Φροντιστήριο 5

Άσκηση 1

Θεωρείστε το αλφάβητο {α,β,γ,δ,ε} και την εξής φράση: «α α β γ γ α α α α α δ β ε β».
α) Βάσει αυτής της φράσης ποια 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 εφαρμόζεται σε μεθόδους κωδικοποίησης και βασίζεται στην πιθανότητα εμφάνισης του κάθε συμβόλου.

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

Το πρώτο πράγμα λοιπόν που πρέπει να κάνουμε είναι να βρούμε τις πιθανότητες εμφάνισης των συμβόλων.
Έτσι έχουμε:

$|\alpha| = 8$ $\quad P_1 = \frac{8}{15} = 0.533$

$|\beta| = 3$ $\quad P_2 = \frac{3}{15} = 0.2$

$|\gamma| = 2$ $\quad P_3 = \frac{2}{15} = 0.133$

$|\delta| = 1$ $\quad P_4 = \frac{1}{15} = 0.067$

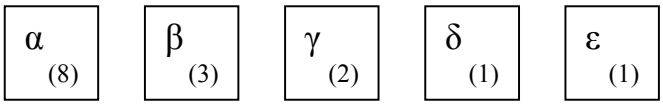
$|\varepsilon| = 1$ $\quad P_5 = \frac{1}{15} = 0.067$

Έτσι σύμφωνα με τον παραπάνω τύπο ο υπολογισμός του Ε είναι :

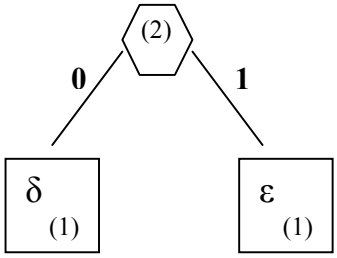
E=0.533*0.907 + 0.2*2.322 + 0.133*2.907 + 0.067*3.907 +0.067*3.907
E=1.858

(β)

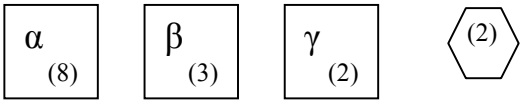
Αρχικά για να υπολογίσουμε τους Huffman κώδικες πρέπει να δημιουργήσουμε ένα κόμβο για κάθε σύμβολο του αλφαβήτου και να υπολογίσουμε ο πλήθος των εμφανίσεων κάθε συμβόλου. (παρακάτω για απλότητα στην απεικόνιση χρησιμοποιείται αντί για την πιθανότητα εμφάνισης το πλήθος των εμφανίσεων).
Έτσι λοιπόν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



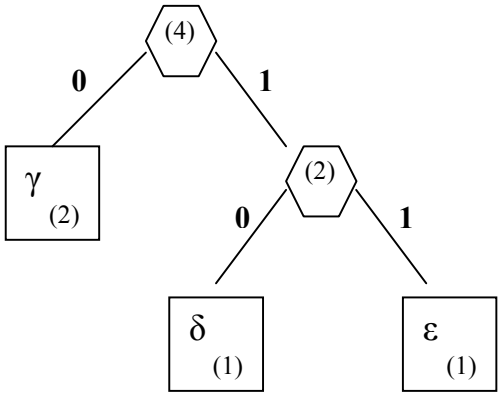
Κατόπιν πρέπει να πάρουμε τους 2 κόμβους με την μικρότερη πιθανότητα εμφάνισης και να τους συνδέσουμε σε ένα κοινό πατρικό ο οποίος ως πιθανότητα εμφάνισης θα έχει το άθροισμα των πιθανοτήτων εμφάνισης των 2 παιδιών του.



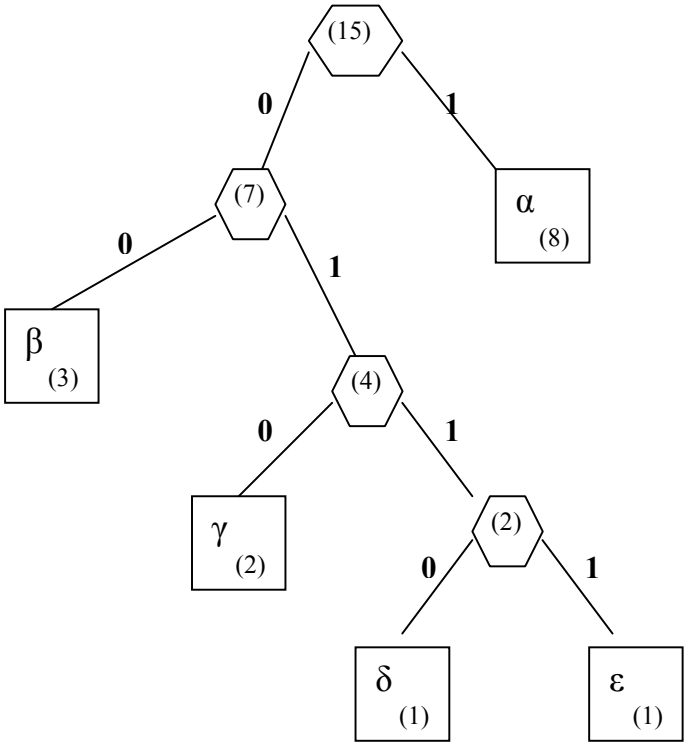
Κατόπιν η διαδικασία αυτή πρέπει να επαναληφθεί για όλους τους κόμβους με τον ίδιο τρόπο αγνοώντας όμως τους κόμβους που ήδη έχουν εισαχθεί αλλά λαμβάνοντας υπ' όψιν τους γονικούς που δεν έχουν άλλους γονείς. Εδώ δηλαδή οι κόμβοι που πρέπει να κοιτάξουμε θα είναι :



Και έτσι οι 2 τελευταίοι θα συνδεθούν με ένα κοινό γονέα.



Τελικά το δένδρο που θα προκύψει θα είναι:



Τέλος το παραπάνω κείμενο κωδικοποιείται :

α	α	β	γ	γ	α	α	α	α	α	α	δ	β	ε	β
1	1	00	010	010	1	1	1	1	1	1	0110	00	0111	00

Άσκηση 2

Θεωρείστε 8 έγγραφα Α, Β, Γ, Δ, Ε, Ζ, Η, Θ, Ι, Κ και έστω ότι οι αποστάσεις μεταξύ τους είναι αυτές του παρακάτω πίνακα. Δώστε το δενδρικό διάγραμμα που προκύπτει εφαρμόζοντας ιεραρχική ομαδοποίηση εγγράφων τύπου: (α) SingleLink και (β) CompleteLink.

Α										
Β	1									
Γ	2	1								
Δ	3	2	1							
Ε	4	3	2	1						
Ζ	2	3	4	5	6					
Η	3	2	3	4	5	1				
Θ	4	3	2	3	4	2	1			
Ι	5	4	3	2	3	3	2	1		
Κ	6	5	4	3	2	4	3	2	1	
	Α	Β	Γ	Δ	Ε	Ζ	Η	Θ	Ι	Κ

Λύση

Τα βήματα για τον υπολογισμό του δενδρικού διαγράμματος που προκύπτει εφαρμόζοντας ιεραρχική ομαδοποίηση εγγράφων τύπου *SingleLink* είναι :

1. Βάζουμε κάθε έγγραφο σε ένα ξεχωριστό Cluster
2. Υπολογίζουμε την ομοιότητα μεταξύ όλων των ζευγαριών Cluster (έχει ήδη γίνει)
   a. $SIM(c,c') = \max\{sim(d,d') \mid d \in c, d' \in c'$
3. Βρίσκουμε το ζεύγος $\{C_u , C_v\}$ με την υψηλότερη (inter-cluster) ομοιότητα
4. Συγχωνεύουμε τα C_u , C_v σε ένα Cluster
5. Εάν έχει μείνει μόνο ένα cluster τελείωσε αλλιώς πήγαινε στο 2.

Στον πίνακα που έχουμε παρακάτω έχουμε την απόσταση μεταξύ των εγγράφων. Όταν λοιπόν αναζητάμε τα 2 πιο όμοια clusters αναζητάμε ουσιαστικά τα 2 clusters με την μικρότερη απόσταση μεταξύ τους.

Αρχικά έχουμε τον πίνακα

Α										
Β	1									
Γ	2	1								
Δ	3	2	1							
Ε	4	3	2	1						
Ζ	2	3	4	5	6					
Η	3	2	3	4	5	1				
Θ	4	3	2	3	4	2	1			
Ι	5	4	3	2	3	3	2	1		
Κ	6	5	4	3	2	4	3	2	1	
	Α	Β	Γ	Δ	Ε	Ζ	Η	Θ	Ι	Κ

Και τα Clusters
$C_A , C_B , C_\Gamma , C_\Delta , C_E , C_Z , C_H , C_\Theta , C_I , C_K , C_\Lambda$

Βρίσκουμε το ζευγάρι με την μεγαλύτερη (inter – cluster) ομοιότητα.
C_A , C_B

Συγχωνεύουμε τα $C_A C_B$
$C_{AB} = \{C_A , C_B\}$

Ξανά υπολογίζουμε την ομοιότητα μεταξύ όλων των ζευγαριών.

AB									

Γ	min{2,1}								
Δ	min{3,2}	1							
E	min{4,3}	2	1						
Z	min{2,3}	4	5	6					
H	min{3,2}	3	4	5	1				
Θ	min{4,3}	2	3	4	2	1			
I	min{5,4}	3	2	3	3	2	1		
K	min{6,5}	4	3	2	4	3	2	1	
	AB	Γ	Δ	E	Z	H	Θ	I	K

Βρίσκουμε το ζευγάρι με την μεγαλύτερη (inter – cluster) ομοιότητα.
$C_Γ$, $C_Δ$

Συγχωνεύουμε τα $C_Γ$ $C_Δ$
$C_{ΓΔ} = \{C_Γ , C_Δ\}$

Ξανά υπολογίζουμε την ομοιότητα μεταξύ όλων των ζευγαριών.

AB								
ΓΔ	min{2,1}							
E	3	min{2,1}						
Z	2	min{4,5}	6					
H	2	min{3,4}	5	1				
Θ	3	min{2,3}	4	2	1			
I	4	min{3,2}	3	3	2	1		
K	5	min{4,3}	2	4	3	2	1	
	AB	ΓΔ	E	Z	H	Θ	I	K

Βρίσκουμε το ζευγάρι με την μεγαλύτερη (inter – cluster) ομοιότητα.
C_Z , C_H

Συγχωνεύουμε τα C_Z C_H
$C_{ZH} = \{C_Z , C_H\}$

Ξανά υπολογίζουμε την ομοιότητα μεταξύ όλων των ζευγαριών.

AB							
ΓΔ	1						
E	3	1					
ZH	min{2,2}	min{4,3}	min{6,5}				
Θ	3	2	4	1			
I	4	2	3	2	1		
K	5	3	2	3	2	1	
	AB	ΓΔ	E	ZH	Θ	I	K

Βρίσκουμε το ζευγάρι με την μεγαλύτερη (inter – cluster) ομοιότητα.
C_{ZH} , C_{Θ}

Συγχωνεύουμε τα C_{ZH} C_{Θ}
$C_{Z H\Theta} = \{C_{ZH} , C_{\Theta}\}$

Ξανά υπολογίζουμε την ομοιότητα μεταξύ όλων των ζευγαριών.

ΑΒ						
ΓΔ	1					
Ε	3	1				
ΖΗΘ	min{2,3}	min{3,2}	min{5,4}			
Ι	4	2	3	min{2,1}		
Κ	5	3	2	min{3,2}	1	
	ΑΒ	ΓΔ	Ε	ΖΗΘ	Ι	Κ

Βρίσκουμε το ζευγάρι με την μεγαλύτερη (inter – cluster) ομοιότητα.
C_{I} , C_{K}

Συγχωνεύουμε τα C_{I} C_{K}
$C_{IK} = \{C_{I} , C_{K}\}$

Ξανά υπολογίζουμε την ομοιότητα μεταξύ όλων των ζευγαριών.

ΑΒ					
ΓΔ	1				
Ε	3	1			
ΖΗΘ	2	2	4		
ΙΚ	min{4,5}	min{2,3}	min{3,2}	min{1,2}	
	ΑΒ	ΓΔ	Ε	ΖΗΘ	ΙΚ

Βρίσκουμε το ζευγάρι με την μεγαλύτερη (inter – cluster) ομοιότητα.
C_{AB} , $C_{\Gamma\Delta}$

Συγχωνεύουμε τα C_{AB} $C_{\Gamma\Delta}$
$C_{AB\Gamma\Delta} = \{C_{AB} , C_{\Gamma\Delta}\}$

Ξανά υπολογίζουμε την ομοιότητα μεταξύ όλων των ζευγαριών.

ΑΒΓΔ				
Ε	min{3,1}			
ΖΗΘ	min{2,2}	4		
ΙΚ	min{4,2}	2	1	
	ΑΒΓΔ	Ε	ΖΗΘ	ΙΚ

Βρίσκουμε το ζευγάρι με την μεγαλύτερη (inter – cluster) ομοιότητα.
$C_{ZH\Theta}$, C_{IK}

Συγχωνεύουμε τα $C_{ZH\Theta}$ C_{IK}
$C_{ZH\Theta IK} = \{C_{ZH\Theta} , C_{IK}\}$

Ξανά υπολογίζουμε την ομοιότητα μεταξύ όλων των ζευγαριών.

ΑΒΓΔ			
Ε	1		
ΖΗΘΙΚ	min{2,2}	min{4,2}	
	ΑΒΓΔ	Ε	ΖΗΘΙΚ

Βρίσκουμε το ζευγάρι με την μεγαλύτερη (inter – cluster) ομοιότητα.
$C_{AB\Gamma\Delta}$, C_E

Συγχωνεύουμε τα $C_{AB\Gamma\Delta}$ C_E
$C_{AB\Gamma\Delta E} = \{C_{AB\Gamma\Delta} , C_E\}$

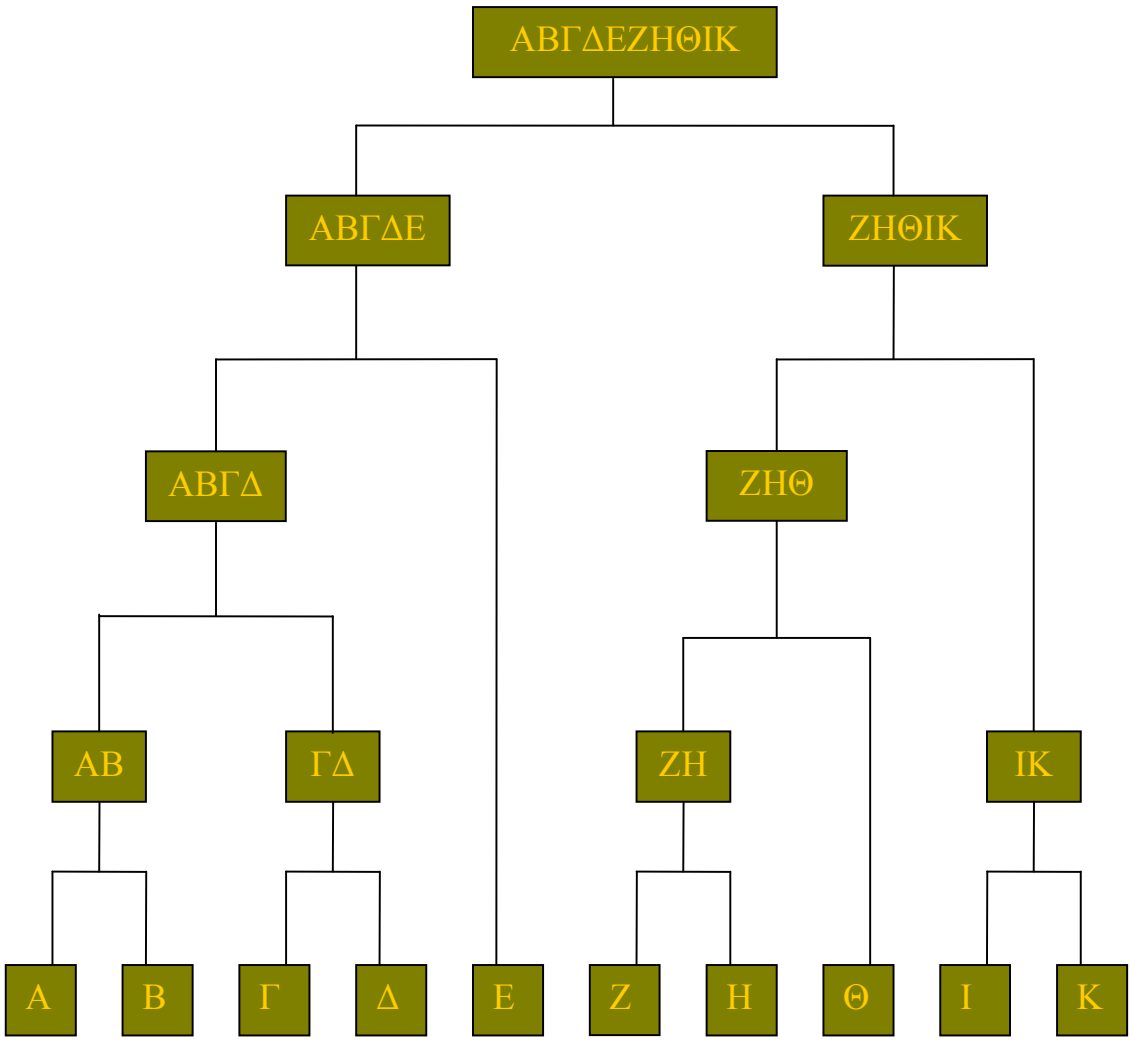
Ξανά υπολογίζουμε την ομοιότητα μεταξύ όλων των ζευγαριών.

ΑΒΓΔΕ		
ΖΗΘΙΚ	min{2,2}	
	ΑΒΓΔΕ	ΖΗΘΙΚ

Και τέλος τα 2 εναπομείναντα clusters συγχωνεύονται σε ένα

$C_{AB\Gamma\Delta EZH\Theta IK} = \{C_{AB\Gamma\Delta E} , C_{ZH\Theta IK}\}$

Τώρα που έχουμε μείνει με ένα μόνο cluster μπορούμε να σχεδιάσουμε το δενδρικό διάγραμμα που προκύπτει.


ΑΒΓΔΕΖΗΘΙΚ
ΑΒΓΔΕ
ΖΗΘΙΚ
ΑΒΓΔ
ΖΗΘ
ΑΒ
ΓΔ
ΖΗ
ΙΚ
Α
Β
Γ
Δ
Ε
Ζ
Η
Θ
Ι
Κ

(β)

Τα βήματα για τον υπολογισμό του δενδρικού διαγράμματος που προκύπτει εφαρμόζοντας ιεραρχική ομαδοποίηση εγγράφων τύπου *CompleteLink* είναι :

1. Βάζουμε κάθε έγγραφο σε ένα ξεχωριστό Cluster
2. Υπολογίζουμε την ομοιότητα μεταξύ όλων των ζευγαριών Cluster (έχει ήδη γίνει)
   a. $SIM(c,c') = \min\{sim(d,d') \mid d \in c, d' \in c'$
3. Βρίσκουμε το ζεύγος $\{C_u , C_v\}$ με την υψηλότερη (inter-cluster) ομοιότητα
4. Συγχωνεύουμε τα C_u , C_v σε ένα Cluster
5. Εάν έχει μείνει μόνο ένα cluster τελείωσε αλλιώς πήγαινε στο 2.

Στον πίνακα που έχουμε παρακάτω έχουμε την απόσταση μεταξύ των εγγράφων. Όταν λοιπόν αναζητάμε τα 2 λιγότερο όμοια clusters αναζητάμε ουσιαστικά τα 2 clusters με την μεγαλύτερη απόσταση μεταξύ τους.

Αρχικά έχουμε τον πίνακα

Α										
Β	1									
Γ	2	1								
Δ	3	2	1							
Ε	4	3	2	1						
Ζ	2	3	4	5	6					
Η	3	2	3	4	5	1				
Θ	4	3	2	3	4	2	1			
Ι	5	4	3	2	3	3	2	1		
Κ	6	5	4	3	2	4	3	2	1	
	Α	Β	Γ	Δ	Ε	Ζ	Η	Θ	Ι	Κ

Και τα Clusters
$C_A , C_B , C_Γ , C_Δ , C_E , C_Z , C_H , C_Θ , C_I , C_K , C_Λ$

Βρίσκουμε το ζευγάρι με την μεγαλύτερη (inter – cluster) ομοιότητα.
C_A , C_B

Συγχωνεύουμε τα C_A C_B
$C_{AB} = \{C_A , C_B\}$

Ξανά υπολογίζουμε την ομοιότητα μεταξύ όλων των ζευγαριών.

AB									
Γ	max{2,1}								
Δ	max{3,2}	1							
E	max{4,3}	2	1						
Z	max{2,3}	4	5	6					
H	max{3,2}	3	4	5	1				
Θ	max{4,3}	2	3	4	2	1			
I	max{5,4}	3	2	3	3	2	1		
K	max{6,5}	4	3	2	4	3	2	1	
	AB	Γ	Δ	E	Z	H	Θ	I	K

Βρίσκουμε το ζευγάρι με την μεγαλύτερη (inter – cluster) ομοιότητα.
$C_Γ$, $C_Δ$

Συγχωνεύουμε τα $C_Γ$ $C_Δ$
$C_{ΓΔ} = \{C_Γ , C_Δ\}$

Ξανά υπολογίζουμε την ομοιότητα μεταξύ όλων των ζευγαριών.

AB								
ΓΔ	max{2,3}							
E	4	max{2,1}						
Z	3	max{4,5}	6					
H	3	max{3,4}	5	1				
Θ	4	max{2,3}	4	2	1			
I	5	max{3,2}	3	3	2	1		
K	6	max{4,3}	2	4	3	2	1	
	AB	ΓΔ	E	Z	H	Θ	I	K

Βρίσκουμε το ζευγάρι με την μεγαλύτερη (inter – cluster) ομοιότητα.
C_Z , C_H

Συγχωνεύουμε τα C_Z C_H
$C_{ZH} = \{C_Z , C_H\}$

Ξανά υπολογίζουμε την ομοιότητα μεταξύ όλων των ζευγαριών.

ΑΒ							
ΓΔ	3						
Ε	2	2					
ΖΗ	max{3,3}	max{5,4}	max{6,5}				
Θ	4	3	4	max{2,1}			
Ι	5	3	3	max{3,2}	1		
Κ	6	4	2	max{4,3}	2	1	
	ΑΒ	ΓΔ	Ε	ΖΗ	Θ	Ι	Κ

Βρίσκουμε το ζευγάρι με την μεγαλύτερη (inter – cluster) ομοιότητα.
C_{Θ} , C_{I}

Συγχωνεύουμε τα C_{Θ} C_{I}
$C_{\Theta I} = \{C_{\Theta} , C_{I}\}$

Ξανά υπολογίζουμε την ομοιότητα μεταξύ όλων των ζευγαριών.

ΑΒ						
ΓΔ	3					
Ε	2	2				
ΖΗ	3	5	6			
ΘΙ	max{4,5}	max{3,3}	max{4,3}	max{2,3}		
Κ	6	4	2	4	max{2,1}	
	ΑΒ	ΓΔ	Ε	ΖΗ	ΘΙ	Κ

Βρίσκουμε το ζευγάρι με την μεγαλύτερη (inter – cluster) ομοιότητα.
C_{AB} , C_{E}

Συγχωνεύουμε τα C_{AB} C_{E}
$C_{ABE} = \{C_{AB} , C_{E}\}$

Ξανά υπολογίζουμε την ομοιότητα μεταξύ όλων των ζευγαριών.

ΑΒΕ					
ΓΔ	max{3,2}				
ΖΗ	max{3,6}	5			
ΘΙ	max{5,4}	3	3		
Κ	max{6,2}	4	4	2	
	ΑΒΕ	ΓΔ	ΖΗ	ΘΙ	Κ

Βρίσκουμε το ζευγάρι με την μεγαλύτερη (inter – cluster) ομοιότητα.
C_{K} , $C_{\Theta I}$

Συγχωνεύουμε τα C_{K} $C_{\Theta I}$
$C_{\Theta IK} = \{C_{K} , C_{\Theta I}\}$

Ξανά υπολογίζουμε την ομοιότητα μεταξύ όλων των ζευγαριών.

ΑΒΕ				
ΓΔ	3			
ΖΗ	6	5		
ΘΙΚ	max{5,6}	max{3,4}	max{3,4}	
	ΑΒΕ	ΓΔ	ΖΗ	ΘΙΚ

Βρίσκουμε το ζευγάρι με την μεγαλύτερη (inter – cluster) ομοιότητα.
$C_{ΓΔ}$, $C_{ΑΒΕ}$

Συγχωνεύουμε τα $C_{ΓΔ}$ $C_{ΑΒΕ}$
$C_{ΑΒΓΔΕ} = \{C_{ΓΔ} , C_{ΑΒΕ}\}$

Ξανά υπολογίζουμε την ομοιότητα μεταξύ όλων των ζευγαριών.

ΑΒΓΔΕ			
ΖΗ	max{6,5}		
ΘΙΚ	max{6,4}	4	
	ΑΒΓΔΕ	ΖΗ	ΘΙΚ

Βρίσκουμε το ζευγάρι με την μεγαλύτερη (inter – cluster) ομοιότητα.
$C_{ΘΙΚ}$, $C_{ΖΗ}$

Συγχωνεύουμε τα $C_{ΘΙΚ}$ $C_{ΖΗ}$
$C_{ΖΗΘΙΚ} = \{C_{ΘΙΚ} , C_{ΖΗ}\}$

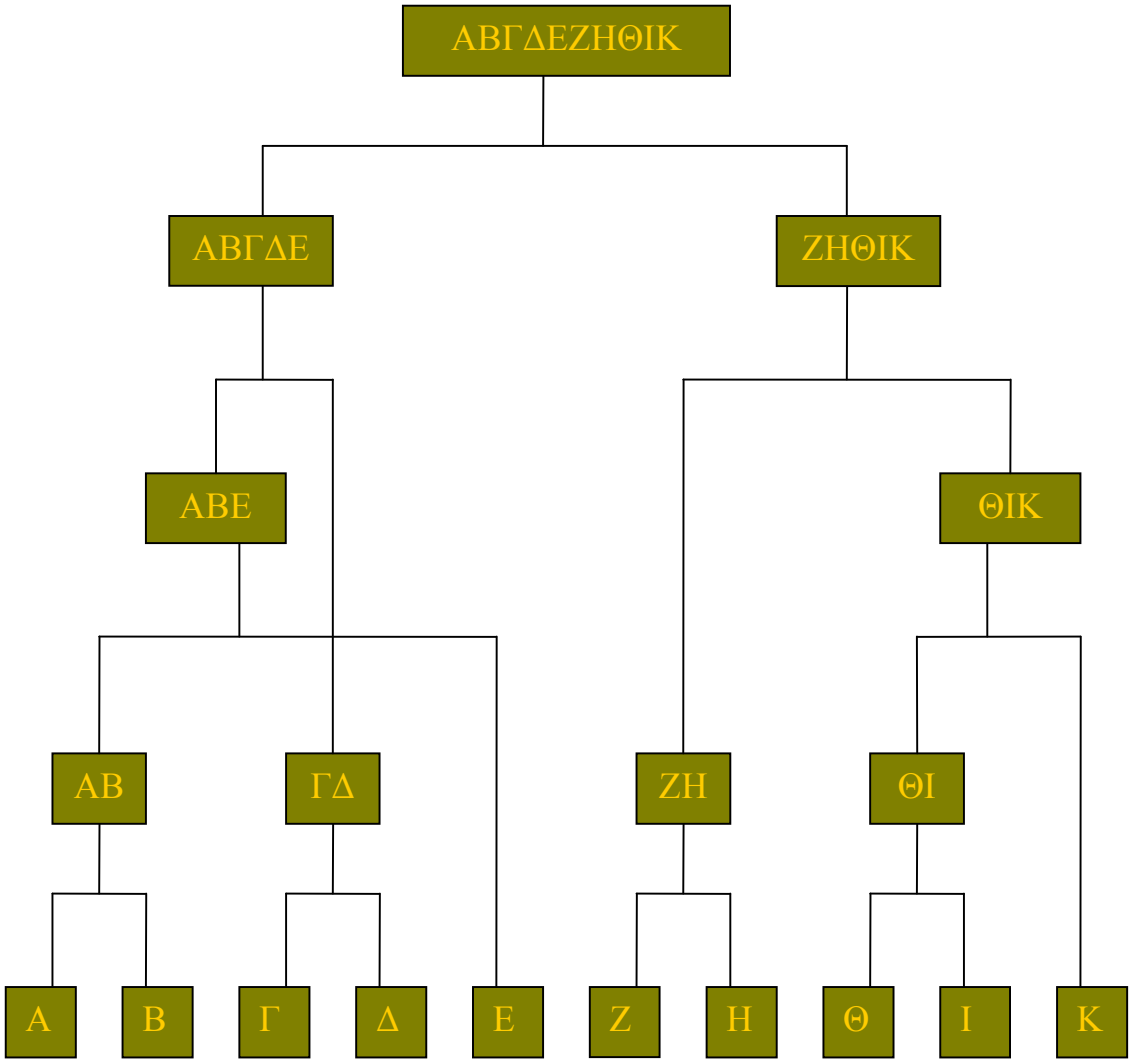
Ξανά υπολογίζουμε την ομοιότητα μεταξύ όλων των ζευγαριών.

ΑΒΓΔΕ		
ΖΗΘΙΚ	max{6,6}	
	ΑΒΓΔΕ	ΖΗΘΙΚ

Και τέλος ομαδοποιούμε τα 2 εναπομείναντα clusters.

$C_{ΑΒΓΔΕΖΗΘΙΚ} = \{C_{ΑΒΓΔΕ} , C_{ΖΗΘΙΚ}\}$

Τώρα που έχουμε μείνει με ένα μόνο cluster μπορούμε να σχεδιάσουμε το δενδρικό διάγραμμα που προκύπτει.


ΑΒΓΔΕΖΗΘΙΚ
ΑΒΓΔΕ
ΖΗΘΙΚ
ΑΒΕ
ΘΙΚ
ΑΒ
ΓΔ
ΖΗ
ΘΙ
Α
Β
Γ
Δ
Ε
Ζ
Η
Θ
Ι
Κ

Άσκηση 3

Θεωρείστε τα ακόλουθα έγγραφα όπου τα γράμματα Α-Ε συμβολίζουν λέξεις.
d1 = «Β Γ Β », d2 = «Β Α Α Β»
d3 = «Α Β », d4 = «Γ Ε Γ Ε»
d5 = «Δ Γ Γ Α», d6 = «Γ Ε»
d7 = «Β Δ Β», d8 = «Ε Β Α»
Έστω ότι τα d1,d5, d6 ανήκουν σε ένα σύστημα S1, τα d2,d4 σε ένα σύστημα S2, και τα υπόλοιπα (d3,d7,d8) σε ένα σύστημα S3. Θέλουμε να φτιάξουμε έναν μεσίτη Μ πάνω από αυτά τα συστήματα.
(α) Για την επιλογή πηγής ο Μ θέλει να περιγράψει τα περιεχόμενα της κάθε πηγής με ένα διάνυσμα. Δώστε τα διανύσματα πηγών των S1, S2 και S3.

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

(γ) Ο Μ λαμβάνει μια επερώτηση, την προωθεί σε όλες τις πηγές, και λαμβάνει τα εξής αποτελέσματα από την κάθε μια:
S1: <d5, d1, d6 >
S2: <d2, d4>
S3: <d7, d8, d3>
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 δυο φάσεις ώστε οι πηγές να λαμβάνουν τα καθολικά στατιστικά που χρειάζονται για τον 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 αυτήν του S1, δηλαδή και αυτό παρέχει πρόσβαση στα έγγραφα d1, d5, d6. Έστω ότι ο Μ προωθεί μια επερώτηση q στα S1 και S4 και λαμβάνει τις εξής απαντήσεις:
S1: <d1, d5, d6>
S4: <d6, d5, d1>
Ποιο είναι το κορυφαίο έγγραφο αν ενοποιήσουμε τις διατάξεις: (i) κατά Borda, (ii) κατά Condorcet;
Ο Μ 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 μια γεύση για το βαθμό συμφωνίας των πηγών). Ποια είναι αυτή η απόσταση στην προκειμένη;

(στ) Τα συστήματα S1, S2, S3 δεν θέλουν πλέον να έχουν ανάγκη τον Μ και αποφασίζουν να «ανεξαρτητοποιηθούν» φτιάχνοντας ένα σύστημα ομοτίμων (P2P), συγκεκριμένα ένα δομημένο σύστημα τύπου Chord. Προσελκύουν μάλιστα άλλα δυο συστήματα S5 και S6 (τα οποία δεν έχουν καμία συλλογή εγγράφων).
Αποφασίζουν να χρησιμοποιήσουν μια συνάρτηση κατακερματισμού h των 3 bits, και έστω ότι
h(IPaddress(S1))=2, h(IPaddress(S2))=5, h(IPaddress(S3))=3,
h(IPaddress(S5))=1, h(IPaddress(S6))=4
Αποφασίζουν να διανείμουν το ανεστραμμένο ευρετήριο θεωρώντας κάθε όρο σαν κλειδί και έστω ότι
h(Α)=2, h(Β)=3, h(Γ)=6

h(Δ)=6, h(E)=5
Δώστε (i) τους πίνακες δρομολόγησης των κόμβων S1 και S3 και (ii) πως θα κατανεμηθεί το ανεστραμμένο ευρετήριο στους κόμβους του δικτύου (δείξτε τι ακριβώς θα έχει κάθε κόμβος)

Λύση

(α)

Το λεξιλόγιο μας αποτελείται από τις λέξεις Α, Β , Γ , Δ , Ε

Θα θεωρήσουμε τις 3 πηγές σαν ένα έγγραφο το οποίο περιέχει τις λέξεις που περιέχουν όλα τα έγγραφα της πηγής .

Έτσι για τις 3 πηγές θα έχουμε:

	Α	**Β**	**Γ**	**Δ**	**Ε**
S1	1	2	4	1	1
S2	2	2	2	0	2
S3	2	4	0	1	1

Για να δώσουμε τα διανύσματα που θα αναπαριστούν τις πηγές πρέπει να κάνουμε βάρυνση με TF-IDF.

	Α	**Β**	**Γ**	**Δ**	**Ε**	**maxk{freqkj}**
S1	1	2	4	1	1	4
S2	2	2	2	0	2	2
S3	2	4	0	1	1	4
DF	3	3	2	2	3	
$\frac{N}{DF}$	$\frac{3}{3}=1$	$\frac{3}{3}=1$	$\frac{3}{2}=1.5$	$\frac{3}{2}=1.5$	$\frac{3}{3}=1$	

Κατόπιν πρέπει να υπολογίσουμε τα βάρη TF-IDF. Γνωρίζουμε ότι
$w_{ij} = tf_{ij} idf_i = \log_2\left(\frac{N}{df_i}\right)$. Για ευκολία στους υπολογισμούς θεωρούμε $IDF = \frac{N}{DF}$

	A	**B**	**Γ**	**Δ**	**E**	**maxk{freqkj}**
S1	$\frac{1}{4}\bullet 1 = 0.25$	$\frac{2}{4}\bullet 1 = 0.5$	$\frac{4}{4}\bullet 1.5 = 1.5$	$\frac{1}{4}\bullet 1.5 = 0.375$	$\frac{1}{4}\bullet 1 = 0.25$	4
S2	$\frac{2}{2}\bullet 1 = 1$	$\frac{2}{2}\bullet 1 = 1$	$\frac{2}{2}\bullet 1.5 = 1.5$	0	$\frac{2}{2}\bullet 1 = 1$	2
S3	$\frac{2}{4}\bullet 1 = 0.5$	$\frac{4}{4}\bullet 1 = 1$	0	$\frac{1}{4}\bullet 1.5 = 0.375$	$\frac{1}{4}\bullet 1 = 0.25$	4
DF	3	3	2	2	3	
$\frac{N}{DF}$	$\frac{3}{3} = 1$	$\frac{3}{3} = 1$	$\frac{3}{2} = 1.5$	$\frac{3}{2} = 1.5$	$\frac{3}{3} = 1$	

Έτσι τα διανύσματα των ε πηγών είναι :

S1: < 0.25 , 0.5 , 1.5 , 0.375 , 0.25 >
S2: < 1 , 1 , 1.5 , 0 , 1 >
S3: < 0.5 , 1 , 0 , 0.375 , 0.25 >

(β)

Δεδομένου ότι για να υπολογίσουμε τα διανύσματα των πηγών θεωρήσαμε τις πηγές ως έγγραφα τα οποία περιέχουν τις λέξεις που υπάρχουν στα περιεχόμενα έγγραφα της κάθε πηγής. Έτσι λοιπόν αρκεί να συγκρίνουμε το διάνυσμα της επερώτησης με τα αντίστοιχα διανύσματα των πηγών. Σε αυτό θα μας βοηθήσει το μέτρο ομοιότητας συνημίτονου.

	A	**B**	**Γ**	**Δ**	**E**	**maxk{freqkj}**
Q	1	0	1	0	0	1

Και επομένως υπολογίζοντας τα TF-IDF βάρη έχουμε :

	A	**B**	**Γ**	**Δ**	**E**	**maxk{freqkj}**
Q	$\frac{1}{1}\bullet 1 = 1$	0	$\frac{1}{1}\bullet 1.5 = 1.5$	0	0	1

Το διάνυσμα της επερώτησης λοιπόν είναι:

S1: < 1 , 0 , 1.5 , 0 , 0 >

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

Πηγή 1:

$$CosSim(S_1, Q) = \frac{0.25 \bullet 1 + 0.5 \bullet 0 + 1.5 \bullet 1.5 + 0.375 \bullet 0 + 0.25 \bullet 0}{\sqrt{(0.0625 + 0.25 + 2.25 + 0.140625 + 0.0625) \bullet (1 + 0 + 2.25 + 0 + 0)}} = \frac{2.5}{\sqrt{8.98828125}} \approx 0.833$$

Πηγή 2:

$$CosSim(S_2, Q) = \frac{1 \bullet 1 + 1 \bullet 0 + 1.5 \bullet 1.5 + 0 \bullet 0 + 1 \bullet 0}{\sqrt{(1 + 1 + 2.25 + 0 + 1) \bullet (1 + 0 + 2.25 + 0 + 0)}} = \frac{3.25}{\sqrt{17.0625}} \approx 0.787$$

Πηγή 3:

$$CosSim(S_3, Q) = \frac{0.5 \bullet 1 + 1 \bullet 0 + 0 \bullet 1.5 + 0.375 \bullet 0 + 0.25 \bullet 0}{\sqrt{(0.25 + 1 + 0 + 0.140625 + 0.0625) \bullet (1 + 0 + 2.25 + 0 + 0)}} = \frac{0.5}{\sqrt{4.722656}} \approx 0.23$$

Επομένως η πηγή την οποία θα στείλει την επερώτηση είναι η S_2.

(γ)

Η ενοποιημένη διάταξη κατά Robin Round Interleaving είναι

ANS(q) = <{d_5, d_2, d_7}, {d_1, d_4, d_8}, {d_6, d_3}>

(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.

Στην $2^{η}$ φάση τα στατιστικά που θα στείλει ο Μ είναι (αγνοώντας για απλότητα τον υπολογισμό του λογαρίθμου)

$$\left\{ A : \frac{8}{4}, B : \frac{8}{5}, \Gamma : \frac{8}{4}, \Delta : \frac{8}{2}, E : \frac{8}{3} \right\}$$

(ε)

S_1 : <d_1, d_5, d_6>
S_4 : <d_6, d_5, d_1>

Σύμφωνα με το *Borda* νικητής στην διάταξη αναδεικνύεται αυτός που έχει καλύτερο άθροισμα βαθμολογίας στην διάταξη. Έτσι κατά Borda έχουμε:

d_1:1+3=4
d_5:2+2=4

d_6:3+1=4

Όλα τα έγγραφα έχουν ίδιες βαθμολογίες οπότε δεν μπορούμε να αποφανθούμε ποιο είναι το κορυφαίο έγγραφο.

Σύμφωνα με τον *Condorset* νικητής στην διάταξη είναι αυτός που έχει περισσότερες νίκες έναντι του αντιπάλου του. Έτσι κατά Condorset έχουμε:

d_1:d_5 1:1
d_1:d_6 1:1
d_5:d_6 1:1

Άρα παρατηρούμε ότι και με τον Condorset δεν μπορεί να υπάρξει νικητής.

Η *Kemeny Distance* μας δίνει το πλήθος των διαφωνιών στην διάταξη των ζευγαριών.
Η απόσταση είναι 3 αφού :

- $d_1 >_{S_1} d_5, d_1 <_{S_4} d_5$
- $d_1 >_{S_1} d_6, d_1 <_{S_4} d_6$
- $d_5 >_{S_1} d_6, d_5 <_{S_4} d_6$

(στ)

(-i-)

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 $n+2^{i-1}$ δηλαδή

finger[i] = successor($n+2^{i-1}$)

οπότε για τον S_1 γνωρίζουμε ότι είναι ο κόμβος με $h(S_1)=2$ και άρα θα έχει πίνακα δρομολόγησης :

finger[1]=successor($2+2^{1-1}$)=successor(3)=3 ⇨ S_3
finger[2]=successor($2+2^{2-1}$)=successor(4)=4 ⇨ S_6
finger[3]=successor($2+2^{3-1}$)=successor(6)=1 ⇨ S_5

όμοια για τον S_3 :

finger[1]=successor($3+2^{1-1}$)=successor(4)=4 ⇨ S_6
finger[2]=successor($3+2^{2-1}$)=successor(5)=5 ⇨ S_2
finger[3]=successor($3+2^{3-1}$)=successor(7)=2 ⇨ S_1

-ii-

Γνωρίζουμε ότι το ευρετήριο κατανέμεται βάσει των όρων του.
Πρώτα να λοιπόν πρέπει να βρούμε το ανεστραμμένο αρχείο βάσει των εμφανίσεων των λέξεων Α, Β, Γ, Δ, Ε στα έγγραφα $d_1, \ldots, d_8$

A: $<d_2,2>,<d_3,1>,<d_5,1>,<d_8,1>$
B: $<d_1,2>,<d_2,2>,<d_3,1>,<d_7,2>,<d_8,1>$
Γ: $<d_1,1>,<d_4,2>,<d_5,2>,<d_6,1>$
Δ: $<d_5,1>,<d_7,1>$
E: $<d_4,2>,<d_6,1>,<d_8,1>$

Επίσης γνωρίζουμε ότι ένα κλειδί **k** εκχωρείται στον πρώτο κόμβο **p** έτσι ώστε

$$h(p) \geq h(k)$$

Επομένως :

- Το κλειδί Α εκχωρείται στον κόμβο S_1 αφού $h(S_1) \geq h(A)$
- Το κλειδί Β εκχωρείται στον κόμβο S_3 αφού $h(S_3) \geq h(B)$
- Το κλειδί Γ εκχωρείται στον κόμβο S_5 αφού $h(S_5) \geq h(\Gamma)$
- Το κλειδί Δ εκχωρείται στον κόμβο S_5 αφού $h(S_5) \geq h(\Delta)$
- Το κλειδί Ε εκχωρείται στον κόμβο S_2 αφού $h(S_2) \geq h(E)$

Έτσι το ανεστραμμένο ευρετήριο θα κατανεμηθεί στους κόμβους ως εξής:

S_1: key(A) $<d_2,2>,<d_3,1>,<d_5,1>,<d_8,1>$
S_3: key(B) $<d_1,2>,<d_2,2>,<d_3,1>,<d_7,2>,<d_8,1>$
S_5: key(Γ) $<d_1,1>,<d_4,2>,<d_5,2>,<d_6,1>$
S_5: key(Δ) $<d_5,1>,<d_7,1>$
S_2: key(E) $<d_4,2>,<d_6,1>,<d_8,1>$

Άσκηση 4

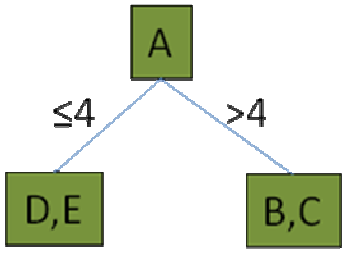
Έστω ότι έχουμε 5 εικόνες A, B, C, D, E των οποίων οι αποστάσεις φαίνονται στον παρακάτω πίνακα. Προκειμένου να μπορούμε να απαντήσουμε επερωτήσεις γρήγορα θέλουμε να φτιάξουμε ένα μετρικό ευρετήριο, συγκεκριμένα ένα Vantage-Point-Tree (VTP).
Σχεδιάστε το VTP που προκύπτει:
α) αν επιλέξουμε την εικόνα A ως κεντρική (pivot),
β) αν επιλέξουμε την εικόνα C ως κεντρική (pivot).

A					
B	6				
C	4.2	5			
D	4	2	5		
E	2	8	5	6	
	A	**B**	**C**	**D**	**E**

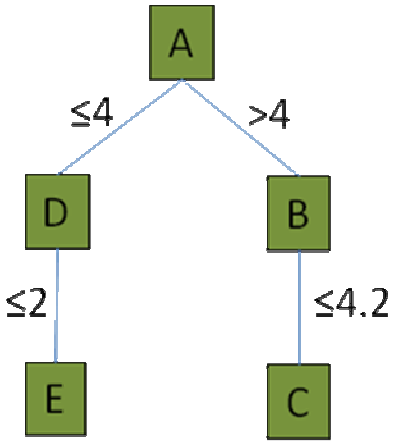
Λύση

(α)
Επιλέγοντας την εικόνα A σαν pivot έχουμε
Median(*d(A,B)* , *d(A,C)* , *d(A,D)* , *d(A,E)*) = Median(*2*, *4* , *4.2* , *6*) =4

Επομένως στο ρίζα με δένδρο το pivot to A θα τοποθετήσουμε αριστερά τις εικόνες με απόσταση μικρότερη ή ίση από την median απόσταση επομένως.

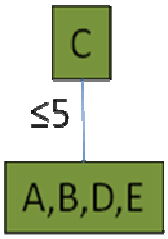


Αναδρομικά υπολογίζουμε τους median για το αριστερό και το δεξιό υποδένδρο αντίστοιχα. Έτσι το τελικό δένδρο που προκύπτει είναι το εξής.



(β)
Επιλέγοντας την εικόνα C σαν pivot έχουμε
Median(*d(C,A)* , *d(C,B)* , *d(C,D)* , *d(C,E)*) = Median(*4.2*, *5* , *5* , *5*) = 5

Επομένως στο ρίζα με δένδρο το pivot to A θα τοποθετήσουμε αριστερά τις εικόνες με απόσταση μικρότερη ή ίση από την median απόσταση επομένως.



Αναδρομικά υπολογίζουμε τους median για το υποδένδρο. Έτσι το τελικό δένδρο που προκύπτει είναι το εξής.