Πανεπιστήμιο Κρήτης, Τμήμα Επιστήμης Υπολογιστών
ΗΥ463 – Συστήματα Ανάκτησης Πληροφοριών
2008-2009 Εαρινό Εξάμηνο

Προτεινόμενες Λύσεις 1ης Σειράς Ασκήσεων

(Αξιολόγηση της Αποτελεσματικότητας της Ανάκτησης & Μοντέλα Ανάκτησης)

Άσκηση 1 (2.5 Βαθμοί)

Θεωρείστε μια συλλογή αξιολόγησης που αποτελείται από 20 έγγραφα $\{d_1,\ldots,d_{20}\}$. 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)= $<d_1, d_2, d_7, d_{11}, d_5, d_{10}, d_{12}, d_{14}, d_3, d_4>$
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)= $<d_1, d_{11}, d_7, d_8, d_{15}>$

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

<u>Λύση (από Ψαράκη Μαρία-Γεωργία)</u>

(a) Precision (Ακρίβεια)
Τα έγγραφα της συλλογής που είναι συναφή με την επερώτηση είναι τα {d2, d6, d7, d11}, συνολικά 4. Οπότε έχουμε:

Precision(S) = (ευρεθέντα και συναφή έγγραφα)/ευρεθέντα

- Για το S1:

Το S1 επιστρέφει 10 έγγραφα, τα 3 συναφή ({d2, d7, d11})
P(S1) = 3/10 = 0,3

- Για το S2:

Το S2 επιστρέφει 10 έγγραφα, τα 3 συναφή ({d2, d6, d11})
P(S2) = 3/10 = 0,3

- Για το S3:

Το S3 επιστρέφει 5 έγγραφα, τα 2 συναφή ({d7, d11})
P(S3) = 2/5 = 0,4

Άρα, βλέπουμε ότι τα S1 και S2 έχουν την ίδια ακρίβεια (παρόλο που το ένα σύστημα επέστρεψε σε πιο πρώτες θέσεις τα συναφή έγγραφα η ακρίβεια δεν επηρεάστηκε), ενώ το σύστημα S3 έχει την μεγαλύτερη ακρίβεια (παρόλο που επιστρέφει τα λιγότερα συναφή έγγραφα από τα άλλα έχει την μεγαλύτερη ακρίβεια καθώς επιστρέφει και τα λιγότερο μη συναφή έγγραφα σε σχέση με τα υπόλοιπα).
Επομένως, από πλευράς ακρίβειας το σύστημα S3 είναι το καλύτερο.

(b) Recall (Ανάκληση)
Για την ανάκληση έχουμε:

Recall(S) = (ευρεθέντα και συναφή έγγραφα)/ συναφή

- Για το S1:

Το S1 επιστρέφει 3 συναφή ({d2, d7, d11}) στα 4 που υπάρχουν
R(S1) = 3/4 = 0,75

- Για το S2:

Το S2 επιστρέφει 3 συναφή ({d2, d6, d11}) στα 4 που υπάρχουν
R(S2) = 3/4 = 0,75

- Για το S3:

Το S3 επιστρέφει 2 συναφή ({d7, d11}) στα 4 που υπάρχουν
R(S3) = 2/4 = 0,5

Άρα, βλέπουμε ότι τα S1 και S2 έχουν την ίδια ανάκληση (αφού επιστρέφουν το ίδιο πλήθος συναφών εγγράφων), ενώ το σύστημα S3 έχει την χαμηλότερη ανάκληση (αφού επιστρέφει τα λιγότερα συναφή έγγραφα από τα υπόλοιπα).
Επομένως, από πλευράς ανάκλησης τα συστήματα S1 και S2 είναι τα καλύτερα.

(c) F-Measure
Για το F-Measure έχουμε:

F-Measure(S) = 2*P*R/(P+R)

- Για το S1:

F-Measure (S1) = 2*P(S1)*R(S1)/(P(S1)+R(S1)) = 2*0,3*0,75/(0,3+0,75) = 0,429

- Για το S2:

F-Measure (S2) = 2*P(S1)*R(S1)/(P(S1)+R(S1)) = 2*0,3*0,75/(0,3+0,75) = 0,429

- Για το S3:

F-Measure (S3) = 2*P(S1)*R(S1)/(P(S1)+R(S1)) = 2*0,4*0,5/(0,4+0,5) = 0,44

Άρα, βλέπουμε ότι το S3 έχει το μεγαλύτερο F-Measure από τα υπόλοιπα, καθώς θέλουμε να έχουμε και υψηλή τιμή και στο P και στο R.
Επομένως, από πλευράς F-Measure το σύστημα S3 είναι το καλύτερο.

(d) R-Precision (R-Ακρίβεια)
Για το R-Precision έχουμε:

R-Precision(S) = Ακρίβεια στην k θέση της διάταξης της απάντησης, όπου k ο αριθμός των συναφών εγγράφων.
Εδώ έχουμε k=4.

- Για το S1:

Το S1 επιστρέφει 3 συναφή έγγραφα στις 4 πρώτες θέσεις
R-Precision(S1) = 3/4 = 0,75

- Για το S2:

Το S2 επιστρέφει 1 συναφή έγγραφο στις 4 πρώτες θέσεις
R-Precision(S2) = 1/4 = 0,25

- Για το S3:

Το S3 επιστρέφει 2 συναφή έγγραφα στις 4 πρώτες θέσεις
R-Precision(S3) = 2/4 = 0,5

Άρα, βλέπουμε ότι το S1 έχει το μεγαλύτερο R-Precision από τα υπόλοιπα, δηλαδή μας επιστρέφει στις πρώτες 4 θέσεις τα περισσότερα συναφή έγγραφα.
Επομένως, από πλευράς R-Precision το σύστημα S1 είναι το καλύτερο.

(e) Fallout
Για το Fallout έχουμε:

Fallout(S) = (μη συναφή έγγραφα που ανακτήθηκαν)/(μη συναφή έγγραφα της συλλογής).
Τα μη συναφή έγγραφα της συλλογής είναι 16 (20-4).

- Για το S1:

Το S1 επιστρέφει 7 μη συναφή έγγραφα
Fallout(S1) = 7/16 = 0,438

- Για το S2:

Το S2 επιστρέφει 7 μη συναφή έγγραφα
Fallout(S2) = 7/16 = 0,438

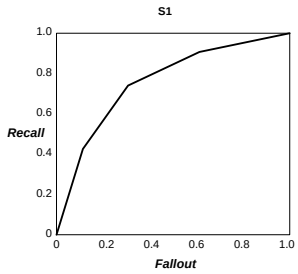
- Για το S3:

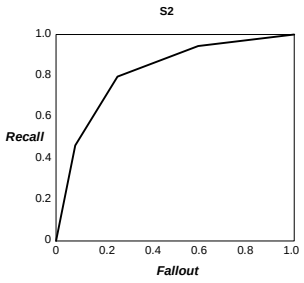
Το S3 επιστρέφει 3 μη συναφή έγγραφα
Fallout(S3) = 3/16 = 0,188

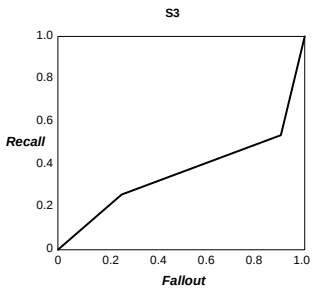
Άρα, βλέπουμε ότι το S3 έχει το μικρότερο Fallout από τα υπόλοιπα, δηλαδή μας επιστρέφει τα λιγότερα μη συναφή έγγραφα.
Επομένως, από πλευράς Fallout το σύστημα S3 είναι το καλύτερο.

Άσκηση 2 (2.5 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 Fallout, ενώ ο Υ 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;







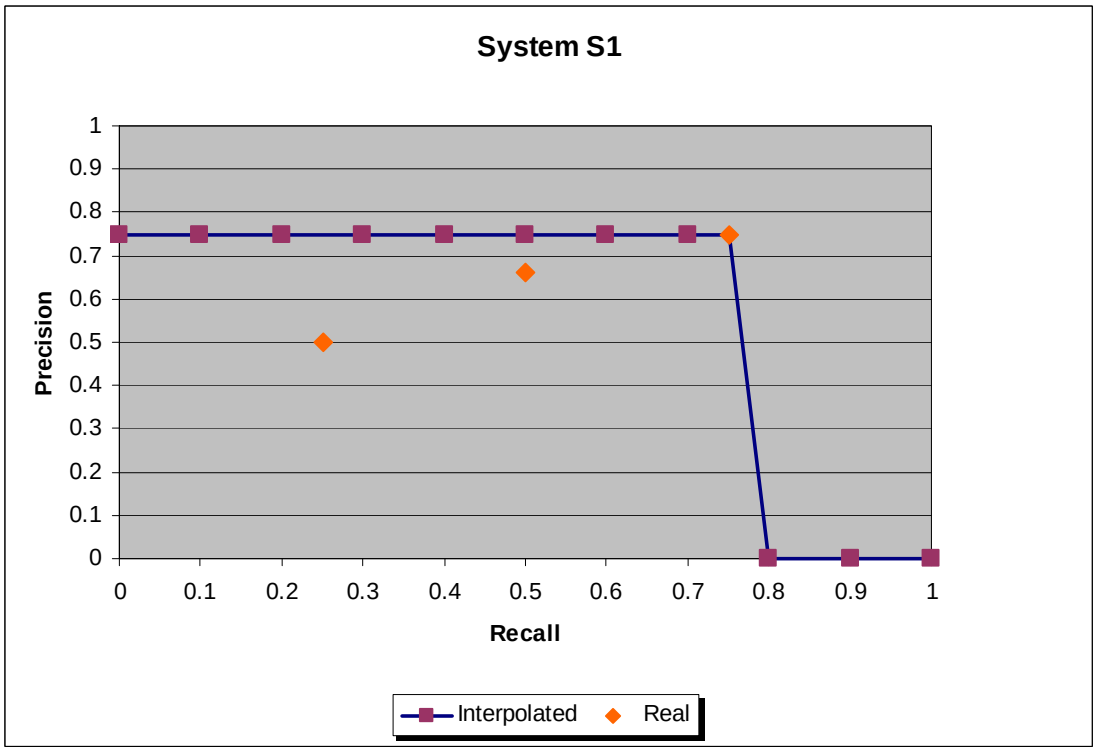
Λύση (από Καμπουράκη Μαίρη)

α) Precision/Recall Curves:

Για την προηγούμενη άσκηση υπολογίζουμε τις τιμές των *precision* και *recall* για κάθε συναφές έγγραφο κάθε συστήματος. Οι τιμές αυτές αποτελούν τα σημεία για τις γραφικές παραστάσεις των συστημάτων. Ισχύουν τα γνωστά επίπεδα ανάκλησης: *{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}*.

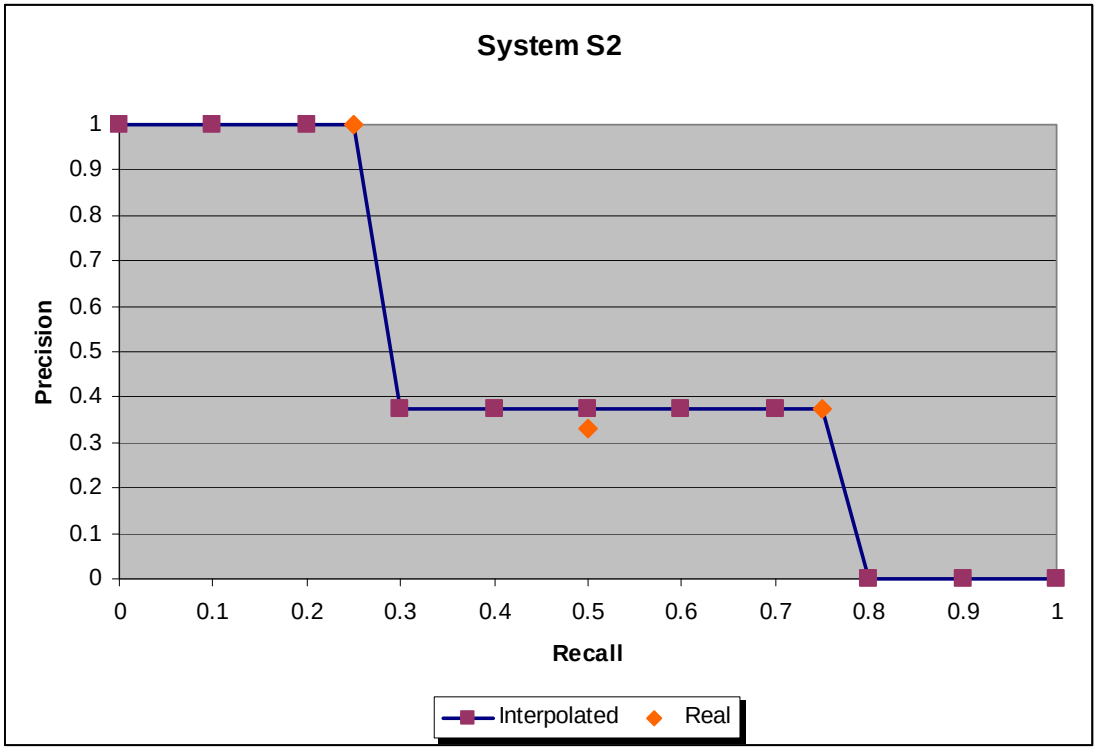
Το S1 επιστρέφει συνολικά 3 έγγραφα:

	Precision (S1)	Recall (S1)
1^o συναφές έγγραφο (d_2)	1 / 2	1 / 4
2^o συναφές έγγραφο (d_7)	2 / 3	2 / 4
3^o συναφές έγγραφο (d_{11})	3 / 4	3 / 4



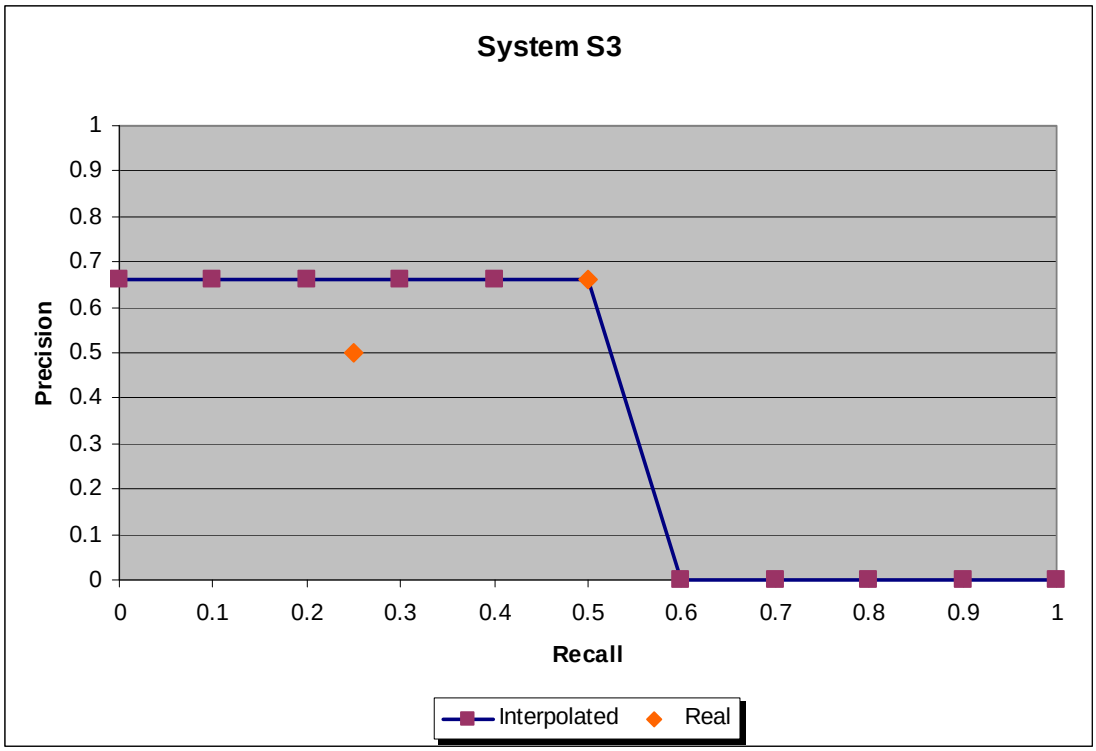
Το S2 επιστρέφει συνολικά 3 έγγραφα:

	Precision (S2)	Recall (S2)
1^{o} συναφές έγγραφο (d_6)	1 / 1	1 / 4
2^{o} συναφές έγγραφο (d_{11})	2 / 6	2 / 4
3^{o} συναφές έγγραφο (d_2)	3 / 8	3 / 4



Το S3 επιστρέφει συνολικά 2 έγγραφα:

	Precision (S3)	Recall (S3)
1^{o} συναφές έγγραφο (d_{11})	1 / 2	1 / 4
2^{o} συναφές έγγραφο (d_7)	2 / 3	2 / 4



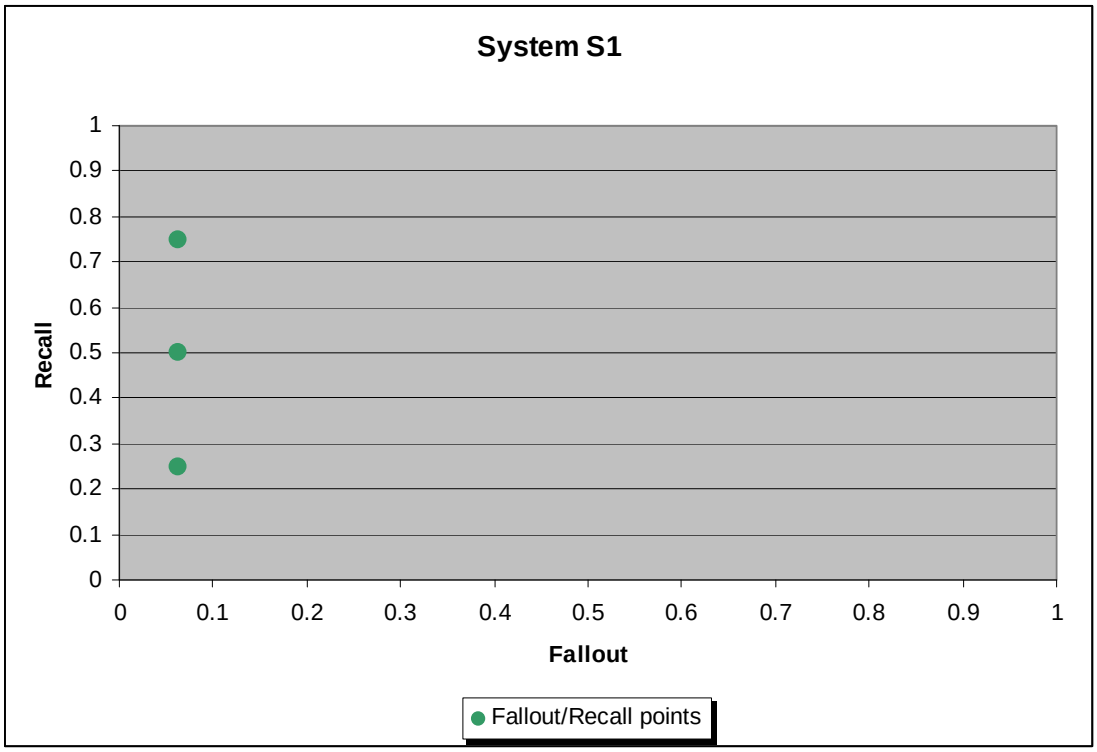
Εάν πρόκειται να διαλέξουμε ένα σύστημα σύμφωνα με τις *precision/recall* καμπύλες τότε θα επιλέξουμε αυτό με τις υψηλότερες τιμές σε *precision* και *recall*. Επομένως ως καλύτερο σύστημα θα επιλέξουμε αυτό του οποίου η καμπύλη τείνει προς την πάνω δεξιά γωνία, όπου *precision,recall = 1,1* και το εμβαδό (η περιοχή κάτω και δεξιά από την καμπύλη) είναι μεγαλύτερο αφού μεγαλύτερο εμβαδό σημαίνει μεγαλύτερες τιμές ακρίβειας και ανάκλησης. Το καλύτερο σύστημα είναι το *S1* όπως φαίνεται και από τη γραφική παράσταση.

β) Recall/Fallout Curves:

Για την προηγούμενη άσκηση υπολογίζουμε τις τιμές των *recall* και *fallout* για κάθε συναφές έγγραφο κάθε συστήματος. Οι τιμές αυτές αποτελούν τα σημεία για τις γραφικές παραστάσεις των συστημάτων. Ισχύουν τα γνωστά επίπεδα ανάκλησης: *{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}*.

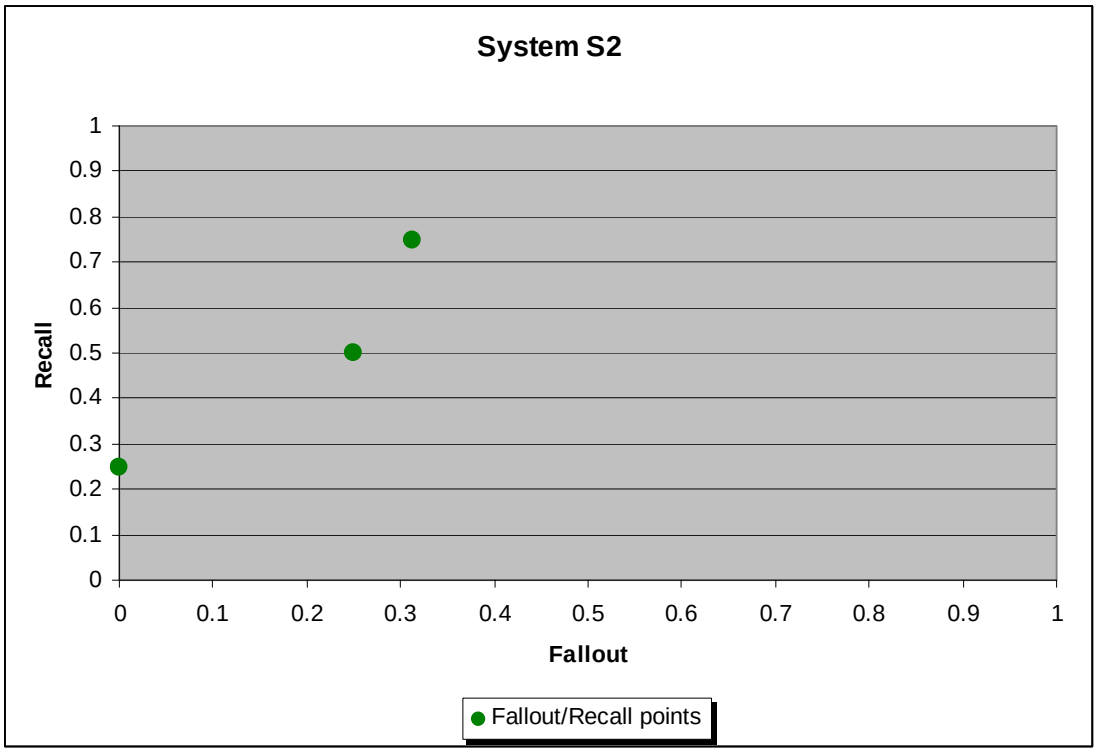
Το S1 επιστρέφει συνολικά 3 έγγραφα:

	Recall (S1)	Fallout (S1)
1^o συναφές έγγραφο (d_2)	1 / 4	1 / 16
2^o συναφές έγγραφο (d_7)	2 / 4	1 / 16
3^o συναφές έγγραφο (d_{11})	3 / 4	1 / 16



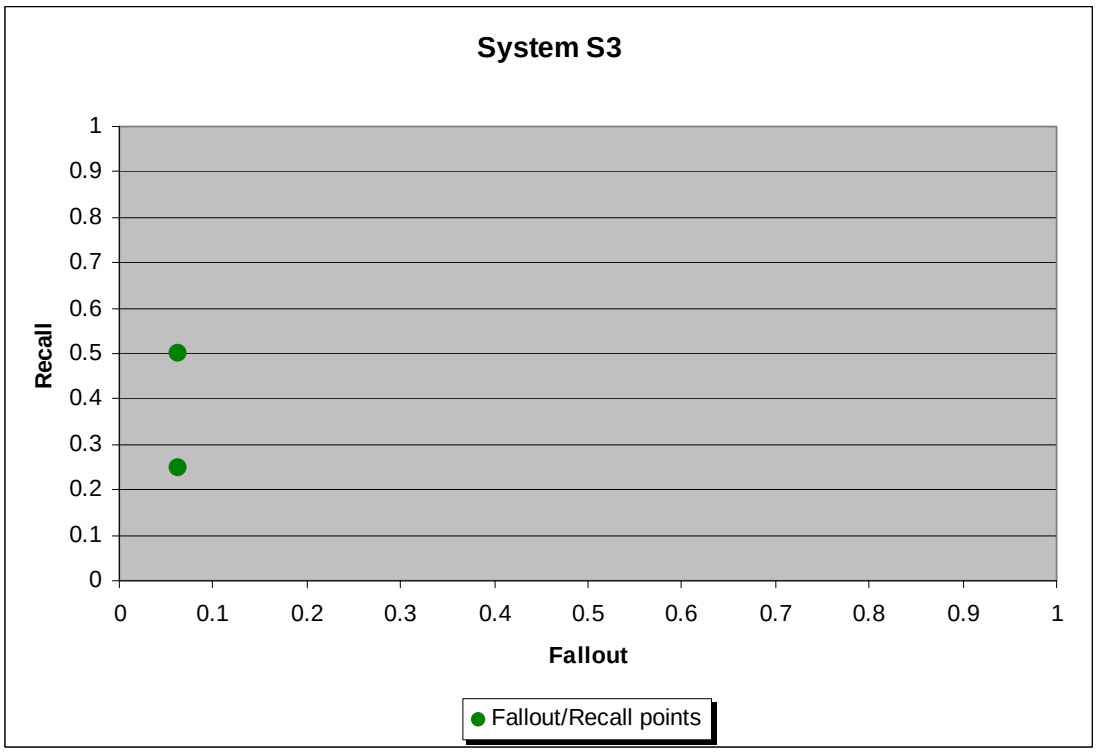
Το S2 επιστρέφει συνολικά 3 έγγραφα:

	Recall (S2)	Fallout (S2)
1^{o} συναφές έγγραφο (d_6)	1 / 4	0 / 16
2^{o} συναφές έγγραφο (d_{11})	2 / 4	4 / 16
3^{o} συναφές έγγραφο (d_2)	3 / 4	5 / 16



Το S3 επιστρέφει συνολικά 2 έγγραφα:

	Recall (S3)	Fallout (S3)
$1^{ο}$ συναφές έγγραφο (d_{11})	1 / 4	1 / 16
$2^{ο}$ συναφές έγγραφο (d_7)	2 / 4	1 / 16

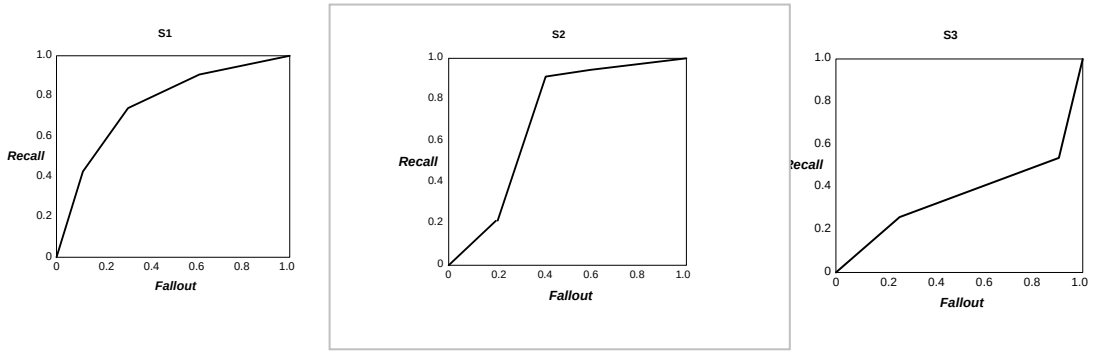


Το καλύτερο σύστημα έχει χαμηλές τιμές *Fallout* για όλες τις τιμές του *Recall*. Άρα επιστρέφει πολύ λίγα (έως καθόλου) μη συναφή έγγραφα στην προσπάθειά του να ανακτήσει τα συναφή έγγραφα ενώ το *Recall* πρέπει να έχει υψηλές τιμές. Ένα ιδανικό σύστημα έχει *Recall = 1* και *Fallout = 0*. Έτσι σύμφωνα με τις καμπύλες *Recall/Fallout* το καλύτερο σύστημα είναι αυτό στο οποίο καταλήξαμε και με τις καμπύλες *Precision/Recall*, δηλαδή το σύστημα *S1*.

Εάν θέλουμε να συντάξουμε την Recall – Fallout καμπύλη για τα 11 επίπεδα Recall (όπως με τις Precision-Recall curves) προκειμένου να μπορούμε να συγκρίνουμε τις επιδόσεις διαφορετικών συστημάτων ακολουθούμε τα εξής βήματα :

- Για τα 11 fallout points $f_0,f_1,f_2,\ldots,f_{11}$ έχουμε
  $R(f_i) = \max R(f_j)$, $f_j \leq f_i$
- Για τα 11 recall points $r_0,r_1,r_2,\ldots,r_{11}$ έχουμε
  $F(r_i) = \max F(r_j)$, $r_j \geq r_i$

γ)



Το σύστημα *S1* προσφέρει πιο αποτελεσματική ανάκτηση πληροφορίας διότι ο δείκτης *fallout* έχει τις πιο χαμηλές τιμές ενώ παράλληλα ο δείκτης *precision* έχει τις πιο υψηλές τιμές σε σχέση με τα άλλα δύο συστήματα *S2*, *S3*. Επίσης, όπως φαίνεται και στις τρεις γραφικές παραστάσεις, το εμβαδό που σχηματίζεται από την καμπύλη *recall-fallout* είναι μεγαλύτερο για το σύστημα *S1* πράγμα που επιβεβαιώνει το αρχικό συμπέρασμα. Πρακτικά αυτό σημαίνει ότι ανακτώνται πολλά συναφή έγγραφα στο σύνολο των ευρεθέντων εγγράφων (υψηλό precision) και παράλληλα ανακτώνται πολύ λίγα μη συναφή έγγραφα στο σύνολο των μη συναφών εγγράφων (χαμηλό *fallout*). Στη σειρά κατάταξης ακολουθούν τα συστήματα *S2* και *S3* με το *S2* να είναι πολύ κοντά στις τιμές του *S1* (εμφανίζει σημεία όπου μπορεί το *fallout* να είναι υψηλότερο, άρα χειρότερο αλλά το *precision* είναι υψηλότερο, άρα καλύτερο). Το *S3* είναι κατά πολύ χειρότερο από τα συστήματα *S1* και *S2*.

Άσκηση 3 (1 βαθμός)

Έστω ότι η συλλογή αξιολόγησης αποτελείται από 50 έγγραφα {d1,…,d50} και γνωρίζουμε ότι υπάρχουν 3 έγγραφα της συλλογής, συγκεκριμένα τα {d1, d2, d3}, 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:

Ans(S1,q) = < d1, {d2, d20-d50}, d3 >

Ans(S2,q) = < d1, d2, d3 >

Ans(S3,q) = < {d1,d8}, d2, d3 >

Η απάντηση < {d1,d8}, d2, d3 > σημαίνει ότι τα d1, d8 ισοβαθμούν στην πρώτη θέση (άρα έλαβαν τον μεγαλύτερο βαθμό συνάφειας). Η απάντηση < d1, {d2, d20-d50}, d3 > σημαίνει ότι το d1 έλαβε το μεγαλύτερο βαθμό , ενώ μετά ακολουθεί μια ομάδα από 32 έγγραφα τα οποία ισοβαθμούν και στο τέλος της κατάταξης βρίσκεται το d3. 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;

<u>Λυση (από Τσικουδη Νικόλαο)</u>

α) R=3;

Για το S1 έχουμε:

Στην 2η και 3η θέση της διάταξης μπορεί είναι 32 διαφορετικά έγγραφα που έχουν λάβει τον ίδιο βαθμό συνάφειας. Μέσα σε αυτά είναι και το d2 το οποίο είναι συναφές για την επερώτηση q. Άρα για διαφορετικές θέσεις του d2 παίρνουμε και διαφορετικές τιμές για το R-Precision.

Το d2 μπορεί να πάρει 32 διαφορετικές θέσεις.
Για τις 2 από αυτές(2η, 3η) έχουμε:
R-Precision(S1) = 2/3 = 0.66
Για τις υπόλοιπες(4η-33η) έχουμε:
R-Precision(S1) = 1/3 = 0.33
Το συνολικό R-precision είναι:
R-Precision (S1) = 2/32*0.66+30/32*0.33 = 0.041+0.31 = 0.351
Για το S2 έχουμε:
R-Precision(S2) = 3/3 = 1
Στις 3 θέσεις τις διάταξης είναι τα 3 συναφή έγγραφα για την επερώτηση q.
Για το S3 έχουμε:
Στις 3 πρώτες θέσεις θα έχουμε πάντα τα d1, d2 και ένα μη-συναφές. Άρα
R-Precision(S3) = 2/3 = 0.66

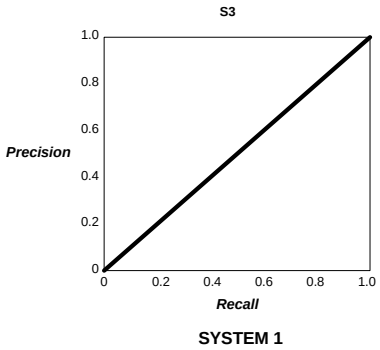
β) Ζητάμε το πλήθος των εγγράφων που πρέπει να αναζητήσουμε για να βρούμε 2 συναφή έγγραφα.
Για το S1 έχουμε:
Το έγγραφο d2 μπορεί να βρίσκεται σε κάποια από τις θέσεις {2η,3η,…,33η}. Ανάλογα με την θέση διαφοροποιείται το μήκος αναζήτησης. Άρα για
d2 στη θέση 2 SearchLength(S1)=2
d2 στη θέση 3 SearchLength(S1)=3
…
d2 στη θέση 33 SearchLength(S1)=33
Άρα SearchLength(S1) = 2+3+…+33/32 = 17.5
Για το S2 έχουμε:
SearchLength(S2) = 2;
Για το S3 έχουμε:
Το d2 που είναι το δεύτερο συναφές έγγραφο το συναντάμε πάντα στη τρίτη θέση άρα
SearchLength(S3) = 3;

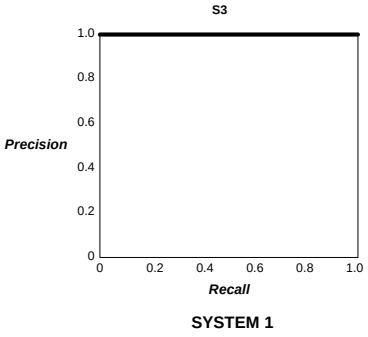
γ) Για το S1 έχουμε:
1ο συναφές SearchLength=1
2ο συναφές SearchLength=17.5
3ο συναφές SearchLength=34
Άρα το μέσο αναμενόμενο μήκος αναζήτησης είναι:
(1/1+17.5/2+34/3)/3 = (1+8.75+11.33)/3 = 7.02
Για το S2 έχουμε:
1ο συναφές SearchLength=1
2ο συναφές SearchLength=2
3ο συναφές SearchLength=3
Άρα το μέσο αναμενόμενο μήκος αναζήτησης είναι:
(1/1+2/2+3/3)/3 = (1+1+1)/3 = 1
Για το S3 έχουμε:
1ο συναφές SearchLength=1.5
2ο συναφές SearchLength=3
3ο συναφές SearchLength=4
Άρα το μέσο αναμενόμενο μήκος αναζήτησης είναι:
(1.5/1+3/2+4/3)/3 = (1.5+1.5+1.33)/3 = 1.44

Άσκηση 4 (0.5 Βαθμοί)

Έστω ένα Σύστημα Ανάκτησης Πληροφοριών που βασίζεται στο λογικό μοντέλο (Boolean Model). Θα είχε νόημα η αξιολόγησή του με
(α) καμπύλες Precision-Recall;

(β) καμπύλες Recall-Fallout;
(γ) αναμενόμενο μήκος αναζήτησης;
(δ) Έστω δυο συστήματα S1 και S2 για τα οποία σας δίδουν τα ακόλουθα διαγράμματα. Σχολιάστε τα. Υπάρχει κάτι που σας παραξενεύει;





Λύση (από Ζαμπετάκη Σταμάτη)

Το κύριο χαρακτηριστικό του boolean μοντέλου είναι ότι τα αποτελέσματα που επιστρέφονται δεν έχουν καμία σχέση διάταξης μεταξύ τους. Η απάντηση δηλαδή είναι απλά ένα σύνολο εγγράφων οπότε οποιαδήποτε μέθοδος αξιολόγησης που λαμβάνει υπόψιν της τη σειρά των αποτελεσμάτων δεν μπορεί να εφαρμοστεί στο boolean μοντέλο.

α)Η αξιολόγηση ενός μοντέλου με καμπύλες Precision-Recall δεν είναι δυνατή και αυτό μπορεί να αποδειχθεί με ένα παράδειγμα. Έστω δύο boolean μοντέλα που ανακτούν τα ίδια ακριβώς έγγραφα αλλά τα επιστρέφουν με διαφορετική σειρά. Εφόσον η σειρά ανάκτησης δεν παίζει ρόλο στα boolean μοντέλο οποιαδήποτε μέθοδος αξιολόγησης θα έπρεπε να τους αποδίδει την ίδια βαθμολογία.Κάνοντας ποιο συγκεκριμένο το παράδειγμα έχουμε τις εξής απαντήσεις για τα δύο μοντέλα.
Ans(S1,q) = <d1,d2,d3,d4,d5,d6>
Ans(S2,q) = <d4,d5,d6,d1,d2,d3>
έστω Σ = {d1,d2,d3}.

Οι καμπύλες Precision-Recall σχεδιάζονται παίρνοντας τις απαντήσεις σαν ένα διατεταγμένο σύνολο εγγράφων και υπολογίζουμε τα Precision και Recall για τα συναφή έγγραφα. Ακολουθόντας αυτή την διαδικασία θα οδηγηθούμε σε δύο διαφορετικές καμπύλες,κάτι δηλαδή που μας λέει ότι το ένα σύστημα είναι καλύτερο από το άλλο, πράγμα που προφανώς δεν ισχύει καθώς υποθέσαμε ότι τα μοντέλα είναι ισοδύναμα και επομένως το αναμενόμενο αποτέλεσμα θα ήταν οι καμπύλες να είναι οι ίδιες.

β)Χρησιμοποιώντας το παραπάνω παράδειγμα και τις παραπάνω υποθέσεις κατασκευάζοντας τις καμπυλές Recall-Fallout θα οδηγηθούμε σε δύο διαφορετικές επομένως η αξιολόγησή του boolean μοντέλου δεν μπορεί να γίνει ούτε με Recall-Fallout καμπύλες.

γ)Εφόσον και στον υπολογισμό του αναμενόμενου μήκους χρησιμοποιούμε τη διάταξη των εγγράφων τα δύο παραπάνω συστήματα θα λάβουν διαφορετική βαθμολόγηση επομένως στο boolean μοντέλο δεν μπορεί να χρησιμοποιηθεί ούτε αυτός ο τρόπος αξιολόγησης.
Καλά μέτρα αξιολόγησης για το boolean μοντέλο μπορούν να θεωρηθούν τα F-Measeure,E-Measure και Fallout που δεν έχουν να κάνουν καθόλου με διάταξη εγγράφων και το πως θα επιστραφούν αυτά από το σύστημα ανάκτησης.

δ)Στο πρώτο σύστημα αρχικά μπορούμε να πούμε ότι δεν έχουν χρησιμοποιηθεί κανονικοποιημένα επίπεδα ανάκλησης γιατί άμα συνέβαινε αυτό θα έπρεπε να είχαμε σταθερές τιμές στα διαστήματα 0.1-0.2,0.2-0.3 κτλ. Επίσης αυτό το σύστημα δεν μπορεί να προκύψει ποτέ στην πράξη καθώς παρατηρούμε ότι έχουμε τιμές ακόμα και στο (0,0) δηλαδή recall= 0 και precision = 0.Όταν παίρνουμε τα σημεία P/R υπολογίζουμε precision και recall στα συναφή έγγραφα επομένως η τιμή (0,0) δεν μπορεί να εμφανιστεί ποτέ. Επίσης παρατηρούμε ότι η συνάρτηση(καμπύλη) κάποια στιγμή φτάνει στο (1,1) κάτι που επίσης δεν μπορεί να συμβεί εκτός και αν το σύστημα μας επιστρέψει μόνο συναφή και μάλιστα όλα αλλά εφόσον υπάρχουν και μικρότερες τιμές precision αυτό σημαίνει ότι το σύστημα μας επέστρεψε και άσχετα έγγραφα επομένως αν το σύστημα κάποια στιγμή επέστρεψε κάποιο άσχετο έγγραφο αυτόματα έχασε την δυνατότητα να φτάσει το precision = 1.Από θεωρία επίσης είναι γνωστό ότι σε κανονικοποιημένα επίπεδα ανάκλησης οι καμπύλες που προκύπτουν είναι φθίνουσες βάση ορισμού.

Τώρα εξετάζοντας το δεύτερο σύστημα μας παρουσιάζετε ένα σύστημα που έχει συνέχεια precision 1 δηλαδή επιστρέφει μόνο συναφή έγγραφα και μάλιστα βλέπουμε ότι φτάνει και το recall 1 δηλάδη σε κάποιο σημείο θα τα επιστρέψει όλα. Επομένως πρόκειται για το ιδανικό σύστημα.

Άσκηση 5(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:

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

<u>Λύση</u>

Μπορούμε να στηρίξουμε την απάντηση μας χρησιμοποιώντας ως παράδειγμα τα έγγραφα της άσκησης 6

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»

και ως επερώτηση την επερώτηση q_1 = «information retrieval»

	R_1	**R_2**	**R_3**	**R_4**
d_1	2/2 = 1	2/2 = 1	2/2 = 1	4/2 = 2
d_2	2/2 = 1	2/2 = 1	2/2 = 1	4/2 = 2

d_3	1/2 = 0.5	1/2 = 0.5	1/3 = 0.33	2/3 = 0.667
d_4	2/2 = 1	2/3 = 0.667	2/3 = 0.667	4/3 = 1.33
d_5	2/2 = 1	2/3 = 0.667	2/3 = 0.667	4/3 = 1.33
Ranking	$<\{d_1, d_2, d_4, d_5\}, d_3>$	$<\{d_1, d_2\},\{d_4, d_5\}, d_3>$	$<\{d_1, d_2\},\{ d_4, d_5\}, d_3>$	$<\{d_1, d_2\},\{ d_4, d_5\}, d_3>$

Συνάρτηση Διαβάθμισης R_1

Έχουμε,

$$R_1(d, q) = \frac{|d \cap q|}{|q \cap d| + |q \setminus d|} = \frac{|d \cap q|}{|q|}$$

Ο παρονομαστής q είναι πάντα ο ίδιος, επομένως δεν λαμβάνονται υπόψη οι λέξεις του κάθε εγγράφου που δεν ταιριάζουν με την επερώτηση. Έτσι τα έγγραφα που εκτός από τις λέξεις της επερώτησης περιέχουν και άλλες θα έχουν την ίδια διαβάθμιση με τα έγγραφα που περιέχουν μόνο τις λέξεις της επερώτησης**.**

Συνάρτηση Διαβάθμισης R_2

Έχουμε,

$$R_2(d, q) = \frac{|d \cap q|}{|q \cap d| + |d \setminus q|} = \frac{|d \cap q|}{|d|}$$

Ο παρονομαστής είναι πάντα διαφορετικός άρα λαμβάνει υπόψη το μέγεθος του κάθε εγγράφου και κατ' επέκταση το ποσοστό του εγγράφου που δεν έχουμε ταίριασμα. Οπότε θα έχουμε διαφορετική διαβάθμιση για τα έγγραφα που περιέχουν μόνο τους όρους της επερώτησης και αυτά που περιέχουν και άλλους. Όμως έστω ότι είχαμε ένα επιπλέον έγγραφο d_6= "information", το οποίο ο μόνος όρος που περιέχει είναι ένας όρος της επερώτησης («information retrieval»), τότε $R_2(d_6, q) = 1$, ενώ $R_2(d_5, q) = 0.667$. Άρα το έγγραφο d_6 έχει μεγαλύτερη συνάφεια από το έγγραφο d_5, παρόλο που το έγγραφο d_5 περιέχει και τους 2 όρους της επερώτησης. Επίσης η R_2 ευνοεί τα μικρά έγγραφα.

Συνάρτηση Διαβάθμισης R_3

Έχουμε,

$$R_3(d, q) = \frac{|d \cap q|}{|d \setminus q| + |q \setminus d| + |d \cap q|} = \frac{|d \cap q|}{|d \cup q|}$$

Άρα η συγκεκριμένη συνάρτηση διαβάθμισης λαμβάνει υπόψη όχι μόνο το ποσοστό του εγγράφου στο οποίο δεν έχουμε ταίριασμα αλλά και το ποσοστό της επερώτησης στο οποίο δεν έγινε ταίριασμα. Η συνάρτηση αυτή επιστρέφει 1, αυστηρά μόνο στην περίπτωση που το έγγραφο είναι απολύτως σχετικό με την επερώτηση, δηλαδή περιέχει μόνο τους όρους της επερώτησης.

Συνάρτηση Διαβάθμισης R_4

Η συνάρτηση διαβάθμισης R_4, ομοίως με την R_4 λαμβάνει υπόψη όχι μόνο το ποσοστό του εγγράφου στο οποίο δεν έχουμε ταίριασμα αλλά και το ποσοστό της επερώτησης στο οποίο δεν έγινε ταίριασμα. Η συνάρτηση αυτή παίρνει τιμές στο διάστημα [0,2] και επιστρέφει 2, αυστηρά μόνο στην περίπτωση που το έγγραφο είναι απολύτως σχετικό με την επερώτηση, δηλαδή περιέχει μόνο τους όρους της επερώτησης.

Με βάσει τις παραπάνω παρατηρήσεις βλέπουμε ότι την καλύτερη κατάταξη εγγράφων δίνουν οι συναρτήσεις R_3 και R_4.

Άσκηση 6(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: «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»

Λύση

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

Term Occurrence Table

	course	**information**	**retrieval**	**MAXk{FREQi,j}**
d_1	0	2	2	2
d_2	0	1	1	1
d_3	1	2	0	2
d_4	1	2	2	2
d_5	1	1	1	1
df	3	5	4	-
idf	5 / 3	5 / 5	5 / 4	-

- *FREQi,j* : το πλήθος των εμφανίσεων του όρου i στο έγγραφο j
- *N* = 5
- *IDF = N / DF*
- *MAXk{FREQi,j}*: συχνότητα της λέξης με τη μέγιστη συχνότητα στο κείμενο

Term Weight Table

	course	**information**	**retrieval**	**MAXk{FREQi,j}**
d_1	0	2/2 * 5/5	2/2 * 5/4	2
d_2	0	1/1 * 5/5	1/1 * 5/4	1
d_3	1/2 * 5/3	2/2 * 5/5	0	2
d_4	1/2 * 5/3	2/2 * 5/5	2/2 * 5/4	2
d_5	1/1 * 5/3	1/1 * 5/5	1/1 * 5/4	1
df	3	5	4	-

idf	5/3	5/5	5/4	-

- *TFi,j = FREQi,j / MAXk{FREQi,j}*
- *Vi,j = TFi,j * IDFi*

Οι διανυσματικές αναπαραστάσεις των εγγράφων d_1,d_2,d_3,d_4,d_5 είναι οι εξής:

$V_1 = \{0, 1, 1.25\}$, $|V_1| = 2.5625$
$V_2 = \{0, 1, 1.25\}$, $|V_2| = 2.5625$
$V_3 = \{0.83, 1, 0\}$, $|V_3| = 1.6889$
$V_4 = \{0.83, 1, 1.25\}$, $|V_4| = 3.2514$
$V_5 = \{1.66, 1, 1.25\}$, $|V_5| = 5.3181$

(β)

	course	**information**	**retrieval**
q_1	0	1/1 * 5/5	1/1 * 5/4
q_2	1/1 * 5/3	0	0
q_3	1/1 * 5/3	1/1 * 5/5	0
idf	5/3	5/5	5/4

$q_1 = \{0, 1, 1.25\}$, $|q_1| = 2.5625$
$q_2 = \{1.66, 0, 0\}$, $|q_2| = 2.7556$
$q_3 = \{1.66, 1, 0\}$, $|q_3| = 3.7556$

Για κάθε έγγραφο υπολογίζουμε το μέτρο ομοιότητας συνημίτονου:

$$R(d_j, q) = CosSim(d_j, q) = \frac{\vec{d_j} \bullet \vec{q}}{|\vec{d_j}| x |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{i,j} x w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} x \sqrt{\sum_{i=1}^{t} w_{i,q}^2}}$$

Άρα για την **επερώτηση q_1** έχουμε:

$V_1 * q_1 = 0 * 0 + 1 * 1 + 1.25 * 1.25 = 2.5625$
$V_2 * q_1 = 0 * 0 + 1 * 1 + 1.25 * 1.25 = 2.5625$
$V_3 * q_1 = 0.83 * 0 + 1 * 1 + 0 * 1.25 = 1$
$V_4 * q_1 = 0.83 * 0 + 1 * 1 + 1.25 * 1.25 = 2.5625$
$V_5 * q_1 = 1.66 * 0 + 1 * 1 + 1.25 * 1.25 = 2.5625$

- $R(d_1, q_1) = \frac{2.5625}{\sqrt{2.5625 * 2.5625}} = \frac{2.5625}{2.5625} = 1$
- $R(d_2, q_1) = \frac{2.5625}{\sqrt{2.5625 * 2.5625}} = \frac{2.5625}{2.5625} = 1$
- $R(d_3, q_1) = \frac{1}{\sqrt{1.6889 * 2.5625}} = 0,480$
- $R(d_4, q_1) = \frac{2.5625}{\sqrt{3.2514 * 2.5625}} = \frac{2.5625}{2.886} = 0,887$
- $R(d_5, q_1) = \frac{2.5625}{\sqrt{5.341389 * 2.5625}} = \frac{2.5625}{3.6996} = 0,693$

Άρα βάσει του διανυσματικού μοντέλου η κατάταξη των εγγράφων για την επερώτηση q_1 είναι η εξής:
< {d_1, d_2}, d_4, d_5, d_3 >

Τα έγγραφα d_1, d_2 βρίσκονται στην 1η θέση επειδή περιέχουν όλους τους όρους της επερώτησης και μόνο αυτούς

Για την **επερώτηση q_2** έχουμε:

$V_1 * q_2 = 0 * 1.66 + 1 * 0 + 1.25 * 0 = 0$
$V_2 * q_2 = 0 * 1.66 + 1 * 0 + 1.25 * 0 = 0$
$V_3 * q_2 = 0.83 * 1.66 + 1 * 0 + 1.25 * 0 = 1.3778$
$V_4 * q_2 = 0.83 * 1.66 + 1 * 0 + 1.25 * 0 = 1.3778$
$V_5 * q_2 = 1.66 * 1.66 + 1 * 0 + 1.25 * 0 = 2.7556$

- $R(d_1, q_2) = 0$
- $R(d_2, q_2) = 0$
- $R(d_3, q_2) = \frac{1.3778}{\sqrt{1.6889 * 2.7556}} = 0{,}638$
- $R(d_4, q_2) = \frac{1.3778}{\sqrt{3.2514 * 2.7556}} = 0{,}460$
- $R(d_5, q_2) = \frac{2.7556}{\sqrt{5.341389 * 2.7556}} = 0{,}718$

Άρα βάσει του διανυσματικού μοντέλου η κατάταξη των εγγράφων για την επερώτηση q_2 είναι η εξής:
< d_5, d_3.d_4 >

Τα έγγραφα d_1,d_2 δεν είναι συναφή με την επερώτηση q_3, αφού δεν περιέχουν καθόλου τον όρο ‘course’ και για αυτό η συνάφεια τους με την επερώτηση αυτή είναι 0.

Για την **επερώτηση q_3** έχουμε:

$V_1 * q_3 = 0 * 1.66 + 1 * 1 + 1.25 * 0 = 1$
$V_2 * q_3 = 0 * 1.66 + 1 * 1 + 1.25 * 0 = 1$
$V_3 * q_3 = 0.83 * 1.66 + 1 * 1 + 1.25 * 0 = 2.3778$
$V_4 * q_3 = 0.83 * 1.66 + 1 * 1 + 1.25 * 0 = 2.3778$
$V_5 * q_3 = 1.66 * 1.66 + 1 * 1 + 1.25 * 0 = 3.7556$

- $R(d_1, q_3) = \frac{1}{\sqrt{2.5625 * 3.7556}} = 0{,}322$
- $R(d_2, q_3) = \frac{1}{\sqrt{2.5625 * 3.7556}} = 0{,}322$
- $R(d_3, q_3) = \frac{2.3778}{\sqrt{1.6889 * 3.7556}} = 0{,}944$
- $R(d_4, q_3) = \frac{2.3778}{\sqrt{3.2514 * 3.7556}} = 0{,}680$
- $R(d_5, q_3) = \frac{3.7556}{\sqrt{5.341389 * 3.7556}} = 0{,}839$

Άρα βάσει του διανυσματικού μοντέλου η κατάταξη των εγγράφων για την επερώτηση q_3 είναι η εξής:
< {d_3, d_5, d_4,{ d_1, d_2} >

(γ) Ανεστραμμένο Ευρετήριο για την συλλογή D

Term	**< Document Frequency, (Document; Position) >**
Course	< 3 (d3;3), (d4;1), (d5;3) >
Information	< 5 (d1;1), (d1;4), (d2;2), (d3;1), (d3;2), (d4;3), (d4;4), (d5;2) >
Retrieval	< 4 (d1;2), (d1;3), (d2;1), (d4;2), (d4;5), (d5;1) >

Άσκηση 7(0.5 βαθμοί)

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 /*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 /2 Microsoft»

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

<u>Λύση</u>

(α) Το σύνολο των συναφών εγγράφων για την επερώτηση «Gates /2 Microsoft» είναι το {1,3}

- Στο έγγραφο 1, η λέξη Gates βρίσκεται στην $3^η$ θέση και η λέξη Microsoft στην $1^η$ θέση, άρα η απόσταση τους είναι 2 θέσεις, οπότε το έγγραφο αυτό ικανοποιεί την επερώτηση «Gates /2 Microsoft».
- Στο έγγραφο 3, λέξη Gates βρίσκεται στην $2^η$ θέση και η λέξη Microsoft στην $3^η$ θέση , άρα η απόσταση τους είναι 1 θέση, οπότε το έγγραφο αυτό ικανοποιεί την επερώτηση «Gates /2 Microsoft».

(β) Εφόσον οι όροι «Gates» και «Microsoft» εμφανίζονται από κοινού μόνο στα έγγραφα 1,2,3, τότε εξετάζουμε μόνο αυτά τα έγγραφα ως υποψήφια. Σε αυτά τα έγγραφα οι πιθανές

τιμές του κ, που έχει νόημα να εξετάσουμε είναι {1,2,...,15}, αφού οι λέξεις αυτές εμφανίζονται σε απόσταση το πολύ 15 λέξεων στα έγγραφα αυτά.

- Για $\kappa = 1$, επιστρέφεται μόνο το έγγραφο 3, αφού είναι το μοναδικό έγγραφο στο οποίο οι λέξεις «Gates», «Microsoft» είναι γειτονικές (βρίσκονται στις θέσεις 2,3 του εγγράφου αντίστοιχα)
- Για $2 \leq \kappa \leq 4$ επιστρέφονται τα έγγραφα {1,3}
- Για $5 \leq \kappa \leq 15$ επιστρέφονται τα έγγραφα {1,2,3}