Πανεπιστήμιο Κρήτης, Τμήμα Επιστήμης Υπολογιστών
HY463 - Συστήματα Ανάκτησης Πληροφοριών
2007 – 2008 Εαρινό Εξάμηνο

# Φροντιστήριο 3

## Retrieval Models

**Άσκηση 1**

Θεωρείστε μια συλλογή κειμένων που περιέχει τα ακόλουθα 5 έγγραφα:

Έγγραφο 1: «Computer Games»
Έγγραφο 2: «Computer Games Computer Games»
Έγγραφο 3: « Games Theory and Computer »
Έγγραφο 4: «Computer for Computer »
Έγγραφο 5: «Cheap Games Computer Games»

1) Δώστε τη διανυσματική παράσταση του κάθε εγγράφου με βάρη TF-IDF. Θεωρείστε ότι η θέση της κάθε λέξης στα διανύσματα γίνεται κατά αλφαβητική σειρά.

2) Θεωρείστε την επερώτηση q1=«Computer Games». Υπολογίστε το TF-IDF διάνυσμα αυτής της επερώτησης και δώστε την διάταξη των εγγράφων που θα επιστρέψει ένα σύστημα που βασίζεται στο διανυσματικό μοντέλο.

Σχεδιάστε το ανεστραμμένο ευρετήριο για αυτή τη συλλογή.

**Λύση**

1)

Έγγραφο 1: «Computer Games»
Έγγραφο 2: «Computer Games Computer Games»
Έγγραφο 3: « Games Theory and Computer »
Έγγραφο 4: «Computer for Computer »
Έγγραφο 5: «Cheap Games Computer Games»

| | And | Cheap | Computer | for | Games | Theory | MAXk{FREQij} |
|---|---|---|---|---|---|---|---|
| $D_1$ | 0 | 0 | 1 | 0 | 1 | 0 | 1 |
| $D_2$ | 0 | 0 | 2 | 0 | 2 | 0 | 2 |
| $D_3$ | 1 | 0 | 1 | 0 | 1 | 1 | 1 |
| $D_4$ | 0 | 0 | 2 | 1 | 0 | 0 | 2 |
| $D_5$ | 0 | 1 | 1 | 0 | 2 | 0 | 2 |
| DF | 1 | 1 | 5 | 1 | 4 | 1 | |
| IDF | 5/1 | 5/1 | 5/5 | 5/1 | 5/4 | 5/1 | |

- $FREQ_{ij}$=το πλήθος των εμφανίσεων του όρου i στο έγγραφο j
- IDF = N / DF
- $MAX_k\{FREQ_{ij}\}$ = συχνότητα της λέξης με τη μέγιστη συχνότητα στο κείμενο

| TF-IDF | | | | | | | |
|---|---|---|---|---|---|---|---|
| | And | Cheap | Computer | for | Games | Theory | MAXk{FREQij} |
| $D_1$ | 0 | 0 | 1/1*5/5 | 0 | 1/1*5/4 | 0 | 1 |
| $D_2$ | 0 | 0 | 2/2*5/5 | 0 | 2/2*5/4 | 0 | 2 |
| $D_3$ | 1/1*5/1 | 0 | 1/1*5/5 | 0 | 1/1*5/4 | 1/1*5/1 | 1 |
| $D_4$ | 0 | 0 | 2/2*5/5 | 1/2*5/1 | 0 | 0 | 2 |
| $D_5$ | 0 | 1/2*5/1 | 1/2*5/5 | 0 | 2/2*5/4 | 0 | 2 |
| DF | 1 | 1 | 5 | 1 | 4 | 1 | |
| IDF | 5/1 | 5/1 | 5/5 | 5/1 | 5/4 | 5/1 | |

- $TF_{ij} = FREQ_{ij} / MAX_k\{FREQ_{ij}\}$
- $V_{ij} = TF_{ij} * IDF_i$

Οι διανυσματικές παραστάσεις των κειμένων είναι :
$V_1 = \{0, 0, 1, 0, 1.25, 0\}$, $|V_1| = 2,5625$
$V_2 = \{0, 0, 1, 0, 1.25, 0\}$, $|V_2| = 2,5625$
$V_3 = \{5, 0, 1, 0, 1.25, 5\}$, $|V_3| = 52,5625$
$V_4 = \{0, 0, 1, 2.5, 0, 0\}$, $|V_4| = 7,25$
$V_5 = \{0, 2.5, 0.5, 0, 1.25, 0\}$, $|V_5| = 8,0625$

2)

| | And | Cheap | Computer | For | Games | Theory |
|---|---|---|---|---|---|---|
| q1=Computer Games | 0 | 0 | 1/1*5/5 | 0 | 1/1*5/4 | 0 |
| IDF | 5/1 | 5/1 | 5/5 | 5/1 | 5/4 | 5/1 |

$q_1 = \{0, 0, 1, 0, 1.25, 0\}$, $|q_1| = 2,5625$
$V_1 * q_1 = 1*1+1,25*1,25=1+1,5625=2,5625$
$V_2 * q_1 = 2,5625$
$V_3 * q_1 = 5*0+1*1+1,25*1,25+5*0=2,5625$
$V_4 * q_1 = 1$
$V_5 * q_1 = 0,5*1+1,25*1,25=2,0625$

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

Με βάση τον παρπάνω τύπο, υπολογίζουμε το μέτρο ομοιότητας συνημίτονου για κάθε έγγραφο Dj

$R(D_{1,} q_1) = 2,5625/(2,5625*2,5625)^{1/2}$ => **$R(D_{1,} q_1) = 1$**

$R(D_{2,} q_1) = 2,5625/(2,5625*2,5625)^{1/2}$ => **$R(D_{2,} q_1) = 1$**

$R(D_{3,} q_1) = 2,5625/(52,5625*2,5625)^{1/2} = 2,5625/(134,69140625)^{½} = 2,5625/11,6056627$ => **$R(D_{3,} q_1) = 0,21928589$**

$R(D_{4,} q_1) = 1/(7,25*2,5625)^{1/2} = 1/(18,578125)^{1/2} = 1/4,31023491$ =>

**$R(D_{4,} q_1) = 0,23200592$**

$R(D_{5,} q_1) = 2,0625/(8,0625*2,5625)^{1/2} = 2,0625/(20,660156)^{1/2} = 2,0625/4,54534443$ => **$R(D_{5,} q_1) = 0,45376099$**

Με βάση το διανυσματικό μοντέλο η διάταξη των εγγράφων είναι :
**$< \{D_{1,} D_2\}, D_{5,} D_{4,} D_3 >$**

Θα περιμέναμε το έγγραφο $D_3$ να έρθει στη σειρά πριν το $D_4$ ,επειδή περιέχει όλους τους όρους της επερώτησης, όμως περιέχει και τον όρο Theory ο οποίος εμφανίζεται μόνο σε αυτό το έγγραφο και αυτό επηρέασε το βάρος του $D_3$.

**Ανεστραμμένο ευρετήριο**
Μία μορφή του ανεστραμμένου ευρετηρίου στο οποίο εμφανίζονται μόνο οι θέσεις των όρων είναι :

| Term | < Document Frequency, (Document; Position) > |
|---|---|
| Computer | < 5 ($D_1$;1), ($D_2$;1), ($D_2$;3), ($D_3$;4), ($D_4$;1), ($D_4$;3), ($D_5$;3)> |
| Games | < 4 ($D_1$;2), ($D_2$;2), ($D_2$;4), ($D_3$;1), ($D_5$;2), ($D_5$;4) > |
| Theory | < 1 ($D_3$;2) > |
| Cheap | < 1 ($D_5$;1) > |
| and | < 1 ($D_3$;3)> |
| for | < 1 ($D_4$;2)> |

Μία άλλη μορφή του ανεστραμμένου ευρετηρίου στο οποίο εμφανίζεται το TF του κάθε όρου σε κάθε έγγραφο είναι :

| Term | < Document : Term Frequency : { Position } > |
|---|---|
| Computer | < $D_1$ : 1 : { 1 } > |
| | < $D_2$ : 1 : { 1, 3 } > |
| | < $D_3$ : 1 : { 4 } > |
| | < $D_4$ : 1 : { 1, 3 } > |
| | < $D_5$ : 0.5 : { 3 } > |
| Games | < $D_1$ : 1.25 : { 2 } > |
| | < $D_2$ : 1.25 : { 2, 4 } > |
| | < $D_3$ : 1.25 : { 1 } > |
| | < $D_5$ : 1.25 : { 2 } > |
| Theory | < $D_3$ : 5 : { 2 } > |
| Cheap | < $D_5$ : 2.5 : { 1 } > |
| and | < $D_3$ : 5 : { 3 } > |
| for | < $D_4$ : 2.5 : { 2 } > |

**Άσκηση 2**

Έστω μια συλλογή από κείμενα D, και έστω A ένα διατεταγμένο υποσύνολο αυτής. Έστω ότι μας δίνουν το A και μας ζητούν να βρούμε αν υπάρχει επερώτηση q τ.ω. η απάντηση της να έχει στην αρχή της το διατεταγμένο σύνολο A. Για παράδειγμα, αν A=<d1, d2, d3> και βρούμε μια επερώτηση q τ.ω. Answer(q) = <d1, d2, d3, d8, …> τότε αυτή είναι μια λύση του προβλήματος μας. Θεωρώντας ότι το σύστημα σας βασίζεται στο διανυσματικό μοντέλο, απαντήστε τα παρακάτω ερωτήματα.

(α) Πως μπορούμε να βρούμε αν υπάρχει τέτοια επερώτηση;

(β) Αν υπάρχει ποια είναι;

(γ) Αν δεν υπάρχει τέτοια επερώτηση, πως θα χαλαρώνατε το πρόβλημα και τι θα μπορούσατε να επιστρέψετε; Μπορείτε να αναπτύξετε τις σκέψεις σας όσο θέλετε.

*Σημείωση*: Προσέξτε ώστε το υπολογιστικό κόστος των λύσεων που θα προτείνετε για τα (α) και (β) να μην είναι απαγορευτικό.

**Λύση**

(α) Για να υπάρχει μία τέτοια επερώτηση q θα πρέπει να ισχύουν οι δύο παρακάτω συνθήκες:

(α1) Έστω q η επερώτηση που ψάχνουμε. Για να επιστρέφει η q όλα τα έγγραφα του A και μάλιστα με την σχετική διάταξη που έχουν στο A, θα πρέπει το μέτρο ομοιότητας του συνημίτονου μεταξύ της q και του πρώτου εγγράφου στο A να είναι μεγαλύτερο από το μέτρο του δευτέρου εγγράφου και εκείνο μεγαλύτερο από του τρίτου εγγράφου κ.ο.κ., και όλα να είναι μεγαλύτερα του μηδενός

Δηλαδή αν A = < d1,d2,d3 > τότε θα έπρεπε

**Sim ($d_1$, q) > Sim ($d_2$, q) > Sim ($d_3$, q) > 0 (1)**

(α2) Για να μας επιστρέφει η q όλα τα έγγραφα που ανήκουν στο A, με την διάταξη που έχουν σε αυτό και πριν από οποιοδήποτε άλλο έγγραφο θα πρέπει όλα τα έγγραφα που μας επιστρέφει η επερώτηση q και δεν ανήκουν στο A, να έχουν μέτρο ομοιότητας του συνημίτονου μικρότερο από την τιμή του τελευταίου εγγράφου που ανήκει στο A.
Δηλαδή αν A = < d1,d2,d3 > τότε θα έπρεπε
**Sim ($d_1$, q) > Sim ($d_2$, q) > Sim ($d_3$, q) > Sim ($d_x$, q)** για κάθε $d_x$
που δεν ανήκει στο A.

(β)
Εύκολα μπορούμε να δείξουμε ότι αν ικανοποιούνται οι παραπάνω συνθήκες τότε το q=d1 είναι μια επιθυμητή απάντηση (από τις ενδεχομένως πολλές). Ο έλεγχος της συνθήκης (α1) είναι απλός και όχι ιδιαίτερα ακριβός. Συγκεκριμένα απαιτεί |A|-1 υπολογισμούς βαθμού ομοιότητας. Ένας εύκολος τρόπος για να δούμε αν ισχύει η συνθήκη (α2) είναι να υπολογίσουμε το Answer(d1) και να δούμε εάν τα πρώτα |A| στοιχεία του είναι τα στοιχεία του A.

(γ)
Μια χαλάρωση του προβλήματος είναι η εξής: Η συνθήκη (α1) ικανοποιείται αλλά δεν ικανοποιείται η συνθήκη (α2) . Και σε αυτήν την περίπτωση το q=d1 θα ήταν μια πιθανή λύση του προβλήματος. Απλά η απάντηση του q θα μπορούσε θα είχε τη μορφή:
A(q) = < **d1**, d5, d6, **d2**, d4, d8, d9, **d3**, d10, d7 >

Μια άλλη χαλάρωση του προβλήματος θα ήταν να μειώσουμε το σύνολο των εγγράφων του συνόλου A αρχίζοντας από το τέλος. Δηλαδή αντί για A = < $d_1$, $d_2$, $d_3$ > να δούμε αν υπάρχει λύση για το σύνολο **A' = < $d_1$, $d_2$ >. Αν δεν υπάρχει ούτε για το A' να δούμε αν υπάρχει για το A''=<d1>.**

Για περισσότερα δείτε τις διαφάνειες του Μαθήματος 11(2007) καθώς και το άρθρο:
http://www.ics.forth.gr/~tzitzik/publications/2007_TzitzikasTheoharisNaming.pdf

**Άσκηση 3**
Στο μάθημα είδαμε δύο μοντέλα ανάκτησης του βασίζονται στη Θεωρία Ασαφών Συνόλων. Το πρώτο θεωρεί βάρυνση TF*IDF, ενώ το δεύτερο είναι εκείνο που προτάθηκε από τους [Ogawa, Morita, Kobayashi, 1991]. Θεωρείστε έναν όρο $t_i$ ενός εγγράφου $d_j$. Συγκρίνετε την συμπεριφορά των δύο αυτών μοντέλων για διάφορες περιπτώσεις, π.χ.: για μικρές και μεγάλες τιμές του $tf_{ij}$, για μικρές και μεγάλες τιμές του $idf_i$, για μικρές και μεγάλες τιμές του $w_{ij}$ αν προκύπτει από tf*idf.

**Λύση**
Για το **πρώτο μοντέλο που βασίζεται σε βάρυνση TF-IDF** ξέρουμε ότι:
$d_j = ( w_{1,j} , w_{2,j} , \dots , w_{t,j} )$ όπου $w_{i,j} \in [0,1]$
$R(d_j, t_i) = \mu_{ti} (d_j) = w_{i,j} = tf_{ij} * idf_i$ όπου $tf_{ij} = freq_{ij} / MAX_k \{ freq_{kj} \}$, $idf_i = \log_2 ( N / df_i )$
και N ο αριθμός των εγγράφων.
Για το μοντέλο **που προτάθηκε από τους [Ogawa, Morita, and Kobayashi,1991]** ξέρουμε ότι: $d_j = ( w_{1,j} , w_{2,j} , \dots , w_{t,j} )$ όπου $w_{i,j} \in \{0,1\}$ και

$w_{i,j} = 1$ όταν ο όρος $t_i$ εμφανίζεται στο κείμενο $d_j$. (αλλιώς $w_{i,j} = 0$)

$R(d_j,t_i) = \mu_{ti}(d_j)$ το οποίο ορίζεται ως εξής:
Αρχικά ορίζεται η εγγύτητα μεταξύ των όρων με τον εξής τύπο:
**c(i,j) = n(i,j) / ni + nj – n(i,j) όπου**
n(i,j) : το πλήθος των εγγράφων που περιέχουν τον όρο $k_i$ και τον $k_l$.
ni : το πλήθος των εγγράφων που περιέχουν τον όρο $k_i$.
nj : το πλήθος των εγγράφων που περιέχουν τον όρο $k_l$.
Κατόπιν θέτουμε $\mu_i(j) = \Sigma\, c(i,w)$ , $t_w \epsilon\, d_j$
**$\mu_i(j) = 1 – \Pi(1- c(i,w))$** , $t_w \epsilon\, d_j$
Έστω ότι θέλουμε να κάνουμε μία επερώτηση σε μία σύλλογη εγγράφων και η επερώτηση αποτελείται από ένα όρο κ, δηλαδή q = " k ".

**Μικρές τιμές $tf_{ij}$**
Έστω ένα έγγραφο j. Αν το **$tf_{ij}$** του όρου **ki** είναι μικρό, αυτό σημαίνει ότι ο όρος **ki** εμφανίζεται λίγες φορές στο έγγραφο αυτό άρα η κατάταξη του εγγράφου θα είναι χαμηλή σύμφωνα με το πρώτο μοντέλο ανάκτησης. Το δεύτερο μοντέλο ([Ogawa, Morita, Kobayashi, 1991]) αγνοεί το πλήθος των εμφανίσεων ενός όρου σε ένα έγγραφο Λαμβάνει όμως υπόψη τον βαθμό συνεμφάνισης των όρων στη συλλογή. Αυτό σημαίνει ότι αν το έγγραφο j περιέχει πολλούς όρους οι οποίοι έχουν μεγάλη εγγύτητα (συνεμφάνιση) με τον όρο ki, τότε το έγγραφο αυτό μπορεί να σταθμιστεί υψηλά.

**Μεγάλες τιμές $tf_{ij}$**
Έστω ένα έγγραφο j. Αν το **$tf_{ij}$** του όρου **ki** είναι μεγάλο, αυτό σημαίνει ότι ο όρος **ki** εμφανίζεται πολλές φορές στο έγγραφο αυτό άρα η κατάταξη του εγγράφου θα είναι υψηλή σύμφωνα με το πρώτο μοντέλο ανάκτησης. Στο δεύτερο μοντέλο (που αγνοεί το tf) το έγγραφο αυτό θα μπορούσε να σταθμιστεί χαμηλά αν περιέχει λίγους όρους που έχουν εγγύτητα (συνεμφάνιση) με τον όρο ki.

**Μικρές τιμές $idf_i$**
Αν το idf του όρου **ki** είναι μικρό αυτό σημαίνει ότι το πλήθος των εγγράφων που περιέχουν τον όρο **ki** είναι μεγάλο (αφού idf = N/df).
Έστω ένα έγγραφο j που περιέχει τον όρο ki.
Αν λαμβάναμε υπόψη μόνο το idf τότε το $1^ο$ μοντέλο, θα έδινε μικρό βαθμό συνάφειας στο έγγραφο j (καθώς και σε όλα τα υπόλοιπα έγγραφα). Αντίθετα το $2^ο$ μοντέλο ενδεχομένως να έδινε μεγαλύτερο βαθμό συνάφειας στο έγγραφο j διότι από τη στιγμή που ο όρος ki εμφανίζεται σε πολλά κείμενα μπορεί να έχει υψηλό βαθμό συνεμφάνισης με άλλους όρους που περιέχει το έγγραφο j.

**Μεγάλες τιμές $idf_i$**
Αν το idf του όρου **ki** είναι μεγάλο αυτό σημαίνει ότι το πλήθος των εγγράφων που περιέχουν τον όρο **ki** είναι μικρό (αφού idf = N/df).
Έστω ένα έγγραφο j που περιέχει τον όρο ki.
Αν λαμβάναμε υπόψη μόνο το idf τότε το $1^ο$ μοντέλο, θα έδινε μεγάλο βαθμό συνάφειας στο έγγραφο j (καθώς και σε όλα τα υπόλοιπα έγγραφα). Αντίθετα το $2^ο$ μοντέλο ενδεχομένως να έδινε μικρότερο βαθμό συνάφειας στο έγγραφο j διότι από τη στιγμή που ο όρος ki εμφανίζεται σε λίγα κείμενα μπορεί να έχει μικρό βαθμό συνεμφάνισης με άλλους όρους που περιέχει το έγγραφο j.

**Μικρές και μεγάλες τιμές του $w_{ij}$**
Θα μπορούσαμε να διακρίνουμε τις παρακάτω περιπτώσεις:

1) Μικρό TF*IDF λόγω πολύ μικρού TF

Έστω ένα έγγραφο j. Αν το $\mathbf{tf_{ij}}$ του όρου **ki** είναι πολύ μικρό, αυτό σημαίνει ότι ο όρος **ki** εμφανίζεται πολύ λίγες φορές στο έγγραφο αυτό άρα η κατάταξη του εγγράφου θα είναι χαμηλή σύμφωνα με το πρώτο μοντέλο ανάκτησης. Το δεύτερο μοντέλο αγνοεί το πλήθος των εμφανίσεων ενός όρου σε ένα έγγραφο Λαμβάνει όμως υπόψη τον βαθμό συνεμφάνισης των όρων στη συλλογή. Αυτό σημαίνει ότι αν το έγγραφο j περιέχει πολλούς όρους οι οποίοι έχουν μεγάλη εγγύτητα (συνεμφάνιση) με τον όρο ki, τότε το έγγραφο αυτό μπορεί να σταθμιστεί υψηλά.

2) Μικρό TF*IDF λόγω πολύ μικρού IDF

Αν το idf του όρου **ki** είναι πολύ μικρό αυτό σημαίνει ότι το πλήθος των εγγράφων που περιέχουν τον όρο **ki** είναι πολύ μεγάλο (αφού idf = N/df).
Έστω ένα έγγραφο j που περιέχει τον όρο ki.
Αν λαμβάναμε υπόψη μόνο το idf τότε το $1^{ο}$ μοντέλο, θα έδινε πολύ μικρό βαθμό συνάφειας στο έγγραφο j (καθώς και σε όλα τα υπόλοιπα έγγραφα). Αντίθετα το $2^{ο}$ μοντέλο ενδεχομένως να έδινε πολύ μεγαλύτερο βαθμό συνάφειας στο έγγραφο j διότι από τη στιγμή που ο όρος ki εμφανίζεται σε πολλά κείμενα μπορεί να έχει υψηλό βαθμό συνεμφάνισης με άλλους όρους που περιέχει το έγγραφο j.

3) Μικρό TF*IDF λόγω μικρού TF και μικρού IDF

Έστω ένα έγγραφο j. Αν το $\mathbf{tf_{ij}}$ του όρου **ki** είναι μικρό, αυτό σημαίνει ότι ο όρος **ki** εμφανίζεται λίγες φορές στο έγγραφο αυτό και αν το idf του όρου **ki** είναι μικρό αυτό σημαίνει ότι το πλήθος των εγγράφων που περιέχουν τον όρο **ki** είναι μεγάλο (αφού idf = N/df). Επομένως, το πρώτο μοντέλο θα έδινε μικρό βαθμό συνάφειας στο έγγραφο j.
Αντίθετα το $2^{ο}$ μοντέλο ενδεχομένως να έδινε πολύ μεγαλύτερο βαθμό συνάφειας στο έγγραφο j διότι από τη στιγμή που ο όρος ki εμφανίζεται σε πολλά κείμενα μπορεί να έχει υψηλό βαθμό συνεμφάνισης με άλλους όρους που περιέχει το έγγραφο j.

4) Μεγάλο TF*IDF λόγω πολύ μεγάλου TF

Έστω ένα έγγραφο j. Αν το $\mathbf{tf_{ij}}$ του όρου **ki** είναι πολύ μεγάλο, αυτό σημαίνει ότι ο όρος **ki** εμφανίζεται πολλές φορές στο έγγραφο αυτό άρα η κατάταξη του εγγράφου θα είναι πολύ υψηλή σύμφωνα με το πρώτο μοντέλο ανάκτησης. Στο δεύτερο μοντέλο (που αγνοεί το tf) το έγγραφο αυτό θα μπορούσε να σταθμιστεί χαμηλά αν περιέχει λίγους όρους που έχουν εγγύτητα (συνεμφάνιση) με τον όρο ki.

5) Μεγάλο TF*IDF λόγω πολύ μεγάλου IDF

Αν το idf του όρου **ki** είναι πολύ μεγάλο αυτό σημαίνει ότι το πλήθος των εγγράφων που περιέχουν τον όρο **ki** είναι πολύ μικρό (αφού idf = N/df).
Έστω ένα έγγραφο j που περιέχει τον όρο ki.
Αν λαμβάναμε υπόψη μόνο το idf τότε το $1^{ο}$ μοντέλο, θα έδινε πολύ μεγάλο βαθμό συνάφειας στο έγγραφο j (καθώς και σε όλα τα υπόλοιπα έγγραφα). Αντίθετα το $2^{ο}$

μοντέλο ενδεχομένως να έδινε πολύ μικρότερο βαθμό συνάφειας στο έγγραφο j διότι από τη στιγμή που ο όρος ki εμφανίζεται σε πολλά κείμενα μπορεί να έχει υψηλό βαθμό συνεμφάνισης με άλλους όρους που περιέχει το έγγραφο j.

6) Μεγάλο TF*IDF λόγω μεγάλου TF και μεγάλου IDF

Έστω ένα έγγραφο j. Αν το $\mathbf{tf_{ij}}$ του όρου **ki** είναι μεγάλο, αυτό σημαίνει ότι ο όρος **ki** εμφανίζεται πολλές φορές στο έγγραφο αυτό και αν το idf του όρου **ki** είναι μεγάλο αυτό σημαίνει ότι το πλήθος των εγγράφων που περιέχουν τον όρο **ki** είναι μικρό (αφού idf = N/df). Επομένως, το πρώτο μοντέλο θα έδινε μεγάλο βαθμό συνάφειας στο έγγραφο j.
Αντίθετα το $2^{ο}$ μοντέλο ενδεχομένως να έδινε πολύ μικρότερο βαθμό συνάφειας στο έγγραφο j διότι από τη στιγμή που ο όρος ki εμφανίζεται σε λίγα κείμενα μπορεί να έχει μικρό βαθμό συνεμφάνισης με άλλους όρους που περιέχει το έγγραφο j.

**Άσκηση 4**
Θεωρείστε μια συλλογή κειμένων που περιέχει τα ακόλουθα 5 έγγραφα:

Έγγραφο 1: «New York Times»
Έγγραφο 2: «New Times»
Έγγραφο 3: «Financial Times»
Έγγραφο 4: «High High Times»
Έγγραφο 5: «New Financial Times»

1) Δώστε τη διανυσματική παράσταση του κάθε εγγράφου με βάρη TF-IDF (για ευκολία θεωρήστε ότι IDF=N/DF και όχι IDF=log(N/DF)). Θεωρείστε ότι η θέση της κάθε λέξης στα διανύσματα γίνεται αλφαβητικά.

2) Θεωρείστε την επερώτηση q1=«high financial». Υπολογίστε το TF-IDF διάνυσμα αυτής της επερώτησης και δώστε την διάταξη των εγγράφων που θα επιστρέψει ένα σύστημα που βασίζεται στο διανυσματικό μοντέλο.

3) Θεωρείστε τις επερωτήσεις q2=«high AND financial» q3=«high OR financial» και δώστε τις απαντήσεις που θα επιστρέψει ένα σύστημα που βασίζεται στο Extended Boolean μοντέλο.

**Λύση**
1)
**Term Occurrence Table**

| | Financial | High | New | Times | York | $MAX_k\{FREQ_{ij}\}$ |
|---|---|---|---|---|---|---|
| $D_1$ | | | 1 | 1 | 1 | 1 |
| $D_2$ | | | 1 | 1 | | 1 |
| $D_3$ | 1 | | | 1 | | 1 |
| $D_4$ | | 2 | | 1 | | 2 |
| $D_5$ | 1 | | 1 | 1 | | 1 |
| DF | 2 | 1 | 3 | 5 | 1 | |
| IDF | 5/2 | 5/1=5 | 5/3 | 5/5=1 | 5/1=5 | |

- $FREQ_{ij}$ = το πλήθος των εμφανίσεων του όρου i στο έγγραφο j
- N = 5
- IDF = N/DF
- $MAX_k\{ FREQ_{ij}\}$ = συχνότητα της λέξης με τη μέγιστη συχνότητα στο κείμενο

**Term Weight Table**

| | Financial | High | New | Times | York | $MAX_k\{FREQ_{ij}\}$ |
|---|---|---|---|---|---|---|
| $D_1$ | | | 1/1*(5/3)=1,66 | 1/1*(5/5)=1 | 1/1*(5/1)=5 | 1 |
| $D_2$ | | | 1/1*(5/3)=1,66 | 1/1*(5/5)=1 | | 1 |
| $D_3$ | 1/1*(5/2)=2,5 | | | 1/1*(5/5)=1 | | 1 |
| $D_4$ | | 2/2*(5/1)=5 | | 1/2*(5/5)=0,5 | | 2 |
| $D_5$ | 1/1*(5/2)=2,5 | | 1/1*(5/3)=1,66 | 1/1*(5/5)=1 | | 1 |
| DF | 2 | 1 | 3 | 5 | 1 | |
| IDF | 5/2=2,5 | 5/1=5 | 5/3=1,66 | 5/5=1 | 5/1=5 | |

*Συχνότητα/MAXk{FREQij}*IDF*

- $TF_{ij} = FREQ_{ij} / MAX_k\{FREQ_{ij}\}$
- $W_{ij} = TF_{ij}*IDF_i$

2)

| | Financial | High | New | Times | York |
|---|---|---|---|---|---|
| Q1 = high financial | 1*5/2=2,5 | 1*5/1=5 | | | |
| IDF | 5/2=2,5 | 5/1=5 | 5/3=1,66 | 5/5=1 | 5/1=5 |

q1*d1 = (5/2,5,0,0,0)*(0,0,5/3,1,5) = 0
q1*d2 = (5/2,5,0,0,0)*(0,0,5/3,1,0) = 0
q1*d3 = (5/2,5,0,0,0)*(5/2,0,0,1,0) = 25/4
q1*d4 = (5/2,5,0,0,0)*(0,5,0,1/2,0) = 25
q1*d5 = (5/2,5,0,0,0)*(5/2,0,5/3,1,0) = 25/4

Άρα η διάταξη των εγγράφων που θα επιστρέψει η ερώτηση Q1 είναι: D4, D3, D5

3)

| | Financial | High | New | Times | York |
|---|---|---|---|---|---|
| $D_1$ | | | 1/1*(5/3)=1,66 | 1/1*(5/5)=1 | 1*(5/1)=5 |
| $D_2$ | | | 1/1*(5/3)=1,66 | 1/1*(5/5)=1 | |
| $D_3$ | 1/1*(5/2)=2,5 | | | 1/1*(5/5)=1 | |
| $D_4$ | | 2/2*(5/1)=5 | | 1/2*(5/5)=0,5 | |
| $D_5$ | 1/1*(5/2)=2,5 | | 1/1*(5/3)=1,66 | 1/1*(5/5)=1 | |

Κανονικοποίηση των διανυσμάτων maxIDFi=5:

- d1'=(0,0,5/3,1,5)/5=(0,0,1/3,1/5,1)
- d2'=(0,0,5/3,1,0)/5=(0,0,1/3,1/5,0)
- d3'=(5/2,0,0,1,0)/5=(1/2,0,0,1/5,0)
- d4'=(0,5,0,1/2,0)/5=(0,1,0,1/10,0)
- d5'=(5/2,0,5/3,1,0)/5=(1/2,0,1/3,1/5,0)

Q2 = "high **AND** financial"

$$sim(q_{AND}, d) = 1 - \sqrt{\frac{(1-x)^2 + (1-y)^2}{2}}$$

Με βάση τον παραπάνω τύπο, υπολογίζουμε την ομοιότητα της επερώτησης Q2 με κάθε έγγραφο.

Sim(q2,d1') = 1-sqrt(((1-0)^2+(1-0)^2)/2) = 0
Sim(q2,d2') = 1-sqrt(((1-0)^2+(1-0)^2)/2) = 0
Sim(q2,d3') = 1-sqrt(((1-1/2)^2+(1-0)^2)/2) = 0.21
Sim(q2,d4') = 1-sqrt(((1-0)^2+(1-1)^2)/2) = 0.29
Sim(q2,d5') = 1-sqrt(((1-1/2)^2+(1-0)^2)/2) = 0.21

Άρα η διάταξη των εγγράφων που θα επιστρέψει η ερώτηση Q2 είναι:
D4, D3, D5

Q3="high **OR** financial"

$$sim(q_{OR}, d) = \sqrt{\frac{x^2 + y^2}{2}}$$

Με βάση τον παραπάνω τύπο, υπολογίζουμε την ομοιότητα της επερώτησης Q3 με κάθε έγγραφο.

Sim(q3,d1') = sqrt((0^2+0^2)/2) = 0
Sim(q3,d2') = sqrt((0^2+0^2)/2) = 0
Sim(q3,d3') = sqrt(((1/2)^2+0^2)/2) = 1/(2sqrt(2))
Sim(q3,d4') = sqrt((0^2+1^2)/2) = 1/(2sqrt(2))
Sim(q3,d5') = sqrt(((1/2)^2+0^2)/2) = 1/(2sqrt(2))

Άρα η διάταξη των εγγράφων που θα επιστρέψει η ερώτηση Q3 είναι:
D4, D3, D5

**Άσκηση 5 (συνάρτηση διαβάθμισης)**
Θεωρείστε ένα Σύστημα Ανάκτησης Πληροφοριών (ΣΑΠ) από μια μεγάλη συλλογή κειμένων. Θέλουμε να δώσουμε τη δυνατότητα χρήσης του ΣΑΠ μέσω κινητού τηλεφώνου. Για το λόγο αυτό θέλουμε να ορίσουμε μια συνάρτηση διαβάθμισης (ranking function) η οποία να ευνοεί τα μικρά κείμενα, αφενός για να κρατήσουμε σε χαμηλά επίπεδα τον όγκο δεδομένων που θα μεταφέρονται και αφετέρου διότι οι χρήστες κινητών τηλεφώνων προτιμούν τα μικρά κείμενα (ένεκα του μικρού μεγέθους της οθόνης). Θεωρείστε ότι οι επερωτήσεις των χρηστών είναι σάκοι λέξεων (bag of words). Σχεδιάστε μια συνάρτηση διαβάθμισης για το σκοπό αυτό για κάθε μια από τις παρακάτω περιπτώσεις
(α) Το ευρετήριο του ΣΑΠ έχει δυαδικά (0,1) βάρη (όπως για παράδειγμα το ευρετήριο του Boolean μοντέλου)
(β) Το ευρετήριο έχει βάρη TF-IDF.
Τεκμηριώστε τις προτάσεις σας (με αποδείξεις ή παραδείγματα).

**Λύση**
α) Θέλουμε να τροποποιήσουμε το Boolean μοντέλο έτσι ώστε να ευνοεί τα μικρότερα κείμενα. Η συνάρτηση που θα χρησιμοποιήσουμε είναι η R(d,q) = |d ∩q|/|d| η οποία κανονικοποιεί την συνάρτηση που εκφράζει τη συσχέτιση ενός κειμένου με μια επερώτηση με βάση το μέγεθος του κειμένου.

Γνωρίζουμε ότι R(d,q) = |d ∩q|/|d| = |d ∩q|/(|d ∩q|+|d\q|).

Αν η τομή $d \cap q$ είναι σταθερή, δηλαδή n έγγραφα έχουν την ίδια συνάφεια, τότε αυτό που έχει μεγαλύτερο μέγεθος θα διαβαθμιστεί πιο χαμηλά, διότι θα μεγαλώσει ο παρανομαστής άρα η συνάρτηση θα επιστρέψει μικρότερη τιμή διαβάθμισης. Στη γενική περίπτωση που έχουμε διαφορετικές συνάφειες τα μικρότερα έγγραφα ευνοούνται έναντι των μεγαλύτερων.

Πειραματικά αυτό σημαίνει:

| | |
|---|---|
| q = “a b” | R(d,q) = \|d ∩q\|/\|d\| |
| d1 = “a” | 1/1=1 |
| d2 = “a c” | 1/2=0.5 |
| d3 = “a c d” | 1/3=0.33 |
| d4 = “a b c” | 2/3=0.66 |
| d5 = “a b c d a“ | 2/5 = 0.4 |
| Διάταξη εγγράφων | <d1, d4, d2, d5, d3> |

β) Γενικεύουμε την ιδέα του (α) για την περίπτωση που έχουμε μη δυαδικά βάρη. Συγκεκριμένα μπορούμε να ορίσουμε: R(d,q) = d*q / ||d|| (όπου το * είναι το

εσωτερικό γινόμενο) όπου το εσωτερικό γινόμενο υπολογίζεται από τον τύπο: $\sum_{i=1}^{t} (w_{ij} \cdot w_{iq})$

και τα $w_{i,j}=tf_{ij}*idf_i$ και $w_{i,q}= tf_{iq}*idf_i$

| | a | **b** | **c** | **D** | **$MAX_k\{FREQ_{ij}\}$** |
|---|---|---|---|---|---|
| **$D_1$** | 1 | 0 | 0 | 0 | 1 |
| **$D_2$** | 1 | 0 | 1 | 0 | 1 |
| **$D_3$** | 1 | 0 | 1 | 1 | 1 |
| **$D_4$** | 1 | 1 | 1 | 0 | 1 |
| **$D_5$** | 2 | 1 | 1 | 1 | 2 |
| **DF** | 5 | 2 | 4 | 2 | |
| **IDF** | 5/5=1 | 5/2=2.5 | 5/4=1.25 | 5/2=2.5 | |

| | **A** | **b** | **C** | **D** | **$MAX_k\{FREQ_{ij}\}$** |
|---|---|---|---|---|---|
| **$D_1$** | 1/1*5/5=1 | 0 | 0 | 0 | 1 |
| **$D_2$** | 1/1*5/5=1 | 0 | 1/1*5/4=1.25 | 0 | 1 |
| **$D_3$** | 1/1*5/5=1 | 0 | 1/1*5/4=1.25 | 1/1*5/2=2.5 | 1 |
| **$D_4$** | 1/1*5/5=1 | 1/1*5/2=2.5 | 1/1*5/4=1.25 | 0 | 1 |
| **$D_5$** | 2/2*5/5=1 | 1/2*5/2=1.25 | 1/2*5/4=0.625 | 1/2*5/2=1.25 | 2 |
| **Q1** | 1/1*5/5=1 | 1/1*5/2=2.5 | 0 | 0 | 1 |
| **DF** | 5 | 2 | 4 | 2 | |
| **IDF** | 5/5=1 | 5/2=2.5 | 5/4=1.25 | 5/2=2.5 | |

$W_1*Q_1$= (1,0,0,0)*(1,2.5,0,0) = 1
$W_2*Q_1$= (1,0,1.25,0)*(1,2.5,0,0) = 1
$W_3*Q_1$= (1,0,1.25,2.5)*(1,2.5,0,0) = 1
$W_4*Q_1$= (1,2.5,1.25,0)*(1,2.5,0,0) =7.25
$W_5*Q_1$= (1,1.25,0.625,1.25)*(1,2.5,0,0) =4.125

Εφαρμόζουμε τη συνάρτηση για τον υπολογισμό της συνάφειας:
$R(D_1,Q_1)$=1/1= $D_1* Q_1 / || D_1||$=1
$R(D_2,Q_1)$=1/2=0.5
$R(D_3,Q_1)$=1/3=0.33
$R(D_4,Q_1)$=7.25/3=2.416
$R(D_5,Q_1)$=4.125/5=0.825

Επομένως η διάταξη που επιστρέφει η συνάρτηση μας είναι η D4, D1, D5, D2, D3. Παρατηρούμε ότι από τα κείμενα που είναι ποιο σχετικά στο ερώτημα μας (δηλαδή περιέχουν και τους δύο όρους) ευνοήθηκε αυτό με το μικρότερο πλήθος όρων.