4/18 = 0.222$
|γ| = 2	$P_3 = 2/18 = 0.111$
|δ| = 2	$P_4 = 2/18 = 0.111$
|ε| = 2	$P_5 = 2/18 = 0.111$
|ζ| = 1	$P_6 = 1/18 = 0.055$

Άρα έχουμε **E** = 0.388 * 1.365 + 0.222 * 2.171 + 0.111 * 3.171 + 0.111 * 3.171 + 0.111 * 3.171 +0.055 * 4.184 = 2.297

(β)
Αρχικά για να υπολογίσουμε τους Huffman κώδικες πρέπει να δημιουργήσουμε ένα κόμβο για κάθε σύμβολο του αλφαβήτου και να υπολογίσουμε το πλήθος των εμφανίσεων κάθε συμβόλου. Παρακάτω για απλότητα στην απεικόνιση χρησιμοποιείται αντί για την πιθανότητα εμφάνισης το πλήθος των εμφανίσεων.
Έτσι λοιπόν :



Κατόπιν πρέπει να πάρουμε τους 2 κόμβους με την μικρότερη πιθανότητα εμφάνισης και να τους συνδέσουμε σε ένα κοινό πατρικό κόμβο, ο οποίος ως πιθανότητα εμφάνισης θα έχει το άθροισμα των πιθανοτήτων εμφάνισης των 2 παιδιών του. Η διαδικασία αυτή πρέπει να επαναληφθεί για όλους τους κόμβους με τον ίδιο τρόπο αγνοώντας όμως τους κόμβους που ήδη έχουν εισαχθεί αλλά λαμβάνοντας υπ' όψιν τους γονικούς που δεν έχουν άλλους γονείς. Ακολουθώντας αυτή τη διαδικασία φθάνουμε στο εξής δένδρο:



Από το παραπάνω δένδρο μπορούμε να κωδικοποιήσουμε τα σύμβολα ως εξής:

α	1
β	00
γ	0100
δ	0101
ε	0110
ζ	0111

Άρα το παραπάνω κείμενο κωδικοποιείται ως εξής

α	α	β	β	γ	α	γ	α	δ	α	α	α	δ	β	ε	β	ε	ζ
1	1	00	00	0100	1	0100	1	0101	1	1	1	0101	00	0110	00	0110	0111

Άσκηση 3 *(1 βαθμός)*

Υπολογίστε την Edit distance μεταξύ των λέξεων paris και alice. Δώστε τον 5x5 πίνακα που περιγράφει τον τρόπο λειτουργίας του σχετικού αλγορίθμου δυναμικού προγραμματισμού.

Λύση:

Χρησιμοποιώντας τον αλγόριθμο του Levenshtein έχουμε:

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

Άρα το edit distance είναι 4.

Υπάρχουν δύο τρόποι για να πάρουμε από τη λέξη paris τη λέξη alice
Ο πρώτος είναι ως εξής:

- Αφαιρούμε το p
- Αντικαθιστούμε το r με l
- Αντικαθιστούμε το s με c
- Προσθέτουμε το e

Ο δεύτερος έχει ως εξής:

- Αφαιρούμε το p
- Αντικαθιστούμε το r με l
- Προσθέτουμε το c
- Αντικαθιστούμε το s με e

Όταν μετακινούμαστε κάθετα έχουμε διαγραφή χαρακτήρα, όταν μετακινούμαστε οριζόντια έχουμε εισαγωγή χαρακτήρα, ενώ όταν μετακινούμαστε διαγώνια και η διαφορά είναι +1 έχουμε αντικατάσταση, αλλιώς δεν γίνεται κάποια λειτουργία.

Άσκηση 4 *(2 βαθμοί)*

Θεωρείστε μια συλλογή εγγράφων στην οποία εμφανίζονται 400 διαφορετικές λέξεις και η συχνότητα της πιο συχνά εμφανιζόμενης ισούται με 900.
(α) Εκτιμήστε τον αριθμό εμφανίσεων της $10_{ης}$ πιο συχνά εμφανιζόμενης λέξης και της $20_{ης}$.
(β) Αποφασίζετε να φτιάξετε ένα ανεστραμμένο ευρετήριο μόνο για τις 200 πιο συχνά εμφανιζόμενες λέξεις. Πόσο μικρότερο να είναι το μέγεθός του σε σχέση με εκείνο για όλες τις λέξεις; Θεωρείστε ότι δεν κάνουμε απαλοιφή λέξεων αποκλεισμού, ούτε στελέχωση.

Λύση:

(α)

Ο νόμος του Zipf, λέει ότι η συχνότητα της i-οστής πιο συχνά εμφανιζόμενης λέξης είναι 1/i φορές η συχνότητα της πιο συχνής. Άρα αν η πιο συχνά εμφανιζόμενη λέξη έχει συχνότητα 900, τότε ο αριθμός εμφανίσεων της $10_{ης}$ πιο συχνά εμφανιζόμενης λέξης θα είναι 1/10 * 900 = 90 φορές και της της $20_{ης}$ 1/20 * 90 = 45 φορές.

(β)

Το να χρησιμοποιήσουμε μόνο τις 200 πιο συχνά εμφανιζόμενες λέξεις από τις 400 διαθέσιμες, σίγουρα θα μειώσει το μέγεθος του ανεστραμμένου ευρετηρίου, αλλά όχι σημαντικά. Ο λόγος είναι ότι αυτές οι λέξεις δεν θα εμφανίζονται συχνά, με αποτέλεσμα το μέγεθος του posting file (που είναι και η μέγιστη επιβάρυνση σε όγκο) να είναι ελάχιστα μεγαλύτερο.

Έστω ότι θεωρούμε το μέσο μέγεθος των λέξεων του λεξιλογίου μας είναι 5 bytes. Τότε το μέξεθος του λεξιλογίου για τις 200 λέξεις είναι VS_{200} = 1000 bytes, ενώ για τις 400 λέξεις είναι VS_{400} = 2000 bytes.

Όσον αφορά τις εμφανίσεις των λέξεων, αν υποθέσουμε ότι κάθε εμφάνιση χρειάζεται 4 bytes για την αποθήκευσή της, τότε το μέγεθος του posting file είναι PS = 4 * OccNum bytes. Οπότε χρειάζεται να υπολογίσουμε το πλήθος των εμφανίσεων των λέξεων για τις 2 περιπτώσεις μας.

Για τις 400 πιο συχνά εμφανιζόμενες λέξεις έχουμε ότι το συνολικό άθροισμα των εμφανίσεών τους είναι:

$$PS_{٤..}=\sum_{i=1}^{400} 900*1/i=\int_{i=1}^{400}(900*1/i)\,di=\ln(x)*\int_{i=1}^{400} di=\ln(400)-\ln(1)=\ln(400)-0=\ln(400)=5991$$

Άρα PS_{400} = 5991 * 4 = 23964 bytes

Άντίστοιχα για τις 200 λέξεις έχουμε:

$$PS_{٢..}=\sum_{i=1}^{200} 900*1/i=\int_{i=1}^{200}(900*1/i)\,di=\ln(x)*\int_{i=1}^{200} di=\ln(200)-\ln(1)=\ln(200)-0=\ln(200)=5298$$

Άρα PS_{200} = 5298 * 4 = 21192 bytes

Άρα το τελικό μέγεθος του κάθε ευρετηρίου είναι:
$S_{400} = VS_{400} + PS_{400}$ = 23964 + 2000 = **25965**
και $S_{200} = VS_{200} + PS_{200}$ = 21192+ 1000 = **22192**

Άρα το μέγεθος του ανεστραμμένου ευρετηρίου είναι 14.5% μικρότερο μόνο.

Άσκηση 5 *(3 βαθμοί)*

Θεωρείστε τα ακόλουθα έγγραφα, όπου τα γράμματα A-E συμβολίζουν λέξεις.

d1 = "A A B"
d2 = "B A A"
d3 = "A B C"
d4 = "C C B"
d5 = "C D A"
d6 = "D E"
d7 = "A A B"
d8 = "E B"

Έστω ότι τα d1, d2, d3 ανήκουν σε ένα σύστημα S1, τα d4, d5 σε ένα σύστημα S2 και τα υπόλοιπα d6, d7, d8 σε ένα σύστημα S3. Θέλουμε να φτιάξουμε ένα μεσίτη M πάνω από αυτά τα συστήματα.

(α) Για την επιλογή της πηγής, ο M θέλει να περιγράψει τα περιεχόμενα της κάθε πηγής με ένα διάνυσμα. Δώστε τα διανύσματα πηγών των S1, S2, και S3.

(β) Έστω ότι ο M έχει ήδη τα δυανύσματα πηγών των S1, S2 και S3 και λαμβάνει την επερώτηση q = "B C". Αν θέλει να προωθήσει την επερώτηση q σε μία μόνο πηγή, ποια θα επιλέξει;

(γ) Ο M λαμβάνει μια επερώτηση, την προωθεί σε όλες τις πηγές και λαμβάνει τα εξής αποτελέσματα από την κάθε μια:
S1: (d1, d2, d3)
S2: (d5, d4)
S3: (d8, d7, d6)
Δώστε την ενοποιημένη διάταξη κατά round robin interleaving

(δ) Προκειμένου ο μεσίτης να λαμβάνει από τις πηγές απαντήσεις με συγκρίσιμα σκορ, αποφασίζει να κάνει αποτίμηση επερωτήσεων σε 2 φάσεις ώστε οι πηγές να λαμβάνουν τα καθολικά στατιστικά που χρειάζονται για το σωστό υπολογισμό των σκορ. Δώστε το idf του κάθε όρου στην καθολική συλλογή εγγράφων.

(ε) Ο Μεσίτης βρίσκει άλλο ένα σύστημα S4 το οποίο έχει την ίδια συλλογή με αυτή του S1, δηλαδή και αυτό παρέχει πρόσβαση στα έγγραφα d1, d2, d3. Έστω ότι ο M προωθεί μια επερώτηση q στα S1και S4 και λαμβάνει τις εξής απαντήσεις:
S1: (d2, d1, d3)
S4: (d2, d3, d1)
Ποιο είναι το κορυφαίο έγγραφο αν ενοποιήσουμε τις διατάξεις: i) κατά Borda και ii) κατά Condorcet. Ο M αποφασίζει να δίνει στο χρήστη όχι μόνο την ενοποιημένη διάταξη, αλλά και την Kemeny distance, μεταξύ των διατάξεων που έλαβε από τα υποσυστήματα (προκειμένου ο χρήστης να παίρνει μία γεύση για το βαθμό συμφωνίας των πηγών). Ποια είναι αυτή η απόσταση στην προκειμένη;

(στ) Τα συστήματα S1, S2, S3 δε θέλουν πλέον να έχουν ανάγκη τον M, και αποφασίζουν να ανεξαρτητοποιηθούν, φτιάχνοντας ένα σύστημα ομοτίμων (P2P), συγκεκριμένα ένα σύστημα τύπου Chord. Προσελκύουν μάλιστα άλλα 2 συστήματα S5 και S6 (τα οποία δεν έχουν καμία συλλογή εγγράφων). Αποφασίζουν να χρησιμοποιήσουν μία συνάρτηση κατακερματισμού h των 3 bits και έστω ότι: h(IPaddress(S1)) = 1, h(IPaddress(S2)) = 2, h(IPaddress(S3)) = 4, h(IPaddress(S5)) = 5, h(IPaddress(S6)) = 6. Αποφασίζουν να διανείμουν το ανεστραμμένο ευρετήριο θεωρώντας κάθε όρο σαν κλειδί και έστω ότι h(A) = 2, h(B) = 3, h(C) = 4, h(D) = 4 και h(E) = 4.
Δώστε i) τους πίνακες δρομολόγησης των κόμβων S1 και S3 και ii) πως θα κατανεμηθεί το

ανεστραμμένο ευρετηρίο στους κόμβους του δικτύου (δείξτε τι ακριβώς θα έχει κάθε κόμβος).

Λύση:

(α)

Το λεξιλόγιό μας αποτελείται από τις λέξεις A, B, C, D και E.
Θεωρούμε την κάθε πηγή σαν ένα έγγραφο, το οποίο περιέχει όλες τις λέξεις που περιέχουν όλα τα έγγραφα της. Άρα για τις 3 πηγές έχουμε και χρησιμοποιώντας βάρυνση TF-IDF, έχουμε (για απλοποίηση IDF = N / DF):

	$Freq_{S1}$	$Freq_{S2}$	$Freq_{S3}$	TF_{S1}	TF_{S2}	TF_{S3}	IDF	$TF\text{-}IDF_{S1}$	$TF\text{-}IDF_{S2}$	$TF\text{-}IDF_{S3}$
A	5	1	2	5/5	1/3	2/2	3/3	1	1/3	1
B	3	1	2	3/5	1/3	2/2	3/3	3/5	1/3	1
C	1	3	0	1/5	3/3	0	3/2	3/10	3/2	0
D	0	1	1	0	1/3	1/2	3/2	0	1/2	3/4
E	0	0	2	0	0	2/2	3/1	0	0	3

Άρα τα διανύσματα των πηγών είναι:
S1: <1, 0.6, 0.3, 0, 0>
S2: <0.333, 0.333, 1.5, 0.5, 0>
S3: <1, 1, 0, 0.75, 3>

(β)

Έστω q = "B C". Μπορούμε να χρησιμοποιήσουμε το μέτρο ομοιότητας συνημιτόνου για να δούμε σε ποια πηγή θα προωθήσει ο Μεσίτης την συγκεκριμένη επερώτηση. Άρα αρκεί να συγκρίνουμε το διάνυσμα της επερώτησης με τα διανύσματα των πηγών.

Υπολογίζουμε το διάνυσμα της επερώτησης χρησιμοποιώντας τα idf των πηγών από τα τον παραπάνω πίνακα και έχουμε q = <(0 * 1), (1*1), (1 * 1.5), (0 * 1.5), (0 * 3)> = <0, 1, 1.5, 0, 0>

Οπότε χρησιμοποιώντας τον παραπάνω πίνακα και τον τύπο:

$$CosSim(d_j, Q) = \frac{\vec{d_j} \bullet \vec{q}}{\left|\vec{d_j}\right| \bullet \left|\vec{q}\right|} = \frac{\sum_{i=1}^{t} \left(w_{ij} \bullet w_{iq}\right)}{\sqrt{\sum_{i=1}^{t} w_{ij}^2 \bullet \sum_{i=1}^{t} w_{iq}^2}}$$

έχουμε:
|S1| = 1.2
|S2| = 1.65
|S3| = 3.4
|q| = 1.8

CosSim(S1, q) = (0.6 + 0.45) / (1.2 * 1.8) = 1.05 / 2.16 = 0.49
CosSim(S2, q) = (0.33 + 2.25) / (1.65 * 1.8) = 0.87
CosSim(S3, q) = (1 * 1) / (3.4 * 1.8) = 0.16

Άρα ο μεσίτης θα προτιμήσει να προωθήσει την επερώτηση στην πηγή S2.

(γ)
Ο Μ λαμβάνει για την επερώτηση q από κάθε πηγή τα εξής:
S1: <d1, d2, d3>
S2: <d5, d4>
S3: <d8, d7, d6>

Άρα η ενοποιημένη διάταξη κατά round robin interleaving είναι:
<{d1, d5, d8}, {d2, d4, d7}, {d3, d6}>

(δ)
Σε πρώτη φάση ο Μ στέλνει όλες τις λέξεις του ευρετηρίου και τις αποτιμά ώστε να υπολογίσει και να στείλει τα καθολικά στατιστικά των όρων. Στη 2η φάση τα στατιστικά που θα στείλει ο Μ (αγνοώντας για απλότητα τον υπολογισμό του αλγορίθμου και γνωρίζοντας ότι ο αριθμός των εγγράφων είναι 8) είναι :
{A: 8/5, B: 8/6, Γ: 8/3, Δ: 8/2, E: 8/2}

(ε)
Έχουμε
S1:<d2, d1, d3>
S4: <d2, d3, d1>

i) Κατά Borda έχουμε:
V(d1) = 2 + 3 = 5
V(d2) = 1 + 1 = 2
V(d3) = 3 + 2 = 5

Άρα το κορυφαίο έγγραφο είναι το d2

ii) Κατά Condorset κοιτάμε ανά ζεύγη ποιος νικά
d1:d2 0:2 (Το d2 νικά 2 φορές το d1)
d1:d3 1:1 (Ισοπαλία)
d2:d3 2:0 (Το d2 νικά 2 φορές το d3)

Άρα το d2 είναι το καλύτερο

iii) Απόσταση κατά Kemeny, το πλήθος των διαφωνιών στη διάταξη ζευγαριών
K(S1, S2) = 1
(S1:d1>d3, S2:d3>d1)

στ) Από εκφώνηση έχουμε ότι:
h(IPaddress(S1)) = 1, h(IPaddress(S2)) = 2, h(IPaddress(S3)) = 4, h(IPaddress(S5)) = 5, h(IPaddress(S6)) = 6 και h(A) = 2, h(B) = 3, h(C) = 4, h(D) = 4 και h(E) = 4.

Ο πίνακας δρομολόγησης ενός κόμβου στο Chord αποτελείται από m εγγραφές (όπου m είναι ο αριθμός των bits στην hash function) και κάθε εγγραφή έχει την διεύθυνση του πρώτου κόμβου κλειδί μεγαλύτερο ή ίσο με n+2i-1 δηλαδή
finger[i] = successor(n+2i-1)

οπότε για τον S1 γνωρίζουμε ότι είναι ο κόμβος με h(S1) = 1 και άρα θα έχει πίνακα δρομολόγησης :
finger[1] = successor(1 + 1) = successor(2) = 2 (S2)
finger[2] = successor(1 + 2) = successor(3) = 4 (S3)
finger[3] = successor(1 + 4) = successor(5) = 5 (S5)

Όμοια για το S3 έχουμε:
finger[1] = successor(4 + 1) = successor(5) = 5 (S5)
finger[2] = successor(4 + 2) = successor(6) = 6 (S6)
finger[3] = successor(4 + 4) = successor(8) = 2 (S2)

(β)
Αρχικά υπολογίζουμε το ανεστραμμένο αρχείο βάσει των εμφανίσεων των λέξεων A, B, C, D, E στα έγγραφά (d1, ..., d8)

A: <d1, 2>, <d2, 2>, <d3, 1>, <d5, 1>, <d7,2>
B: <d1, 1>, <d2, 1>, <d3, 1>, <d4, 1>, <d7, 1>, <d8, 1>
C: <d3, 1>, <d4, 2>, <d5, 1>
D: <d5, 1>, <d6, 1>
E: <d6, 1>, <d8, 1>

Γνωρίζουμε ότι ένα κλειδί k εκχωρείται στον πρώτο κόμβο p, τέτοιο ώστε h(p) >= h(k). Από τις δοσμένες τιμές της άσκησης προκύπτει ότι:

Το κλειδί A εκχωρείται στον κόμβο S2, αφού h(S2) >= h(A)
Το κλειδί B εκχωρείται στον κόμβο S3, αφού h(S3) >= h(B)
Το κλειδί C εκχωρείται στον κόμβο S3, αφού h(S3) >= h(C)
Το κλειδί D εκχωρείται στον κόμβο S3, αφού h(S3) >= h(D)
Το κλειδί E εκχωρείται στον κόμβο S3, αφού h(S3) >= h(E)

Επομένως το ανεστραμμένο ευρετήριο θα κατανεμηθεί στους κόμβους ώς εξής:
S2: key (A) <d1, 2>, <d2, 2>, <d3, 1>, <d5, 1>, <d7, 2>
S3: key (B) <d1, 1>, <d2, 1>, <d3, 1>, <d4, 1>, <d7, 1>, <d8, 1>
S3: key (C) <d3, 1>, <d4, 2>, <d5, 1>
S3: key (D) <d5, 1>, <d6, 1>
S3: key (E) <d6, 1>, <d8, 1>