ΜΥΕ003: Ανάκτηση Πληροφορίας

Διδάσκουσα: Ευαγγελία Πιτουρά

Αξιολόγηση στην Ανάκτηση Πληροφορίας.

Ακαδημαϊκό Έτος 2022-2023

Τι θα δούμε σήμερα;

- Πως μπορούμε να *αξιολογήσουμε* ένα σύστημα ανάκτησης;
- Ποια τεχνική ή ποιο σύστημα ανάκτησης πληροφορίας είναι *καλύτερο*;

- Αξιολόγηση συστημάτων ανάκτησης πληροφορίας και μηχανών αναζήτησης:
  (1) μεθοδολογία και
  (2) μέτρα

Αξιολόγηση συστήματος

Αποδοτικότητα (Performance)

- Πόσο *μεγάλο* είναι το ευρετήριο (*αποθήκευση*);
- Πόσο *γρήγορη* είναι η *κατασκευή του ευρετηρίου*;
  - Αριθμός εγγράφων την ώρα (throughput)
- Πόσο *γρήγορη* είναι η *αναζήτηση*;
  - π.χ., latency (χρόνος απόκρισης) ή throughput (ρυθμό-απόδοση) ως συνάρτηση των ερωτημάτων ανά δευτερόλεπτο ή του μεγέθους του ευρετηρίου

Εκφραστικότητα της γλώσσας ερωτημάτων

επιτρέπει τη διατύπωση περίπλοκων αναγκών πληροφόρησης;

Ποιο είναι το κόστος ανά ερώτημα;

- Π.χ., σε δολάρια

Μέτρα για μηχανές αναζήτησης

- Όλα αυτά τα κριτήρια είναι μετρήσιμα (measurable):

  μπορούμε να ποσοτικοποιήσουμε την ταχύτητα/μέγεθος/χρήματα και να κάνουμε την εκφραστικότητα συγκεκριμένη

- Ωστόσο μια βασική μέτρηση για μια μηχανή αναζήτησης είναι η ικανοποίηση των χρηστών (user happiness)

Μέτρα για μηχανές αναζήτησης

- *Τι κάνει ένα χρήστη χαρούμενο;*
  - Οι παράγοντες περιλαμβάνουν:
    - Ταχύτητα απόκρισης (Speed of response)
    - Μέγεθος/κάλυψη ευρετηρίου
    - *Εύχρηστη διεπαφή* (Uncluttered UI)
    - Χωρίς κόστος (free)
  - Συνάφεια (relevance): *Κανένα από αυτά δεν αρκεί*: εξαιρετικά γρήγορες αλλά άχρηστες απαντήσεις δεν ικανοποιούν ένα χρήστη
- Θα επικεντρωθούμε στο πως «μετράμε» τη συνάφεια;
- Effectiveness (αποτελεσματικότητα) vs Efficiency (αποδοτικότητα)

Βασικό κριτήριο: Συνάφεια

Η ικανοποίηση του χρήστη συνήθως εξισώνεται με τη συνάφεια (relevance) των αποτελεσμάτων της αναζήτησης με το ερώτημα

Μα πως θα μετρήσουμε τη συνάφεια;

Συνάφεια και Ανάγκη Πληροφόρησης

- Συνάφεια ως προς τι;
  - Συνάφεια ως προς το ερώτημα ή ως προς την ανάγκη πληροφόρησης (information need)

Παράδειγμα

Ανάγκη Πληροφόρησης *i*: «Ψάχνω για πληροφορία σχετικά με το αν το κόκκινο κρασί είναι πιο αποτελεσματικό από το λευκό κρασί για τη μείωση του ρίσκου για καρδιακή προσβολή»

Μεταφράζεται στην ερώτημα:

Ερώτημα *q*: [red wine white wine heart attack]

Έγγραφο *d*: At heart of his speech was an attack on the wine industry lobby for downplaying the role of red and white wine in drunk driving.

- *d* άριστο ταίριασμα στο ερώτημα *q*
- *d* δεν είναι συναφές με την ανάγκη πληροφόρησης *i*

Συνάφεια και Ανάγκη Πληροφόρησης

- Η ικανοποίηση του χρήστη σχετίζεται με τη συνάφεια ως *προς την ανάγκη πληροφόρησης* και όχι ως προς το *ερώτημα*

- Το ακριβές είναι *συνάφεια έγγραφου-ανάγκης πληροφόρησης* αν και συνήθως χρησιμοποιούμε *συνάφεια εγγράφου-ερωτήματος*.

Μεθοδολογία: Benchmarks

Η καθιερωμένη μεθοδολογία στην Ανάκτηση Πληροφορίας αποτελείται από τρία στοιχεία:

1. Μία πρότυπη *συλλογή εγγράφων* (benchmark document collection)
2. Μια πρότυπη *ομάδα ερωτημάτων* (benchmark suite of queries)
3. Μια *αποτίμηση της συνάφειας* για κάθε ζεύγος ερωτήματος-εγγράφου, συνήθως δυαδική: συναφής (R) - μη συναφής (N) – (*gold standard/ground truth*) που μας λέει αν το έγγραφο είναι συναφές ως προς το ερώτημα

Μέτρα Συνάφειας

Απλό παράδειγμα

q1	d1 (0) d2 (0) d3 (5) ….	d1000(1)
q2	d1 (1) d2 (0) d3 (1) ….	d1000(0)
..		
qn	d1 (4) d2 (0) d3 (0) ….	d1000(2)

Ερώτημα

Έγγραφα, σε παρένθεση ο βαθμός συνάφειας σε κλίμακα από 0 έως 5

Μέτρα Συνάφειας

- Δεδομένης της αποτίμησης των αποτελεσμάτων ενός συστήματος (ground truth) πως εκτιμάμε τη συνάφεια του συστήματος;

q1 d209 (5) d63 (5) d5 (4) …
q2 d12 (5) d140 (5) d245 (5) …
..
qn d109 (5) d134 (5) d15 (5) …

Ground thruth: Έγγραφα σε διάταξη με βάση τη συνάφεια τους στο ερώτημα - σε παρένθεση ο βαθμός συνάφεια

Ένα σύστημα ανάκτησης πληροφορίας επιστρέφει
για το q1 d209 d5 d12 d299 d63
για το q2 d12 d140 d19 d120
..
για το qn d209 d15 d134 d109 d22

Πόσο καλό είναι αυτό το σύστημα;

Μέτρα Συνάφειας

- Δεδομένης της αποτίμησης των αποτελεσμάτων ενός συστήματος (ground truth) πως εκτιμάμε τη συνάφεια του συστήματος;
- Θα ορίσουμε σχετικά μέτρα
- Το μέτρο υπολογίζεται *για κάθε ερώτημα* και παίρνουμε το *μέσο όρο* για το σύνολο των ερωτημάτων
- Αρχικά, θα θεωρήσουμε δυαδικές αξιολογήσεις: Συναφές/Μη Συναφές

Μέτρα Συνάφειας

Δυο κατηγορίες μέτρων:

- Μέτρα που αγνοούν τη διάταξη
- Μέτρα που λαμβάνουν υπ’ όψιν τη διάταξη

Θα δούμε στην αρχή *μέτρα που αγνοούν τη διάταξη*

Όχι διάταξη
Δυαδική συνάφεια

- **Δυαδική συνάφεια** (Συναφές, Μη συναφές)
  - Ground truth: για κάθε ερώτημα q, λίστα με τα συναφή έγγραφα

- **Αγνοούμε τη διάταξη**
  - για παράδειγμα, το αποτέλεσμα d15 d22 d30 θεωρείτε ίδιο με πχ d22 d15 d30

Μέτρα Συνάφειας χωρίς Διάταξη

Τα αποτελέσματα μιας ερώτησης θεωρούνται ως *σύνολο*, δηλαδή αξιολογούμε τη συνάφεια ενός συνόλου (δεν υπάρχει διάταξη)

Παράδειγμα:

Έστω μια συλλογή με *1,000,120* έγγραφα, και μια *ερώτηση q* για την οποία οι οποία έχει *80 συναφή* έγγραφα.

Η απάντηση που μας δίνει το ΣΑΠ έχει *60* έγγραφα. Από τα 60 αυτά έγγραφα τα *20* είναι συναφή (τα υπόλοιπα *40* δεν είναι).

- Πόσο «καλό» είναι;
- Πως θα μετρήσουμε τη συνάφεια του;

Πίνακας Ενδεχομένων
Confusion/Incidence Matrix

Ground truth

Predicted

	Συναφή (relevant)	**Μη συναφή (not relevant)**	
Ανακληθέντα (retrieved)	20 **(TP)**	40 **(FP)**	60
Μη ανακληθέντα (not retrieved)	60 **(FN)**	1,000,000 **(TN)**	1,000,060
	80	1,000,040	1,000,120

Παράδειγμα

Το διαγνωστικό τεστ για ένα άτομο που πάσχει από Covid19 είναι αρνητικό. Αυτό είναι:

A. TN

B. TP

C. FN

D. FP

E. Τίποτα από τα παραπάνω

Παράδειγμα

ΣΑΠ με 5000 έγγραφα. Σε ερώτημα για το οποίο το συναφή έγγραφα είναι 100, επιστρέφει 20 από τα οποία συναφή είναι 15.

Ακρίβεια και Ανάκληση

- Precision (P) – Ακρίβεια είναι το ποσοστό των ανακτημένων εγγράφων που είναι συναφή

$$\text{Precision} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} = P(\text{relevant}|\text{retrieved})$$

- Recall (R) – Ανάκληση είναι το ποσοστό των συναφών εγγράφων που ανακτώνται

$$\text{Recall} = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})} = P(\text{retrieved}|\text{relevant})$$

Ακρίβεια και Ανάκληση

Πίνακας Ενδεχόμενων (Incidence Matrix)

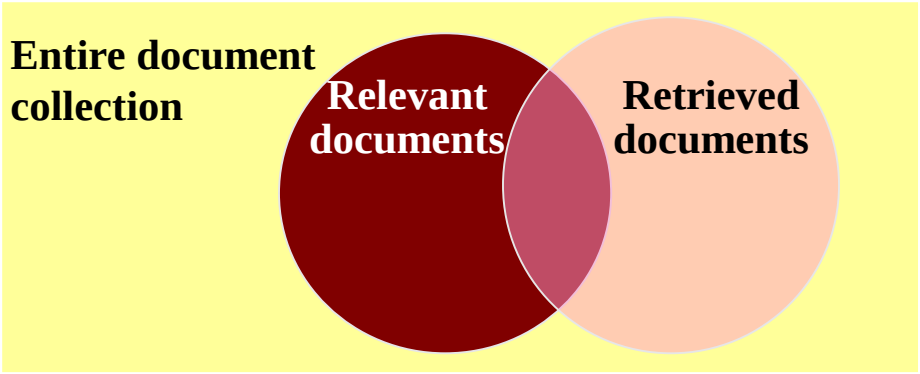
πραγματικά

αποτέλεσμα

	Relevant	Nonrelevant
Retrieved	true positives (TP)	false positives (FP)
Not retrieved	false negatives (FN)	true negatives (TN)

$P = TP / (TP + FP)$

$R = TP / (TP + FN)$



Πίνακας Ενδεχομένων (Incidence Matrix)

	Συναφή (relevant)	**Μη συναφή (not relevant)**	
Ανακληθέντα (retrieved)	20 **(TP)**	40 **(FP)**	60
Μη ανακληθέντα (not retrieved)	60 **(FN)**	1,000,000 **(TN)**	1,000,060
	80	1,000,040	1,000,120

Precision = 20/60 = 1/3

Recall = 20/80 = 1/4

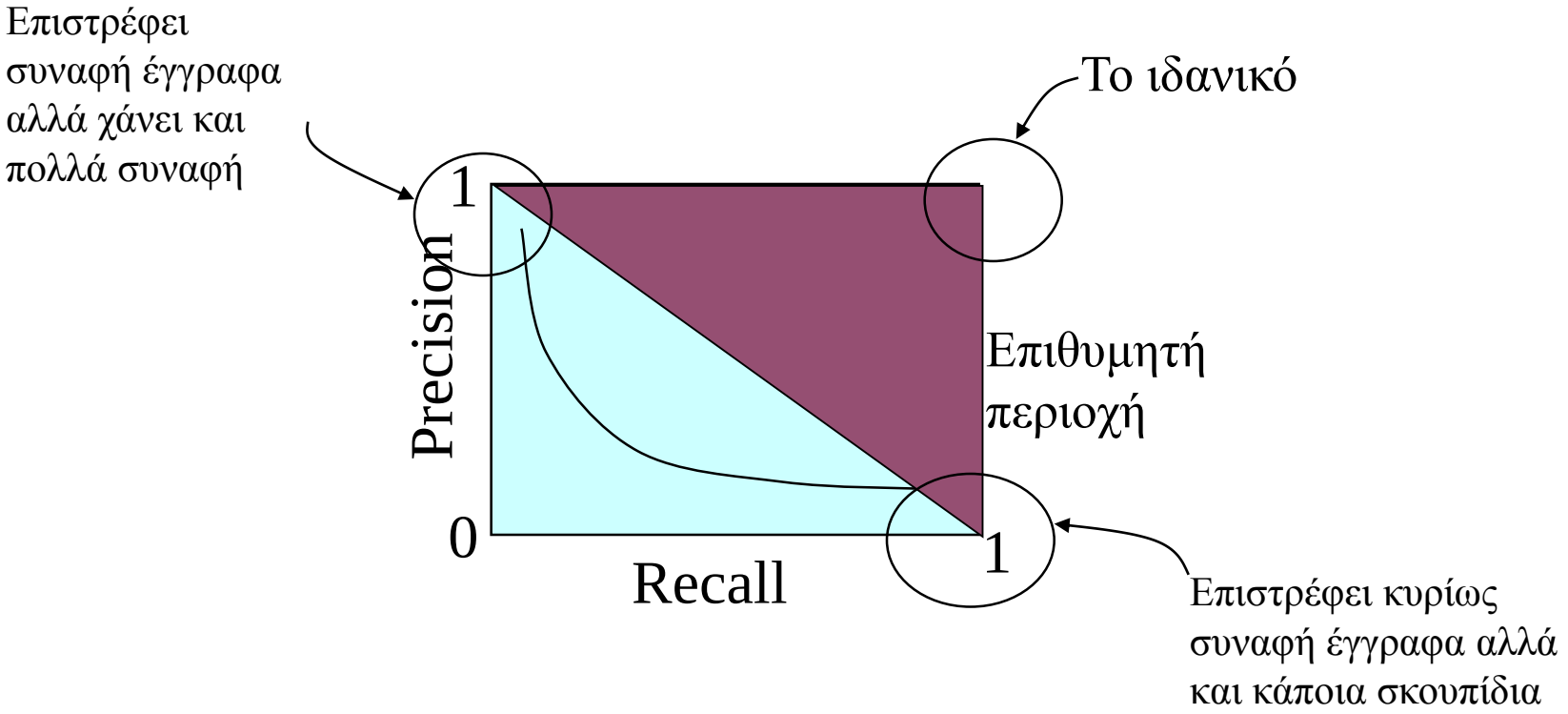
Ακρίβεια vs Ανάκληση

- Η ανάκληση μπορεί να *αυξηθεί* με το να επιστρέψουμε *περισσότερα έγγραφα*
  - Η ανάκληση είναι μια μη-φθίνουσα συνάρτηση των εγγράφων που ανακτώνται (Ένα σύστημα που επιστρέφει όλα τα έγγραφα έχει ποσοστό ανάκλησης 100%!)
- Το αντίστροφο ισχύει για την ακρίβεια (συνήθως)
  - *Είναι εύκολο να πετύχεις μεγάλη ακρίβεια με πολύ μικρή ανάκληση* (Έστω ότι το έγγραφο με το μεγαλύτερο βαθμό είναι συναφές. Πως μπορούμε να μεγιστοποιήσουμε την ακρίβεια;)

Σε ένα καλό σύστημα η ακρίβεια ελαττώνεται όσο περισσότερα έγγραφα ανακτούμε ή με την αύξηση της ανάκλησης

Το τι από τα δύο μας ενδιαφέρει περισσότερα εξαρτάται και από την εφαρμογή (π.χ., web vs email/desktop search)

Ακρίβεια και Ανάκληση



Αρμονικό Μέσο

Πως θα συνδυάσουμε το *P* και *R*;

- Το *αριθμητικό μέσο* (arithmetic mean) (P+R)/2
  - Το απλό αριθμητικό μέσο της ακρίβειας και ανάκλησης μιας μηχανής αναζήτησης που επιστρέφει *τα πάντα* είναι
    A. ~50%
    B. 0%
    C. ~100%
    D. ~25%

  Θα θέλαμε με κάποιο τρόπο να τιμωρήσουμε *την πολύ κακή συμπεριφορά* σε οποιοδήποτε από τα δύο μέτρα.

Αρμονικό Μέσο

Θα θέλαμε με κάποιο τρόπο να τιμωρήσουμε *την πολύ κακή συμπεριφορά* σε οποιοδήποτε από τα δύο μέτρα.

- Αυτό επιτυγχάνεται παίρνοντας το *ελάχιστο*
  - Αλλά το ελάχιστο είναι λιγότερο ομαλό (smooth) και είναι δύσκολο να σταθμιστεί
- Γεωμετρικό μέσο (geometric mean): (ρίζα του) γινόμενου
  - $\sqrt{PR}$
- Το F (αρμονικό μέσο) είναι ένα είδος ομαλού ελάχιστου

Το μέτρο F_1

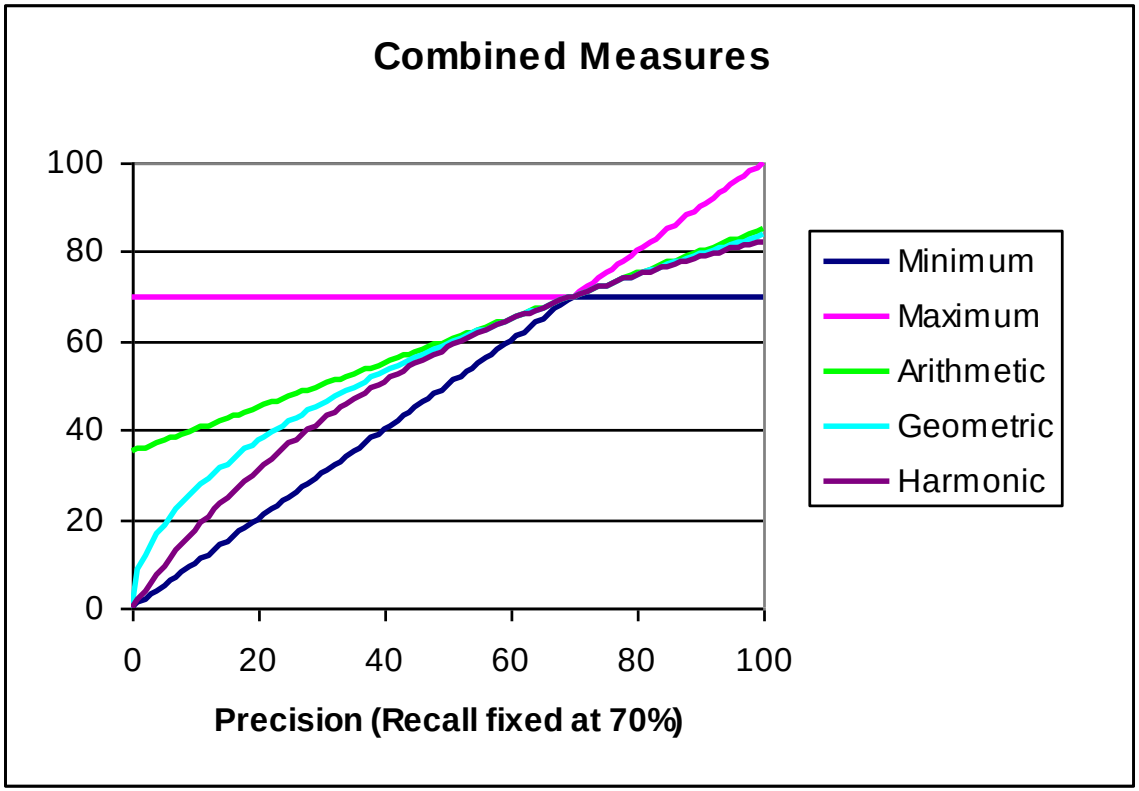
Συνήθως ισορροπημένο (balanced) F_1

- Αρμονικό μέσο των P και R

$$F_1 = \frac{1}{\frac{1}{2}\left(\frac{1}{P}+\frac{1}{R}\right)} = \frac{2\,PR}{P+R}$$

✓ Πιο κοντά στη μικρότερη από δύο τιμές

Αρμονικό Μέσο



Τιμές στο 0-1, αλλά συνήθως σε ποσοστά

Το μέτρο *F*

Το μέτρο ***F*** επιτρέπει μια αντιστάθμιση (trade off) της ακρίβεια και της ανάκλησης.

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R}$$

$$F_1 = \frac{1}{\frac{1}{2P}+\frac{1}{2R}}$$

όπου $\beta^2 = \frac{1-\alpha}{\alpha}$ α ϵ [0, 1] and thus b^2 ϵ [0,∞]

α = 1, F = P

α = 0, F = R

Συνήθως ισορροπημένο (balanced) F_1 με *α* = 0.5 και *β* = 1

- Αυτό είναι το αρμονικό μέσο των P και R

- Για ποια περιοχή τιμών του *β* η ανάκληση σταθμίζεται περισσότερο από την ακρίβεια; Συχνές τιμές, *β* = 0.5 και *β* = 2

Πίνακας Ενδεχομένων

	Συναφή (relevant)	**Μη συναφή (not relevant)**	
Ανακληθέντα (retrieved)	20 **(TP)**	40 **(FP)**	60
Μη ανακληθέντα (not retrieved)	60 **(FN)**	1,000,000 **(TN)**	1,000,060
	80	1,000,040	1,000,120

Precision = 20/60 = 1/3
Recall = 20/80 = 1/4

$$F_1 = 2\frac{1}{\frac{1}{\frac{1}{3}}+\frac{1}{\frac{1}{4}}} = 2/7$$

Ορθότητα (Accuracy)

- Γιατί να χρησιμοποιούμε περίπλοκα μέτρα όπως ακρίβεια, ανάκληση και F?
- Γιατί όχι κάτι πιο απλό;

Ορθότητα (Accuracy): το ποσοστό των αποφάσεων (συναφή/μη συναφή) που είναι *σωστές* (ως πρόβλημα ταξινόμησης σε δύο κλάσεις).

Με βάση τον πίνακα ενδεχομένων:

accuracy = (TP + TN)/(TP + FP + FN + TN).

Γιατί αυτό δεν είναι χρήσιμο στην ΑΠ;

Ορθότητα

Η μηχανή αναζήτησης snoogle επιστρέφει πάντα 0 αποτελέσματα (“0 matching results found”), ανεξάρτητα από το ερώτημα. Τι μας λέει όμως η ορθότητα (accuracy);

Ορθότητα

Παράδειγμα

	relevant	not relevant
retrieved	18	2
not retrieved	82	1,000,000,000

Ορθότητα

- Απλό κόλπο για τη μεγιστοποίηση της ορθότητας στην ΑΠ: *πες πάντα όχι* και μην επιστρέφεις κανένα έγγραφο
  - Αυτό έχει ως αποτέλεσμα 99.99% ορθότητα στα περισσότερα ερωτήματα

Αναζητήσεις στο web (και γενικά στην ΑΠ) θέλουν να βρουν κάτι και έχουν κάποια ανεκτικότητα στα «σκουπίδια»

Καλύτερα να επιστρέφεις κάποια κακά hits αρκεί να επιστέφεις κάτι

→ Για την αποτίμηση, χρησιμοποιούμε την ακρίβεια, ανάκληση και F

Δυσκολίες στη χρήση P/R

- Πρέπει να υπολογιστούν *μέσοι όροι* για μεγάλες ομάδες συλλογών εγγράφων/ερωτημάτων
- Χρειάζονται *εκτιμήσεις συνάφειας από ανθρώπους*
  - Οι χρήστες γενικά δεν είναι αξιόπιστοι αξιολογητές
- Οι εκτιμήσεις πρέπει να είναι *δυαδικές*
  - Ενδιάμεσες αξιολογήσεις;
- Εξαρτώνται από τη συλλογή/συγγραφή
  - Τα αποτελέσματα μπορεί να διαφέρουν από το ένα πεδίο στο άλλο
  - Development test collection (tune το σύστημα για μια συλλογή και εκτίμησε την απόδοση του σε αυτήν)

Μη γνωστή ανάκληση

- Ο συνολικός αριθμός των συναφών εγγράφων δεν είναι πάντα γνωστός:
  - Δειγματοληψία – πάρε έγγραφα από τη συλλογή και αξιολόγησε τη συνάφεια τους.
  - Εφάρμοσε *διαφορετικούς αλγόριθμους* για την ίδια συλλογή και την ίδια ερώτηση και χρησιμοποίησε το *άθροισμα των συναφών εγγράφων*

Μέτρα Συνάφειας χωρίς Διάταξη (περίληψη)

Τα αποτελέσματα μιας ερώτησης θεωρούνται *σύνολο*, δηλαδή αξιολογούμε τη συνάφεια ενός συνόλου

Πίνακας Ενδεχομένων Incident/Confusion Matrix)

	relevant	**not relevant**	
retrieved	**TP**	**FP**	All retrieved
not retrieved	**FN**	**TN**	All non retrieved
	All relevant	All non relevant	All

- Ακρίβεια (precision): $P = TP / (TP + FP)$
- Ανάκληση (recall): $R = TP / (TP + FN)$
- Μέτρο F: $F = 2PR/ P + R$
- Ορθότητα (accuracy) $A = (TP + TN)/(TP + FP + FN + TN)$.

ΜΕΤΡΑ ΠΟΥ ΘΕΩΡΟΥΝ ΤΗ ΔΙΑΤΑΞΗ

Αξιολόγηση Καταταγμένης Ανάκτησης

Στην πραγματικότητα, ο χρήστης δε βλέπει όλη την απάντηση, αντίθετα αρχίζει από την κορυφή της λίστας των αποτελεσμάτων

Θεωρείστε την περίπτωση που:

Answer(System1,q) = <N N N N N N N R R R>

Answer(System2,q) = <R R R N N N N N N N>

Όπου R: συναφές, N: μη συναφές

- Είναι σημαντικό τα συναφή αποτελέσματα να είναι στην κορυφή

  ✓ *Η ακρίβεια, ανάκληση και το F είναι μέτρα για μη καταταγμένα (unranked) σύνολα .*

  Πως μπορούμε να τα τροποποιήσουμε τα μέτρα για λίστες με διάταξη;

Καμπύλη Ακρίβειας/Ανάκλησης

Πως μπορούμε να τα τροποποιήσουμε τα μέτρα για λίστες με διάταξη;

- Απλώς υπολόγισε το μέτρο συνόλου *για κάθε πρόθεμα*: το κορυφαίο 1, κορυφαία 2, κορυφαία 3, κορυφαία 4 κλπ αποτελέσματα

Με αυτόν τον τρόπο παίρνουμε μια καμπύλη ακρίβειας-ανάκλησης (precision-recall curve).

RRNRNRNNNNNNNRN

Παράδειγμα I

n	doc #	relevant
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	x
14	990	

Συνολικός # από συναφή έγγραφα = 6

Έλεγχος σε κάθε ***νέο σημείο recall***:

R = 1/6 P = 1/1
R = 2/6 P = 2/2
R = 2/ 6 P = 2/3
R = 3/6 P = 3/4
R = 3/6 P = 3/5
R = 4/6 P = 4/6
R = 4/6 P = 4/7
R = 4/6 P = 4/8
R = 4/6 P – 4/9
R = 4/6 P = 4/10
R = 4/6 P = 4/11
R = 4/6 P = 4/12
R = 5/6 P = 5/13
R = 5/6 P=5/14

Λείπει ένα έγγραφο, άρα ποτέ 100% recall

Παράδειγμα I

Συνολικός # από συναφή έγγραφα = 6

Έλεγχος σε κάθε ***νέο σημείο recall***:

n	doc #	relevant	
1	588	x	R = 1/6 P = 1/1
2	589	x	R = 2/6 P = 2/2
3	576		R = 2/ 6 P = 2/3
4	590	x	R = 3/6 P = 3/4
5	986		R = 3/6 P = 3/5
6	592	x	R = 4/6 P = 4/6
7	984		R = 4/6 P = 4/7
8	988		R = 4/6 P = 4/8
9	578		R = 4/6 P – 4/9
10	985		R = 4/6 P = 4/10
11	103		R = 4/6 P = 4/11
12	591		R = 4/6 P = 4/12
13	772	x	R = 5/6 P = 5/13
14	990		R = 5/6 P=5/14

Το precision πέφτει μέχρι το επόμενο σημείο που αλλάζει/αυξάνεται το recall

Παράδειγμα I

Συνολικός # από συναφή έγγραφα = 6

Έλεγχος σε κάθε ***νέο σημείο recall***:

n	doc #	relevant
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	x
14	990	

R=1/6=0.167; P=1/1=1

R=2/6=0.333; P=2/2=1

R=3/6=0.5; P=3/4=0.75

R=4/6=0.667; P=4/6=0.667

R=5/6=0.833; P=5/13=0.38

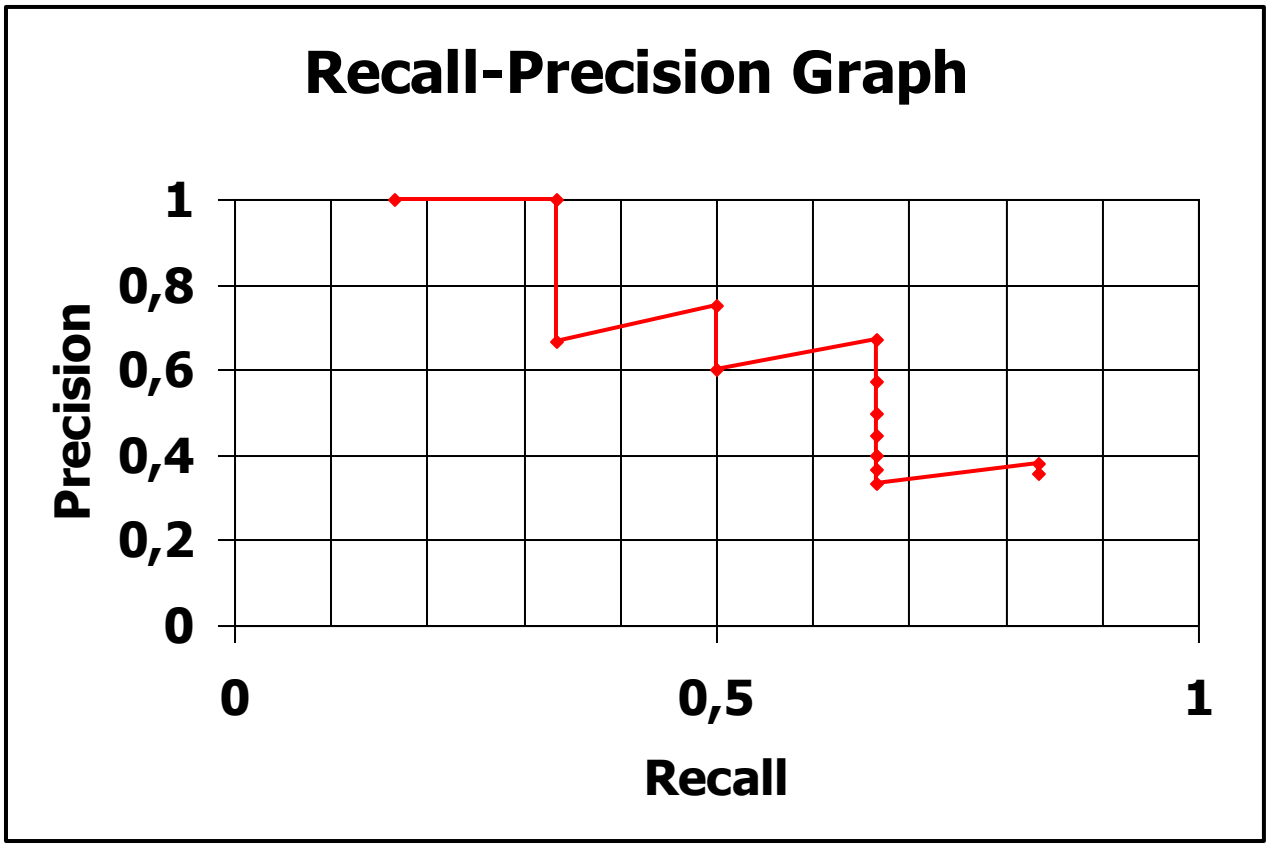
Συνολικός # από συναφή έγγραφα = 6

Έλεγχος σε κάθε ***νέο σημείο recall***:

n	doc #	relevant
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	x
14	990	

R = **0.167** P = 1
R = **0.33** P = 1
R = 0.33 P = 0.67
R = **0.5** P = 0.75
R = 0.5 P = 0.6
R = 0.667 P = 0.667
R = 0.667 P = 0.57
R P = 0.5
R P = 0.44
R P = 0.4
R P = 0.36
R P = 0.33
R = 0.833 P = 0.38
R P = 0.38

Παράδειγμα Ι (συνέχεια)



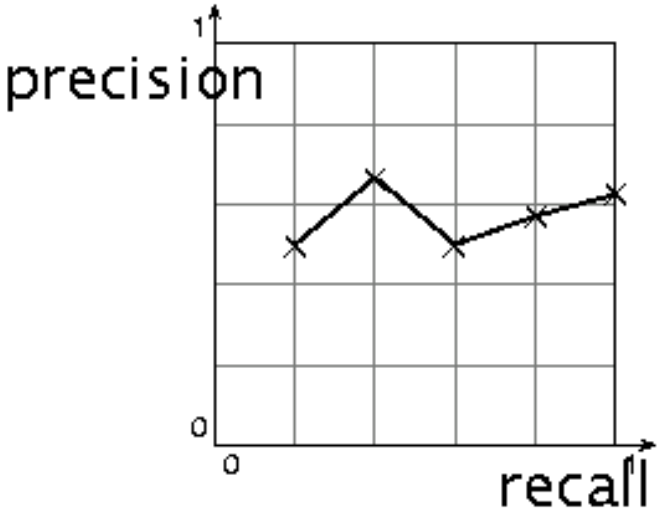
Πριονωτή – το precision ελαττώνεται για το ίδιο recall μέχρι να βρεθεί το επόμενο συναφές έγγραφο

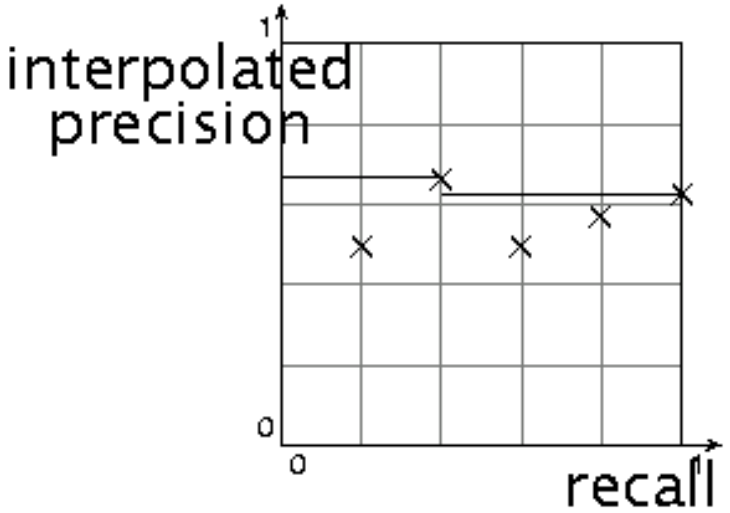
Ακρίβεια εκ παρεμβολής (Interpolated precision)

- Αν η ακρίβεια αλλάζει τοπικά με την αύξηση της ανάκλησης, το λαμβάνουμε υπ' όψιν – *ο χρήστης θέλει να δει και άλλα έγγραφα αν αυξάνεται και η ακρίβεια και η ανάκληση*

Ακρίβεια με παρεμβολή σε επίπεδο ανάκλησης r *είναι η μεγαλύτερη ακρίβεια για επίπεδο ανάκλησης r' >= r.*

- Παίρνουμε τη μέγιστη τιμή της ακρίβειας στα δεξιά της τιμής





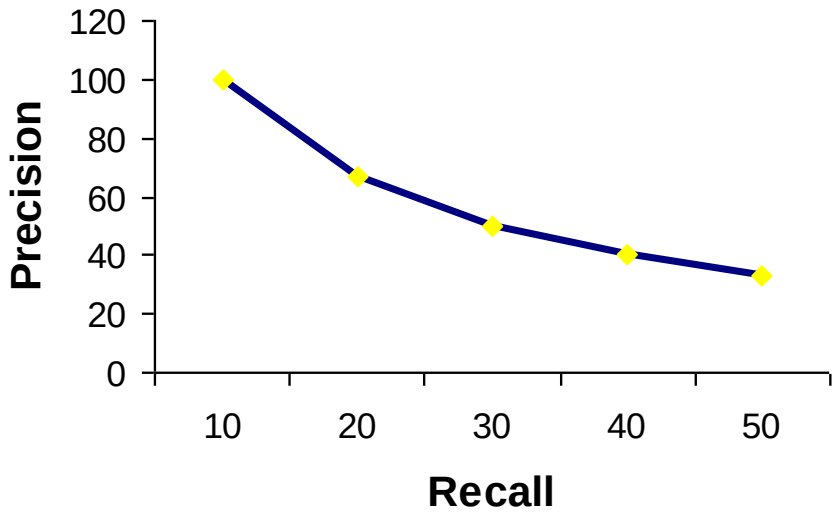
Παράδειγμα II

$$\text{Relevant} = \{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}\}$$

$$\text{Retrieved} = d_{123}, d_{84}, d_{56}, d_6, d_{85}, d_9, d_{511}, d_{129}, d_{187}, d_{25}, d_{38}, d_{48}, d_{250}, d_{113}, d_3$$

Παράδειγμα II

$$\text{Relevant} = \left\{ \begin{matrix} d_3, d_5, d_9, d_{25}, d_{39}, \\ d_{44}, d_{56}, d_{71}, d_{89}, d_{123} \end{matrix} \right\}$$



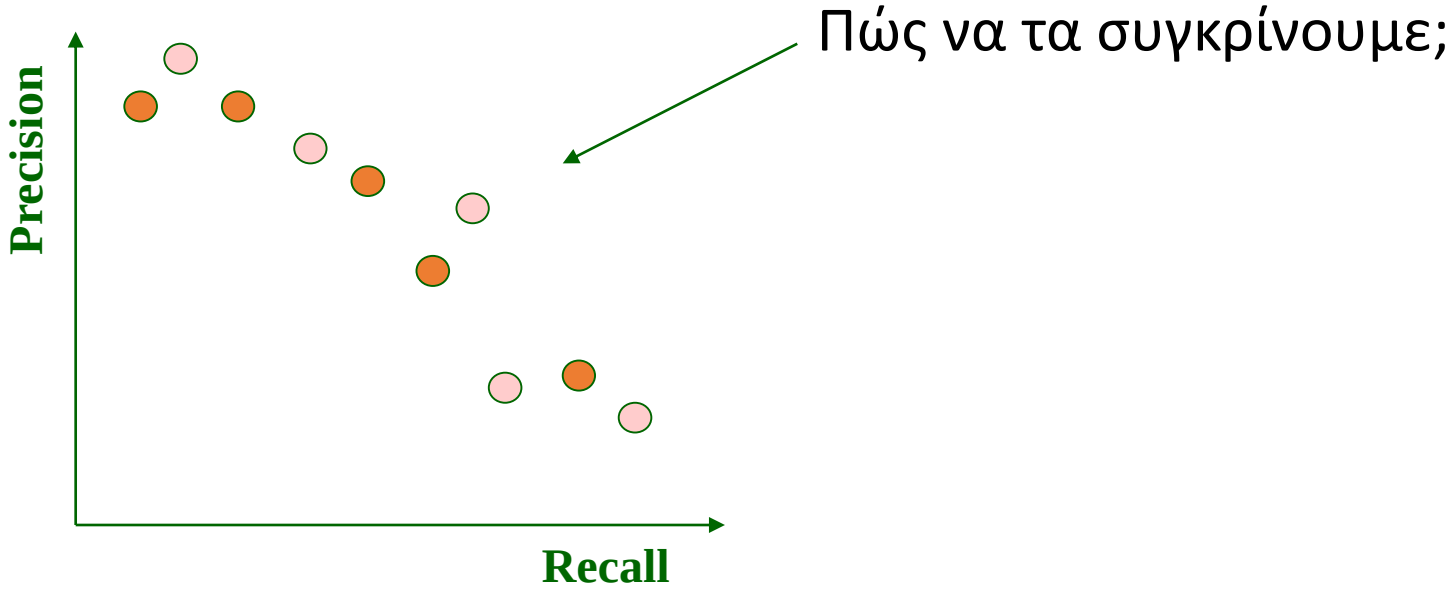
Rank	Doc	Rel	Recall	Precision
0			0%	0%
1	d_{123}	✔	10%	100%
2	d_{84}		10%	50%
3	d_{56}	✔	20%	67%
4	d_{6}		20%	50%
5	d_{85}		20%	40%
6	d_{9}	✔	30%	50%
7	d_{511}		30%	43%
8	d_{129}		30%	38%
9	d_{187}		30%	33%
10	d_{25}	✔	40%	40%
11	d_{38}		40%	36%
12	d_{48}		40%	33%
13	d_{250}		40%	31%
14	d_{113}		40%	29%
15	d_{3}	✔	50%	33%

Μέσοι όροι από πολλά ερωτήματα

- Το γράφημα για ένα ερώτημα δεν αρκεί
- Χρειαζόμαστε *τη μέση απόδοση σε αρκετά ερωτήματα*.
- Αλλά:
  - Οι υπολογισμοί ακρίβειας-ανάκλησης τοποθετούν κάποια σημεία στο γράφημα
  - Πως καθορίζουμε μια τιμή ανάμεσα στα σημεία;

Σύγκριση Συστημάτων

- Σύστημα 1
- Σύστημα 2



Κανονικοποιημένα επίπεδα ανάκλησης

Σκοπός: Δυνατότητα σύγκρισης διαφορετικών συστημάτων (ή μέσων όρων σε διαφορετικά ερωτήματα)

Πως; Χρήση *κανονικοποιημένων επιπέδων ανάκλησης* (standard recall levels)

Παράδειγμα κανονικοποιημένων επιπέδων ανάκλησης (πλήθος επιπέδων: 11):

Standard Recall levels at 0%, 10%, 20%, …, 100%

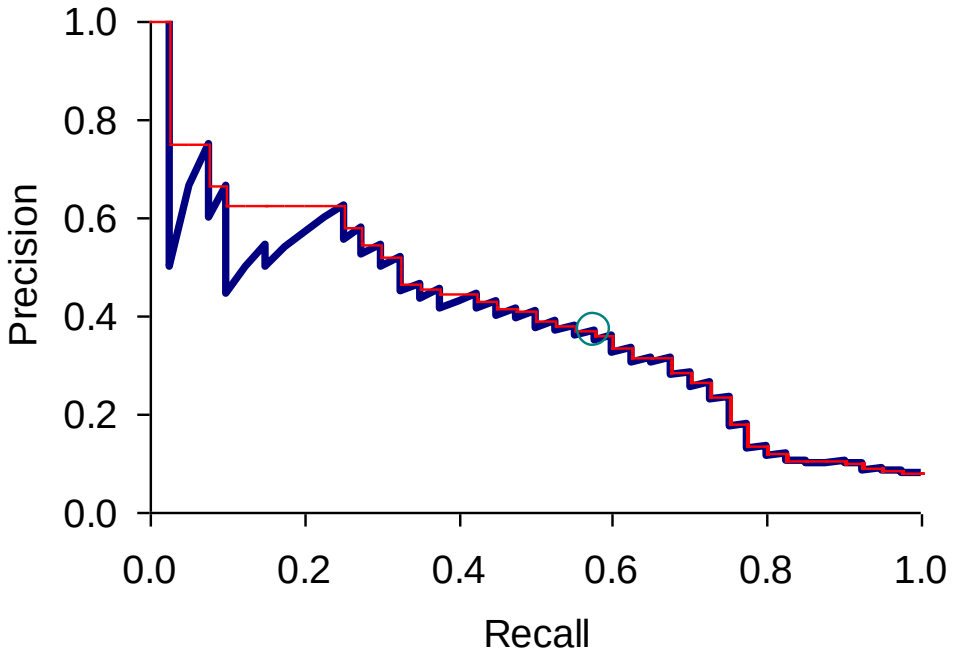
$r_j \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$

$r_0 = 0.0, r_1 = 0.1, \ldots, r_{10}=1.0$

Μέση ακρίβεια 11-σημείων με παρεμβολή (11-point interpolated average precision)

- Υπολόγισε την ακρίβεια *με παρεμβολή* στα επίπεδα ανάκτησης 0.0, 0.1, 0.2, . . .
- Επανέλαβε το για όλα τα ερωτήματα στο evaluation benchmark και πάρε το μέσο όρο
- Αυτό το μέτρο μετρά την απόδοση *σε όλα τα επίπεδα ανάκλησης* (at all recall levels).

Καμπύλη Ακρίβειας/Ανάκλησης



Recall	Interp. Precision
0.0	1.00
0.1	0.67
0.2	0.63
0.3	0.55
0.4	0.45
0.5	0.41
0.6	0.36
0.7	0.29
0.8	0.13
0.9	0.10
1.0	0.08

Κάθε σημείο αντιστοιχεί σε ένα αποτέλεσμα για τα κορυφαία *k* έγγραφα (*k* = 1, 2, 3, 4, . . .).

Παρεμβολή (με κόκκινο): μέγιστο των μελλοντικών σημείων

Μέση ακρίβεια 11-σημείων με παρεμβολή (11-point interpolated average precision)

Recall	Interpolated Precision
0.0	1.00
0.1	0.67
0.2	0.63
0.3	0.55
0.4	0.45
0.5	0.41
0.6	0.36
0.7	0.29
0.8	0.13
0.9	0.10
1.0	0.08

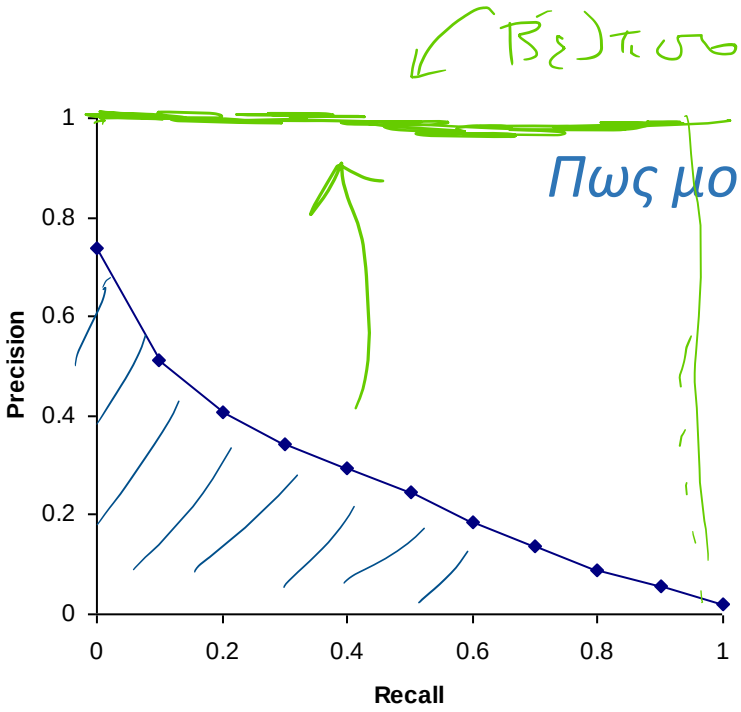
- Υπολόγισε την ακρίβεια με παρεμβολή στα επίπεδα ανάκτησης 0.0, 0.1, 0.2, .
- Επανέλαβε το για όλα τα ερωτήματα στο evaluation benchmark και πάρε το μέσο όρο

$$\overline{P}(r) = \sum_{i=1}^{N_q} \frac{P_i(r)}{N_q}$$

N_q – πλήθος ερωτημάτων
$P_i(r)$ - precision at recall level r for i^{th} query

Τυπική (καλή;) ακρίβεια 11-σημείων

- SabIR/Cornell 8A1 11pt precision from TREC 8 (1999)



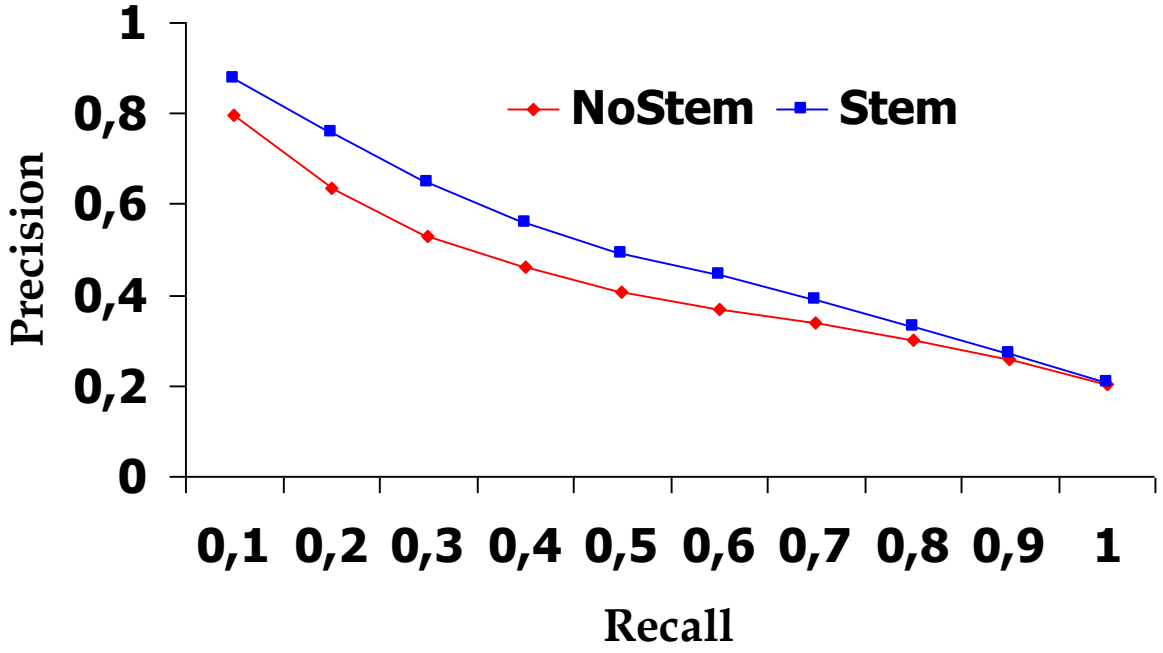
Πως μοιάζει το optimal;

AUC

Measure: PR AUC (Area Under the Curve) (χρήσιμη όταν πάρα πολλά True Negative)

Σύγκριση Συστημάτων

- Η καμπύλη που είναι πιο κοντά στη πάνω δεξιά γωνία του γραφήματος υποδηλώνει και καλύτερη απόδοση



Μέτρα Συνάφειας με Διάταξη

Η καμπύλη ανάκλησης-ακρίβειας υποθέτει ότι *έχουμε όλο το αποτέλεσμα*

Σε πολλές μηχανές αναζήτησης

- Το αποτέλεσμα είναι *πολύ μεγάλο*
- Ο χρήστης ενδιαφέρεται *μόνο για τα πρώτα αποτελέσματα*

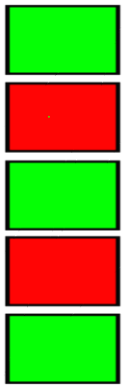
Ακρίβεια στα *k (precision@k)*

Ακρίβεια-στα-k (Precision-at-k): *Η ακρίβεια των κορυφαίων* *k* αποτελεσμάτων

Πχ ακρίβεια-στα-10, αγνοεί τα έγγραφα μετά το 10ο

Πχ

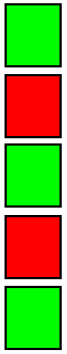
- Prec@3 2/3
- Prec@4 2/4
- Prec@5 3/5



Πράσινο συναφές
Κόκκινο μη συναφές

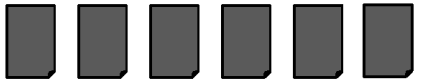
- Πιθανώς κατάλληλο για τις περισσότερες αναζητήσεις στο web: οι χρήστες θέλουν καλά αποτελέσματα στις πρώτες μία ή δύο σελίδες
- Αντίστοιχα ανάκληση στα *k*

MAP

- Θεωρείστε τη θέση διάταξης (rank position) κάθε ***συναφούς εγγράφου***
  - $K_1, K_2, \ldots K_R$
- Υπολογισμός του Precision@K για κάθε $K_1, K_2, \ldots K_R$
- Μέση ακρίβεια = average of P@K
- Π.χ.,:  έχει AvgPrec $\frac{1}{3}\cdot\left(\frac{1}{1}+\frac{2}{3}+\frac{3}{5}\right)\approx 0.76$
- Mean Average Precision (MAP) *Μέση αντιπροσωπευτική ακρίβεια*: η μέση ακρίβεια για πολλαπλά ερωτήματα

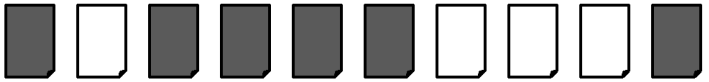
MAP

Σύγκριση διατάξεων



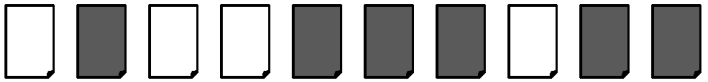
= the relevant documents

Ranking #1



Recall	0.17	0.17	0.33	0.5	0.67	0.83	0.83	0.83	0.83	1.0
Precision	1.0	0.5	0.67	0.75	0.8	0.83	0.71	0.63	0.56	0.6

Ranking #2



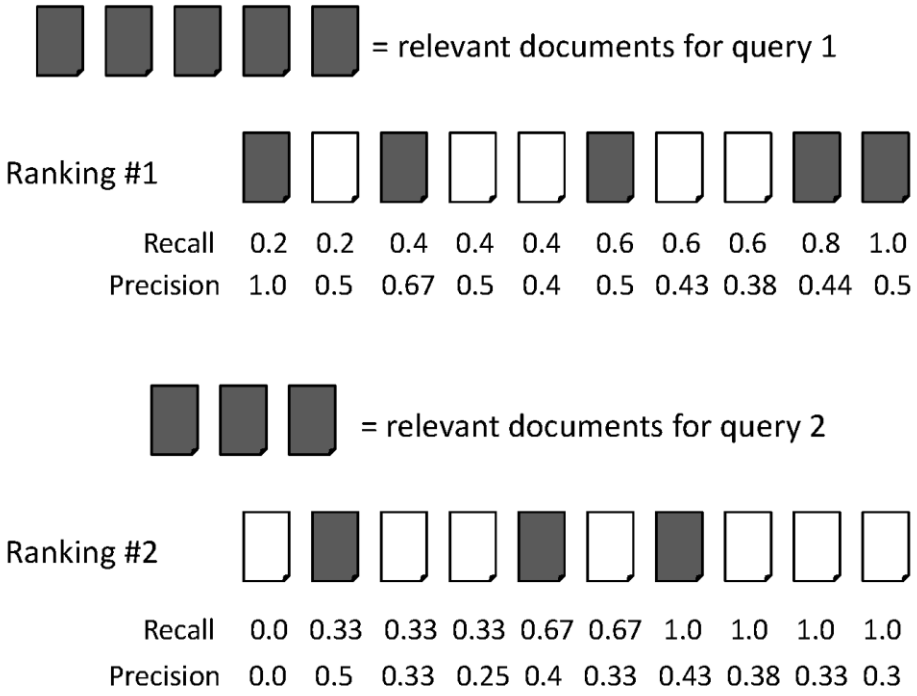
Recall	0.0	0.17	0.17	0.17	0.33	0.5	0.67	0.67	0.83	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.5	0.57	0.5	0.56	0.6

$$\text{Ranking } \#1\text{: } (1.0 + 0.67 + 0.75 + 0.8 + 0.83 + 0.6)/6 = 0.78$$

$$\text{Ranking } \#2\text{: } (0.5 + 0.4 + 0.5 + 0.57 + 0.56 + 0.6)/6 = 0.52$$

MAP

Πολλά ερωτήματα:



$$average\ precision\ query\ 1 = (1.0 + 0.67 + 0.5 + 0.44 + 0.5)/5 = 0.62$$
$$average\ precision\ query\ 2 = (0.5 + 0.4 + 0.43)/3 = 0.44$$

$$mean\ average\ precision = (0.62 + 0.44)/2 = 0.53$$

MAP

- Μέσος όρος της τιμής της ακρίβειας των κορυφαίων *k* εγγράφων, κάθε φορά που επιστρέφεται ένα σχετικό έγγραφο
- Αποφεύγει την παρεμβολή και τη χρήση προκαθορισμένων επιπέδων ανάκλησης
- MAP για μια *συλλογή ερωτημάτων* είναι το *αριθμητικό μέσο*.
  - Macro-averaging: κάθε ερώτημα μετράει το ίδιο

$$\mathrm{MAP}(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j}\mathrm{Precision}(R_{jk})$$

Q σύνολο ερωτημάτων, $\mathbf{q_j}$ ένα από τα ερωτήματα, $\mathbf{\{d_1, d_2, \ldots, d_{mj}\}}$ είναι τα συναφή έγγραφα και $\mathbf{R_{jk}}$ είναι ο αριθμός των εγγράφων στο αποτέλεσμα μέχρι να φτάσουμε στο d_{jk} (0 αν το d_{jk} δεν ανήκει στο αποτέλεσμα)

MAP

- Συχνά οι τιμές της MAP για το *ίδιο ερώτημα* σε διαφορετικά συστήματα διαφέρουν λιγότερο από τις τιμές τις MAP για διαφορετικά ερωτήματα στο ίδιο σύστημα (υπάρχουν «δύσκολα» ερωτήματα)

- MAP δίνει μια προσέγγιση του AUC της καμπύλης Ακρίβειας-Ανάκλησης

Επανάληψη και Ασκήσεις

Βασικό κριτήριο: Συνάφεια

Η ικανοποίηση του χρήστη συνήθως εξισώνεται με τη συνάφεια (relevance) των αποτελεσμάτων της αναζήτησης με το ερώτημα

Μα πως θα μετρήσουμε τη συνάφεια;

Μεθοδολογία: Benchmarks

Η καθιερωμένη μεθοδολογία στην Ανάκτηση Πληροφορίας αποτελείται από τρία στοιχεία:

1. Μία πρότυπη *συλλογή εγγράφων* (benchmark document collection)
2. Μια πρότυπη *ομάδα ερωτημάτων* (benchmark suite of queries)
3. Μια *αποτίμηση της συνάφειας* για κάθε ζεύγος ερωτήματος-εγγράφου, συνήθως δυαδική: συναφής (R) - μη συναφής (N) – (*gold standard/ground truth*) που μας λέει αν το έγγραφο είναι συναφές ως προς το ερώτημα

Μέτρα Συνάφειας

Δυο κατηγορίες μέτρων:

- Μέτρα που αγνοούν τη διάταξη
- Μέτρα που λαμβάνουν υπ’ όψιν τη διάταξη

Θα δούμε στην αρχή *μέτρα που αγνοούν τη διάταξη*

Όχι διάταξη
Δυαδική συνάφεια

Πίνακας Ενδεχομένων
Confusion/Incidence Matrix

Ground truth

Predicted	Συναφή (relevant)	Μη συναφή (not relevant)	
Ανακληθέντα (retrieved)	**TP**	**FP**	retrieved
Μη ανακληθέντα (not retrieved)	**FN**	**TN**	
	relevant		all

Ακρίβεια και Ανάκληση

Πίνακας Ενδεχόμενων (Incidence Matrix)

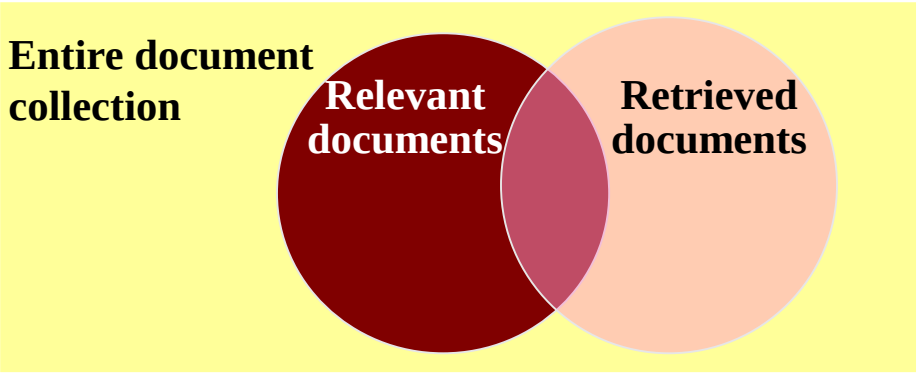
πραγματικά

αποτέλεσμα

	Relevant	Nonrelevant
Retrieved	true positives (TP)	false positives (FP)
Not retrieved	false negatives (FN)	true negatives (TN)

$$P = TP / (TP + FP)$$

$$R = TP / (TP + FN)$$



Το μέτρο F_1

Συνήθως ισορροπημένο (balanced) F_1

- Αρμονικό μέσο των P και R

$$F_1 = \frac{1}{\frac{1}{2}\left(\frac{1}{P}+\frac{1}{R}\right)} = \frac{2\,PR}{P+R}$$

✓ Πιο κοντά στη μικρότερη από δύο τιμές

Ορθότητα (Accuracy)

- Γιατί να χρησιμοποιούμε περίπλοκα μέτρα όπως ακρίβεια, ανάκληση και F?
- Γιατί όχι κάτι πιο απλό;

Ορθότητα (Accuracy): το ποσοστό των αποφάσεων (συναφή/μη συναφή) που είναι *σωστές* (ως πρόβλημα ταξινόμησης σε δύο κλάσεις).

Με βάση τον πίνακα ενδεχομένων:

accuracy = (TP + TN)/(TP + FP + FN + TN).

Γιατί αυτό δεν είναι χρήσιμο στην ΑΠ;

Confusion matrix, precision, recall, F1, accuracy

Άσκηση
Θεωρείστε ένα σύστημα ανάκτησης για μια συλλογή 10000 εγγράφων.

Έστω μια ερώτηση q για την οποία τα συναφή έγγραφα είναι τα
{d1, d33, d50, d99, d121, d317, d590, d690, d2000, d3010, d3196, d3412, d5555, d6661, d7671, d8032, d9099, d9234, d9325}

Ένα σύστημα ανάκτησης πληροφορίας επιστρέφει την παρακάτω απάντηση
d50 d2 d8032 d99 d7898 d121

Διάταξη
Δυαδική συνάφεια

Καμπύλη Ακρίβειας/Ανάκλησης

Πως μπορούμε να τα τροποποιήσουμε τα μέτρα για λίστες με διάταξη;

- Απλώς υπολόγισε το μέτρο συνόλου *για κάθε πρόθεμα*: το κορυφαίο 1, κορυφαία 2, κορυφαία 3, κορυφαία 4 κλπ αποτελέσματα

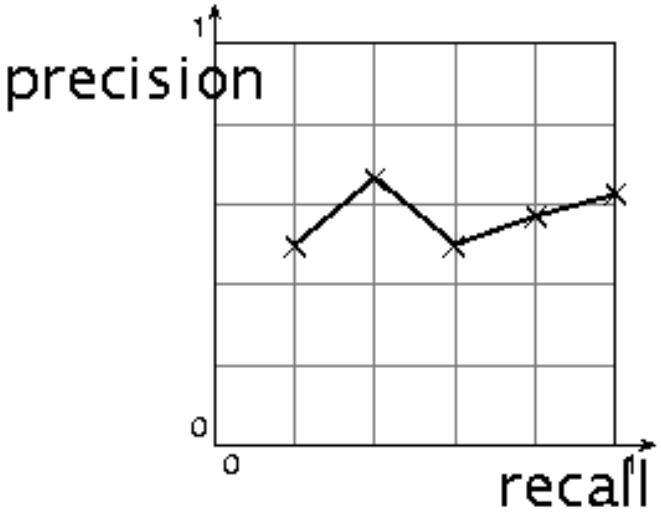
Με αυτόν τον τρόπο παίρνουμε μια καμπύλη ακρίβειας-ανάκλησης (precision-recall curve).

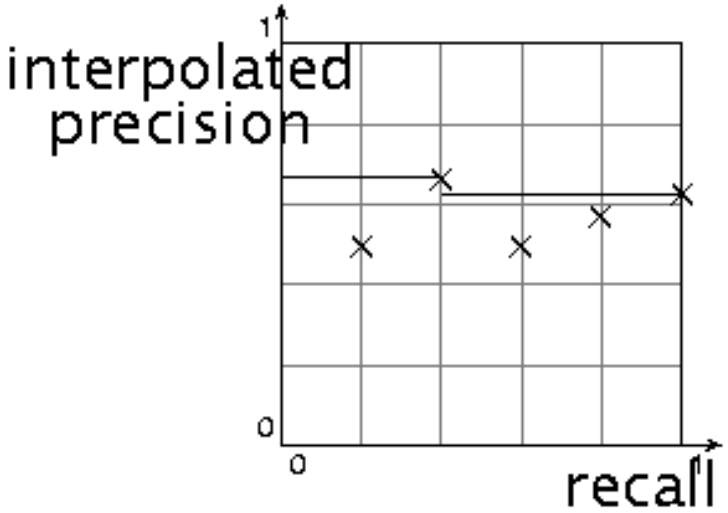
Ακρίβεια εκ παρεμβολής (Interpolated precision)

- Αν η ακρίβεια αλλάζει τοπικά με την αύξηση της ανάκλησης, το λαμβάνουμε υπ' όψιν – *ο χρήστης θέλει να δει και άλλα έγγραφα αν αυξάνεται και η ακρίβεια και η ανάκληση*

Ακρίβεια με παρεμβολή σε επίπεδο ανάκλησης r *είναι η μεγαλύτερη ακρίβεια για επίπεδο ανάκλησης r' >= r.*

- Παίρνουμε τη μέγιστη τιμή της ακρίβειας στα δεξιά της τιμής





Precision Recall curve

Άσκηση

Ένα ΣΑΠ επιστρέφει την απάντηση R N R R N N N R R R

Συνολικά υπάρχουν 5000 έγγραφα από τα οποία τα συναφή είναι 8.

Καμπύλη ανάκλησης-ακρίβειας – και κανονικοποιημένη 11-σημείων με παρεμβολή

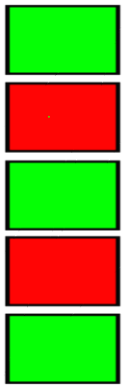
Ακρίβεια στα *k (precision@k)*

Ακρίβεια-στα-k (Precision-at-k): *Η ακρίβεια των κορυφαίων* *k* αποτελεσμάτων

Πχ ακρίβεια-στα-10, αγνοεί τα έγγραφα μετά το 10ο

Πχ

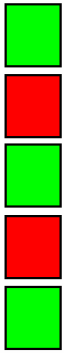
- Prec@3 2/3
- Prec@4 2/4
- Prec@5 3/5



Πράσινο συναφές
Κόκκινο μη συναφές

- Πιθανώς κατάλληλο για τις περισσότερες αναζητήσεις στο web: οι χρήστες θέλουν καλά αποτελέσματα στις πρώτες μία ή δύο σελίδες
- Αντίστοιχα ανάκληση στα *k*

MAP

- Θεωρείστε τη θέση διάταξης (rank position) κάθε ***συναφούς εγγράφου***
  - K_1, K_2, … K_R
- Υπολογισμός του Precision@K για κάθε K_1, K_2, … K_R
- Μέση ακρίβεια = average of P@K
- Π.χ.,:  έχει AvgPrec $\frac{1}{3}\cdot\left(\frac{1}{1}+\frac{2}{3}+\frac{3}{5}\right)\approx 0.76$
- Mean Average Precision (MAP) *Μέση αντιπροσωπευτική ακρίβεια*: η μέση ακρίβεια για πολλαπλά ερωτήματα

MAP

Άσκηση

Ένα ΣΑΠ επιστρέφει την απάντηση R N R R N N N R R R

Συνολικά υπάρχουν 5000 έγγραφα από τα οποία τα συναφή είναι 8.

Επανάληψη και Ασκήσεις (τέλος)

R-ακρίβεια

- Αν έχουμε ένα γνωστό (πιθανών μη πλήρες) σύνολο από συναφή έγγραφα μεγέθους *Rel,* τότε υπολογίζουμε την ακρίβεια των κορυφαίων *Rel* εγγράφων που επιστρέφει το σύστημα (ακρίβεια@Rel)

  Αν υπάρχουν *r* συναφή στα *Rel* κορυφαία, τότε *r/Rel*

- Το τέλειο σύστημα μπορεί να πετύχει βαθμό 1.0

R-ακρίβεια

Ακρίβεια-στο-*Rel*, όπου *Rel* ο αριθμός των συναφών εγγράφων της συλλογής

n	doc #	relevant
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	x
14	990	

R = # of relevant docs = 6

R-Precision = 4/6 = 0.67

Μόνο ένα έγγραφο

- Έστω ότι υπάρχει *μόνο ένα συναφές* έγγραφο
- Περιπτώσεις:
  - Αναζήτηση γνωστού στοιχείου
  - navigational queries
  - Αναζήτηση γεγονότος (fact) πχ πληθυσμός μιας χώρας
- Διάρκεια αναζήτησης ~θέση (rank) της απάντησης
  - Μετρά την προσπάθεια του χρήστη

MRR: Mean Reciprocal Rate

- Θεωρούμε *τη θέση Κ του πρώτου σχετικού εγγράφου*
  - Μπορεί να είναι το μόνο που έκανε click ο χρήστης (δε χρειάζεται ground truth)
- Reciprocal Rank score = $\frac{1}{K}$
- MRR το μέσο RR για πολλαπλές ερωτήσεις

Επανάληψη

Μετρικές για την αξιολόγηση της συνάφειας

	Όχι διάταξη	Διάταξη
Δυαδικές	Ακρίβεια (Precision) Ανάκληση (Recall) F1	Καμπύλη Ακρίβειας-Ανάκλησης (με παρεμβολή, 11-σημείων) Ακρίβεια@k MAP R-Precision MRR **ROC**
Μη δυαδικές	**GC**	**DGC, NDGC**

ROC (Receiver Operating Characteristic Curve)

Καμπύλη χαρακτηριστικής λειτουργίας δέκτη

- Αναπτύχθηκε στη δεκαετία του 1950 για την ανάλυση θορύβου στα σήματα
  - Χαρακτηρίζει το trade-off μεταξύ positive hits και false alarms
  - Συχνά σε ταξινομητές (classifiers)
- Η καμπύλη ROC
  - (y-άξονας) TPR [TruePositiveRate] ή ευαισθησία (sensitivity) (άλλο όνομα του recall)
  - (x-άξονας) FPR [FalsePositiveRate] ή fall out ή 1 – specificity (εξειδίκευση, προσδιοριστικότητα,) (true negative rate)
  - Η απόδοση αναπαρίσταται ως ένα σημείο στην καμπύλη ROC

True Positive Rate - Recall, Sensitivity (ευαισθησία)

[πόσα από τα συναφή (θετικά) ταξινομεί σωστά]

Το ποσοστό των θετικών που επιστρέφει/ανακτά

$$\mathrm{TPR} = \frac{TP}{\underbrace{TP + FN}_{\text{Συναφή}}}$$

True Negative Rate – Specificity (εξειδίκευση)

[πόσα από τα αρνητικά ταξινομεί σωστά]

Το ποσοστό των αρνητικών που δεν επιστρέφει ως θετικά/δεν ανακτά

$$\mathrm{TNR} = \frac{TN}{TN + FP}$$

Θα προτιμούσατε (είναι πιο ασφαλές) ένα διαγνωστικό τεστ με

A. Καλή ευαισθησία και κακή εξειδίκευση

B. Κακή ευαισθησία και καλή εξειδίκευση

False Positive Rate (fall out)

[πόσα από τα αρνητικά ταξινομεί λάθος]

Το ποσοστό των αρνητικών που δεν επιστρέφει ως θετικά/δεν ανακτά

False alarms

$$\mathrm{FPR} = \frac{FP}{\underbrace{TN + FP}_{\text{Μη συναφή}}}$$

$$\mathrm{FPR} = 1 - \mathrm{TNR}$$

ROC (Receiver Operating Characteristic Curve)

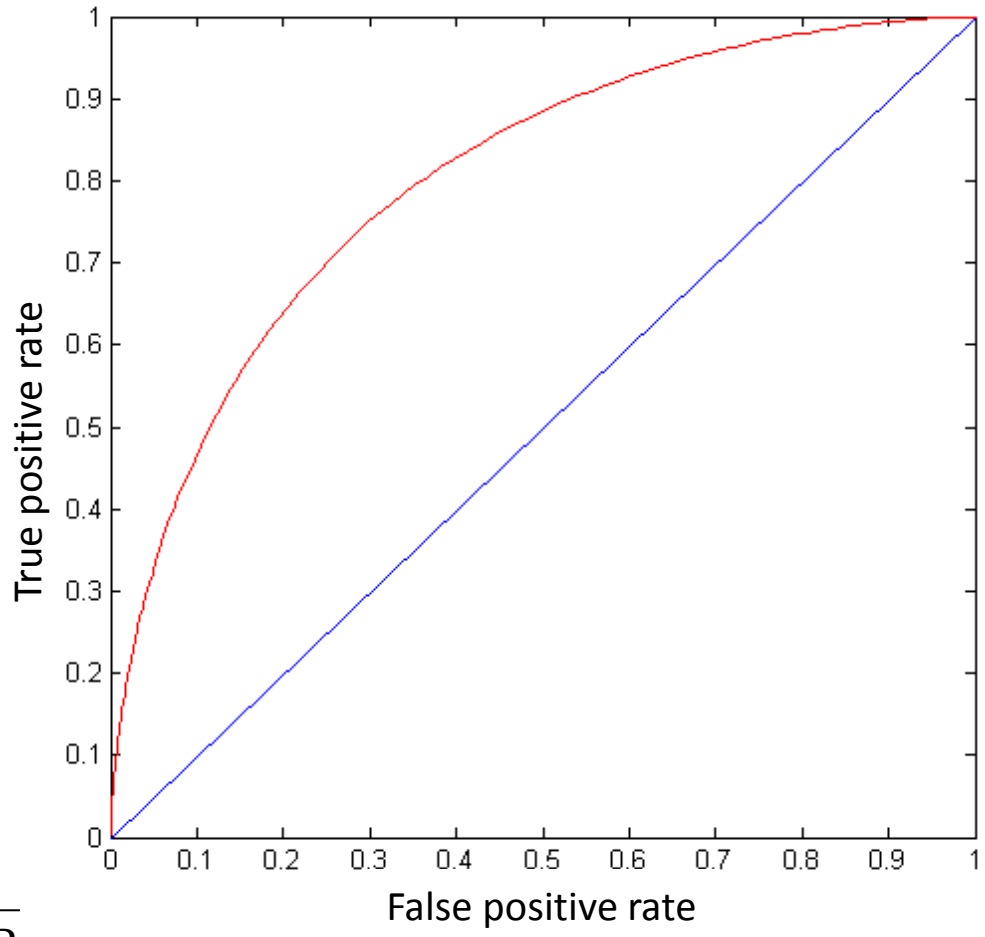
Συναφή -> κατηγορία +
Μη συναφή -> κατηγορία -

(0,0): το μοντέλο προβλέπει τα πάντα ως αρνητική κατηγορία

(1,1): το μοντέλο προβλέπει τα πάντα ως θετική κατηγορία

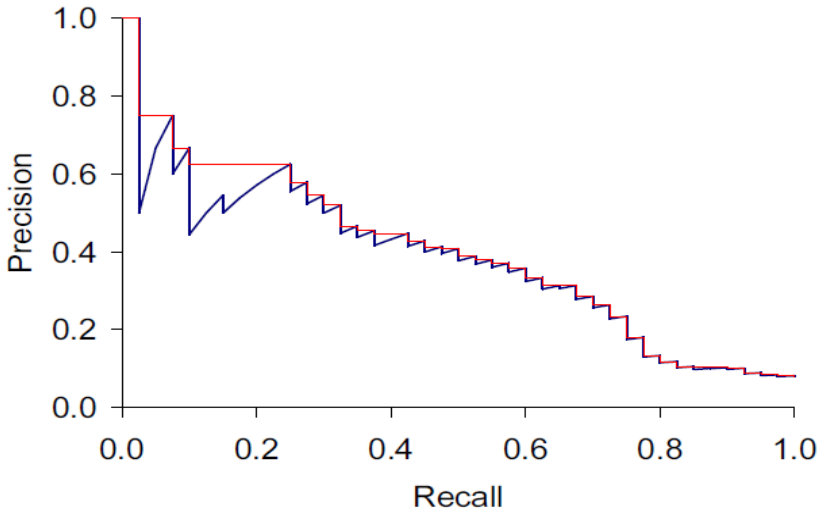
(0,1): ιδανικό
Το ιδανικό στην άνω αριστερή γωνία

Διαγώνια γραμμή: Random guessing



$$\mathrm{TPR} = \frac{TP}{TP + FN} \qquad \mathrm{FPR} = \frac{FP}{TN + FP}$$

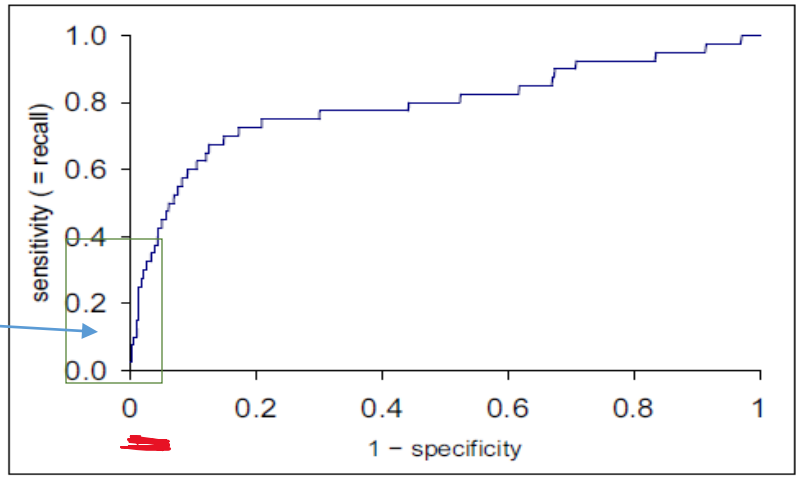
ROC (Receiver Operating Characteristic Curve)



AUC measure (area under the ROC curve)

Γιατί όχι ROC;

- Συχνά TN μεγάλη τιμή (FPR κοντά στο 0)
- Επίσης, συχνά μικρό recall



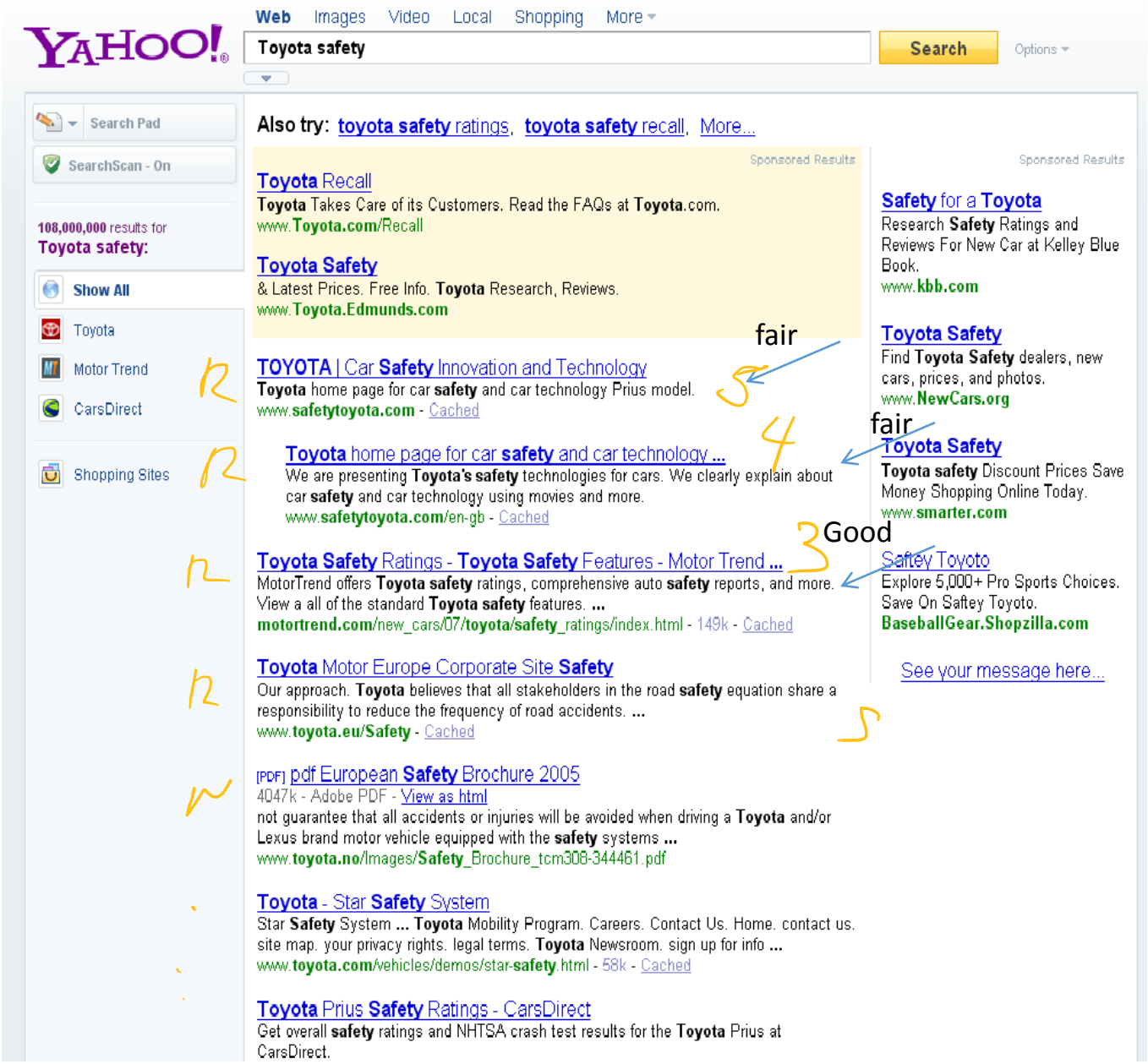
$$\text{TPR} = \frac{TP}{TP + FN}$$

$$\text{FPR} = \frac{FP}{TN + FP}$$

Μη δυαδικές αποτιμήσεις συνάφειας

Μη δυαδικές αποτιμήσεις

- Μέχρι στιγμής δυαδικές αποτιμήσεις συνάφειας (συναφές ή μη συναφές)
- Ας υποθέσουμε ότι τα έγραφα βαθμολογούνται για το «πόσο» συναφή είναι σε κάποια βαθμολογική κλίμακα [0, r], r>2


YAHOO!
Web Images Video Local Shopping More
Toyota safety
Search
Options
Search Pad
SearchScan - On
108,000,000 results for
Toyota safety:
Show All
Toyota
Motor Trend
CarsDirect
Shopping Sites
Also try: toyota safety ratings, toyota safety recall, More...
Sponsored Results
Toyota Recall
Toyota Takes Care of its Customers. Read the FAQs at Toyota.com.
www.Toyota.com/Recall
Toyota Safety
& Latest Prices. Free Info. Toyota Research, Reviews.
www.Toyota.Edmunds.com
fair
TOYOTA | Car Safety Innovation and Technology
Toyota home page for car safety and car technology Prius model.
www.safetytoyota.com - Cached
Toyota home page for car safety and car technology ...
We are presenting Toyota's safety technologies for cars. We clearly explain about car safety and car technology using movies and more.
www.safetytoyota.com/en-gb - Cached
Toyota Safety Ratings - Toyota Safety Features - Motor Trend ...
MotorTrend offers Toyota safety ratings, comprehensive auto safety reports, and more. View a all of the standard Toyota safety features. ...
motortrend.com/new_cars/07/toyota/safety_ratings/index.html - 149k - Cached
Toyota Motor Europe Corporate Site Safety
Our approach. Toyota believes that all stakeholders in the road safety equation share a responsibility to reduce the frequency of road accidents. ...
www.toyota.eu/Safety - Cached
[PDF] pdf European Safety Brochure 2005
4047k - Adobe PDF - View as html
not guarantee that all accidents or injuries will be avoided when driving a Toyota and/or Lexus brand motor vehicle equipped with the safety systems ...
www.toyota.no/Images/Safety_Brochure_tcm308-344461.pdf
Toyota - Star Safety System
Star Safety System ... Toyota Mobility Program. Careers. Contact Us. Home. contact us. site map. your privacy rights. legal terms. Toyota Newsroom. sign up for info ...
www.toyota.com/vehicles/demos/star-safety.html - 58k - Cached
Toyota Prius Safety Ratings - CarsDirect
Get overall safety ratings and NHTSA crash test results for the Toyota Prius at CarsDirect.
Sponsored Results
Safety for a Toyota
Research Safety Ratings and Reviews For New Car at Kelley Blue Book.
www.kbb.com
Toyota Safety
Find Toyota Safety dealers, new cars, prices, and photos.
www.NewCars.org
fair
Toyota Safety
Toyota safety Discount Prices Save Money Shopping Online Today.
www.smarter.com
Good
Saftey Toyoto
Explore 5,000+ Pro Sports Choices. Save On Saftey Toyoto.
BaseballGear.Shopzilla.com
See your message here...

Discounted Cumulative Gain (DCG)

- Δημοφιλές μέτρο για την αποτίμηση της αναζήτησης στο web και σε παρόμοιες εφαρμογές

- Δύο κριτήρια:
  - (βαθμός συνάφειας) Έγγραφα με μεγάλη συνάφεια είναι πιο χρήσιμα από οριακά συναφή έγγραφα
  - (θέση στη διάταξη) Όσο πιο χαμηλά στη διάταξη εμφανίζεται ένα έγγραφο, τόσο λιγότερο χρήσιμο είναι για ένα χρήστη, αφού είναι λιγότερο πιθανό να το εξετάσει

Discounted Cumulative Gain

- Έστω αξιολογήσεις συνάφειας στην κλίμακα [0, r], r>2 και ότι οι αξιολογήσεις των *n* πρώτων εγγράφων είναι $r_1, r_2, \ldots r_n$ (σε σειρά διάταξης)

- Χρήση βαθμιδωτής (graded) συνάφειας ως μέτρου της χρησιμότητας ή του κέρδους (gain) από την εξέταση ενός εγγράφου
- Το κέρδος
  - συγκεντρώνεται/αθροίζεται ξεκινώντας από την κορυφή της διάταξης

    Cumulative Gain (CG) στη θέση διάταξης (rank) *n*

    $CG = r_1+r_2+\ldots r_n$

Θέση	Δυαδική	Βαθμωτή
1	R	3
2	R	2
3	N	0
4	R	3
5	N	0
6	R	1
7	R	2
8	R	1
9	R	1
10	N	0
11	N	0
12	N	0
13	R	3
14	N	0

Discounted Cumulative Gain

- μειώνεται ή γίνεται έκπτωση (discounted) στα χαμηλότερα επίπεδα
  - Discounted Cumulative Gain (DCG) στη θέση διάταξης *n*
  - $DCG = r_1 + r_2/\log_2 2 + r_3/\log_2 3 + \ldots r_n/\log_2 n$
    - Η σχετική μείωση είναι *1/log (rank)*
    - Για βάση 2, η μείωση του κέρδους, στο επίπεδο 4 είναι 1/2 και στο επίπεδο 8 είναι 1/3
    - Χρησιμοποιούνται και άλλες βάσεις εκτός του 2 για το λογάριθμο

Θέση	Δυαδική	Βαθμωτή	Discount
1	**R**	3	1
2	**R**	2	1
3	N	0	1.59
4	**R**	3	2
5	N	0	2.32
6	**R**	1	2.59
7	**R**	2	2.81
8	**R**	1	3
9	**R**	1	3.17
10	N	0	3.32
11	N	0	3.46
12	N	0	3.58
13	**R**	3	3.7
14	N	0	3.8

Discounted Cumulative Gain

Θέση	Δυαδική	Βαθμωτή	Discount
1	**R**	3	1
2	**R**	2	1
3	N	0	1.59
4	**R**	3	2
5	N	0	2.32
6	**R**	1	2.59
7	**R**	2	2.81
8	**R**	1	3
9	**R**	1	3.17
10	N	0	3.32
11	N	0	3.46
12	N	0	3.58
13	**R**	3	3.7
14	N	0	3.8

Discounted Cumulative Gain

Παράδειγμα

- 10 διατεταγμένα έγγραφα σε κλίμακα συνάφειας 0-3:
  3, 2, 3, 0, 0, 1, 2, 2, 3, 0
- Cumulative gain
  3, 5, 8, 8, 8, 9, 11, 13, 16, 16
- Discounted gain:
  3, 2/1, 3/1.59, 0, 0, 1/2.59, 2/2.81, 2/3, 3/3.17, 0
  = 3, 2, 1.89, 0, 0, 0.39, 0.71, 0.67, 0.95, 0
- DCG:
  3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61

Discounted Cumulative Gain

- DCG το ολικό κέρδος που συγκεντρώνεται σε μια συγκεκριμένη θέση διάταξης ***p***:

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}$$

- Εναλλακτική διατύπωση:

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log(1+i)}$$

  - Χρησιμοποιείται από κάποιες μηχανές
  - Μεγαλύτερη έμφαση στην ανάκτηση *πολύ σχετικών* εγγράφων

Κανονικοποιημένο DCG (NDCG)

- Normalized Discounted Cumulative Gain (NDCG) στη θέση διάταξης *n*

  Κανονικοποιούμε το DCG στη θέση διάταξης *n* με την DGG τιμή στη θέση διάταξης *n* *για την ιδανική διάταξη*

  - *Ιδανική διάταξη*: επιστρέφει πρώτα τα έγγραφα που έχουν τον υψηλότερο βαθμό συνάφειας, μετά τα έγγραφα με τον αμέσως υψηλότερο βαθμό, κοκ

- Χρήσιμο για αντιπαράθεση ερωτημάτων με διαφορετικό αριθμό συναφών αποτελεσμάτων
- Ιδιαίτερα δημοφιλές μέτρο στην αναζήτηση στο web

Κανονικοποιημένο DCG (NDCG)

Παράδειγμα

4 έγγραφα: d_1, d_2, d_3, d_4

i	Ground Truth (optimal)		Ranking Function$_1$		Ranking Function$_2$	
	Document Order	r_i	Document Order	r_i	Document Order	r_i
1	d4	2	d3	2	d3	2
2	d3	2	d4	2	d2	1
3	d2	1	d2	1	d4	2
4	d1	0	d1	0	d1	0
	$NDCG_{GT}$=1.00		$NDCG_{RF1}$=1.00		$NDCG_{RF2}$=0.9203	

$$DCG_{GT} = 2 + \left(\frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.6309$$

$$DCG_{RF1} = 2 + \left(\frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.6309$$

$$DCG_{RF2} = 2 + \left(\frac{1}{\log_2 2} + \frac{2}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.2619$$

$$MaxDCG = DCG_{GT} = 4.6309$$

Ποια είναι η ιδανική διάταξη αν θεωρήσουμε ότι τα μόνα συναφή είναι αυτά τα 8 έγγραφα, και ποιο το DCG της;

Θέση	Δυαδική	Βαθμωτή	Discount
1	**R**	3	1
2	**R**	2	1
3	N	0	1.59
4	**R**	3	2
5	N	0	2.32
6	**R**	1	2.59
7	**R**	2	2.81
8	**R**	1	3
9	**R**	1	3.17
10	N	0	3.32
11	N	0	3.46
12	N	0	3.58
13	**R**	3	3.7
14	N	0	3.8

NDCG_{14}

NDCG_{4}

Παρατήρηση: Διασπορά (Variance)

- Για μια συλλογή ελέγχου, συχνά ένα σύστημα έχει *κακή* απόδοση *σε κάποιες* πληροφοριακές ανάγκες (π.χ., MAP = 0.1) και *άριστη* σε άλλες (π.χ., MAP = 0.7)
- Συχνά, η διασπορά στην απόδοση είναι πιο μεγάλη για *διαφορετικά ερωτήματα* του ίδιου συστήματος από τη διασπορά στην απόδοση διαφορετικών συστημάτων για την ίδια ερώτηση

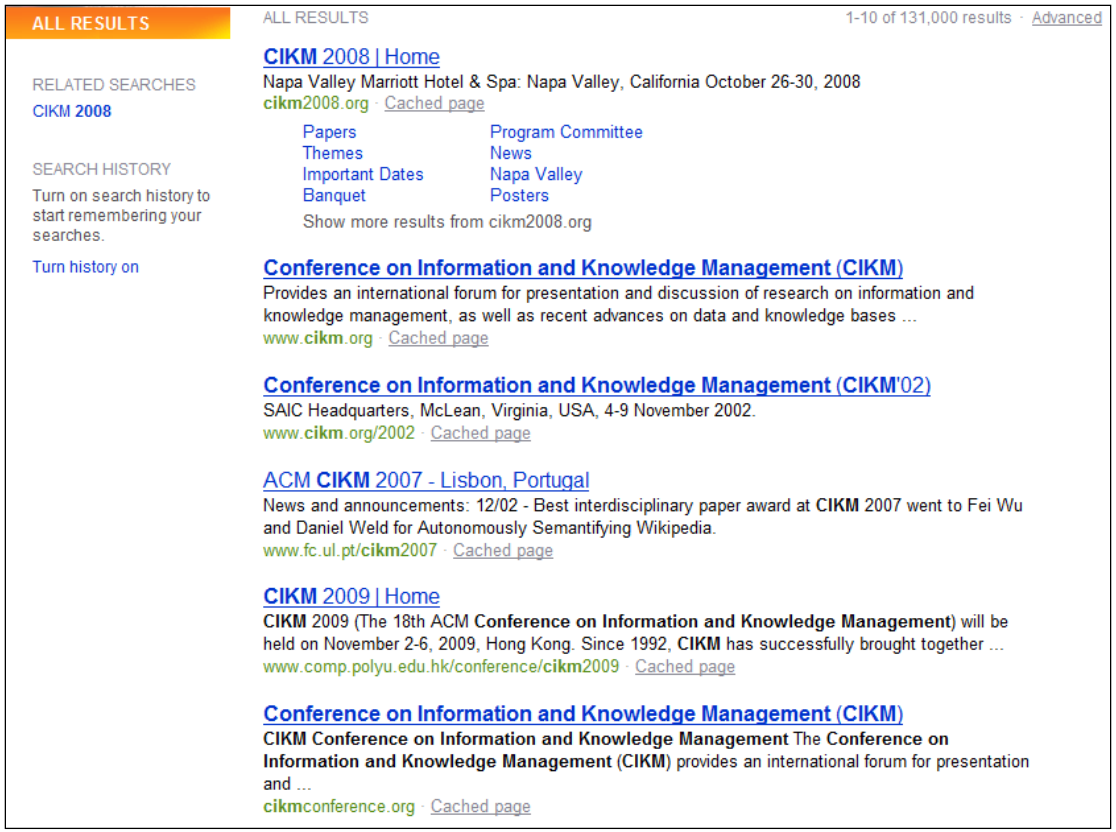
- Δηλαδή, υπάρχουν εύκολες ανάγκες πληροφόρησης και δύσκολες ανάγκες πληροφόρησης!

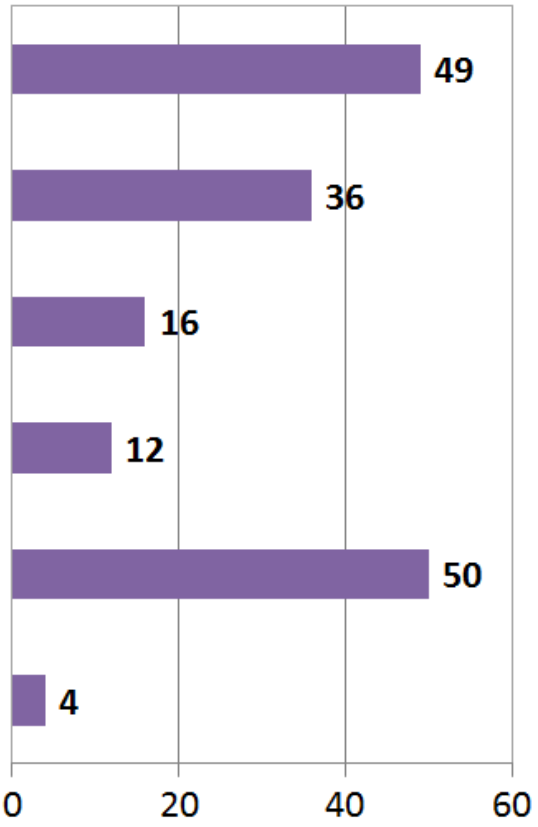
Με χρήση clickthrough

Χρήση δεδομένων clickthrough

- Περιορισμένος αριθμός αξιολογήσεων συνάφειας
- Χρήση του log (ιστορίας) των ερωτήσεων - δισεκατομμύρια από ερωτήσεις και αποτελέσματα
- Ανάλυση της πληροφορίας από το clickthrough data: πόσο συχνά κάποιος χρήστης επιλέγει (clicks) σε ένα συγκεκριμένο έγγραφο, όταν αυτό εμφανίζεται στο αποτέλεσμα μιας ερώτησης

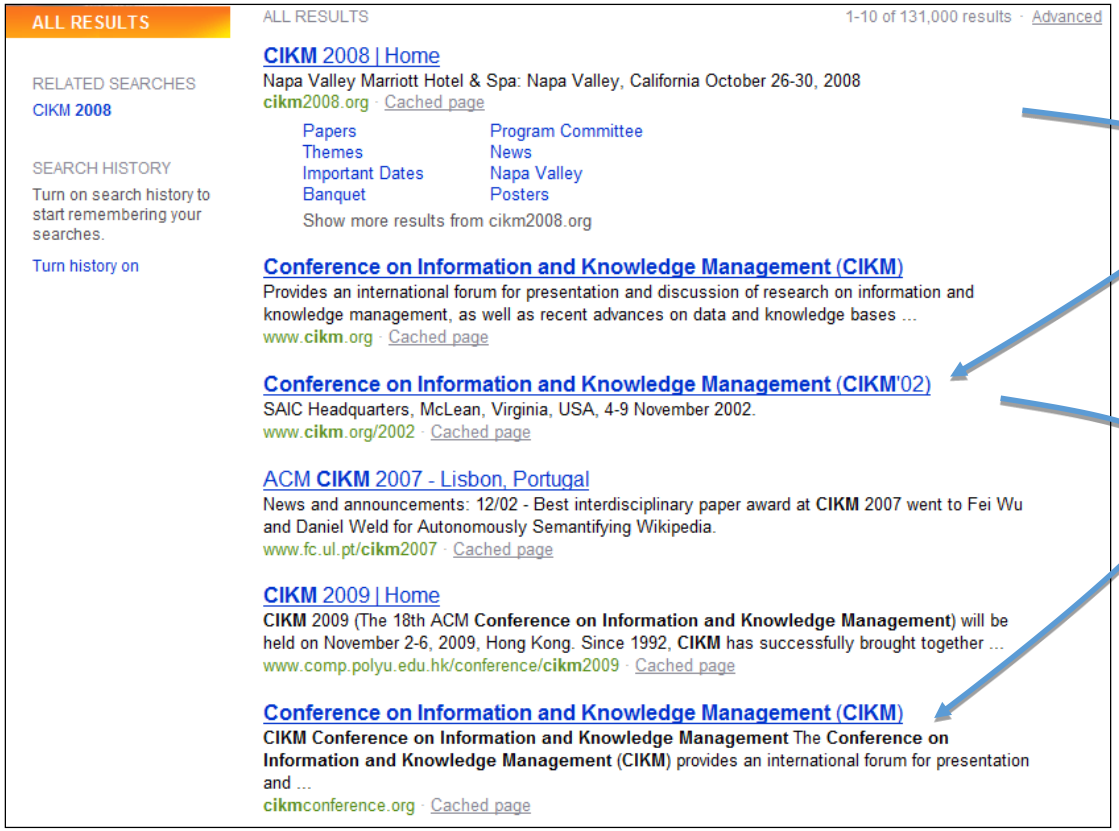
Τι μας λένε οι αριθμοί;





Έχει μεγάλη σημασία η θέση, απόλυτοι αριθμοί όχι ιδιαίτερα αξιόπιστοι
Επηρεάζει η διάταξη (πόσα clicks ένα έγγραφο στη θέση 10,000;)

Σχετική και απόλυτη διάταξη



User's click sequence

Δύσκολο να αποφασίσουμε αν Result1 > Result3

Πιθανών να μπορούμε να πούμε ότι Result3 > Result2

Pairwise relative ratings

- *Ζεύγη* της μορφής: DocA καλύτερο του DocB για μια ερώτηση
  - Δε σημαίνει (απαραίτητα) ότι το DocA είναι συναφές με το ερώτημα
- Αντί για αξιολογήσεις μιας διάταξης εγγράφων συγκεντρώνουμε ένα *ιστορικό από ζεύγη προτιμήσεων* με βάση τα clicks των χρηστών
- Αξιολόγηση με βάση το πόσο *«συμφωνεί»* το αποτέλεσμα με τα ζεύγη των διατάξεων
  - (επίσης) χρήση για μάθηση διάταξης (learn to rank)
- Με βάση διαφορετικές μηχανές-αλγορίθμους διάταξης

Πως θα συγκρίνουμε ζεύγη προτιμήσεων;

Έστω δύο σύνολα *P* και *A* από ζεύγη προτιμήσεων. Θέλουμε ένα μέτρο εγγύτητας (proximity measure) που να λέει πόσο μοιάζουν

- Το μέτρο πρέπει να ανταμείβει τις συμφωνίες και να τιμωρεί τις διαφωνίες

Απόσταση Kendall tau

- Έστω X ο αριθμός των συμφωνιών και Y ο αριθμός των διαφωνιών η Kendall tau distance μεταξύ *A* και *P* είναι (X-Y)/(X+Y)

Παράδειγμα

P = {(1,2), (1,3), (1,4), (2,3), (2,4), (3,4)}

A = (1, 3, 2, 4)

- X=5, Y=1

Ποια είναι η μέγιστη και ποια η ελάχιστη τιμή;

Άσκηση

απόσταση **Kendal tau** μεταξύ δυο ταξινομημένων λιστών

Λέγεται και απόσταση bubble-sort = ο αριθμός των ανταλλαγών (swaps) από τη μια λίστα στην άλλη

Άσκηση

(α) Υπολογίστε την απόσταση Kendal tau μεταξύ

1 2 3 4 5

3 4 1 2 5

(β) Ποια λίστα έχει τη μεγαλύτερη απόσταση με την πρώτη; με τη δεύτερη;

Χρήση clicktrough για σύγκριση ΣΑΠ

Έστω δύο μηχανές αναζήτησης (ή αλγόριθμοι διάταξης) που θέλουμε να συγκρίνουμε με χρήση clickthrough

R_A
Clicks στο 2, 3, 4
R_B
Clicks στο 1 και 5
Καλύτερο?

- Τα έγγραφα στις απαντήσεις μπορεί να είναι διαφορετικά
- Click σε ένα έγγραφο μπορεί να σημαίνει καλύτερο από τα άλλα στην απάντηση

Mix των αποτελεσμάτων: δημιούργησε μια απάντηση που να περιέχει και τα δυο, μέτρα clicks για το A και το B

Επανάληψη

Μετρικές για την αξιολόγηση της συνάφειας

	Όχι διάταξη	Διάταξη
Δυαδικές	Ακρίβεια (Precision) Ανάκληση (Recall) F1	Καμπύλη Ακρίβειας-Ανάκλησης (με παρεμβολή, 11-σημείων) Ακρίβεια@k MAP R-Precision MRR ROC
Μη δυαδικές	GC	DGC, NDGC

Τεχνικές βασισμένες σε ζεύγη: Kendall tau

Αξιοπιστία χρηστών

Αξιολογήσεις από ανθρώπους

- Ακριβές
- Μη συνεπείς
  - Ανάμεσα στους αξιολογητές, ή
  - Και σε διαφορετικές χρονικές στιγμές
- Όχι πάντα αντιπροσωπευτικές των πραγματικών χρηστών
  - Αξιολόγηση με βάση το ερώτημα και όχι την ανάγκη

Αξιοπιστία των αξιολογήσεων των κριτών

- Οι αξιολογήσεις συνάφειας είναι χρήσιμες αν είναι συνεπής (consistent).
- Πως μπορούμε να μετρήσουμε τη συνέπεια ή τη *συμφωνία ανάμεσα στους κριτές*

Μέτρο Kappa της διαφωνίας (συμφωνίας) (disagreement) μεταξύ των κριτών

- Συμφωνία μεταξύ των κριτών
- Αφορά κατηγορική κρίση (δυαδική αξιολόγηση)
- Λαμβάνει υπό όψιν την συμφωνία από τύχη

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

P(A): ποσοστό των περιπτώσεων που οι κριτές συμφωνούν *(observed agreement)*

P(E): τι συμφωνία θα είχαμε από τύχη *(expected agreement)*

κ = 1 για πλήρη συμφωνία, 0 για τυχαία συμφωνία, αρνητική για μικρότερη της τυχαίας

Kappa: παράδειγμα

Συνολικά 400 έγγραφα

Αριθμός εγγράφων	ΚΡΙΤΗΣ 1	ΚΡΙΤΗΣ 2
300	Relevant	Relevant
70	Nonrelevant	Nonrelevant
20	Relevant	Nonrelevant
10	Nonrelevant	Relevant

Κριτής 1: 320 R, 80 N
Κριτής 2: 310 R, 90 N

Kappa: παράδειγμα

Αριθμός εγγράφων	ΚΡΙΤΗΣ 1	ΚΡΙΤΗΣ 2
300	Relevant	Relevant
70	Nonrelevant	Nonrelevant
20	Relevant	Nonrelevant
10	Nonrelevant	Relevant

Συνολικά 400 έγγραφα

P(A) = 370/400 = 0.925

P(nonrelevant) = 80/400 * 90/400 = 0.045

P(relevant) = 320/400 * 310/400 = 0.62

P(E) = 0.045 + 0.62 = 0.665

Kappa = (0.925 – 0.665)/(1 - 0.665) = 0.776

Kappa

- Kappa > 0.8 = καλή συμφωνία

0.67 < Kappa < 0.8 -> “tentative conclusions”

- Εξαρτάται από το στόχο της μελέτης

- Για >2 κριτές: μέσοι όροι ανά-δύο κλπ

Kappa: παράδειγμα

Information need	number of docs judged	disagreements
51	211	6
62	400	157
67	400	68
95	400	110
127	400	106

Συμφωνία κριτών στο TREC

Επίπτωση της Διαφωνίας

- Επηρεάζει την απόλυτη (absolute) μέτρηση απόδοσης αλλά όχι τη σχετική απόδοση ανάμεσα σε συστήματα

- Μπορούμε να αποφύγουμε τις κρίσεις από χρήστες
  - Όχι
- Αλλά μπορούμε να τις επαναχρησιμοποιήσουμε

Αξιολόγηση σε μεγάλες μηχανές αναζήτησης

- Οι μηχανές αναζήτησης διαθέτουν συλλογές ελέγχου ερωτημάτων και αποτελέσματα διατεταγμένα με το χέρι (hand-ranked)

- Στο web είναι δύσκολο να υπολογίσουμε την ανάκληση

  Συνήθως οι μηχανές αναζήτησης χρησιμοποιούν την ακρίβεια στα κορυφαία *k* π.χ., *k* = 10

  Επίσης το MAP, NDCG στα κορυφαία

Αξιολόγηση σε μεγάλες μηχανές αναζήτησης

Οι μηχανές αναζήτησης χρησιμοποιούν επίσης και άλλα μέτρα εκτός της συνάφειας

- *Clickthrough on first result*
  - Όχι πολύ αξιόπιστο όταν ένα clickthrough (μπορεί απλώς η περίληψη να φάνηκε χρήσιμη αλλά όχι το ίδιο το έγγραφο) αλλά αρκετά αξιόπιστο συναθροιστικά
- Μετρήσεις σε εργαστήριο
- Έλεγχος A/B

A/B testing

Στόχος: έλεγχος μιας νέας ιδέας (a single innovation)

Προϋπόθεση: Υπάρχει μια μεγάλη μηχανή αναζήτησης σε λειτουργία

- Οι πιο πολλοί χρήστες χρησιμοποιούν τα παλιό σύστημα
- Παράκαμψε ένα μικρό ποσοστό της κυκλοφορίας (π.χ., 1%) στο νέο σύστημα που χρησιμοποιεί την καινούργια
- Αξιολόγησε με ένα αυτόματο μέτρο όπως το clickthrough στα πρώτα αποτελέσματα

Crowdsourcing

Σημαντικός ο σχεδιασμός του πειράματος

Το Mechanical Truck της Amazon

https://www.mturk.com/mturk/welcome

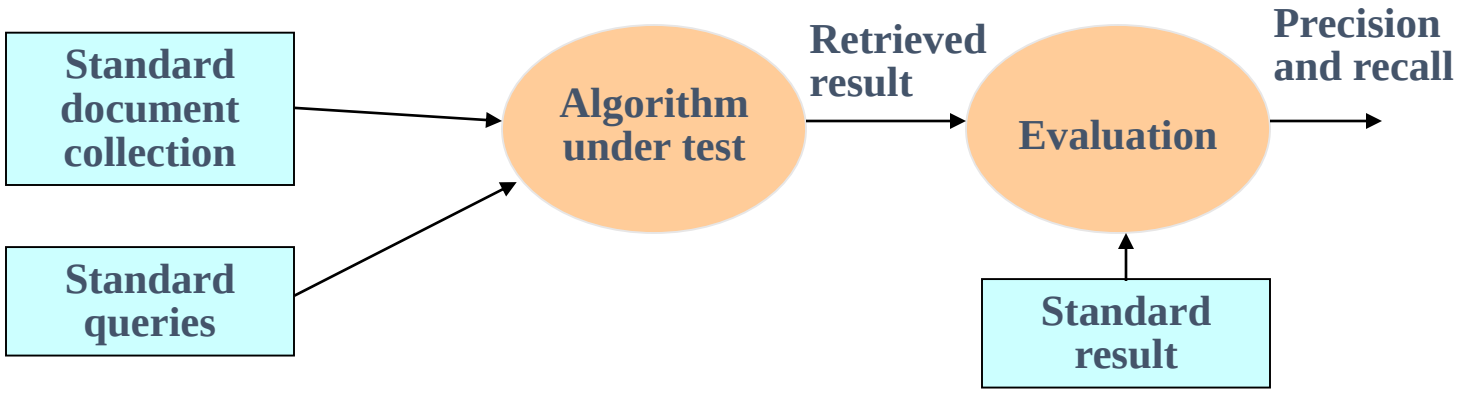
- HITs - *Human Intelligence Tasks*
- Workers

Μεθοδολογία – πρότυπες συλλογές (benchmarks)

Απαιτήσεις από ένα πρότυπο (benchmark)

1. Ένα σύνολο από έγγραφα
   - Τα έγγραφα πρέπει να είναι αντιπροσωπευτικά των πραγματικών εγγράφων
2. Μια συλλογή από ανάγκες πληροφόρησης
   - (ή, καταχρηστικά, ερωτημάτων)
   - Να σχετίζονται με τα διαθέσιμα έγγραφα
   - Οι ανάγκες πληροφόρησης πρέπει να είναι αντιπροσωπευτικές των πραγματικών - τυχαίοι όροι δεν είναι καλή ιδέα
   - Συχνά από ειδικούς της περιοχής
3. Εκτιμήσεις συνάφειας από χρήστες (Human relevance assessments)
   - Χρειάζεται να προσλάβουμε/πληρώσουμε κριτές ή αξιολογητές.
   - Ακριβό χρονοβόρο
   - Οι κριτές πρέπει να είναι αντιπροσωπευτικοί των πραγματικών χρηστών

Benchmarks


Standard document collection
Standard queries
Algorithm under test
Retrieved result
Evaluation
Precision and recall
Standard result

Standard benchmarks συνάφειας

TREC - National Institute of Standards and Technology (NIST) τρέχει ένα μεγάλο IR test bed εδώ και πολλά χρόνια

- Χρησιμοποιεί το Reuters και άλλες πρότυπες συλλογές εγγράφων
- Καθορισμένα "Retrieval tasks"
  - Μερικές φορές ως ερωτήματα
- Ειδικοί (Human experts) βαθμολογούν κάθε ζεύγος ερωτήματος, εγγράφου ως Συναφές Relevant ή μη Συναφές Nonrelevant
  - Ή τουλάχιστον ένα *υποσύνολο των εγγράφων* που επιστρέφονται για κάθε ερώτημα

Standard benchmarks συνάφειας

Cranfield

Πρωτοπόρο: το πρώτο testbed που επέτρεπε ακριβή ποσοτικοποιημένα μέτρα της αποτελεσματικότητας της ανάκτησης

Στα τέλη του 1950, UK

- 1398 abstracts (περιλήψεις) από άρθρα περιοδικών αεροδυναμικής, ένα σύνολο από 225 ερωτήματα, εξαντλητική κρίση συνάφειας όλων των ζευγών
- Πολύ μικρό, μη τυπικό για τα σημερινά δεδομένα της ΑΠ

TREC

TREC Ad Hoc task από τα πρώτα 8 TRECs είναι ένα standard task, μεταξύ του 1992-1999

- *1.89 εκατομμύρια έγγραφα*, κυρίως newswire άρθρα
- *50* λεπτομερείς ανάγκες πληροφόρησης το χρόνο (σύνολο 450)
- Επιστρέφετε η *αξιολόγηση χρηστών σε pooled αποτελέσματα* (δηλαδή όχι εξαντλητική αξιολόγηση όλων των ζευγών)
- και Web track

A TREC query (TREC 5)

<top>

<num> Number: 225

<desc> Description:

What is the main function of the Federal Emergency Management Agency (FEMA) and the funding level provided to meet emergencies? Also, what resources are available to FEMA such as people, equipment, facilities?

</top>

TREC

TREC-COVID

The document set used in the challenge is the COVID-19 Open Research Dataset (CORD-19): a collection of biomedical literature articles

In rounds, participants created ranked lists of documents for each topic ("runs") and submitted their runs to NIST.

Based on the collective set of participants' runs, NIST created sets of documents to be assessed for relevance by human annotators with biomedical expertise.

The results of the human annotation, known as relevance judgments, were then used to score the submitted runs.

The final document and topic sets together with the cumulative relevance judgments comprise a COVID test collection called TREC-COVID Complete.

Incremental nature of the collection as viewed through the successive rounds supports research on search systems for *dynamic environments*.

https://ir.nist.gov/covidSubmit/data.html

Άλλα benchmarks

- GOV2
  - Ακόμα μια TREC/NIST συλλογή
  - 25 εκατομμύρια web σελίδες
  - Αλλά ακόμα τουλάχιστον τάξης μεγέθους μικρότερη από το ευρετήριο της Google/Yahoo/MSN
- NTCIR
  - Ανάκτηση πληροφορίας για τις γλώσσες της Ανατολικής Ασίας και cross-language ανάκτηση
- Cross Language Evaluation Forum (CLEF)
  - Το ίδιο για Ευρωπαϊκές γλώσσες

Συλλογές ελέγχου

TABLE 4.3 Common Test Corpora

Collection	*NDocs*	*NQrys*	*Size (MB)*	*Term/Doc*	*Q-D RelAss*
ADI	82	35			
AIT	2109	14	2	400	>10,000
CACM	3204	64	2	24.5	
CISI	1460	112	2	46.5	
Cranfield	1400	225	2	53.1	
LISA	5872	35	3		
Medline	1033	30	1		
NPL	11,429	93	3		
OSHMED	34,8566	106	400	250	16,140
Reuters	21,578	672	28	131	
TREC	740,000	200	2000	89-3543	» 100,000

Κριτική της Συνάφειας

- Οριακή Συνάφεια (Marginal Relevance)
  - «νέα» έγγραφα
- Και άλλα κριτήρια όπως
  - Novelty
  - Coverage

Ασκήσεις

Άσκηση 8.5

Υπάρχει ή όχι πάντα ένα σημείο (θέση στη διάταξη) στο οποίο *η ακρίβεια είναι ίση με την ανάκληση* (break-even point – σημείο εξισορρόπησης).

Αν ναι, αποδείξτε το, αν όχι, δώστε αντιπαράδειγμα

Με ποιο μέτρο έχει σχέση;

Precision Recall curve

Άσκηση 8.4

Ποιες είναι οι πιθανές τιμές της ακρίβειας με παρεμβολή στο επίπεδο ανάκλησης 0;

πότε παίρνει την τιμή 1 και πότε την τιμή 0;

MAP I

Άσκηση 8.8

Έστω μια ανάγκη πληροφόρησης για την οποία υπάρχουν 4 συναφή έγγραφα. Δύο ΣΑΠ δίνουν τα παρακάτω αποτελέσματα:

ΣΥΣΤΗΜΑ 1: **R** N **R** N N N N N **R R**

ΣΥΣΤΗΜΑ 2: N **R** N **R R R** N N N N

- Υπολογίστε το MAP. Ποιο σύστημα είναι καλύτερο; Είναι διαισθητικά σωστό; Τι μας λέει για το τι είναι σημαντικό για ένα καλό MAP; Βέλτιστη απάντηση;

Άσκηση 8.8 (συνέχεια)

Έστω μια ανάγκη πληροφόρησης για την οποία υπάρχουν 4 συναφή έγγραφα. Δύο ΣΑΠ δίνουν τα παρακάτω αποτελέσματα:

ΣΥΣΤΗΜΑ 1: **R** N **R** N N N N N **R R**

ΣΥΣΤΗΜΑ 2: N **R** N **R R R** N N N N

- Υπολογίστε την R-ακρίβεια. Ποιο σύστημα είναι καλύτερο;

MAP II

Έστω μια ανάγκη πληροφόρησης για την οποία υπάρχουν 2 συναφή έγγραφα. Έχετε την αρχή της απάντησης δύο συστημάτων

ΣΥΣΤΗΜΑ 1: **R** …

ΣΥΣΤΗΜΑ 2: N **R R** ...

Το MAP του Συστήματος 2 είναι

A. 1.167

B. 0.583

C. Δεν μπορεί να υπολογιστεί γιατί δεν βλέπουμε όλο το αποτέλεσμα

D. 0.393

Άσκηση 8.8 (επέκταση)

Έστω μια ανάγκη πληροφόρησης για την οποία υπάρχουν 2 συναφή έγγραφα. Έχετε την αρχή της απάντησης δύο συστημάτων

ΣΥΣΤΗΜΑ 1: **R** …

ΣΥΣΤΗΜΑ 2: N **R R** ...

- Για να είναι το Σύστημα 2 καλύτερο (ως αναφορά το MAP) σε ποια θέση θα πρέπει να εμφανίζεται το επόμενο συναφές έγγραφο στο Σύστημα 1;
- Τι ισχύει για την R-ακρίβεια;

Άσκηση 8.9: Συλλογή από *10.000 έγγραφα*

Μια ερώτηση για την οποία υπάρχουν συνολικά *8 συναφή έγγραφα*

Τα πρώτα 20 αποτελέσματα:

R R N N N N N N **R** N **R** N N N **R** N N N N **R**

Υπολογίστε:

- Την ακρίβεια στα πρώτα 20
- Το F1 στα πρώτα 20?
- Ποια είναι η ακρίβεια χωρίς παρεμβολή για 50% ανάκληση;
- Αν θεωρείστε ότι αυτή είναι όλη η απάντηση, ποια είναι η ακρίβεια με παρεμβολή για 60% ανάκληση;

Άσκηση 8.9

Επιστρέφει ***όλα τα 10.000*** έγγραφα και αυτά είναι *τα πρώτα 20* αποτελέσματα:

R R N N N N N N **R** N **R** N N N **R** N N N N **R**

- Ποια είναι η μεγαλύτερη δυνατή MAP τιμή και ποια η μικρότερη δυνατή MAP τιμή

R R N N N N N N **R** N **R** N N N **R** N N N N **R R R** .. N N N .... N N

2

R R N N N N N N **R** N **R** N N N **R** N N N N **R** N N .. N N N ... **R R**

2

Kappa, precision, recall

Άσκηση 8.10 (kappa)

Στον πίνακα βλέπουμε τις αξιολογήσεις 2 χρηστών σχετικά με τη συνάφεια 12 εγγράφων ως προς ένα ερώτημα. Έστω ότι το αποτέλεσμα είναι τα έγγραφα 4, 5, 6, 7, 8

docID	Judge 1	Judge 2
1	0	0
2	0	0
3	1	1
4	1	1
5	1	0
6	1	0
7	1	0
8	1	0
9	0	1
10	0	1
11	0	1
12	0	1

(α) Υπολογίστε το kappa
(β) Υπολογίστε την ακρίβεια και την ανάκληση αν υποθέσουμε ότι ένα έγγραφο είναι συναφές μόνο αν συμφωνούν και οι δύο κριτές
(γ) Υπολογίστε την ακρίβεια και την ανάκληση αν υποθέσουμε ότι ένα έγγραφο είναι συναφές αρκεί ένας από τους κριτές να το θεωρεί συναφές.

NDCG

Υποθέστε ότι έχουμε μη δυαδικές αποτιμήσεις συνάφειας με βαθμολογίες από 0 έως 5.

Συγκεκριμένα, όλα τα μη συναφή έγγραφα έχουν βαθμό 0, ενώ τα συναφή έγγραφα έχουν τους εξής βαθμούς

d10 βαθμός 4,
d25 βαθμός 5,
d190 βαθμός 3,
d350 βαθμός 4,
d400 βαθμός 2,
d434 βαθμός 5,
d700 βαθμός 1,
d701 βαθμός 3,
d900 βαθμός 2,
d990 βαθμός 5.

Απάντηση: d701 d190 d350 d100 d206 d990 d10 d890

- Υπολογίστε το Discounted Cumulative Gain (DCG) στη θέση διάταξης 5
- Υπολογίστε το κανονικοποιημένο Discounted Cumulative Gain (NDCG) στη θέση διάταξης 5

ΤΕΛΟΣ 8ου Κεφαλαίου

Ερωτήσεις?

Χρησιμοποιήθηκε κάποιο υλικό από:

- ✓ *Pandu Nayak and Prabhakar Raghavan, CS276:Information Retrieval and Web Search (Stanford)*
- ✓ *Hinrich Schütze and Christina Lioma, Stuttgart IIR class*
- ✓ *διαφάνειες του καθ. Γιάννη Τζίτζικα (Παν. Κρήτης)*