

# Introduction to Information Retrieval

ΜΥΕΕ03-ΠΛΕ70: Ανάκτηση Πληροφορίας

*Διδάσκουσα: Ευαγγελία Πιτουρά*

Διάλεξη 8: Αξιολόγηση στην Ανάκτηση Πληροφορίας.

# Τι είδαμε στο προηγούμενο μάθημα

---

- Βαθμολόγηση και κατάταξη εγγράφων
  - Στάθμιση όρων (term weighting)
  - Αναπαράσταση εγγράφων και ερωτημάτων ως διανύσματα
- Θέματα Υλοποίησης

# Τι θα δούμε σήμερα;

---

- Ποια τεχνική ή ποιο σύστημα ανάκτησης πληροφορίας είναι *καλύτερο*;
- Πως μπορούμε να το *αξιολογήσουμε* ένα σύστημα ανάκτησης;
  - *Αξιολόγηση* συστημάτων ανάκτησης πληροφορίας και μηχανών αναζήτησης: μεθοδολογία και μέτρα

# Αξιολόγηση συστήματος

---

## Αποδοτικότητα (Performance)

- Πόσο *γρήγορη* είναι η κατασκευή του ευρετηρίου;
  - Αριθμός εγγράφων την ώρα (throughput)
- Πόσο μεγάλο είναι το ευρετήριο (*αποθήκευση*);
- Πόσο *γρήγορη* είναι η αναζήτηση;
  - π.χ., latency (χρόνος απόκρισης) ή throughput (ρύθμο-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)

# Ποιοι είναι οι χρήστες σε μια μηχανή αναζήτησης;

---

*Ποιος είναι ο χρήστης που προσπαθούμε να ικανοποιήσουμε;  
Εξαρτάται από την εφαρμογή*

- *Μηχανές αναζήτησης στο Web: Searcher.* Επιτυχία: Ο χρήστης βρίσκει αυτό που ψάχνει. Μέτρο: ρυθμός επιστροφής στη συγκεκριμένη μηχανή αναζήτησης
- *Μηχανές αναζήτησης στο Web: Διαφημιστής.* Επιτυχία: Searcher «κλικάρει» στη διαφήμιση. Μέτρο: clickthrough rate
- *Ecommerce: Αγοραστής.* Επιτυχία: Ο αγοραστής αγοράζει κάτι. Μέτρο: χρόνος για την αγορά, ποσοστό των searchers που γίνονται αγοραστές
- *Ecommerce: Πωλητής.* Επιτυχία: Ο πωλητής πουλάει κάτι. Μέτρο: κέρδος ανά πώληση.
- *Επιχείρηση: CEO.* Επιτυχία: Οι εργαζόμενοι γίνονται πιο αποδοτικοί (λόγω αποτελεσματικής αναζήτησης). Μέτρο: κέρδος της εταιρείας.

# 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;*



# 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 συναφείας* 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

# Συνάφεια και Ανάγκη Πληροφόρησης

- Συνάφεια ως προς τι;

Συνάφεια ως προς το **ερώτημα  $q$**  ή ως προς **την ανάγκη πληροφόρησης**

- Ανάγκη Πληροφόρησης (Information need  $i$ ) : «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]

Έστω το 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 μπορεί να μετρηθεί μόνο 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*.

# 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 **δυναδικές 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*

# 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 για την οποία υπάρχουν **80 συναφή** έγγραφα.

Η απάντηση που μας δίνει το ΣΑΠ έχει **60 έγγραφα** από τα οποία τα **20 είναι συναφή** και τα **40 μη συναφή**.

Πόσο «καλό» είναι;

Πως θα μετρήσουμε τη συνάφεια του;

# 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

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

# Πίνακας Ενδεχομένων

## Πίνακας Ενδεχόμενων (Incidence Matrix)

	Συναφή (relevant)	Μη συναφή (not relevant)	
αποτέλεσμα (retrieved)	20 (TP)	40 (FP)	60
Όχι στο αποτέλεσμα (not retrieved)	60 (FN)	1,000,000 (TN)	1,000,060
	80	1,000,040	1,000,120



# Ακρίβεια και Ανάκληση

## Πίνακας Ενδεχόμενων (Incidence Matrix)

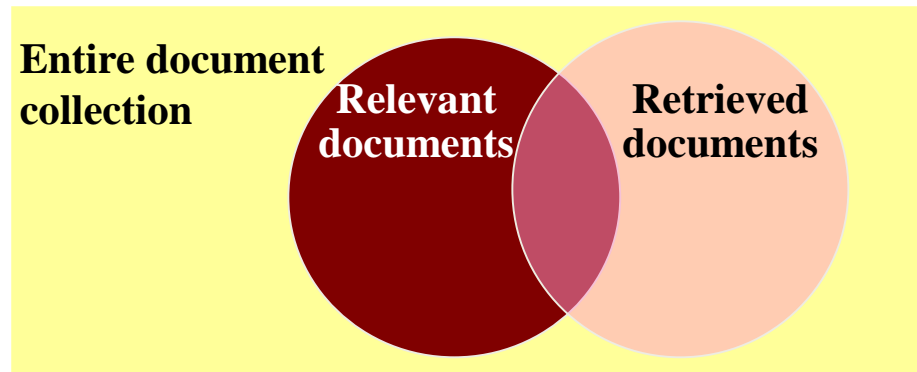
*πραγματικά*

*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

## Πίνακας Ενδεχόμενων (Incidence Matrix)

	relevant	not relevant	
retrieved	20 (TP)	40 (FP)	60
not retrieved	60 (FN)	1,000,000 (TN)	1,000,060
	80	1,000,040	1,000,120

$$\text{Precision} = 20/60 = 1/3$$

$$\text{Recall} = 20/80 = 1/4$$

# Ακρίβεια vs Ανάκληση

---

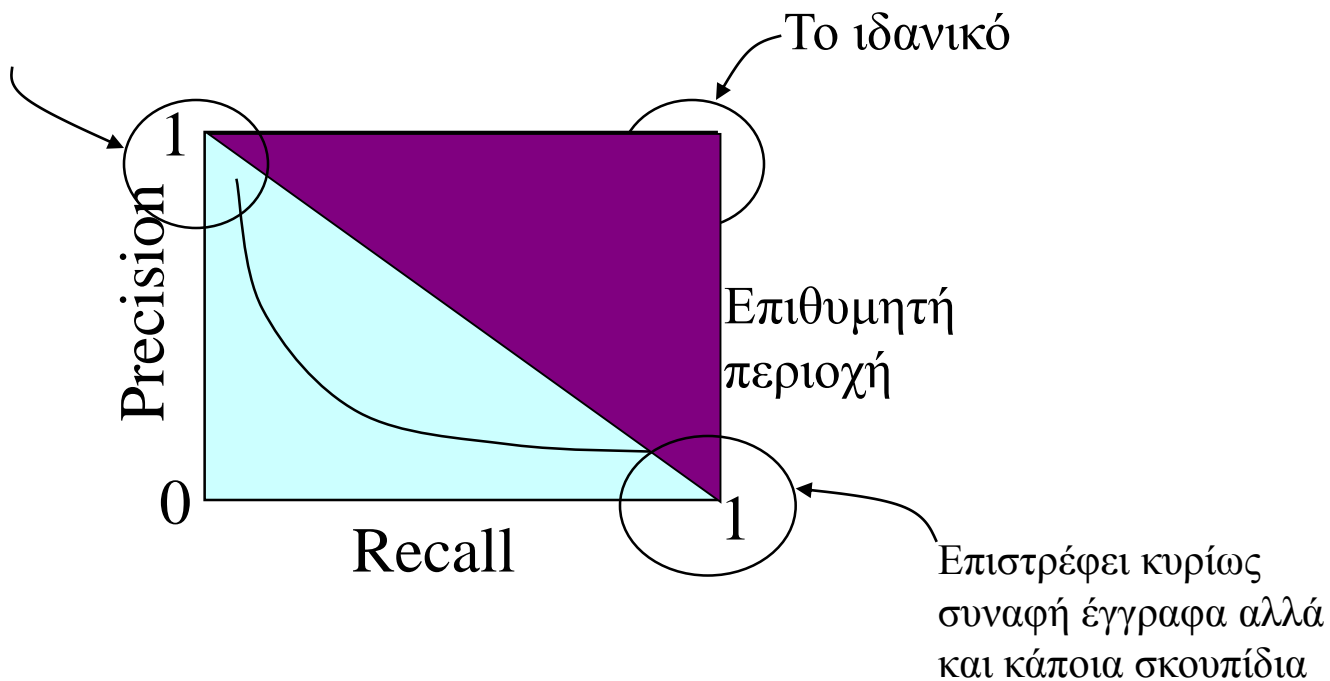
- Η ανάκληση μπορεί να αυξηθεί με το να επιστρέψουμε *περισσότερα έγγραφα*
  - Η ανάκληση είναι μια μη-φθίνουσα συνάρτηση των εγγράφων που ανακτώνται (Ένα σύστημα που επιστρέφει όλα τα έγγραφα έχει ποσοστό ανάκλησης 100%!)
- Το αντίστροφο ισχύει για την ακρίβεια (συνήθως):
  - *Είναι εύκολο να πετύχεις 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 περισσότερο εξαρτάται και από την εφαρμογή (π.χ., web vs email search)

# Ακρίβεια και Ανάκληση

Επιστρέφει συναφή έγγραφα αλλά χάνει και πολλά συναφή



# Αρμονικό Μέσο

---

Πως θα συνδυάσουμε το  $P$  και  $R$ ;

Π.χ., το **αριθμητικό μέσο** (arithmetic mean)

❖ Το απλό αριθμητικό μέσο μιας μηχανής αναζήτησης που επιστρέφει τα πάντα είναι 50%, που είναι πολύ υψηλό

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

Το  $F$  (**αρμονικό μέσο**) είναι ένα είδος ομαλού ελάχιστου

# Ένα συνδυαστικό μέτρο $F$

---

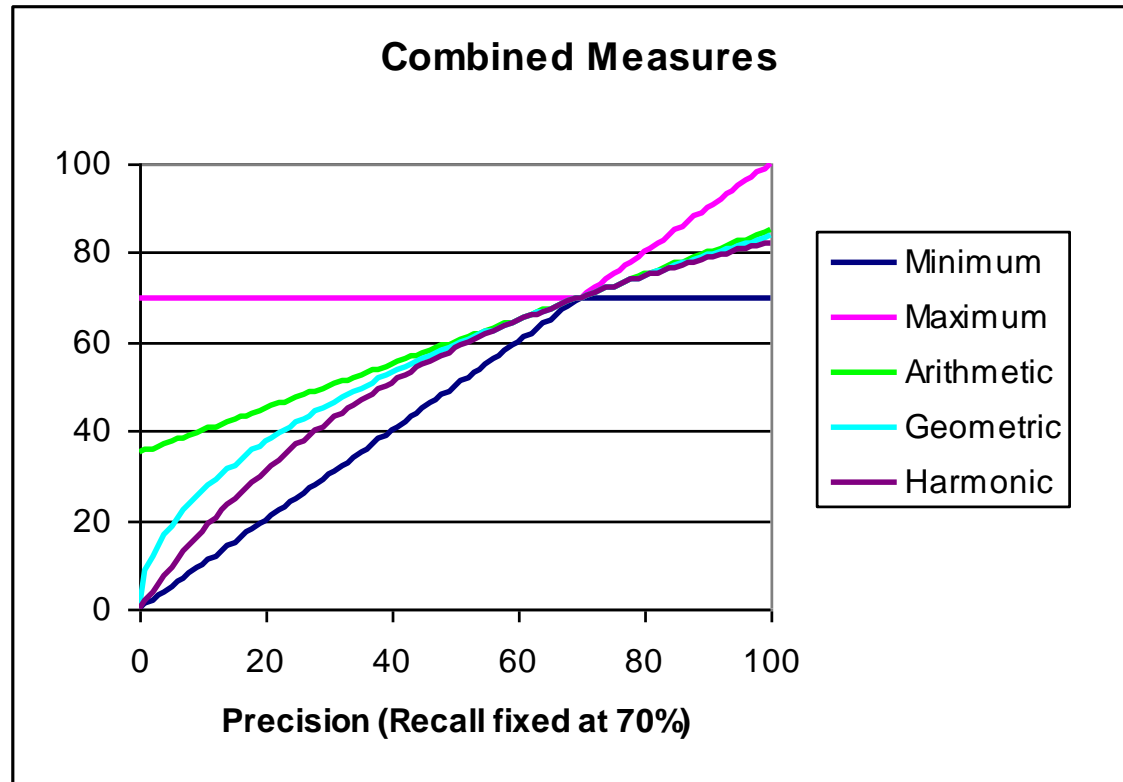
Συνήθως ισοροπημένο (balanced)  $F_1$

- **Αρμονικό μέσο** των  $P$  και  $R$

$$F_1 = 1 / [(1/2)1/P + (1/2)1/R] = 2PR/P+R$$

- ✓ Πιο κοντά στη μικρότερη από δύο τιμές

# Αρμονικό Μέσο



Τιμές στο 0-1, αλλά συνήθως σε ποσοστά

# Ένα συνδυαστικό μέτρο $F$

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

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

όπου  $\beta^2 = \frac{1 - \alpha}{\alpha}$   $\alpha \in [0, 1]$  and thus  $\beta^2 \in [0, \infty]$

Συνήθως ισορροπημένο (balanced)  $F_1$  με  $\alpha = 0.5$  και  $\beta = 1$

▪ Αυτό είναι το **αρμονικό μέσο** των  $P$  και  $R$   $\frac{1}{F} = \frac{1}{2} \left( \frac{1}{P} + \frac{1}{R} \right)$

- Για ποια περιοχή τιμών του  $\beta$  η ανάκληση σταθμίζεται περισσότερο από την ακρίβεια;



# Παράδειγμα

	relevant	not relevant	
retrieved	20	40	60
not retrieved	60	1,000,000	1,000,060
	80	1,000,040	1,000,120

$$P = 20 / (20 + 40) = 1/3$$

$$R = 20 / (20 + 60) = 1/4$$

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

# Ορθότητα (Accuracy)

---

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

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

Με βάση τον πίνακα ενδεχομένων:

$$\text{accuracy} = (TP + TN) / (TP + FP + FN + TN).$$

Γιατί αυτό δεν είναι χρήσιμο στην ΑΠ;

# Ορθότητα

---

## Παράδειγμα

---

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

---

# 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

---

- 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

Searchers 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 *δυναμικές*
  - 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 (επανάληψη)

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

### Πίνακας Ενδεχομένων

	relevant	not relevant	
retrieved	20 <b>TP</b>	40 <b>FP</b>	60
not retrieved	60 <b>FN</b>	1,000,000 <b>TN</b>	1,000,060
	80	1,000,040	1,000,120

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



# Μέτρα Συνάφειας χωρίς Διάταξη (επανάληψη)

---

## Άσκηση 8.1

Ένα ΣΑΠ επιστρέφει 8 συναφή και 10 μη συναφή έγγραφα.

Συνολικά υπάρχουν 20 συναφή έγγραφα.

Τι γίνεται όταν υπάρχει διάταξη των  
αποτελεσμάτων;

# Αξιολόγηση Καταταγμένης Ανάκτησης

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>

✓ 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)**.

# 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*:

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

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

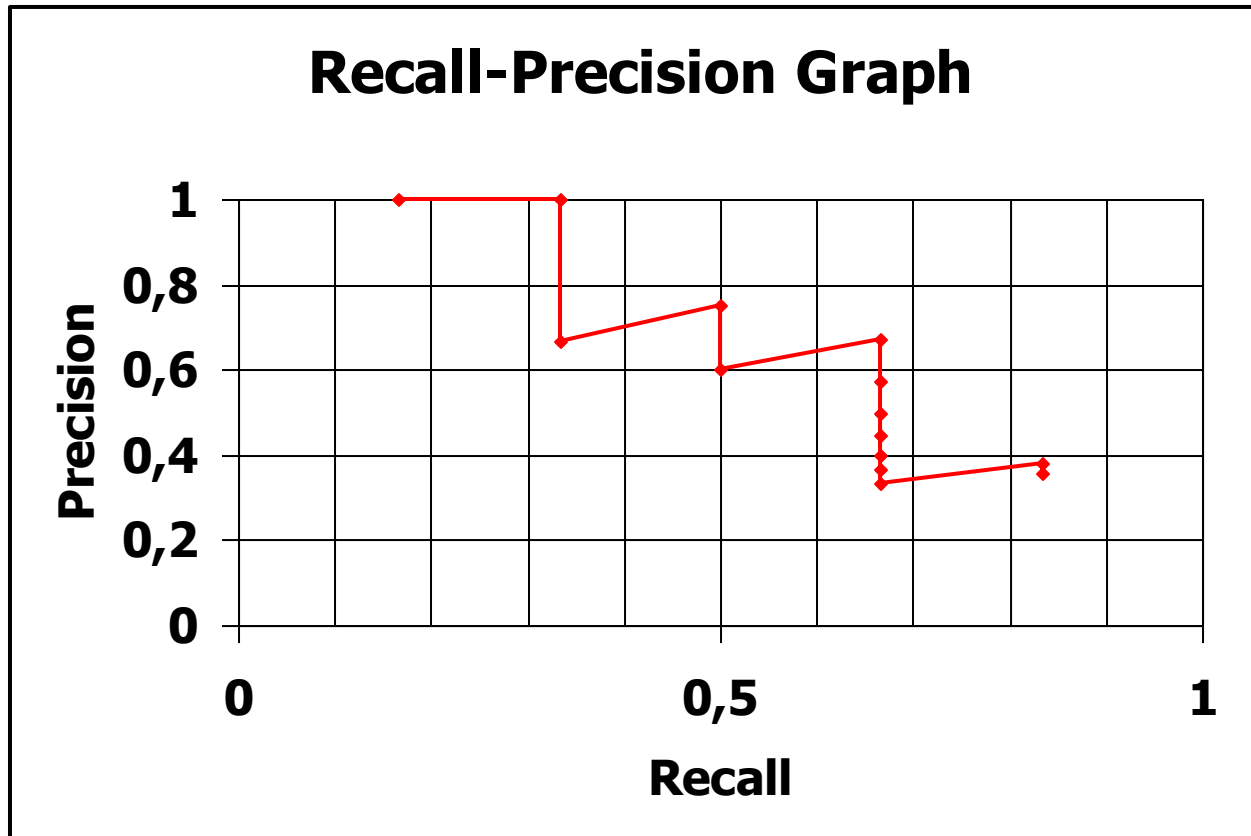
$$R=3/6=0.5; P=3/4=0.75$$

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

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

Missing one  
relevant document.  
Never reach  
100% recall

# Παράδειγμα Ι (συνέχεια)

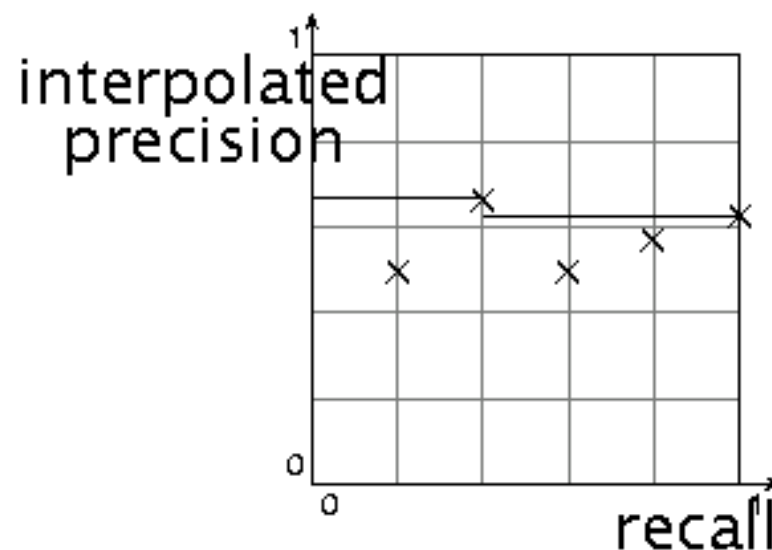
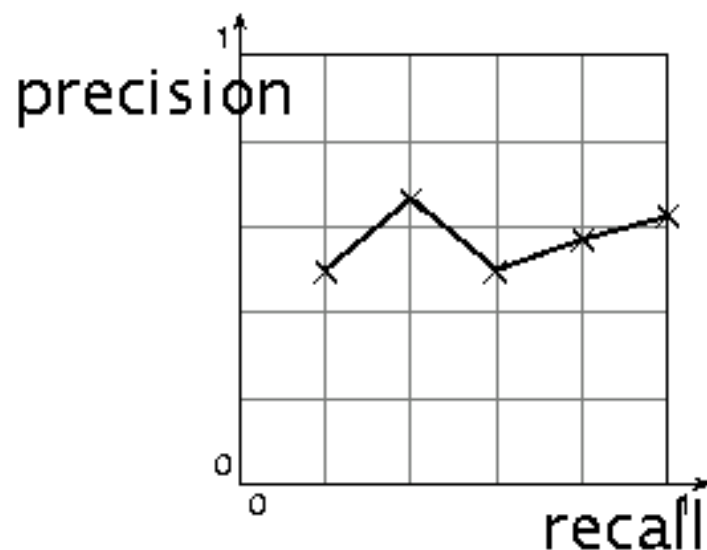


Πριονωτή – το precision ελαττώνεται για το ίδιο recall μέχρι να βρεθεί το επόμενο συναφές έγγραφο

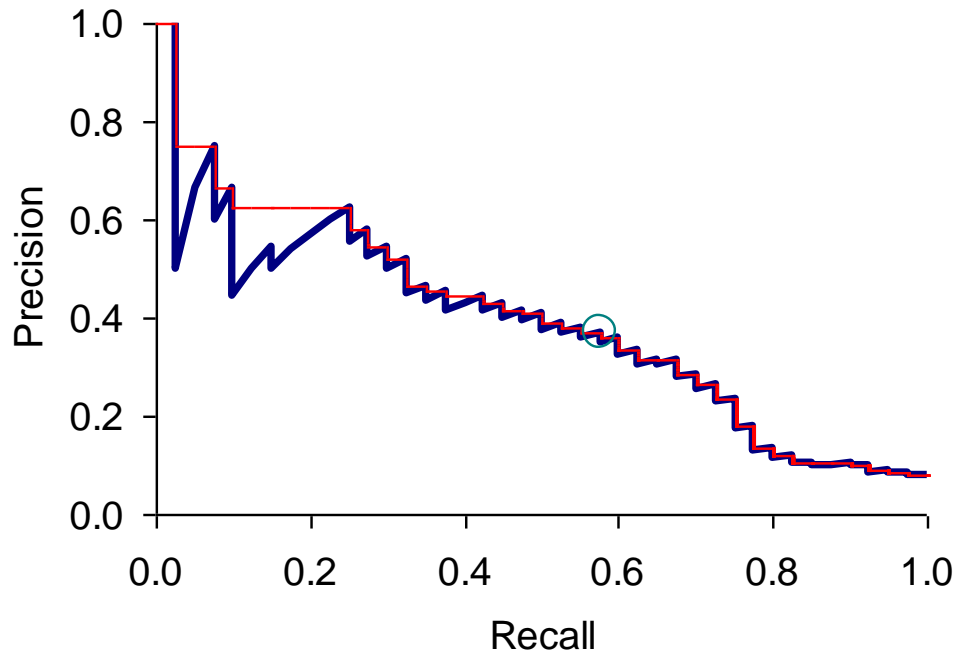
# Ακρίβεια εκ παρεμβολής (Interpolated precision)

- 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*
- Παίρνουμε τη μέγιστη τιμή της ακρίβειας στα δεξιά της τιμής

$$P(r_j) = \max_{r_j \leq r \leq r_{j+1}} P(r)$$



# Καμπύλη Ακρίβειας/Ανάκλησης



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

Παρεμβολή (με κόκκινο): μέγιστο των μελλοντικών σημείων



# Παράδειγμα II

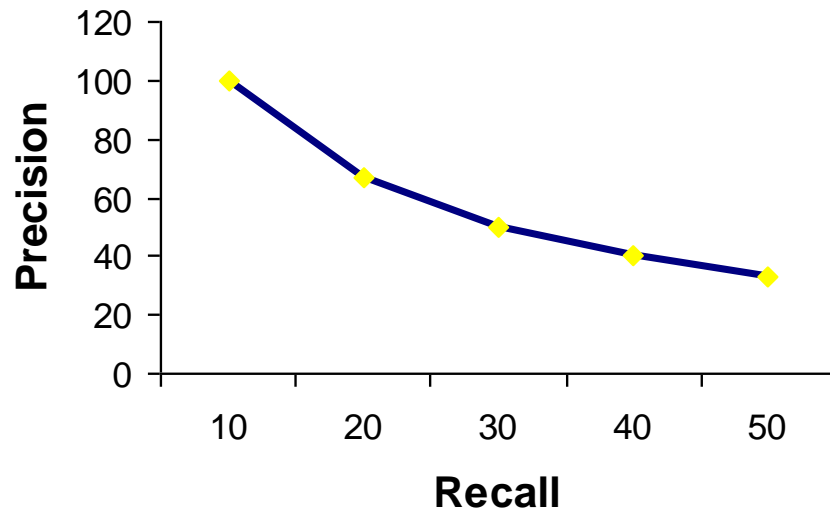
---

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

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

# Παράδειγμα II

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



Rank	Doc	Rel	R <sub>recall</sub>	P <sub>recision</sub>
0			0 %	0 %
1	$d_{123}$	✓	10 %	100 %
2	$d_{84}$		10 %	50 %
3	$d_{56}$	✓	20 %	67 %
4	$d_6$		20 %	50 %
5	$d_{84}$		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 %

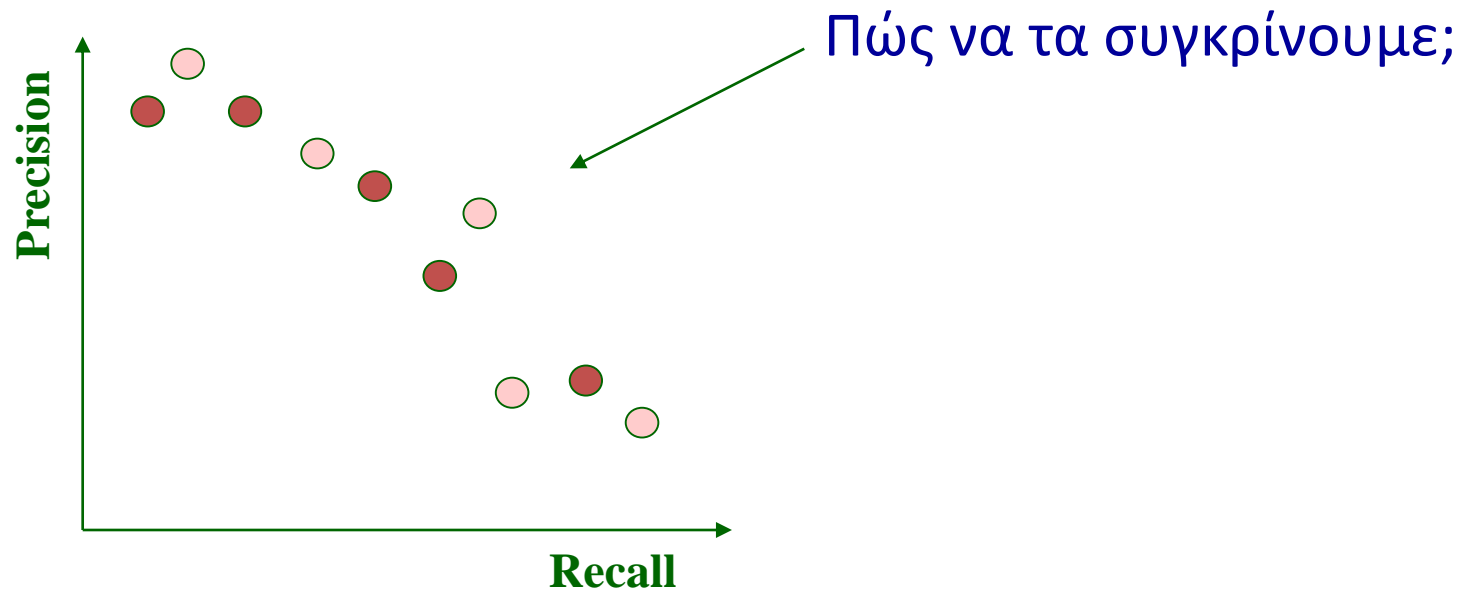
# Μέσοι όροι από πολλά ερωτήματα

---

- 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



# Σύγκριση Συστημάτων

---

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

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

Παράδειγμα καθιερωμένων επιπέδων ανάκλησης (πλήθος επιπέδων: 11):

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

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

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

# 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, . . .
- Επανάλαβε 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)*.

# 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, .
- Επανάλαβε το για όλα τα ερωτήματα στο evaluation benchmark και πάρε το μέσο όρο

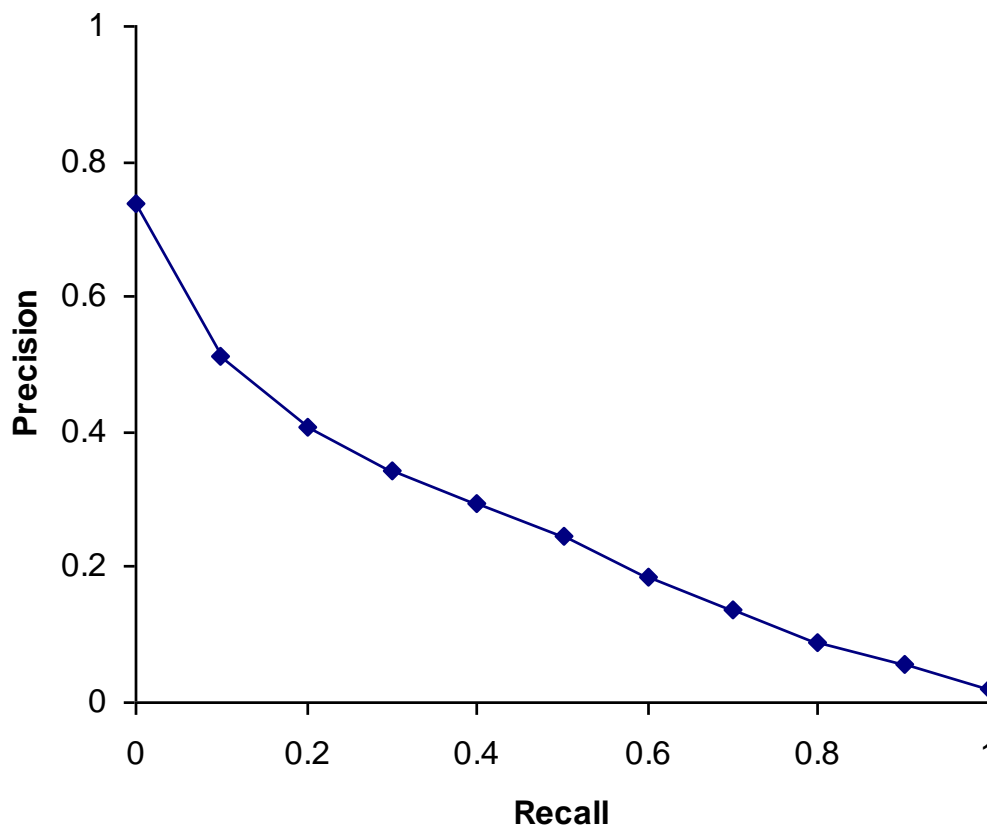
$$\bar{P}(r) = \sum_{i=1}^{N_q} \frac{P_i(r)}{N_q}$$

$N_q$  – πλήθος ερωτημάτων

$P_i(r)$  - precision at recall level  $r$  for  $i^{\text{th}}$  query

# Τυπική (καλή;) ακρίβεια 11-σημείων

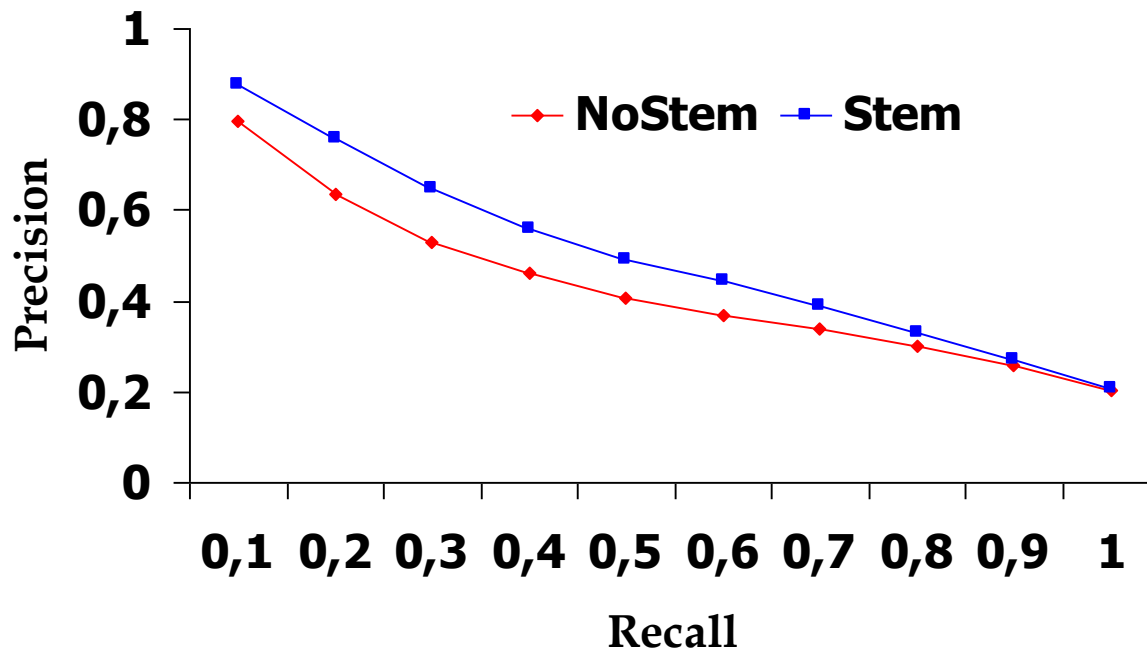
- SabIR/Cornell 8A1 11pt precision from TREC 8 (1999)





# 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*

# Άσκηση 8.5 + 8.6

---

## Άσκηση 8.5

Υπάρχει ή όχι πάντα ένα σημείο στο οποίο *η ακρίβεια είναι ίση με την ανάκληση* (break-even point). Αν ναι, αποδείξτε το, αν όχι, δώστε αντιπαράδειγμα

## Άσκηση 8.6

Ποια είναι η σχέση της τιμής του F1 και του break-even point;

# Άσκηση 8.7

---

## DICE COEFFICIENT

Ο συντελεστής Dice δυο συνόλων είναι μια μέτρηση της τομής του σε σχέση με το μέγεθος τους

$$\text{Dice}(X, Y) = 2|X \cap Y| / (|X| + |Y|)$$

- Δείξτε ότι το F ισούται με το συντελεστή Dice του συνόλου των ανακτημένων εγγράφων και του συνόλου των συναφών εγγράφων

## Άσκηση 8.8

---

Έστω μια ανάγκη πληροφόρησης για την οποία υπάρχουν 4 συναφή έγγραφα.

- ΣΥΣΤΗΜΑ 1: **R N R N N N N N R R**
- ΣΥΣΤΗΜΑ 2: **N R N N R R R 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;
- Υπολογίστε την R-ακρίβεια. Ποιο σύστημα είναι καλύτερο;

# 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 10<sup>ο</sup>

Πχ

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



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

# MAP

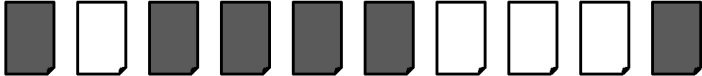

- Θεωρείστε τη θέση διάταξης (rank position) κάθε **συναφούς εγγράφου**
  - $K_1, K_2, \dots, K_R$
- Υπολογισμός του Precision@K για κάθε  $K_1, K_2, \dots, K_R$
- Μέση ακρίβεια = average of P@K

- 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

 = 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

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











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




# MAP











 = relevant documents for query 1

Ranking #1

										
Recall	0.2	0.2	0.4	0.4	0.4	0.6	0.6	0.6	0.8	1.0
Precision	1.0	0.5	0.67	0.5	0.4	0.5	0.43	0.38	0.44	0.5

 = relevant documents for query 2

Ranking #2

										
Recall	0.0	0.33	0.33	0.33	0.67	0.67	1.0	1.0	1.0	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.33	0.43	0.38	0.33	0.3

*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 (περίληψη)

- 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

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

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

Συχνά οι τιμές της ΜΑΠ για το ίδιο ερώτημα σε διαφορετικά συστήματα διαφέρουν λιγότερο από τις τιμές τις ΜΑΠ για διαφορετικά ερωτήματα στο ίδιο σύστημα

# R-ακρίβεια

---

## 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
- Το τέλειο σύστημα μπορεί να πετύχει βαθμό 1.0

Αν υπάρχουν  $r$ , τότε  $r/Rel$

# 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	

$R = \# \text{ of relevant docs} = 6$

$R\text{-Precision} = 4/6 = 0.67$

# Αν τα αποτελέσματα δεν είναι σε λίστα;

---

- 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  $K$  του πρώτου σχετικού εγγράφου*
  - Μπορεί να είναι το μόνο που έκανε click ο χρήστης
- Reciprocal Rank score =  $\frac{1}{K}$
- MRR το μέσο RR για πολλαπλές ερωτήσεις

# ROC (Receiver Operating Characteristic Curve)

## Καμπύλη χαρακτηριστικής λειτουργίας δέκτη

- Αναπτύχθηκε στη δεκαετία 1950 για την ανάλυση θορύβου στα σήματα
  - Χαρακτηρίζει το trade-off μεταξύ positive hits και false alarms
- Η καμπύλη ROC δείχνει τα
  - TPR [TruePositiveRate] ή ευαισθησία (sensitivity) (άλλο όνομα του recall) (στον άξονα των y) προς τα
  - FPR [FalsePositiveRate] ή fall out ή 1 - specificity (στον άξονα των x)
- Η απόδοση αναπαρίσταται ως ένα σημείο στην καμπύλη ROC

True Positive Rate  
Πόσα από τα θετικά (συναφή) βρίσκει  
[πόσα από τα θετικά ταξινομεί σωστά]

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

False Positive Rate

Πόσα από τα αρνητικά θεωρεί θετικά

[πόσα από τα αρνητικά κατηγοριοποιεί λάθος]

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

# ROC

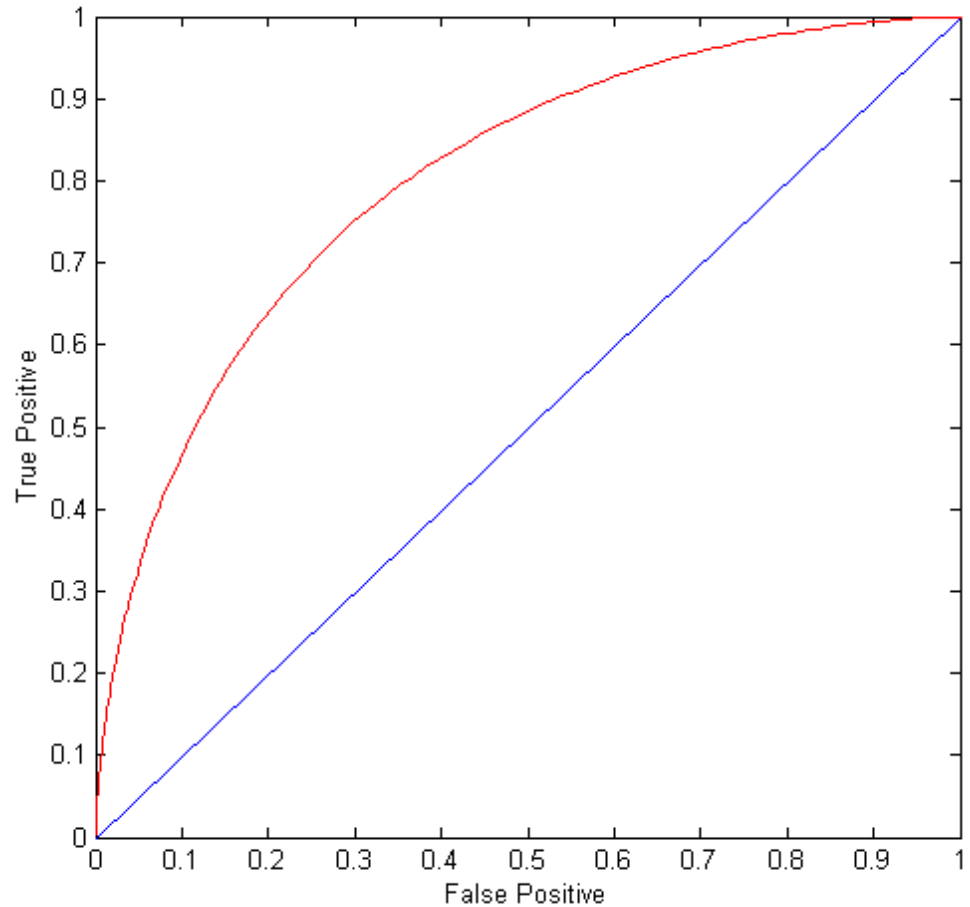
(0,0): το μοντέλο προβλέπει τα πάντα ως αρνητική κατηγορία

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

(0,1): ιδανικό

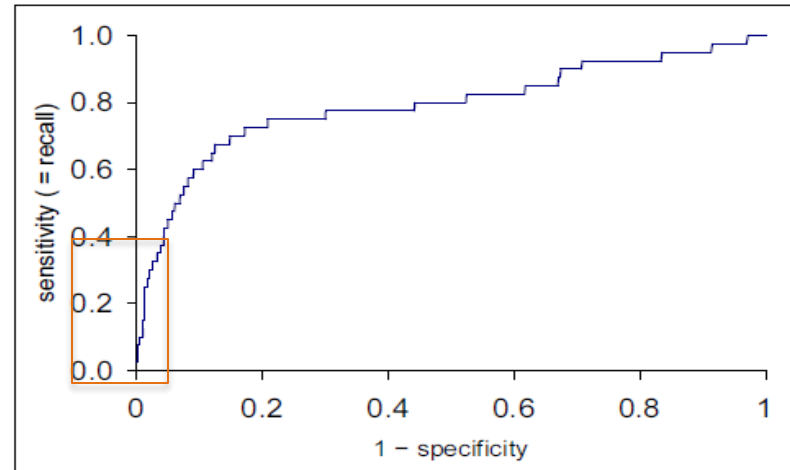
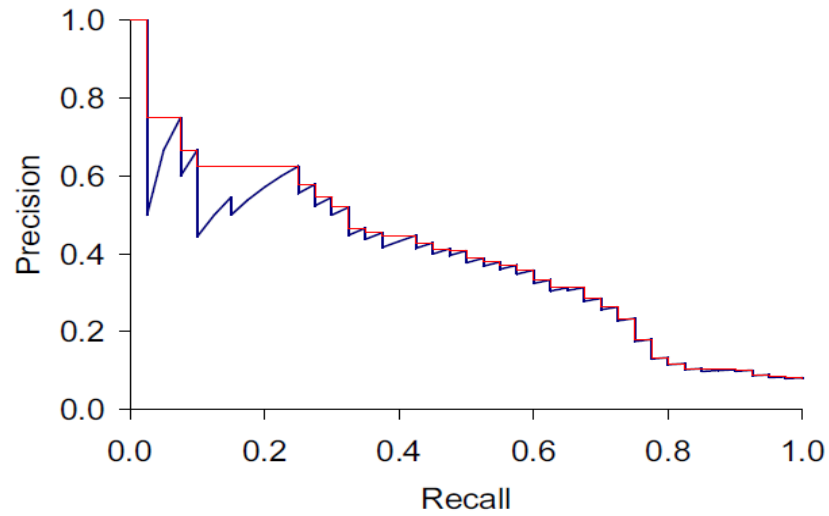
Το ιδανικό στην άνω αριστερή γωνία

Διαγώνια γραμμή: Random guessing





# ROC



# Μη δυαδικές αποτιμήσεις

---

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



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...](#)

### 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](#)

### 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 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

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](#)

### Toyota Safety

[Toyota safety](#) Discount Prices Save Money Shopping Online Today.  
[www.smarter.com](#)

### Safety Toyoto

Explore 5,000+ Pro Sports Choices. Save On Safety Toyota.  
[BaseballGear.Shopzilla.com](#)

[See your message here...](#)

fair

fair

Good

# Discounted Cumulative Gain

---

- Δημοφιλές μέτρο για αποτίμηση της αναζήτησης στο web και σε παρόμοιες εφαρμογές
- Δύο υποθέσεις:
  - (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

---

- 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 και
  - μειώνεται ή γίνεται έκπτωση (discounted) στα χαμηλότερα επίπεδα
- Η σχετική μείωση είναι  $1/\log(\text{rank})$ 
  - Για βάση 2, η μείωση του κέρδους στο επίπεδο 4 είναι  $1/2$  και στο επίπεδο 8 είναι  $1/3$

# Discounted Cumulative Gain

---

- Έστω αξιολογήσεις συνάφειας στη κλίμακα  $[0, r]$ ,  $r > 2$  και ότι οι αξιολογήσεις των  $n$  πρώτων εγγράφων είναι  $r_1, r_2, \dots, r_n$  (σε σειρά διάταξης)
- **Cumulative Gain (CG)** στη θέση διάταξης (rank)  $n$ 
  - $CG = r_1 + r_2 + \dots + r_n$
- **Discounted Cumulative Gain (DCG)** στη θέση διάταξης  $n$ 
  - $DCG = r_1 + r_2 / \log_2 2 + r_3 / \log_2 3 + \dots + r_n / \log_2 n$ 
    - Χρησιμοποιούνται και άλλες βάσεις εκτός του 2 για το λογάριθμο

# 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

# Παράδειγμα

---

- 10 διατεταγμένα έγγραφα σε κλίμακα συνάφειας 0-3:  
3, 2, 3, 0, 0, 1, 2, 2, 3, 0
- 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



# Κανονικοποίηση του DCG

---

- **Normalized Discounted Cumulative Gain (NDCG)** στη θέση διάταξης  $n$ 
  - Κανονικοποιούμε το DCG στη θέση διάταξης  $n$  με την DCG 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, κ.ο.κ
- Χρήσιμο για αντιπαράθεση ερωτημάτων με διαφορετικό αριθμό συναφών αποτελεσμάτων
- Ιδιαίτερα δημοφιλές μέτρο στην αναζήτηση στο web

# Παράδειγμα

4 έγγραφα:  $d_1, d_2, d_3, d_4$

i	Ground Truth (optimal)		Ranking Function <sub>1</sub>		Ranking Function <sub>2</sub>	
	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 <sub>GT</sub> =1.00		NDCG <sub>RF1</sub> =1.00		NDCG <sub>RF2</sub> =0.9203	

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

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

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

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

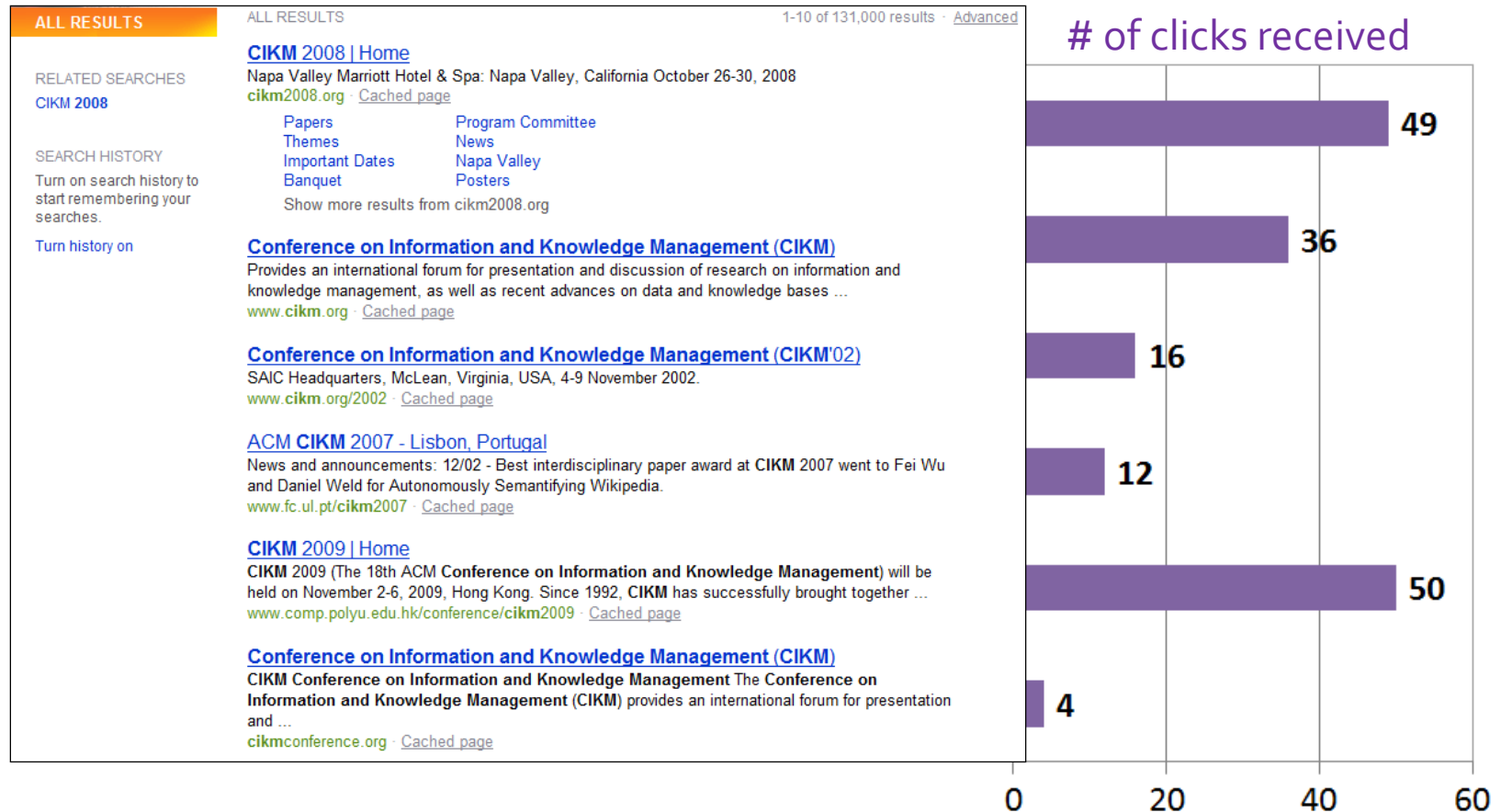
# Αξιολογήσεις από ανθρώπους

---

- 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
- Εναλλακτικές;

Με χρήση clickthrough

# Τι μας λένε οι αριθμοί;



Έχει μεγάλη σημασία η θέση, απόλυτοι αριθμοί όχι ιδιαίτερα αξιόπιστοι

# Σχετική και απόλυτη διάταξη

ALL RESULTS 1-10 of 131,000 results · [Advanced](#)

**ALL RESULTS**

RELATED SEARCHES  
[CIKM 2008](#)

SEARCH HISTORY  
Turn on search history to start remembering your searches.  
[Turn history on](#)

[CIKM 2008 | Home](#)  
Napa Valley Marriott Hotel & Spa: Napa Valley, California October 26-30, 2008  
[cikm2008.org](#) · [Cached page](#)

<a href="#">Papers</a>	<a href="#">Program Committee</a>
<a href="#">Themes</a>	<a href="#">News</a>
<a href="#">Important Dates</a>	<a href="#">Napa Valley</a>
<a href="#">Banquet</a>	<a href="#">Posters</a>

[Show more results from cikm2008.org](#)

[Conference on Information and Knowledge Management \(CIKM\)](#)  
Provides an international forum for presentation and discussion of research on information and knowledge management, as well as recent advances on data and knowledge bases ...  
[www.cikm.org](#) · [Cached page](#)

[Conference on Information and Knowledge Management \(CIKM'02\)](#)  
SAIC Headquarters, McLean, Virginia, USA, 4-9 November 2002.  
[www.cikm.org/2002](#) · [Cached page](#)

[ACM CIKM 2007 - Lisbon, Portugal](#)  
News and announcements: 12/02 - Best interdisciplinary paper award at CIKM 2007 went to Fei Wu and Daniel Weld for Autonomously Semantifying Wikipedia.  
[www.fc.ul.pt/cikm2007](#) · [Cached page](#)

[CIKM 2009 | Home](#)  
CIKM 2009 (The 18th ACM Conference on Information and Knowledge Management) will be held on November 2-6, 2009, Hong Kong. Since 1992, CIKM has successfully brought together ...  
[www.comp.polyu.edu.hk/conference/cikm2009](#) · [Cached page](#)

[Conference on Information and Knowledge Management \(CIKM\)](#)  
CIKM Conference on Information and Knowledge Management The Conference on Information and Knowledge Management (CIKM) provides an international forum for presentation and ...  
[cikmconference.org](#) · [Cached page](#)

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
- 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;

# 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 παρά η διασπορά στην απόδοση διαφορετικών συστημάτων στην ίδια ερώτηση
- Δηλαδή, υπάρχουν εύκολες ανάγκες πληροφόρησης και δύσκολες ανάγκες πληροφόρησης!

# 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

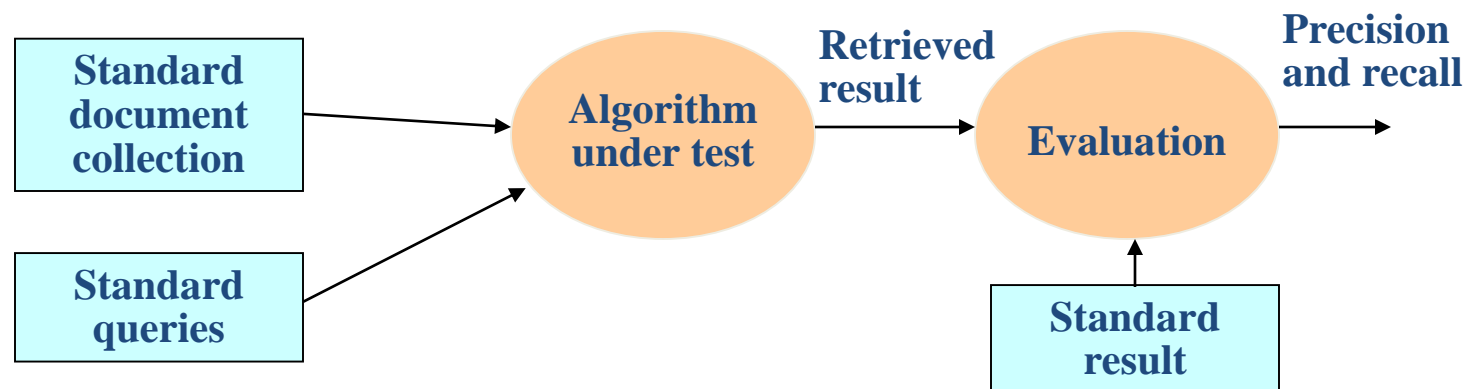
- (ή, καταχρηστικά ερωτημάτων)
- Να σχετίζονται με τα διαθέσιμα έγγραφα
- Οι ανάγκες πληροφόρησης πρέπει να είναι αντιπροσωπευτικές των πραγματικών - τυχαίοι όροι δεν είναι καλή ιδέα
- Συχνά από ειδικούς της περιοχής

## 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 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

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

Στα τέλη του 1950, UK

- 1398 abstracts 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 *rooted* 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>



# 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

<i>Collection</i>	<i>NDocs</i>	<i>NQrys</i>	<i>Size (MB)</i>	<i>Term/Doc</i>	<i>Q-D RelAss</i>
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

# 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) (dis-agreement) μεταξύ των κριτών

---

## Μέτρο Kappa

- Συμφωνία μεταξύ των κριτών
- Αφορά κατηγορική κρίση
- 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

$P(E)$ : τι συμφωνία θα είχαμε από τύχη

*$\kappa = 1$  Για πλήρη συμφωνία,  $0$  για τυχαία συμφωνία, αρνητική για μικρότερη της τυχαίας*

# Καρρα: παράδειγμα

Number of docs	ΚΡΙΤΗΣ 1	ΚΡΙΤΗΣ 2
300	Relevant	Relevant
70	Nonrelevant	Nonrelevant
20	Relevant	Nonrelevant
10	Nonrelevant	Relevant

$$P(A) = 370/400 = 0.925$$

$$P(\text{nonrelevant}) = (70+10+70+20)/800 = 0.2125$$

$$P(\text{relevant}) = (300+20+300+10)/800 = 0.7878$$

$$P(E) = 0.2125^2 + 0.7878^2 = 0.665$$

$$\text{Kappa} = (0.925 - 0.665)/(1 - 0.665) = 0.776$$

# Καρρα

---

- $\text{Καρρα} > 0.8$  = καλή συμφωνία
- $0.67 < \text{Καρρα} < 0.8$  -> “tentative conclusions”
- Εξαρτάται από το στόχο της μελέτης
  - Για  $>2$  κριτές: μέσοι όροι ανά-δύο κλπ

# Καρρα: 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



# Crowdsourcing

---

- To Mechanical Truck της Amazon

# Αξιολόγηση σε μεγάλες μηχανές αναζήτησης

---

- Οι μηχανές αναζήτησης διαθέτουν συλλογές ελέγχου ερωτημάτων και αποτελέσματα καταταγμένα 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

# 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 αξιόπιστα συναθροιστικά ή με χρήση ζευγών
- 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

# Κριτική της Συνάφειας

---

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

# Άσκηση 8.5 + 8.6

---

## Άσκηση 8.5

Υπάρχει ή όχι πάντα ένα σημείο στο οποίο *η ακρίβεια είναι ίση με την ανάκληση* (break-even point). Αν ναι, αποδείξτε το, αν όχι, δώστε αντιπαράδειγμα

## Άσκηση 8.6

Ποια είναι η σχέση της τιμής του F1 και του break-even point;

# Άσκηση 8.7

---

## DICE COEFFICIENT

Ο συντελεστής Dice δυο συνόλων είναι μια μέτρηση της τομής του σε σχέση με το μέγεθος τους

$$\text{Dice}(X, Y) = 2|X \cap Y| / (|X| + |Y|)$$

- Δείξτε ότι το F ισούται με το συντελεστή Dice του συνόλου των ανακτημένων εγγράφων και του συνόλου των συναφών εγγράφων

## Άσκηση 8.8

---

Έστω μια ανάγκη πληροφόρησης για την οποία υπάρχουν 4 συναφή έγγραφα.

- ΣΥΣΤΗΜΑ 1: **R N R N N N N N R R**
- ΣΥΣΤΗΜΑ 2: **N R N N R R R 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;
- Υπολογίστε την R-ακρίβεια. Ποιο σύστημα είναι καλύτερο;



## Άσκηση 8.8 (επέκταση)

---

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

- ΣΥΣΤΗΜΑ 1: **R** ...
- ΣΥΣΤΗΜΑ 2: N **R R** ...
- Για να είναι το Σύστημα 2 καλύτερο (ως αναφορά το MAP) τότε 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 25% ανάκληση;
- Ποια είναι η ακρίβεια με παρεμβολή για 33% ανάκληση;
- Υπολογίστε το ΜΑΠ

## Άσκηση 8.9

---

Επιστρέφει **όλα τα 10.000** έγγραφα και αυτά (όπως πριν) είναι τα πρώτα 20 αποτελέσματα:

**R R** N 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
- Υποθέστε ότι σε κάποιο πείραμα αξιολογούμε μόνο τα 20 πρώτα αποτελέσματα. Φράξτε το λάθος στον υπολογισμό.

# ΤΕΛΟΣ 8<sup>ου</sup> Μαθήματος

## 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 Τζιτζικα (Παν. Κρήτης)*