Πανεπιστήμιο Κρήτης, Τμήμα Επιστήμης Υπολογιστών
ΗΥ463 – Συστήματα Ανάκτησης Πληροφοριών
2009-2010 Χειμερινό Εξάμηνο

1η Σειρά Ασκήσεων

Αξία: 15% του τελικού σας βαθμού
(Αξιολόγηση της Αποτελεσματικότητας της Ανάκτησης & Μοντέλα Ανάκτησης)

Άσκηση 1 (2 Βαθμοί)

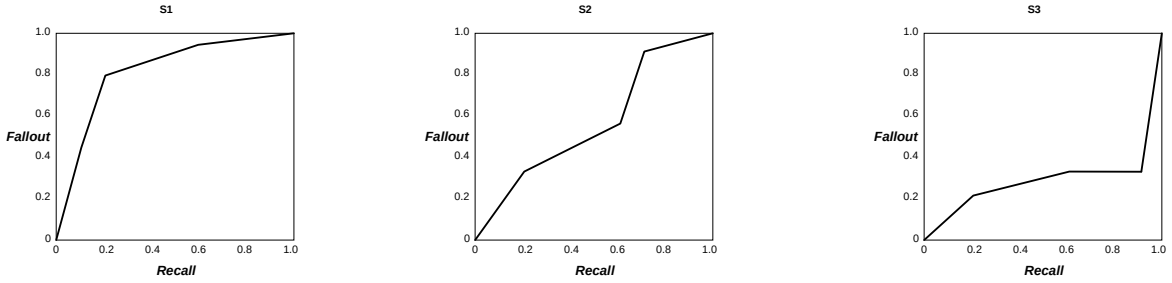
Θεωρείστε μια συλλογή αξιολόγησης που αποτελείται από 25 έγγραφα $\{d_1,\ldots,d_{25}\}$. Η συλλογή αξιολόγησης περιλαμβάνει μια επερώτηση q για την οποία γνωρίζουμε ότι τα έγγραφα της συλλογής που είναι συναφή με αυτήν είναι 7, συγκεκριμένα τα $\{d_1,d_2,d_3,d_5,d_6,d_7,d_{11}\}$. Θέλουμε να αξιολογήσουμε 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)= $<d_1, d_2, d_7, d_{11}, d_5, d_{10}, d_{12}, d_{14}, d_3, d_4>$
Ans(S2,q)= $<d_6, d_9, d_{10}, d_1, d_8, d_{11}, d_{12}, d_2, d_{17}, d_{15}>$
Ans(S3,q)= $<d_1, d_{11}, d_7, d_8, d_{15}, d_2, d_4>$

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:
α) Precision (Ακρίβεια)
β) Recall (Ανάκληση)
γ) F-Measure
δ) R-Precision (R-Ακρίβεια)
ε) Fallout

Άσκηση 2 (2 Βαθμοί)

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 Χ έχει τις τιμές του Recall, ενώ ο Υ τις τιμές του Fallout. 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 (1.5 Βαθμοί)

Έστω ότι η συλλογή αξιολόγησης αποτελείται από 50 έγγραφα $\{d_1,\ldots,d_{50}\}$ και γνωρίζουμε ότι υπάρχουν 4 έγγραφα της συλλογής, συγκεκριμένα τα $\{d_1, d_2, d_3, d_5\}$, που είναι συναφή με την επερώτηση q. Θέλουμε να αξιολογήσουμε την αποτελεσματικότητα τριών συστημάτων S1, S2 και S3 τα οποία επιστρέφουν ως απάντηση έγγραφα συνοδευμένα από ένα βαθμό συνάφειας.

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

$\text{Ans}(S1,q) = < d_1, d_5, \{d_2, d_{20}\text{-}d_{50}\}, d_3 >$

$\text{Ans}(S2,q) = < d_1, d_2, d_3, d_4, d_5 >$

$\text{Ans}(S3,q) = < d_5, \{d_1, d_8\}, d_2, d_3 >$

Η απάντηση $< d_5, \{d_1,d_8\}, d_2, d_3 >$ σημαίνει ότι το d_5 είναι στη πρώτη θέση, ενώ τα d_1, d_8 ισοβαθμούν στη δεύτερη θέση, ακολουθούμενα από τα d_2 και d_3. Η απάντηση $< d_1, d_5, \{d_2, d_{20}\text{-}d_{50}\}, d_3 >$ σημαίνει ότι το d_1 έλαβε το μεγαλύτερο βαθμό, το d_5 το δεύτερο υψηλότερο βαθμό και έπειτα ακολουθεί μια ομάδα από 32 έγγραφα τα οποία ισοβαθμούν και στο τέλος της κατάταξης βρίσκεται το d_3. για κάθε ένα από τα 3 συστήματα απαντήστε τα ακόλουθα ερωτήματα:

α) Ποια είναι η R-Ακρίβεια (R-Precision) ;

β) Ποιο είναι το αναμενόμενο μήκος αναζήτησης για να βρούμε 2 συναφή;

γ) Ποιο είναι το μέσο αναμενόμενο μήκος αναζήτησης;

Άσκηση 4 (1.5 Βαθμοί)

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

$R_1(d, q) = \frac{\lvert d \setminus q \rvert}{\lvert q \cap d \rvert}$	$R_3(d, q) = \frac{\lvert d \setminus q \rvert + \lvert q \setminus d \rvert + \lvert d \cap q \rvert}{\lvert d \cap q \rvert}$
$R_2(d, q) = \frac{\lvert q \setminus d \rvert}{\lvert d \rvert + \lvert q \rvert}$	$R_4(d, q) = \frac{\lvert d \cup q \rvert - \lvert d \setminus q \rvert - \lvert q \setminus d \rvert}{\lvert d \cup q \rvert}$

Άσκηση 5 (1.5 Βαθμοί)

(α) Δώστε την διανυσματική αναπαράσταση των εγγράφων $d_1,..,d_5$ 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: «retrieval information retrieval information»
d_2: «information course»
d_3: «course retrieval retrieval»
d_4: «course information information retrieval»
d_5: «course information retrieval information course»

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

Για ευκολία πράξεων το idf ενός token θεωρούμε ότι είναι N/df και όχι log(N/df))

Άσκηση 6 (1.5 Βαθμοί)
Θεωρείστε το παρακάτω τμήμα ενός ανεστραμμένου ευρετηρίου όπου αποθηκεύονται και οι θέσεις εμφάνισης των λέξεων στα έγγραφα(positional index) με την ακόλουθη μορφή:

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

Gates:	1: (2, 8); 2: (6, 10); 3: (2,17); 4: (1); 5: (4,43,); 7: (20);
IBM:	1: (7); 2: (8); 4: (4); 7: (14); 8: (20, 40);
Microsoft:	1: (1); 2: (1,21); 3: (3); 5: (16,22,51); 8: (15, 20)

Θεωρείστε τον τελεστή επερώτησης /*κ* ο οποίος έχει την εξής σύνταξη και ερμηνεία: μια επερώτηση «word1 /*k* word2 » απαιτεί εκείνα τα έγγραφα στα οποία βρίσκει η λέξη word1 εμφανίζεται σε απόσταση το πολύ *κ* λέξεων από μια εμφάνιση της λέξης word2 , όπου το κ είναι ένας θετικός ακέραιος. Άρα για *κ=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 /3 IBM»

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

(c) Περιγράψτε (με λόγια) ποιο είναι το σύνολο των συναφών εγγράφων για την επερώτηση «Gates /3 IBM» AND «Microsoft / 1 Gates»

(d) Για ποια τιμή του κ θα μπορούσατε να υποστηρίξετε phrase queries? Ποια ιδιαιτερότητα θα είχαν? Υποθέστε ότι η προτεραιότητα του τελεστή είναι από αριστερά προς τα δεξιά.