



## 1<sup>η</sup> Σειρά Ασκήσεων

Αξία: 10% του τελικού σας βαθμού  
(Αξιολόγηση της Αποτελεσματικότητας της Ανάκτησης & Μοντέλα Ανάκτησης)

### Άσκηση 1

Θεωρείστε μια συλλογή αξιολόγησης που αποτελείται από 20 έγγραφα  $\{d_1, \dots, d_{20}\}$ . Η συλλογή αξιολόγησης περιλαμβάνει μια επερώτηση  $q$  για την οποία γνωρίζουμε ότι τα έγγραφα της συλλογής που είναι συναφή με αυτήν είναι 4, συγκεκριμένα τα  $\{d_2, d_6, d_7, d_{11}\}$ . Θέλουμε να αξιολογήσουμε την αποτελεσματικότητα τριών συστημάτων  $S1$ ,  $S2$  και  $S3$ .

Για το λόγο αυτό υποβάλλουμε σε κάθε σύστημα την επερώτηση  $q$  και λαμβάνουμε τις εξής απαντήσεις :

$Ans(S1, q) = \langle d_1, d_2, d_7, d_{11}, d_5, d_{10}, d_{12}, d_{14}, d_3, d_4 \rangle$

$Ans(S2, q) = \langle d_6, d_9, d_{10}, d_1, d_8, d_{11}, d_{12}, d_2, d_{17}, d_{15} \rangle$

$Ans(S3, q) = \langle d_1, d_{11}, d_7, d_8, d_{15} \rangle$

Το αριστερότερο στοιχείο της κάθε απάντησης παριστάνει το υψηλότερα διαβαθμισμένο έγγραφο, αυτό που το σύστημα υπολόγισε ως το πιο συναφές με την επερώτηση  $q$ . Συγκρίνετε τα τρία αυτά συστήματα ως προς τα εξής μέτρα:

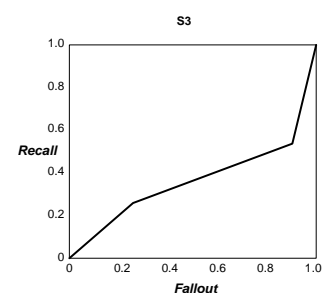
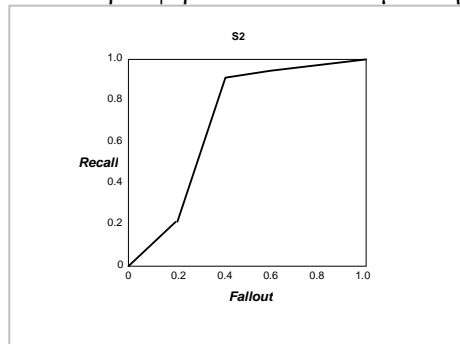
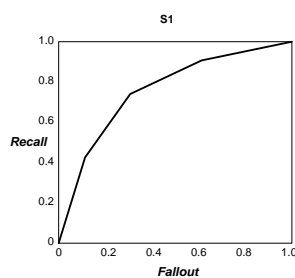
- Precision (Ακρίβεια)
- Recall (Ανάκληση)
- F-Measure
- R-Precision (R-Ακρίβεια)
- Fallout

### Άσκηση 2

α) Σχεδιάστε τις καμπύλες ακρίβειας/ανάκλησης (P/R curves) των συστημάτων της προηγούμενης άσκησης. Για κάθε σύστημα δώστε 2 γραφήματα: ένα που να απεικονίζει τα P/R σημεία όπως προκύπτουν από τις απαντήσεις, και ένα χρησιμοποιώντας κανονικοποιημένα επίπεδα ανάκλησης (standard recall levels). Αν βλέπατε μόνο αυτά τα γραφήματα (και όχι τις απαντήσεις) θα μπορούσατε να επιλέξετε το καλύτερο σύστημα;

β) Ένας εναλλακτικός τρόπος αξιολόγησης της αποτελεσματικότητας ενός συστήματος είναι οι καμπύλες Recall-Fallout. Ορίζονται ανάλογα με τις καμπύλες Precision-Recall, μόνο που τώρα ο άξονας X έχει τις τιμές του Fallout, ενώ ο Y τις τιμές του Recall. Σχεδιάστε τις καμπύλες Recall-Fallout των συστημάτων της προηγούμενης άσκησης.

γ) Θεωρείστε τρία συστήματα με τις καμπύλες Recall-Fallout που ακολουθούν. Παρατηρώντας αυτές τις καμπύλες, ποιο σύστημα θα κρίνατε ότι προσφέρει πιο αποτελεσματική ανάκτηση πληροφορίας;



### Άσκηση 3

Έστω ότι η συλλογή αξιολόγησης αποτελείται από 50 έγγραφα  $\{d_1, \dots, d_{50}\}$  και γνωρίζουμε ότι υπάρχουν 3 έγγραφα της συλλογής, συγκεκριμένα τα  $\{d_1, d_2, d_3\}$ , που είναι συναφή με την επερώτηση  $q$ . Θέλουμε να αξιολογήσουμε την αποτελεσματικότητα τριών συστημάτων  $S_1, S_2$  και  $S_3$  τα οποία επιστρέφουν ως απάντηση έγγραφα συνοδευμένα από ένα βαθμό συνάφειας.

Υποβάλλουμε σε κάθε σύστημα την επερώτηση  $q$  και λαμβάνουμε τις εξής απαντήσεις:

$$\text{Ans}(S_1, q) = \langle d_1, \{d_2, d_{20-d50}\}, d_3 \rangle$$

$$\text{Ans}(S_2, q) = \langle d_1, d_2, d_3 \rangle$$

$$\text{Ans}(S_3, q) = \langle \{d_1, d_8\}, d_2, d_3 \rangle$$

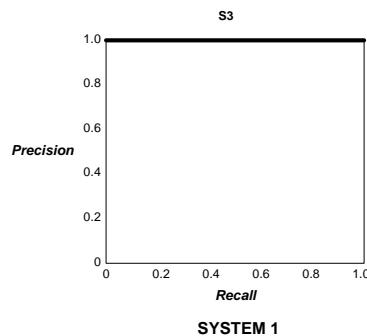
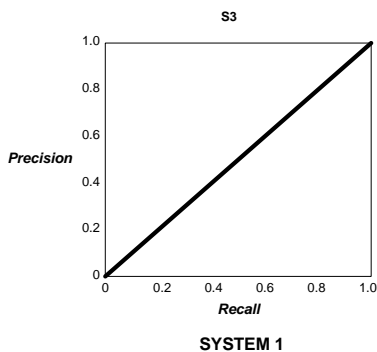
Η απάντηση  $\langle \{d_1, d_8\}, d_2, d_3 \rangle$  σημαίνει ότι τα  $d_1, d_8$  ισοβαθμούν στην πρώτη θέση (άρα έλαβαν τον μεγαλύτερο βαθμό συνάφειας). Η απάντηση  $\langle d_1, \{d_2, d_{20-d50}\}, d_3 \rangle$  σημαίνει ότι το  $d_1$  έλαβε το μεγαλύτερο βαθμό, ενώ μετά ακολουθεί μια ομάδα από 31 έγγραφα τα οποία ισοβαθμούν και στο τέλος της κατάταξης βρίσκεται το  $d_3$ . για κάθε ένα από τα 3 συστήματα απαντήστε τα ακόλουθα ερωτήματα:

- Ποια είναι η R-Ακρίβεια (R-Precision);
- Ποιο είναι το αναμενόμενο μήκος αναζήτησης για να βρούμε 2 συναφή;
- Ποιο είναι το μέσο αναμενόμενο μήκος αναζήτησης;

### Άσκηση 4

Έστω ένα Σύστημα Ανάκτησης Πληροφοριών που βασίζεται στο λογικό μοντέλο (Boolean Model). Θα είχε νόημα η αξιολόγησή του με

- καμπύλες Precision-Recall;
- καμπύλες Recall-Fallout;
- αναμενόμενο μήκος αναζήτησης;
- Έστω δυο συστήματα  $S_1$  και  $S_2$  για τα οποία σας δίδουν τα ακόλουθα διαγράμματα. Σχολιάστε τα. Υπάρχει κάτι που σας παραξενεύει;



### Άσκηση 5

Έστω ότι έχουμε ένα μοντέλο ανάκτησης το οποίο βλέπει τα έγγραφα και τις επερωτήσεις ως σύνολα όρων. Συγκρίνετε τις ακόλουθες συναρτήσεις διαβάθμισης:

$R_1(d, q) = \frac{ d \cap q }{ q \cap d  +  q \setminus d }$	$R_3(d, q) = \frac{ d \cap q }{ d \setminus q  +  q \setminus d  +  d \cap q }$
$R_2(d, q) = \frac{ d \cap q }{ q \cap d  +  d \setminus q }$	$R_4(d, q) = 2 * \frac{ d \cap q }{ d \cup q }$

### Άσκηση 6

- (α) Δώστε την διανυσματική αναπαράσταση των εγγράφων  $d_1, \dots, d_5$  με βάρη TF-IDF. Θεωρείστε ότι η θέση της κάθε λέξης στα διανύσματα δίνεται κατά αλφαβητική σειρά.
- (β) Δώστε την απάντηση που θα έχει η κάθε επερώτηση  $q_1, q_2, q_3$  βάσει του διανυσματικού μοντέλου.
- (γ) Σχεδιάστε την μορφή που θα έχει το ανεστραμμένο ευρετήριο για την συλλογή D.

Documents
$d_1$ : «information retrieval retrieval information»
$d_2$ : «retrieval information»
$d_3$ : «information information course»
$d_4$ : «course retrieval information information retrieval»
$d_5$ : «retrieval information course»

Queries
$q_1$ : «information retrieval»
$q_2$ : «course»
$q_3$ : «information course»

### Άσκηση 7

Θεωρείστε το παρακάτω τμήμα ενός ανεστραμμένου ευρετήριο όπου αποθηκεύονται και οι θέσεις εμφάνισης των λέξεων στα έγγραφα (positional index) με την ακόλουθη μορφή:

word: document: (position, position, . . .); document: (position, . . .)

. . .

Gates:	1: (3); 2: (6); 3: (2,17); 4: (1);
IBM:	4: (3); 7: (14);
Microsoft:	1: (1); 2: (1,21); 3: (3); 5: (16,22,51);

Θεωρείστε τον τελεστή επερώτησης  $/k$  ο οποίος έχει την εξής σύνταξη και ερμηνεία: μια επερώτηση «word1  $/k$  word2» απαιτεί εκείνα τα έγγραφα στα οποία βρίσκεται η λέξη word1 εμφανίζεται σε απόσταση το πολύ  $k$  λέξεων από μια εμφάνιση της λέξης word2, όπου το  $k$  είναι ένας θετικός ακέραιος. Άρα για  $k=1$ , απαιτείται η word1 να είναι γειτονική με την word2 (αλλά όχι απαραίτητα σε αυτή την σειρά).

α) Περιγράψτε (με λόγια) ποιο είναι το σύνολο των συναφών εγγράφων για την επερώτηση «Gates  $/2$  Microsoft»

β) Περιγράψτε κάθε σύνολο των τιμών του  $k$  για το οποίο η επερώτηση «Gates  $/k$  Microsoft», επιστρέφει ένα διαφορετικό σύνολο εγγράφων ως απάντηση (θεωρώντας το παραπάνω ευρετήριο).