ΗΥ-463

Συστήματα Ανάκτησης Πληροφορίας
Information Retrieval Systems

Tutorial on Retrieval Models

ΒΕΛΕΓΡΑΚΗΣ ΔΗΜΗΤΡΙΟΣ

Άσκηση 1

Θεωρείστε μια συλλογή κειμένων που περιέχει τα ακόλουθα 4 έγγραφα:

Έγγραφο 1: «New Year»

Έγγραφο 2: « New Year New Year »

Έγγραφο 3: «Financial New Times»

Έγγραφο 4: «Financial Year»

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 κατά αλφαβητική σειρά.

2) Θεωρείστε την επερώτηση q1=«new 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) Σχεδιάστε το ανεστραμμένο αρχείο για αυτή τη συλλογή.

Άσκηση 1 (α' ερώτημα)

	Financial	New	Times	Year	MAXk {FREQij}
D_1	0	1	0	1	1
D_2	0	2	0	2	2
D_3	1	1	1	0	1
D_4	1	0	0	1	1
DF	2	3	1	3	
IDF	4/2	4/3	4/1	4/3	

- $FREQ_{ij}$=το πλήθος των εμφανίσεων του όρου i στο έγγραφο j
- IDF = N / DF
- $MAX_k\{FREQ_{ij}\}$ = η μέγιστη συχνότητα ενός όρου i σε ένα έγγραφο j

Άσκηση 1 (α' ερώτημα)

	Financial	New	Times	Year	MAXk {FREQij}
D_1	0	1/1*4/3	0	1/1*4/3	1
D_2	0	2/2*4/3	0	2/2*4/3	2
D_3	1/1*4/2	1/1*4/3	1/1*4/1	0	1
D_4	1/1*4/2	0	0	1/1*4/3	1
DF	2	3	1	3	
IDF	4/2=2	4/3	4/1=4	4/3	

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

Άσκηση 1 (α' ερώτημα)

- Οι διανυσματικές παραστάσεις των κειμένων είναι

$W_1=\{0, 1.33, 0, 1.33\}$, $|W_1| = 3.5378$

$W_2=\{0, 1.33, 0, 1.33\}$, $|W_2| = 3.5378$

$W_3=\{2, 1.33, 4, 0\}$, $|W_3| = 21.7689$

$W_4=\{2, 0, 0, 1.33\}$, $|W_4| = 5.7689$

Άσκηση 1 (β' ερώτημα)

	Financial	New	Times	Year
Q1 = New Financial	1/1 * 4/2	1/1 * 4/3	0	0
IDF	4/2 = 2	4/3	4/1 = 4	4/3

$Q_1 = \{2, 1.33, 0, 0\}, \ |Q_1| = 5.7689$

$W_1 * Q_1 = 1.7689$

$W_2 * Q_1 = 1.7689$

$W_3 * Q_1 = 5.7689$

$W_4 * Q_1 = 4$

Άσκηση 1 (β' ερώτημα)

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

- $R(D_1,Q_1) = 1.7689/(\text{sqrt}(3.5378 * 5.7689)) = 0.392$
- $R(D_2,Q_1) = 1.7689/(\text{sqrt}(3.5378 * 5.7689)) = 0.392$
- $R(D_3,Q_1) = 5.7689/(\text{sqrt}(21.7689 * 5.7689)) = 0.515$
- $R(D_4,Q_1) = 4/(\text{sqrt}(5.7689 * 5.7689)) = 0.693$

- **Η διάταξη των κειμένων θα είναι :**
  D_4, D_3, {D_1, D_2}.
- **Το αναμενόμενο αποτέλεσμα θα ήταν να έρθει πρώτο στη διάταξη μας το D_3 όμως αυτό δεν συνέβη διότι το βάρος του εγγράφου επηρεάστηκε από τον όρο Times ο οποίος εμφανίζεται μόνο μια φορά στη συλλογή των κειμένων μας.**

Άσκηση 1 (γ' ερώτημα)

Term	<Frequency, (Document; Position)
financial	<2 (D_3;1), (D_4;1)>
new	<4 (D_1;1), (D_2;1), (D_2;3), (D_3;2)>
times	<1 (D_3;3)>
year	<4 (D_1;2), (D_2;2), (D_2;4), (D_4;2)>

Άσκηση 2

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).

Άσκηση 2 (α' ερώτημα)

Θέλουμε να τροποποιήσουμε το Boolean μοντέλο έτσι ώστε να ευνοεί τα μικρότερα κείμενα.

Η συνάρτηση που θα χρησιμοποιήσουμε είναι η $R(d,q) = |d \cap q|/|d|$ η οποία κανονικοποιεί την συνάρτηση που εκφράζει τη συσχέτιση ενός κειμένου με μια επερώτηση με βάση το μέγεθος του κειμένου.
Γνωρίζουμε ότι $R(d,q) = |d \cap q|/|d| = |d \cap q|/(|d \cap q|+|d \setminus q|)$.

Αν η τομή $d \cap q$ *είναι σταθερή, δηλαδή η 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.

Άσκηση 2 (α' ερώτημα)

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>

Άσκηση 2 (β' ερώτημα)

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)

Το εσωτερικό γινόμενο υπολογίζεται από τον τύπο:

$$\sum_{i=1}^{t} (w_{ij} \cdot w_{iq})$$

$w_{i,j} = tf_{ij} * idf_i$ **και**

$w_{i,q} = tf_{iq} * idf_i$

Άσκηση 2 (β' ερώτημα)

	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	

Άσκηση 2 (β' ερώτημα)

	a	b	C	d	$MAX_k \{FREQ_{ij}\}$
d_1	1/1 * 5/5 = 1	0/1 * 5/2 = 0	0/1*5/4= 0	0/1*5/2= 0	1
d_2	1/1 * 5/5 = 1	0/1 * 5/2 = 0	1/1*5/4= 1.25	0/1* 5/2= 0	1
d_3	1/1 * 5/5 = 1	0/1 * 5/2 = 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= 1	1/1*5/4= 1.25	0/1* 5/2= 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
q	1/1 * 5/5 = 1	1/1 * 5/2 = 2.5	0/1*5/4= 0	0/1* 5/2 = 0	1
DF	5	2	4	2	
IDF	5/5 = 1	5/2 =2.5	5/4 = 1.25	5/2 = 2.5	

$W_1 * Q_1 = (1,0,0,0) * (1,2.5,0,0) = 1$

$W_2 * Q_1 = (1,0,1.25,0) * (1,2.5,0,0) = 1$

$W_3 * Q_1 = (1,0,1.25,2.5) * (1,2.5,0,0) = 1$

$W_4 * Q_1 = (1,1,1.25,0) * (1,2.5,0,0) = 3.5$

$W_5 * Q_1 = (1,1.25,0.625,1.25) * (1,2.5,0,0) = 4.125$

Άσκηση 2 (β' ερώτημα)

Εφαρμόζουμε τη συνάρτηση για τον υπολογισμό της συνάφειας:

$R(D_1,Q_1) = 1/1=1$

$R(D_2,Q_1) = 1/2=0.5$

$R(D_3,Q_1) = 1/3=0.33$

$R(D_4,Q_1) = 3.5/3=1.16$

$R(D_5,Q_1) = 4.125/5=0.825$

Επομένως η διάταξη που επιστρέφει η συνάρτηση μας είναι η:

D_4, D_1, D_5, D_2, D_3

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.