Introduction to

# Information Retrieval

ΠΛΕ70: Ανάκτηση Πληροφορίας

*Διδάσκουσα: Ευαγγελία Πιτουρά*

Διάλεξη 9: Αξιολόγηση στην Ανάκτηση Πληροφοριών.

# Τι είδαμε στο προηγούμενο μάθημα

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

# Μοντέλα διαβαθμισμένης ανάκτησης

- Διαβαθμισμένη ανάκτηση (ranked retrieval): αντί ενός **συνόλου** εγγράφων που ικανοποιούν το ερώτημα, μια **διάταξη** των (κορυφαίων) για την ερώτηση εγγράφων της συλλογής
- Διάταξη με βάση ένα βαθμό, **score(*d*, *q*)**, μετρά πόσο καλά το έγγραφο *d* "ταιριάζει" (match) με το ερώτημα *q*
- Επιστρέφουμε τα κορυφαία-*k* (**top-*k***)
- *Συνήθως μαζί με ερωτήματα **ελεύθερου κειμένου** (Free text queries)*

# Διαβαθμισμένη ανάκτηση: θέματα

1. Πως ορίζουμε το score;

2. Πως υπολογίζουμε αποδοτικά τα *k* καλύτερα έγγραφα όταν η συλλογή είναι μεγάλη;

*Θα δούμε αρχικά μεθόδους που βασίζονται στο «κείμενο» για το ερώτημα 1*

# Μοντέλα διαβαθμισμένης ανάκτησης

## Πως ορίζουμε το βαθμό;

Έστω ένα ερώτημα και $t_i$ οι όροι του

## Δυο βασικά κριτήρια βασισμένα στο «κείμενο»

- Συχνότητα εμφάνισης του όρου $t_i$ στο κείμενο (tf)
- Συχνότητα εμφάνισης του όρου $t_i$ στη συλλογή (idf)

# tf και στάθμιση με log tf

- Η συχνότητα $tf_{t,d}$ του όρου $t$ σε ένα έγγραφο $d$ ορίζεται ως αριθμός των φορών που το $t$ εμφανίζεται στο $d$.
- Συχνά στάθμιση με χρήση του λογάριθμου της συχνότητας (log frequency weight) του όρου $t$ στο $d$ είναι

$$w_{t,d} = \begin{cases} 1 + \log_{10} \mathrm{tf}_{t,d}, & \text{if } \mathrm{tf}_{t,d} > 0 \\ 0, & \text{otherwise} \end{cases}$$

  - 0 → 0, 1 → 1, 2 → 1.3, 10 → 2, 100->3, 1000 → 4, κλπ.

- Ο βαθμός για ένα ζεύγος εγγράφου-ερωτήματος: άθροισμα των βαρών όλων των κοινών όρων:

$$score = \sum_{t \in q \cap d} (1 + \log \mathrm{tf}_{t,d})$$

# idf και στάθμιση

- $\mathrm{df}_t$ είναι η συχνότητα εγγράφων του $t$: ο αριθμός (πλήθος) των εγγράφων της συλλογής που περιέχουν το $t$ ($\mathrm{df}_t \leq N$)
  - $\mathrm{df}_t$ είναι η αντίστροφη μέτρηση της πληροφορίας που παρέχει ο όρος $t$
- Ορίζουμε την αντίστροφη συχνότητα εγγράφων idf (inverse document frequency) του $t$ ως

$$\mathrm{idf}_t = \log_{10}(N/\mathrm{df}_t)$$

  - Χρησιμοποιούμε $\log(N/\mathrm{df}_t)$ αντί για $N/\mathrm{df}_t$ για να «ομαλοποιήσουμε» την επίδραση του idf.

# Στάθμιση με tf-idf

Το tf-idf βάρος ενός όρου είναι το γινόμενο του βάρους tf και του βάρους idf.

$$w_{t,d} = (1 + \log(\mathrm{tf}_{t,d})) \times \log(N / \mathrm{df}_t)$$

- Αυξάνει με τον αριθμό εμφανίσεων του όρου στο έγγραφο
- Αυξάνει με τη σπανιότητα του όρου στη συλλογή

$$\mathrm{Score}(q,d) = \sum_{t \in q \cap d} \mathrm{tf.idf}_{t,d}$$

❖ Έχουν προταθεί και πολλοί άλλοι τρόποι συνδυασμού του tf και idf

# Διανυσματική αναπαράσταση

| | Antony and Cleopatra | Julius Caesar | The Tempest | Hamlet | Othello | Macbeth |
|---|---|---|---|---|---|---|
| Antony | 1 | 1 | 0 | 0 | 0 | 1 |
| Brutus | 1 | 1 | 0 | 1 | 0 | 0 |
| Caesar | 1 | 1 | 0 | 1 | 1 | 1 |
| Calpurnia | 0 | 1 | 0 | 0 | 0 | 0 |
| Cleopatra | 1 | 0 | 0 | 0 | 0 | 0 |
| mercy | 1 | 0 | 1 | 1 | 1 | 1 |
| worser | 1 | 0 | 1 | 1 | 1 | 0 |

Κάθε έγγραφο αναπαρίσταται ως ένα *δυαδικό διάνυσμα* $\in \{0,1\}^{|V|}$ (την αντίστοιχη στήλη)

# Ο πίνακας με μετρητές

| | Antony and Cleopatra | Julius Caesar | The Tempest | Hamlet | Othello | Macbeth |
|---|---|---|---|---|---|---|
| Antony | 157 | 73 | 0 | 0 | 0 | 0 |
| Brutus | 4 | 157 | 0 | 1 | 0 | 0 |
| Caesar | 232 | 227 | 0 | 2 | 1 | 1 |
| Calpurnia | 0 | 10 | 0 | 0 | 0 | 0 |
| Cleopatra | 57 | 0 | 0 | 0 | 0 | 0 |
| mercy | 2 | 0 | 3 | 5 | 5 | 1 |
| worser | 2 | 0 | 1 | 1 | 1 | 0 |

Κάθε έγγραφο είναι ένα διάνυσμα μετρητών (συχνότητα εμφάνισης του όρου στο έγγραφο) στο $\mathbb{N}^{|v|}$: μια στήλη παρακάτω

# Ο πίνακας με βάρη

| | Antony and Cleopatra | Julius Caesar | The Tempest | Hamlet | Othello | Macbeth |
|---|---|---|---|---|---|---|
| Antony | 5.25 | 3.18 | 0 | 0 | 0 | 0.35 |
| Brutus | 1.21 | 6.1 | 0 | 1 | 0 | 0 |
| Caesar | 8.59 | 2.54 | 0 | 1.51 | 0.25 | 0 |
| Calpurnia | 0 | 1.54 | 0 | 0 | 0 | 0 |
| Cleopatra | 2.85 | 0 | 0 | 0 | 0 | 0 |
| mercy | 1.51 | 0 | 1.9 | 0.12 | 5.25 | 0.88 |
| worser | 1.37 | 0 | 0.11 | 4.15 | 0.25 | 1.95 |

Θεωρούμε το tf-idf βάρος του όρου:

- Κάθε έγγραφο είναι ένα *διάνυσμα tf-idf βαρών* στο $R^{|v|}$

# Τα έγγραφα ως διανύσματα

Έχουμε ένα |V|-διάστατο διανυσματικό χώρο

- Οι όροι είναι οι άξονες αυτού του χώρου
- Τα **έγγραφα** και οι **ερωτήσεις** είναι σημεία ή διανύσματα σε αυτόν τον χώρο

- Πολύ μεγάλη διάσταση: δεκάδες εκατομμύρια διαστάσεις στην περίπτωση της αναζήτησης στο web
- Πολύ αραιά διανύσματα – οι περισσότεροι όροι είναι 0

- Το score($q$, $d$) ως το συνημίτονο της γωνίας των q και d

# Βαθμολόγηση στο διανυσματικό χώρο

1. Αναπαράσταση του ερωτήματος ως ένα διαβαθμισμένο tf-idf διάνυσμα
2. Αναπαράσταση κάθε εγγράφου ως ένα διαβαθμισμένο tf-idf διάνυσμα
3. Υπολόγισε το συνημίτονο για κάθε ζεύγος ερωτήματος, εγγράφου
4. Διάταξε τα έγγραφα με βάση αυτό το βαθμό
5. Επέστρεψε τα κορυφαία K (π.χ., K =10) έγγραφα στο χρήστη

# Παραλλαγές της tf-idf στάθμισης

| Term frequency | | Document frequency | | Normalization | |
|---|---|---|---|---|---|
| n (natural) | $\mathrm{tf}_{t,d}$ | n (no) | 1 | n (none) | 1 |
| l (logarithm) | $1 + \log(\mathrm{tf}_{t,d})$ | t (idf) | $\log \frac{N}{\mathrm{df}_t}$ | c (cosine) | $\frac{1}{\sqrt{w_1^2+w_2^2+\ldots+w_M^2}}$ |
| a (augmented) | $0.5 + \frac{0.5 \times \mathrm{tf}_{t,d}}{\max_t(\mathrm{tf}_{t,d})}$ | p (prob idf) | $\max\{0, \log \frac{N-\mathrm{df}_t}{\mathrm{df}_t}\}$ | u (pivoted unique) | $1/u$ |
| b (boolean) | $\begin{cases} 1 & \text{if } \mathrm{tf}_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$ | | | b (byte size) | $1/CharLength^{\alpha}$, $\alpha < 1$ |
| L (log ave) | $\frac{1+\log(\mathrm{tf}_{t,d})}{1+\log(\mathrm{ave}_{t \in d}(\mathrm{tf}_{t,d}))}$ | | | | |

Augmented: θεωρούμε τη συχνότητα του πιο συχνού όρου στο έγγραφο και κανονικοποιούμε με αυτήν

Το 0.5 είναι ένας τελεστές στάθμισης (εξομάλυνσης) – smoothing factor (συχνά και 0.4 αντί 0.5) για να αντιμετωπίσουμε το γεγονός ότι μεγάλα έγγραφα μπορεί να έχουν μεγάλα tf τιμές γιατί επαναλαμβάνουν πληροφορία

# Διαβαθμισμένη ανάκτηση

Αρκεί το tf.id;

Εξαρτάται από τη συλλογή και την εφαρμογή, συνήθως *συνδυασμός πολλών σταθμισμένων όρων*

# Παράδειγμα: διαβαθμισμένης ανάκτηση στο google

SEO: search engine optimization

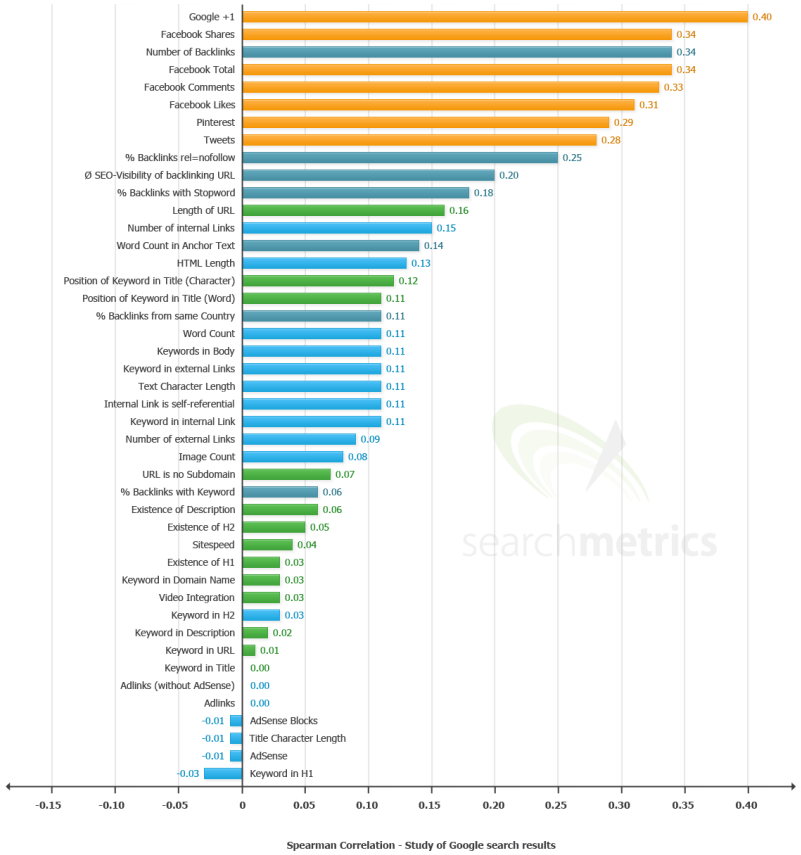
Google 200 ranking factors (*) – δείτε τους στο τέλος των διαφανιών

(*) http://backlinko.com/google-ranking-factors

*Ή δείτε και αυτό*
*http://moz.com/search-ranking-factors*

SERP: search engine result page

# Διαβαθμισμένης ανάκτηση google




social

on page (content)

on page (technical)

backlinks

http://www.searchmetrics.com/en/knowledge-base/ranking-factors-us-2013/

# An example: Facebook News Feed

Edge rank used to have three primary factors:

- ✓ Affinity — i.e., how close is the relationship between the user and the content/source?
- ✓ Weight — i.e., what type of action was taken on the content?
- ✓ Decay — i.e., how recent/current is the content?

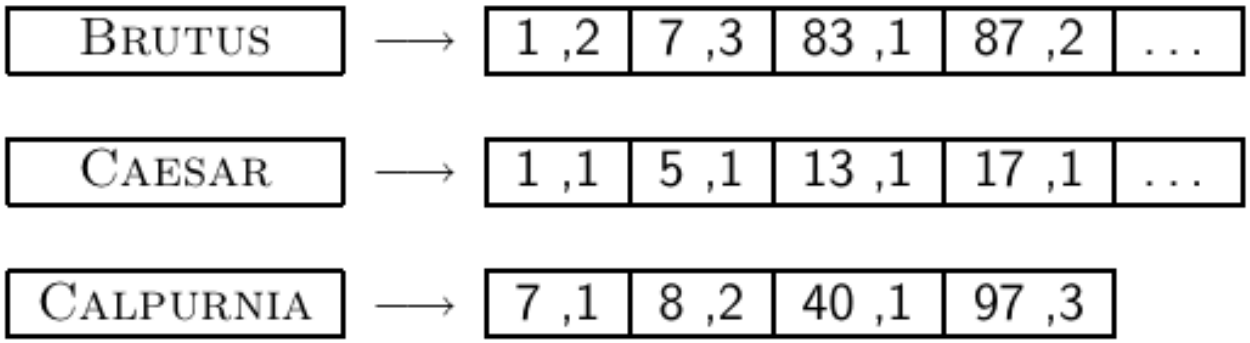
New algorithm uses close to **100K** weight factors

Source: http://marketingland.com/, Aug 16, 2013

# Τι είδαμε στο προηγούμενο μάθημα

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

# Επέκταση καταχωρήσεων




- Συχνότητες όρων

Σε κάθε καταχώρηση, αποθήκευση του $tf_{t,d}$ επιπρόσθετα του docID

- *Η συχνότητα $idf_t$ αποθηκεύεται στο λεξικό μαζί με τον όρο t*

# Υπολογισμός βαθμού

Υπολογισμός ανά-όρο (ένας-όρος-τη-φορά - a-term-at-a-time)

- Η απλούστερη περίπτωση είναι να επεξεργαστούμε όλη τη λίστα καταχωρήσεων για τον πρώτο όρο του ερωτήματος
- Δημιουργούμε ένα συσσωρευτή των βαθμών για κάθε docID εγγράφου που βρίσκουμε
- Μετά επεξεργαζόμαστε πλήρως τη λίστα καταχωρήσεων για τον δεύτερο όρο κοκ

# Υπολογισμός βαθμού

- Υπολογισμός ανά έγγραφο (ένα έγγραφο τη φορά document-at-a-time) Μπορούμε να διατρέχουμε τις λίστες των όρων του ερωτήματος παράλληλα όπως στην περίπτωση της Boolean ανάκτησης (merge sort)
  - Αυτό έχει ως αποτέλεσμα λόγω της διάταξης των εγγράφων στις λίστες καταχωρίσεων τον υπολογισμό του βαθμού ανά έγγραφο

# Υπολογισμός *k*-κορυφαίων αποτελεσμάτων

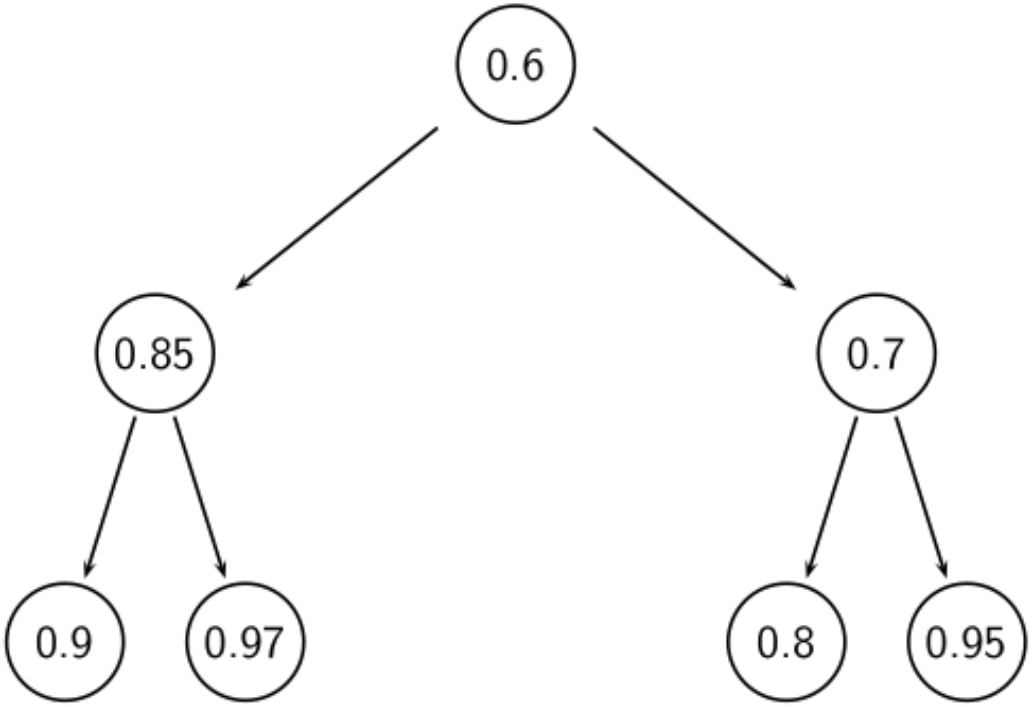
Σε πολλές εφαρμογές, δε χρειαζόμαστε την πλήρη κατάταξη, αλλά μόνο τα κορυφαία *k (top-k)*, για κάποιο μικρό *k*, π.χ., *k* = 100

- Απλοϊκός τρόπος:
  - Υπολόγισε τους βαθμούς για όλα τα N έγραφα
  - Sort
  - Επέστεψε τα κορυφαία *k*

Αν δε χρειαζόμαστε όλη τη διάταξη, υπάρχει πιο αποδοτικός τρόπος να υπολογίσουμε *μόνο* τα κορυφαία *k*;

# Χρήση min-heap

- Χρήση δυαδικού min heap
- Ένα δυαδικό min heap είναι ένα δυαδικό δέντρο που η τιμή ενός κόμβου είναι μικρότερη από την τιμή των δύο παιδιών του.



# Επιλογή των κορυφαίων $k$ σε $O(N \log k)$

Στόχος: Διατηρούμε τα καλύτερα $k$ που έχουμε δει μέχρι στιγμής

- Χρήση δυαδικού min heap
- Για την επεξεργασία ενός νέου εγγράφου $d'$ με score $s'$:
  - Get *current minimum* $h_m$ of heap ($O(1)$)
  - If $s' < h_m$ skip to next document /* υπάρχουν $k$ καλύτερα */
  - If $s' > h_m$ heap-delete-root ($O(\log k)$) /* καλύτερο, σβήσε τη ρίζα
    heap-add $d'/s'$ ($O(\log k)$) και βάλτο στο heap */

# Γενική προσέγγιση «ψαλιδίσματος»

- Βρες ένα σύνολο *A* από υποψήφια έγγραφα (*contenders*), όπου *K* < |*A*| << *N*
  - *Το A δεν περιέχει απαραίτητα όλα τα* top *K, αλλά περιέχει αρκετά καλά έγγραφα και πολλά από τα* top *K*
- Επέστρεψε τα top K έγγραφα του A

Το A είναι ένα ψαλίδισμα (pruning) των μη υποψηφίων

✓ *Έτσι και αλλιώς το συνημίτονο είναι μόνο μια «εκτίμηση» της συνάφειας*

*Θα δούμε σχετικούς ευριστικούς*

# Περιορισμός του ευρετηρίου

- Ο βασικός αλγόριθμος υπολογισμού του συνημίτονου θεωρεί έγγραφα που περιέχουν *τουλάχιστον έναν όρο του ερωτήματος*

- Μπορούμε να επεκτείνουμε αυτήν την ιδέα;
  - Εξετάζουμε μόνο τους όρους του ερωτήματος με *μεγάλο idf*
  - Εξετάζουμε μόνο έγγραφα που περιέχουν *πολλούς από τους όρους του ερωτήματος*

# Λίστες πρωταθλητών

- **Προ-υπολογισμός** για κάθε όρο *t* του λεξικού, των ***r*** εγγράφων με το μεγαλύτερο βάρος ανάμεσα στις καταχωρήσεις του *t* -> *λίστα πρωταθλητών* (champion list , fancy list or top docs for *t*)
  - Αν tf.idf, είναι αυτά με το καλύτερο tf
- *Κατά την ώρα του ερωτήματος*, πάρε ως Α την ένωση των λιστών πρωταθλητών για τους όρους του ερωτήματος, υπολόγισε μόνο τους βαθμούς για τα έγγραφα της Α και διάλεξε τα Κ ανάμεσα τους
- Το *r* πρέπει να επιλεγεί κατά τη διάρκεια της κατασκευής του ευρετηρίου
  - Έτσι, είναι πιθανόν ότι $r < K$

# idf-διατεταγμένοι όροι

Κατά την επεξεργασία των όρων του ερωτήματος

- Τους εξετάζουμε με φθίνουσα διάταξη ως προς idf
  - Όροι με μεγάλο idf πιθανών να συνεισφέρουν περισσότερο στο βαθμό
- Καθώς ενημερώνουμε τη συμμετοχή στο βαθμό κάθε όρου
  - Σταματάμε αν ο βαθμός των εγγράφων δεν μεταβάλλεται πολύ

# Επιπρόσθετοι ευριστικοί για βελτίωση του χρόνου

✓ *Μέχρι στιγμής, θεωρούμε διάταξη των λιστών καταχωρήσεων με βάση το DocID – θα δούμε εναλλακτικές διατάξεις*

- Μέθοδος 1: Με βάση την ποιότητα των εγγράφων
- Μέθοδος 2: Με βάση τη σχετικότητα του εγγράφου με τον όρο, δηλαδή το tf

# Με βάση την «ποιότητα» του εγγράφου (*g*(d))

- ✓ Συχνά υπάρχει ένας *ανεξάρτητος του ερωτήματος (στατικός)* χαρακτηρισμός της καταλληλότητας ("goodness", authority) του εγγράφου – έστω *g*(d)

Για παράδειγμα:

- Στις μηχανές αναζήτησης (στο Google) το PageRank *g*(*d*), wikipedia σελίδες ή άρθρα σε μια συγκεκριμένη εφημερίδα, κλπ

# Με βάση την «ποιότητα» του εγγράφου ($g$(d))

Σε αυτήν την περίπτωση, ο **συγκεντρωτικός βαθμός (net-score) ενός εγγράφου *d*** είναι ένας συνδυασμός της ποιότητας του εγγράφου (που έστω ότι δίνεται από μια συνάρτηση $g$ στο [0, 1]) και της συνάφειας του με το ερώτημα $q$ (που εκφράζεται από το συνημίτονο) π.χ.:

$$\text{net-score}(q, d) = g(d) + \cos(q, d)$$

*Θέλουμε να επιλέξουμε σελίδες που είναι και γενικά σημαντικές (authoritative) και συναφείς ως προς την ερώτηση (το οποίο μας δίνει το συνημίτονο)*

- Πως μπορούμε να επιτύχουμε γρήγορο τερματισμό (early termination); Δηλαδή να μην επεξεργαστούμε όλη τη λίστα καταχωρήσεων για να βρούμε τα καλύτερα $k$;

# Με βάση την «ποιότητα» του εγγράφου ($g$(d))

- Διατάσουμε τις λίστες καταχωρήσεων με βάση την ποιότητα (π.χ., PageRank) των εγγράφων:

$$g(d_1) > g(d_2) > g(d_3) > \ldots$$

Η διάταξη των εγγράφων είναι ίδια για όλες τις λίστες καταχωρήσεων

✓ Τα «καλά» έγγραφα στην αρχή της κάθε λίστας, οπότε αν θέλουμε να βρούμε γρήγορα καλά αποτελέσματα μπορούμε να δούμε μόνο την αρχή της λίστας

# Με βάση την «ποιότητα» του εγγράφου (*g*(d))

*Επεξεργαζόμαστε ένα έγγραφο τη φορά* – δηλαδή, για κάθε έγγραφο υπολογίζουμε πλήρως το net-score του (για όλους τους όρους του ερωτήματος)

- Έστω $g \rightarrow [0, 1]$,

  το τελευταίο *k*-κορυφαίο έγγραφο έχει βαθμό **1.2**

  και για το έγγραφο *d* που επεξεργαζόμαστε $g(d) < 0.1$, άρα και για όλα τα υπόλοιπα συνολικός βαθμός < 1.1 (στην καλύτερη περίπτωση έχουν cos ίσο με 1 που δεν αρκεί όμως για να «ξεπεράσει» το *k*-οστό καλύτερο).

  => δε χρειάζεται να επεξεργαστούμε το υπόλοιπο των λιστών

# Διάταξη καταχωρήσεων του $t$ με βάση το $tf_{t,d}$

Ιδέα: δεν επεξεργαζόμαστε τις καταχωρήσεις που θα συνεισφέρουν λίγο στον τελικό βαθμό

*Διάταξη των εγγράφων με βάση το βάρος (weight)* $w_{t,d}$

✓ *Όχι κοινή διάταξη των εγγράφων σε όλες τις λίστες*

- Η απλούστερη περίπτωση, normalized tf-idf weight
- Τα κορυφαία k έγγραφα είναι πιθανόν να βρίσκονται *στην αρχή* αυτών των ταξινομημένων λιστών.

→ γρήγορος τερματισμός ενώ επεξεργαζόμαστε τις λίστες καταχωρήσεων *μάλλον* δε θα αλλάξει τα κορυφαία k έγγραφα

# Πρόωρος τερματισμός

- Κατά τη διάσχιση των καταχωρήσεων ενός όρου *t*, σταμάτα νωρίς αφού:
  - Δεις ένα προκαθορισμένο αριθμό *r* από έγγραφα
  - Το $wf_{t,d}$ πέφτει κάτω από κάποιο κατώφλι
- Πάρε την ένωση του συνόλου των εγγράφων που προκύπτει
  - Ένα σύνολο για κάθε όρο
- Υπολόγισε τους βαθμούς μόνο αυτών των εγγράφων

# Επιπρόσθετοι ευριστικοί (περίληψη)

- Μέθοδος 1 (με βάση την ποιότητα των εγγράφων): Συχνά υπάρχει μια διαβάθμιση των εγγράφων με βάση κάποια κριτήρια
  - Αντί να διατάσουμε με βάση το docID, *διατάσουμε με βάση κάποια μέτρηση «αναμενόμενης συνάφειας»*
  - Βασισμένο σε *επεξεργασία ανά έγγραφο*
- Μέθοδος 2: Ευριστικό για prune του search space
  - Δεν υπάρχει εγγύηση της ορθότητας του, δηλαδή, μπορεί να μας δώσει έγγραφα που αν και αρκετά καλά, δεν είναι στα top-k
  - Στην πράξη σχεδόν σταθερό χρόνο (constant time).
  - Βασισμένο σε *επεξεργασία ανά όρο*

✓ *Και στις δύο περιπτώσεις διατάσουμε τις λίστες καταχωρήσεων με ειδικό τρόπο*

# Κλάδεμα συστάδων

Προ-επεξεργασία - συσταδοποίηση (clustering) εγγράφων

Για κάθε ερώτημα *q*

- Βρες τον πιο κοντινό αντιπρόσωπο της συστάδας (ηγέτη) *L*.
- Ψάξε για τα *K* πλησιέστερα έγγραφα ανάμεσα *στη συστάδα του L*.

# Σύνοψη: Πιθανοί τρόποι διάταξης της posting list

Τα έγγραφα στη λίστα ενός όρου *t* διατάσσονται είτε:

- Με βάση το *document-id*
- Με βάση τη *σημαντικότητα του εγγράφου* (πχ PageRank), *πρώτα τα πιο σημαντικά έγγραφα*
- Με βάση τη *συχνότητα εμφάνισης του όρου t* στο έγγραφο, *πρώτα τα έγγραφα με πολλές εμφανίσεις του όρου*

# Σύνοψη: Τρόποι υπολογισμού συνάφειας εγγράφου-ερώτησης

- *Document-at-a-time*: για κάθε έγγραφο *d* της συλλογής, υπολογίζουμε τη συνολική συνάφεια του με την ερώτηση (το βαθμό του για όλους του όρους)

- *Term-at-a-time*: για κάθε όρο *t* της ερώτησης, υπολογίζουμε το βαθμό του για όλα τα έγγραφα της συλλογής

- *Score-at-a-time*: ξεκινάμε με τα postings που επηρεάζουν περισσότερο το βαθμό

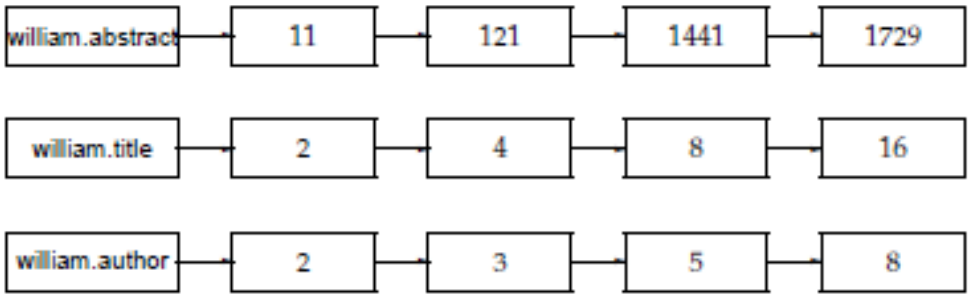
# Βαθμιδωτά ευρετήρια

- Η χρήση βαθμιδωτών ευρετηρίων θεωρείται ως ένας από τους λόγους που η ποιότητα των αποτελεσμάτων του Google ήταν αρχικά σημαντικά καλύτερη (2000/01) από αυτήν των ανταγωνιστών τους.
- μαζί με το PageRank, τη χρήση του anchor text και περιορισμών θέσεων (proximity constraints)

# Παραμετρική αναζήτηση

# Παραμετρική αναζήτηση

Βασικό ευρετήριο ζώνης στο λεξικό:




Η πληροφορία ζώνης στις λίστες καταχώρησης:

# Τι είδαμε στο προηγούμενο μάθημα

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

# Περιλήψεις αποτελεσμάτων

- Η περιγραφή του εγγράφου είναι κρίσιμη γιατί συχνά οι χρήστες βασίζονται σε αυτήν για να αποφασίσουν αν το έγγραφο είναι σχετικό
  - Δε χρειάζεται να διαλέξουν ένα-ένα τα έγγραφα με τη σειρά

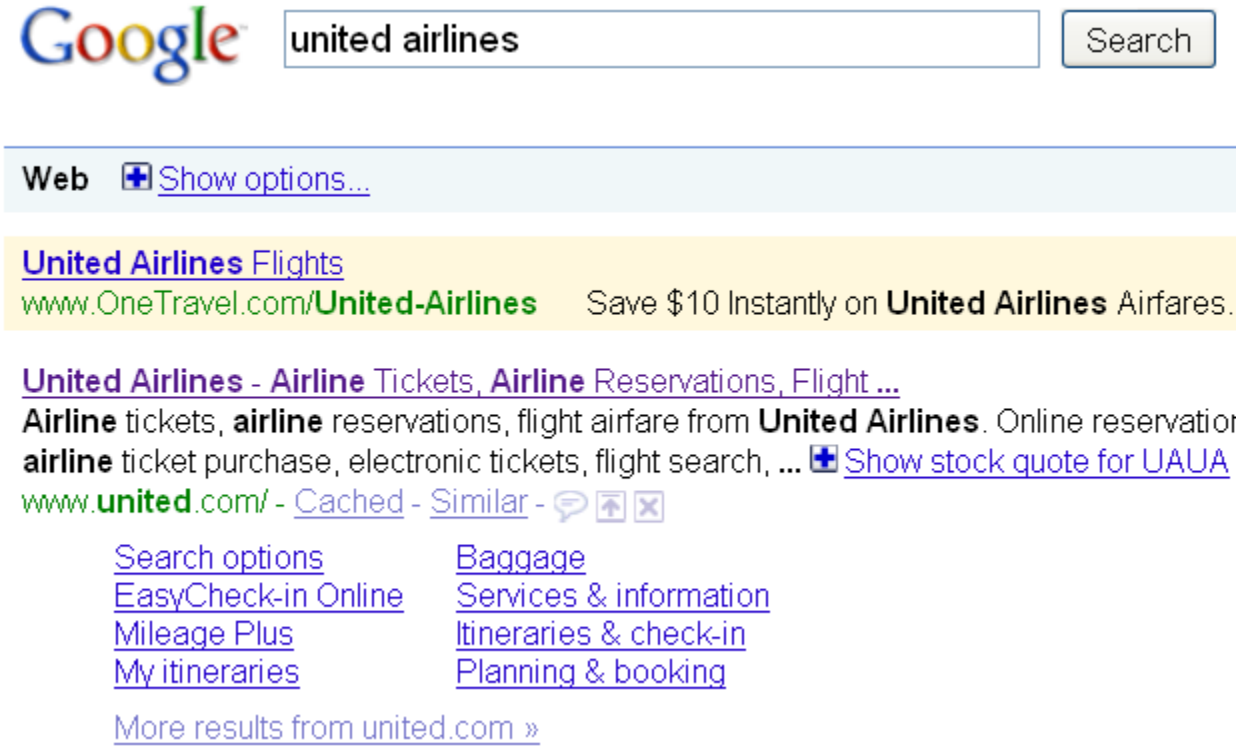
*Ο τίτλος αυτόματα από μεταδεδομένα, αλλά πώς να υπολογίσουμε τις περιλήψεις;*

Δύο βασικά είδη περιλήψεων

- Μια στατική περίληψη (static summary) ενός εγγράφου είναι πάντα η ίδια ανεξάρτητα από το ερώτημα που έθεσε ο χρήστης
- Μια δυναμική περίληψη (dynamic summary) εξαρτάται από το ερώτημα (query-dependent). Προσπαθεί να εξηγήσει γιατί το έγγραφο ανακτήθηκε για το *συγκεκριμένο* κάθε φορά ερώτημα

# Quicklinks

- Για *navigational query* όπως ***united airlines*** οι χρήστες πιθανόν να ικανοποιούνται από τη σελίδα www.united.com
- Quicklinks παρέχουν navigational cues σε αυτή τη σελίδα



# Τι άλλο θα δούμε σήμερα;

- Πως ξέρουμε αν τα αποτελέσματα είναι καλά
  - Αξιολόγηση μηχανών αναζήτησης: μεθοδολογία και μέτρα

# Αξιολόγηση συστήματος

## Αποδοτικότητα (Performance)

- Πόσο γρήγορη είναι η κατασκευή του ευρετηρίου;
  - Αριθμός εγγράφων την ώρα
  - Μέγεθος ευρετηρίου
- Πόσο γρήγορη είναι η αναζήτηση;
  - π.χ., latency 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)**
- *Τι κάνει ένα χρήστη χαρούμενο;* Οι παράγοντες περιλαμβάνουν:
  - Ταχύτητα απόκρισης (Speed of response)
  - Μέγεθος/κάλυψη ευρετηρίου
  - Εύχρηστη διεπαφή (Uncluttered UI)
  - Χωρίς κόστος (free)
- *Κανένα από αυτά δεν αρκεί*: εξαιρετικά γρήγορες αλλά άχρηστες απαντήσεις δεν ικανοποιούν ένα χρήστη **(συνάφεια-relevance)**
- *Πως μπορούμε να μετρήσουμε τη συνάφεια;*

# Ποιοι είναι οι χρήστες

***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.

# Συνήθης ορισμός: Συνάφεια

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

***Μα πως θα μετρήσουμε τη συνάφεια***;

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. Ένα σύνολο αποτίμησης της συνάφειας κάθε ζεύγους ερωτήματος-εγγράφου (συνήθως δυαδικές αποτιμήσεις: συναφής-μη συναφής) - gold standard/ground truth

# Συνάφεια και Ανάγκη Πληροφόρησης

- Συνάφεια ως προς τι;

*Συνάφεια ως προς την ερώτηση*

- 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 σε 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 χρησιμοποιούμε συνάφεια έγγραφου-ερωτήματος.

# 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)

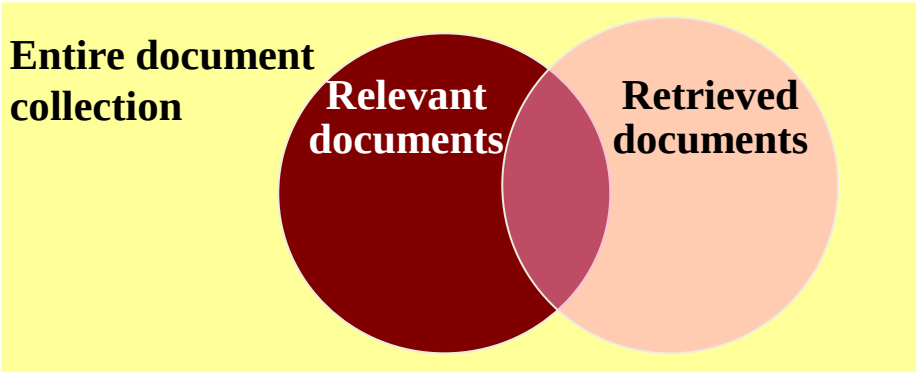
*πραγματικά*

*αποτέλεσμα*

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

$P = TP / ( TP + FP )$

$R = TP / ( TP + FN )$



# 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 ανάκτησης 100%!
- Το αντίστροφο ισχύει επίσης (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

# 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

*Γεωμετρικό μέσο* (geometric mean) γινόμενο

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

Το F (αρμονικό μέσο) είναι ένα είδος ομαλού ελάχιστου

# Ένα συνδυαστικό μέτρο *F*

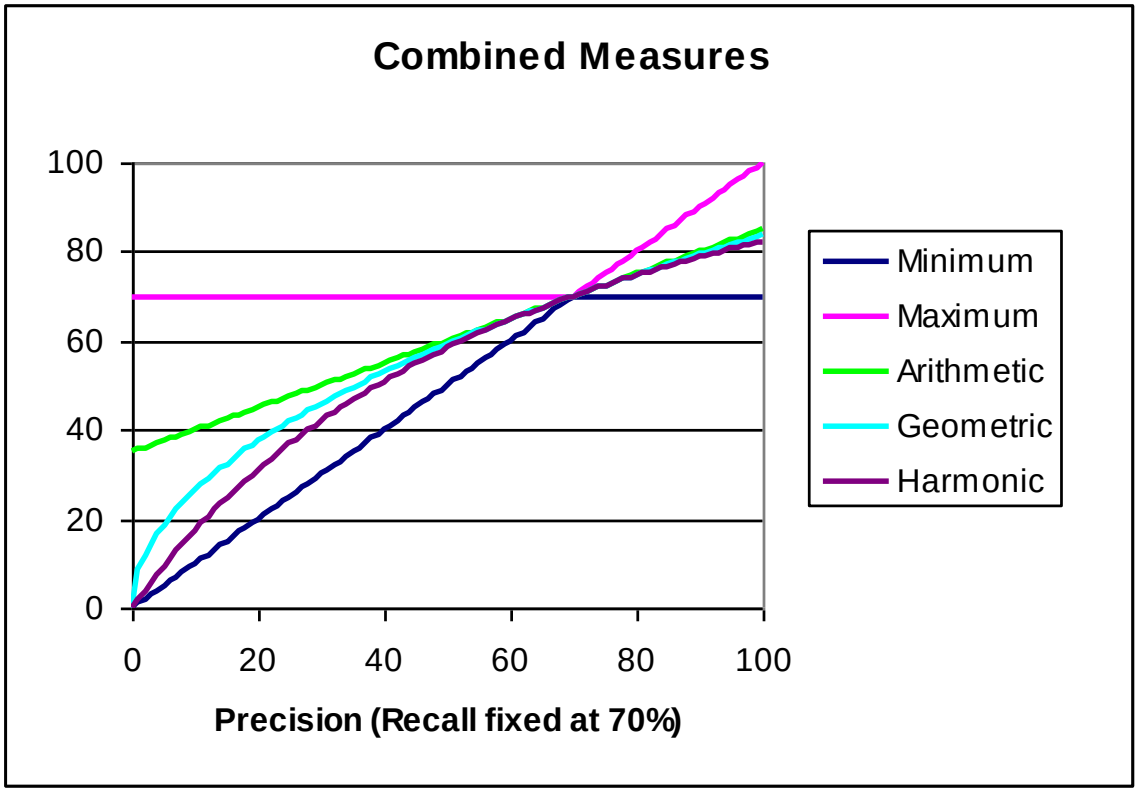
Συνήθως ισορροπημένο (balanced) $F_1$

- Αρμονικό μέσο των P και R

$$F1 = 1/ [(1/2)1/P + (1/2)1/R] = 2PR/P+R$$

✓ 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 $b^2 \in [0,\infty]$

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

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

- Για ποια περιοχή τιμών του β 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 |

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

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

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

# Ορθότητα (Accuracy)

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

**Ορθότητα (Accuracy)** είναι το ποσοστό των αποφάσεων (συναφή/μη συναφή) που είναι σωστές.

Με βάση τον πίνακα ενδεχομένων:

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

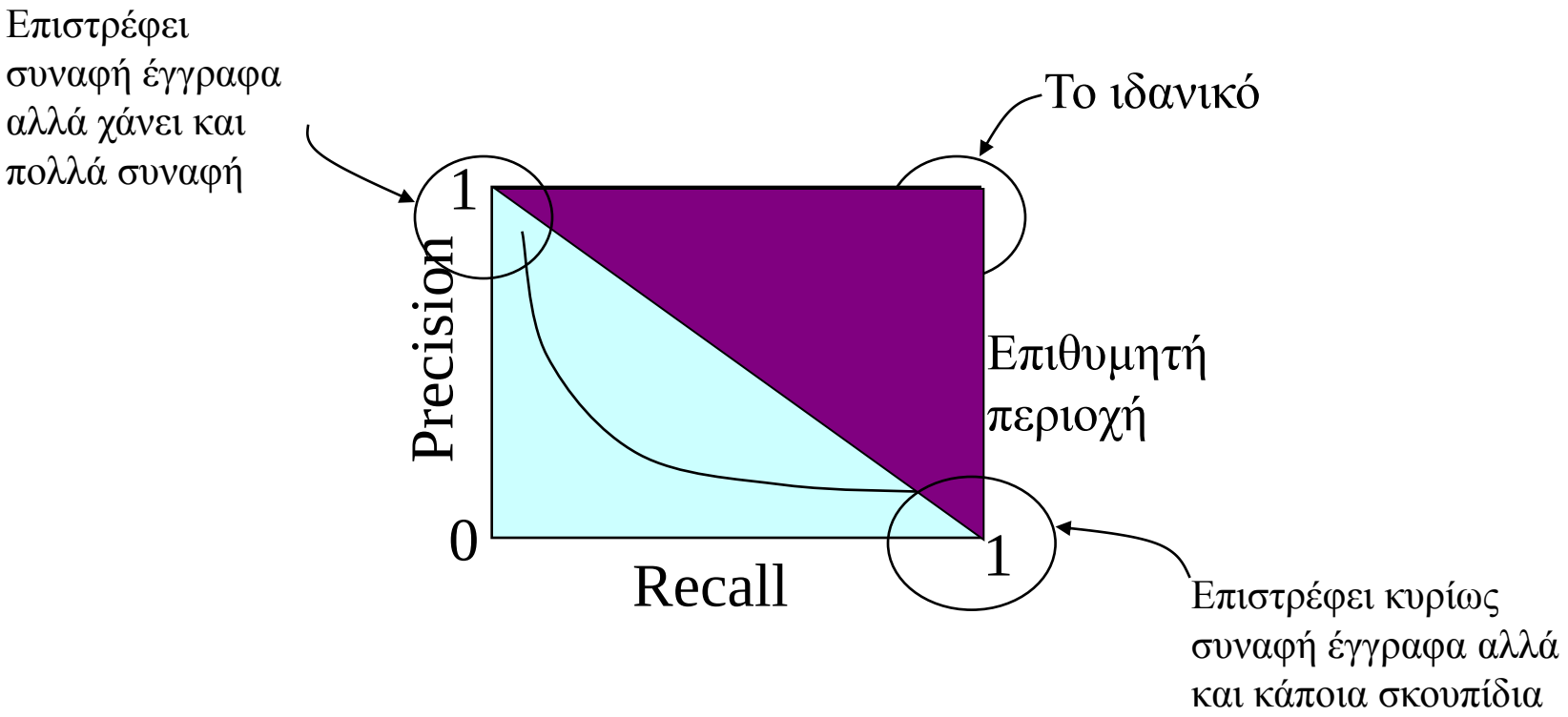
# Δυσκολίες στη χρήση P/R

- 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 δυαδικές
  - Ενδιάμεσες αξιολογήσεις;
- Εξαρτώνται από τη συλλογή/συγγραφή
  - Τα αποτελέσματα μπορεί να διαφέρουν από το ένα πεδίο στο άλλο

# 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

# Ακρίβεια και Ανάκληση



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 | |

total # of relevant docs = 6
Check each new recall point:

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

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

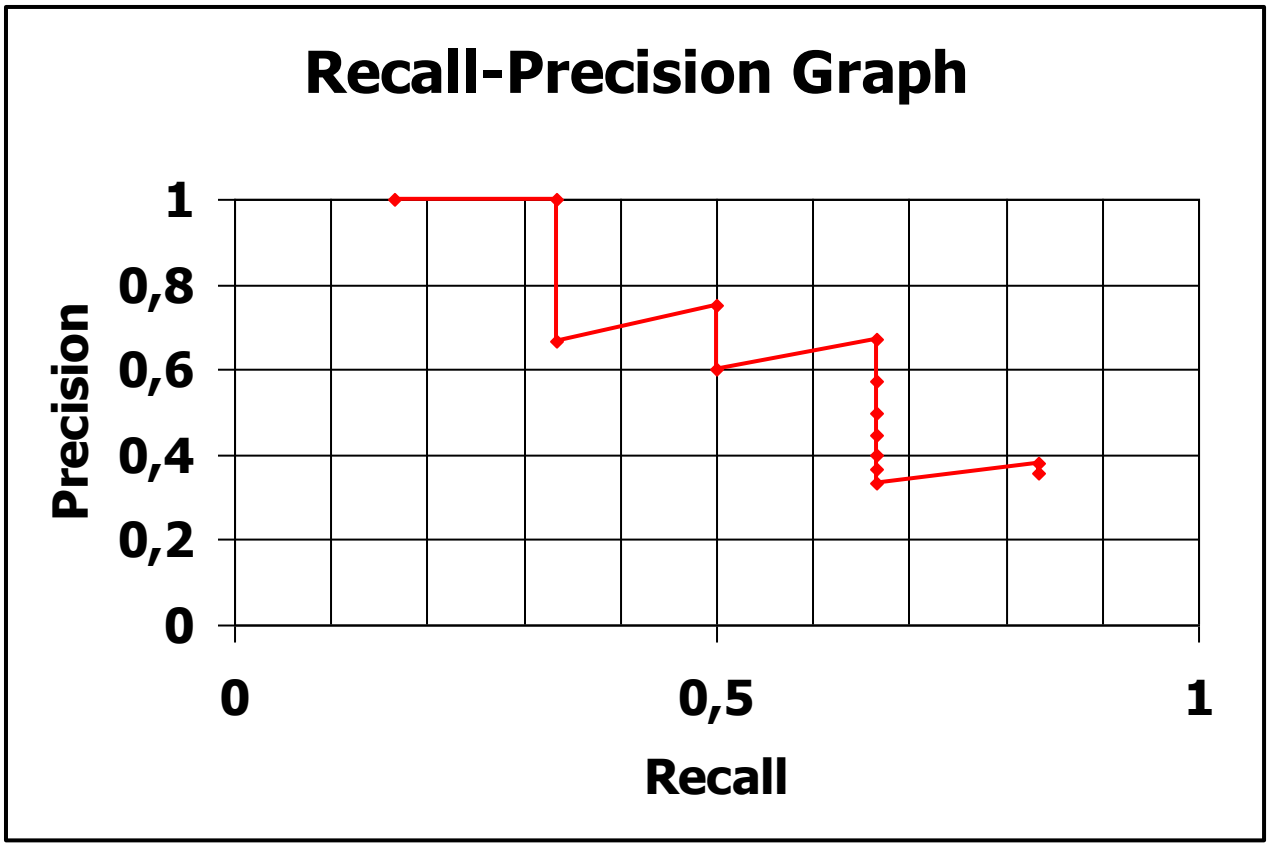
R=3/6=0.5; P=3/4=0.75

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 I (συνέχεια)

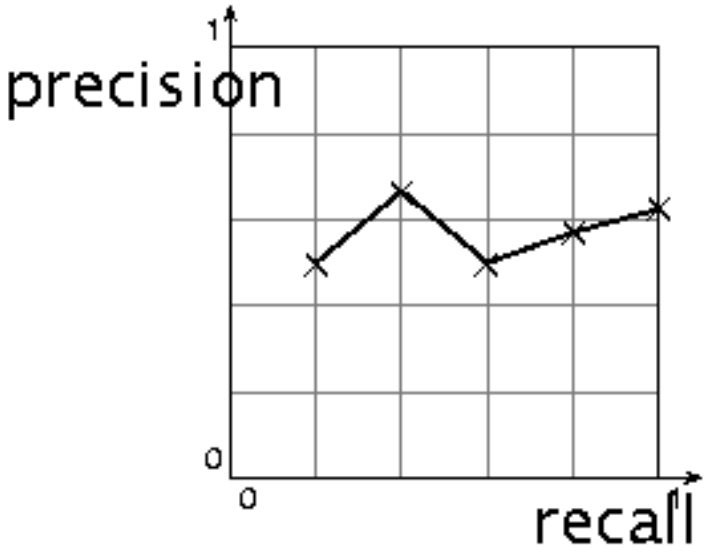


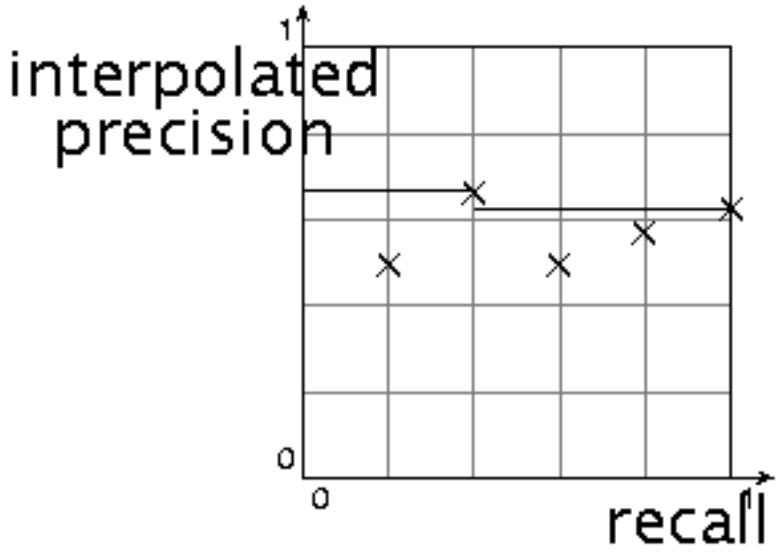

Πριονωτή – το precision ελαττώνεται για το ίδιο recall μέχρι να βρεθεί το επόμενο συναφές έγγραφο

# Ακρίβεια εκ παρεμβολής (Interpolated precision)

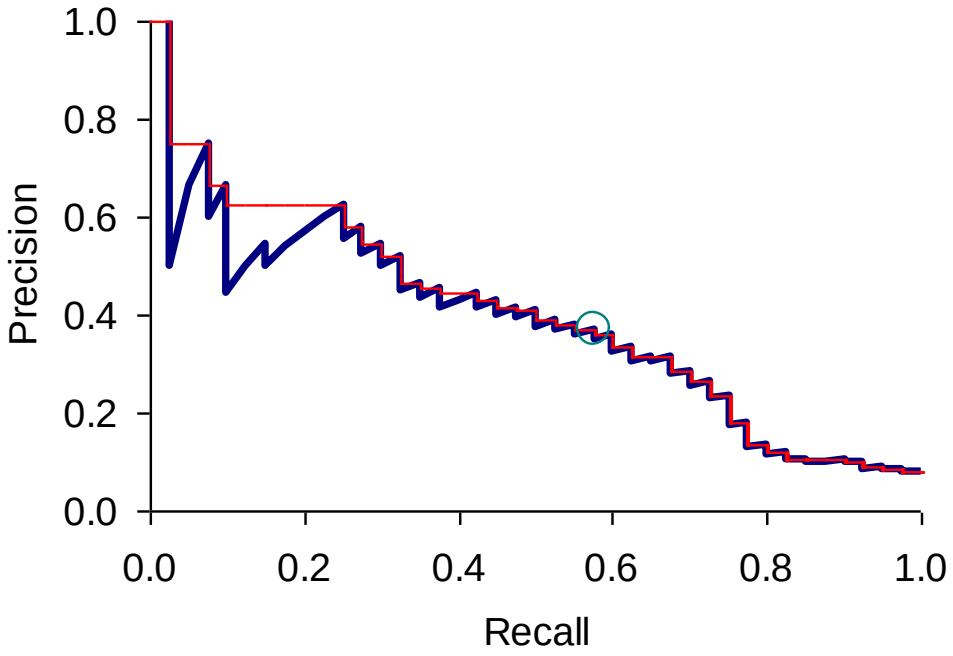
- Αν η ακρίβεια αλλάζει τοπικά με την αύξηση της ανάκλησης, το λαμβάνουμε υπ' όψιν – *ο χρήστης θέλει να δει και άλλα έγγραφα αν αυξάνεται και η ακρίβεια και η ανάκληση*
- Παίρνουμε τη μέγιστη τιμή της ακρίβειας στα δεξιά της τιμής

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



# Καμπύλη Ακρίβειας/Ανάκλησης

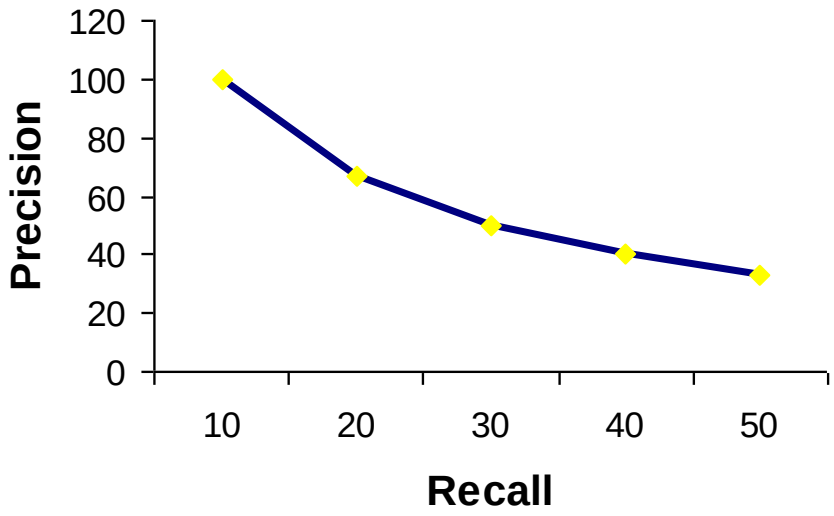



Κάθε σημείο αντιστοιχεί σε ένα αποτέλεσμα για τα κορυφαία *k* έγγραφα (*k* = 1, 2, 3, 4, . . .).

Παρεμβολή (με κόκκινο): μέγιστο των μελλοντικών σημείων

# Παράδειγμα II

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



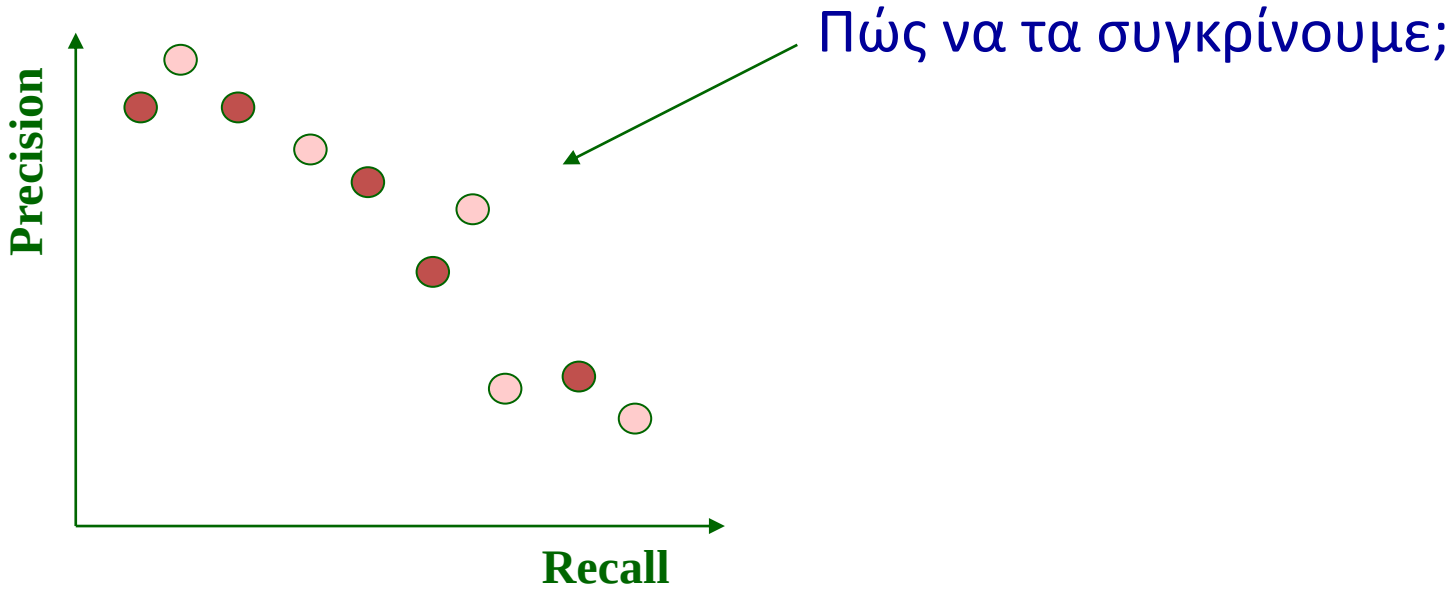

| Rank | Doc | Rel | $R_{ecall}$ | $P_{recision}$ |
|---|---|---|---|---|
| 0 | | | 0 % | 0 % |
| 1 | $d_{123}$ | ✔ | 10 % | 100 % |
| 2 | $d_{84}$ | | 10 % | 50 % |
| 3 | $d_{56}$ | ✔ | 20 % | 67 % |
| 4 | $d_{6}$ | | 20 % | 50 % |
| 5 | $d_{84}$ | | 20 % | 40 % |
| 6 | $d_{9}$ | ✔ | 30 % | 50 % |
| 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

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

Παράδειγμα καθιερωμένων επιπέδων ανάκλησης (πλήθος επιπέδων: 11):

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

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

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

# 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).

# 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 |

11-point average: ≈ 0.425

Γενικά υπολόγισε το μέσο precision recall για ένα σύνολο από ερωτήματα

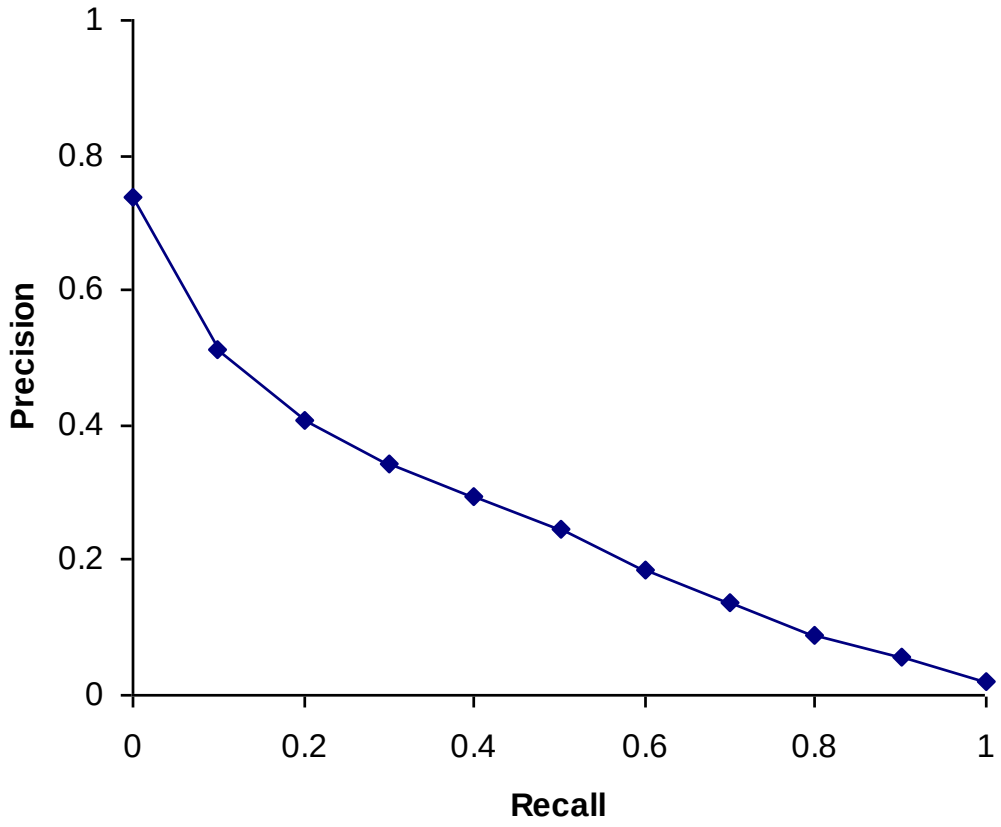
$N_q$ – πλήθος ερωτημάτων

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

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

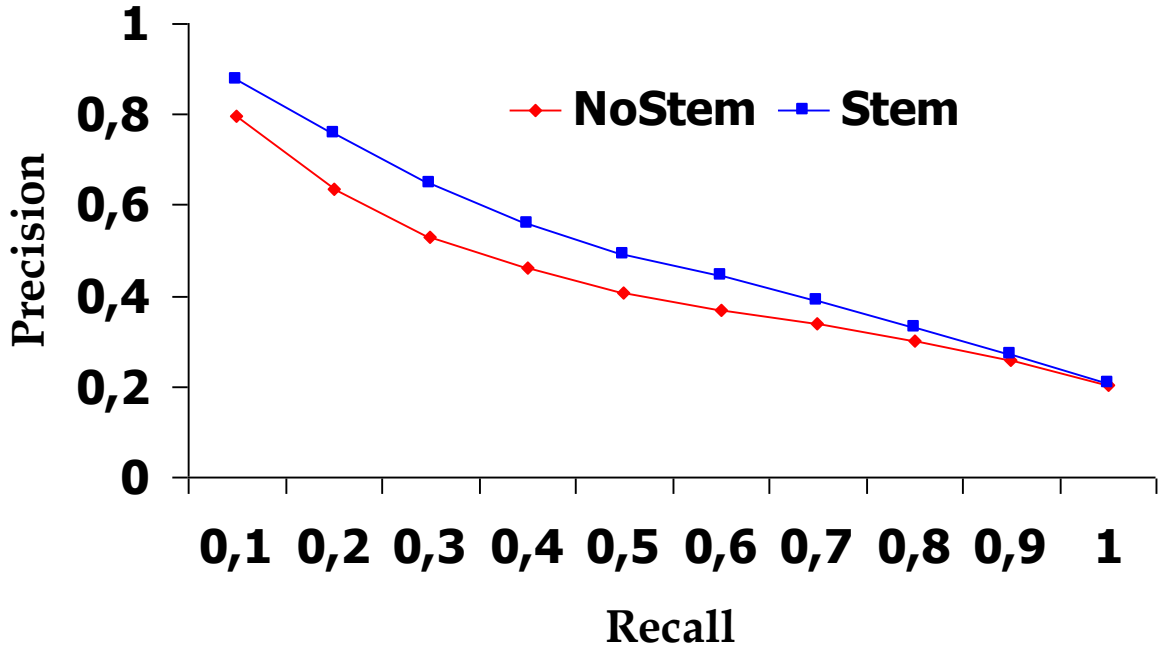
# Τυπική (καλή;) ακρίβεια 11-σημείων

- 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



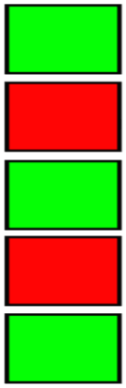
# Ακρίβεια στα *k (precision@k)*

- ***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



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

# ΤΕΛΟΣ 9ου 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 Τζίτζικα (Παν. Κρήτης)*

Οι επόμενες σελίδες δείχνουν αυτούς τους 200 όρους που **μπορεί** να λαμβάνονται υπόψη στη διαβάθμιση της google

# Domain Factors

1. Domain *Age*
2. *Keyword* Appears in Top Level Domain
3. Keyword As First Word in Domain
4. Domain registration (time) length
5. Keyword in Subdomain Name

   http://www.yoursite.com -> http://subdomain.yoursite.com
6. Domain History
7. Exact Match Domain (reduced in 2013)

   http://howtoinvestinstocks.com
8. Public vs. Private *Whois*
9. Penalized Whois Owner (if identified as a spammer)
10. *Country TLD extension* (having a Country Code Top Level Domain (.cn, .pt, .ca) helps the site rank for that particular countr but limits the site's ability to rank globally.

# Page-Level Factors

1. Keyword in *Title Tag*
2. Title Tag Starts with Keyword
3. Keyword in Description Tag
4. Keyword Appears in *H1 Tag*
5. Keyword is the Most Frequently Used Phrase in Document

6. Keyword *Prominence*: Having a keyword appear in the first 100-words
7. Keyword *in H2, H3 Tags*: Having your keyword appear as a subheading in H2 or H3 format may be another weak relevancy signal.
8. Keyword *Word Order*: An exact match of a query keyword in a page content will generally rank better than the same keyword phrase in a different order.

9. Content Length: *Content with more words* can cover a wider breadth and are likely preferred
10. Keyword Density

# Page-Level Factors (II)

11. Page **Loading Speed** via HTML
12. Page Loading Speed via Chrome: Google may also use Chrome user data as this takes into account server speed, CDN usage and other non HTML-related speed signals.
13. **Duplicate Content**: Identical content on the same site (even slightly modified) can negatively influence a site's search engine visibility.
14. Rel=Canonical: When used properly, use of this tag may prevent Google from considering pages duplicate content.
15. Syndicated Content: Is the content on the page original?
16. **Image Optimization**: Images on-page send search engines important relevancy signals through their file name, alt text, title, description and caption.
17. **Recency of Content Updates**: Caffeine favors recently updated content, especially for time-sensitive searches. Google shows the date last update for certain pages
18. Magnitude of Content Updates: The significance of edits and changes is also a freshness factor. Adding or removing entire sections is a more significant than switching around the order of a few words.
19. Historical Updates Page Updates: How often has the page been updated over time?

# Page-Level Factors (III)

20. Outbound Link Quality: linking out to authority sites
21. Outbound Link Theme: search engines may use the content of the pages you link to as a relevancy signal. For example, if you have a page about cars that links to movie-related pages, this may tell Google that your page is about the movie Cars, not the automobile.
22. Number of Outbound Links: Too many dofollow OBLs may "leak" PageRank, which can hurt search visibility.
23. Helpful *Supplementary Content*: helpful supplementary content is an indicator of a quality, e.g., currency converters, loan interest calculators and interactive recipes.
24. References and Sources: Citing references and sources, like research papers do, may be a sign of quality.
25. *Latent Semantic Indexing* Keywords in Content (LSI): LSI keywords help search engines extract meaning from words with more than one meaning (Apple the computer company vs. the fruit). The presence/absence of LSI probably also acts as a content quality signal.
26. LSI Keywords in Title and Description Tags: As with webpage content, LSI keywords in page meta tags probably help Google discern between synonyms. May also act as a relevancy signal.

# Page-Level Factors (IV)

27. Multimedia: e.g, Images, videos
28. *Number of **Internal Links*** Pointing to Page: The number of internal links to a page indicates its importance relative to other pages on the site.
29. Quality of Internal Links Pointing to Page: Internal links from authoritative pages on domain have a stronger effect than pages with no or low PR.

30. *Broken Links*: too many broken links on a page may be a sign of a neglected or abandoned site.
31. *Grammar and Spelling:*
32. *Reading Level:* estimates the reading level (basic, intermediate, advanced) of webpages but what they do with that information is up for debate.
33. *HTML errors/WC3 validation*: Lots of HTML errors or sloppy coding may be a sign of a poor quality site.
34. *Bullets and Numbered Lists*
35. User *Friendly Layout*

# Page-Level Factors (IV)

36. Page Host's Domain Authority: All things being equal a page on an authoritative domain will rank higher than a page on a domain with less authority.
37. Page's PageRank: Not perfectly correlated. In general higher PR pages rank better than low PR pages.
38. URL Length: excessively long URLs may hurt search visibility.
39. URL Path: A page closer to the homepage may get a slight authority boost.
40. Keyword in URL: Another important relevancy signal.
41. URL String: The categories in the URL string are read by Google and may provide a thematic signal to what a page is about
42. Page Category: The category the page appears on is a relevancy signal
43. WordPress Tags: Tags are WordPress-specific relevancy signal.
44. Affiliate Links: Affiliate links themselves probably won't hurt rankings. But if too many, Google's algorithm may pay closer attention to other quality signals to make sure not a "thin affiliate site".

# Page-Level Factors (V)

45. *Human Editors*: Although never confirmed, Google has filed a patent for a system that allows human editors to influence the SERPs.

46. Priority of Page in Sitemap: The priority a page is given via the sitemap.xml file may influence ranking.

47. Quantity of Other Keywords Page Ranks : If the page ranks for several other keywords it may give Google an internal sign of quality.

48. *Page Age*: Although Google prefers fresh content, an older page that's regularly updated may outperform a newer page.
49. Parked Domains: decreased search visibility of parked domains. *Domain parking refers to the registration of an internet domain name without that domain being associated with any services such as e-mail or a website.*
50. Useful Content:

# Site-Level Factors

1. Content Provides Value and Unique Insights
2. Contact Us Page: prefer sites with an "appropriate amount of contact information", e.g., if contact information matches your whois info.
3. **Domain Trust/TrustRank**: Site trust — measured by how many links away a site is from highly-trusted seed sites —*massively important* ranking factor.
4. Site Architecture: A well put-together site architecture (especially a silo structure) helps Google thematically organize your content.
5. Site Updates: How often a site is updated — and especially when new content is added to the site — is a site-wide freshness factor.
6. Number of Pages
7. Presence of Sitemap: A sitemap helps search engines index your pages easier and more thoroughly, improving visibility.
8. Site Uptime: Lots of downtime from site maintenance or server issues may hurt ranking (and even result in deindexing if not corrected).
9. Server Location: may influence where site ranks in **different geographical regions**. Especially important for geo-specific searches.

# Site-Level Factors (II)

10. SSL Certificate (Ecommerce Sites)
11. Terms of Service and Privacy Pages: These two pages help tell Google that a site is a trustworthy member of the internet.
12. Duplicate Content On-Site: Duplicate pages and meta information across your site may bring down all of your page's visibility.
13. Breadcrumb Navigation: This is a style of user-friendly site-architecture that helps users (and search engines) know where they are on a site:
14. Mobile Optimized: create a responsive site; likely that responsive sites get an edge in searches from a mobile device.
15. YouTube: YouTube videos are given preferential treatment in the SERPs
16. Site Usability: A site difficult to use or to navigate can hurt ranking by reducing time on site, pages viewed and bounce rate. This may be an independent algorithmic factor gleaned from massive amounts of user data.

# Site-Level Factors (III)

17. Use of Google Analytics and Google Webmaster Tools: Some think that having these two programs installed on your site can improve your page's indexing. May also directly influence rank by giving Google more data to work with

18. User reviews/Site reputation: A site on review sites like Yelp.com and RipOffReport.com likely play an important role.

# Backlinks Factors

1. Linking *Domain Age*: Backlinks from aged domains may be more powerful
2. **Number of Linking Root Domains**: The number of referring domains is one of the most important ranking factors in Google's algorithm
3. Number of Links *from Separate C-Class IPs*
4. Number of *Linking Pages*: The total number of linking pages — even if some on the same domain — is a ranking factor.
5. Alt Tag (for Image Links): Alt text is an image version of anchor text.
6. Links from .edu or .gov Domains
7. **PR of Linking Page**: The PageRank of the referring page is an extremely important ranking factor.
8. Authority of Linking Domain: The referring domain authority may play an independent role in link importance (ie. a PR2 page link from a site with a homepage PR3 may be worth less than a PR2 page link from PR8 Yale.edu).
9. *Links From Competitors*: Links from other pages ranking in the same SERP (search engine result page) may be more valuable for a page's rank for that particular keyword.

# Backlinks Factors (II)

10. Social Shares of Referring Page: The amount of page-level social shares may influence the link value.
11. Links from Bad Neighborhoods e.g., link farms
12. Guest Posts:
13. Links to Homepage Domain that Page Sits On: Links to a referring page homepage may play special importance in evaluating a site — and therefore a link — weight.
14. Nofollow Links: Google's official word on the matter is: "In general, we don't follow them."
15. Diversity of Link Types
16. "Sponsored Links" Or Other Words Around Link: Words like "sponsors", "link partners" and "sponsored links" may decrease a link's value.

# Backlinks Factors (III)

17. *Contextual Links*: Links embedded inside a page content more powerful than links on an empty page or found elsewhere on the page.
18. Excessive 301 Redirects to Page
19. Backlink Anchor Text
20. Internal Link Anchor Text
21. Link Title Attribution: The link title (the text that appears when you hover over a link)
22. *Country TLD* of Referring Domain: Getting links from country-specific top level domain extensions (.de, .cn, .co.uk) may help rank better in that country.
23. Link Location In Content: Links the beginning of a piece of content carry slight more weight than links placed at the end of the content.
24. Link Location on Page: Where a link appears on a page is important. Generally, links embedded in a page content are more powerful than links in the footer or sidebar area.

# Backlinks Factors (IV)

25. Linking Domain *Relevancy:*
26. Page Level Relevancy: link from a page closely tied to page content more powerful than a link from an unrelated page.

27. *Text Around Link Sentiment*: whether or not a link to your site is a recommendation or part of a negative review. Links with positive sentiments likely carry more weight.

28. Keyword in Title
29. Positive Link Velocity
30. Negative Link Velocity

31. Links from "Hub" Pages: getting links from pages that are considered top resources (or hubs) on a certain topic are given special treatment.
32. Link from Authority Sites: A link from a site considered an "authority site" likely more important than a link from a small, microniche site.

# Backlinks Factors (IV)

33. Linked to as Wikipedia Source: Although the links are nofollow, getting a link from Wikipedia may give a little added trust and authority
34. Co-Occurrences: The words that tend to appear around your backlinks
35. Backlink Age: older links have more ranking power than newly minted backlinks.
36. Links from Real Sites vs. Splogs: probably more weight to links from "real sites" than from fake blogs.
37. Natural Link Profile: A site with a "natural" link profile is going to rank highly and be more durable to updates.
38. Reciprocal Links: _"Excessive link exchanging" as a link scheme to avoid.
39. User Generated Content Links: Google is able to identify links generated from UGC vs. the actual site owner.

# Backlinks Factors (IV)

40. Links from 301: Links from 301 redirects may lose compared to a direct link.
41. Schema.org Microformats: Pages that support microformats may rank above pages without it.
42. DMOZ Listed: Many believe that Google gives DMOZ listed sites a little extra trust.
43. Yahoo! Directory Listed:
44. Number of Outbound Links on Page:
45. Forum Profile Links: Because of industrial-level spamming, Google may significantly devalue links from forum profiles.
46. Word Count of Linking Content: A link from a 1000-word post is more valuable than a link inside of a 25-word snippet.
47. Quality of Linking Content:
48. Sitewide Links: sitewide links are “compressed” to count as a single link.

# User Interaction

1. Organic ***Click Through*** Rate for a Keyword
2. Organic CTR for All Keywords

3. **Bounce Rate**: pages where people quickly bounce is probably not very good
4. **Direct Traffic**: It's confirmed that Google uses data from Google Chrome to determine whether or not people visit a site (and how often). Sites with lots of direct traffic are likely higher quality than sites that get very little direct traffic.
5. **Repeat Traffic**: whether or not users go back to a page or site after visiting.

6. Blocked Sites: Google has discontinued this feature in Chrome. However, Panda used this feature as a quality signal.
7. Chrome **Bookmarks**
8. Google Toolbar Data: besides page loading speed and malware, not know what kind of data they glean from the toolbar.

# User Interaction (II)

9. Number of Comments:

10. Dwell Time: Google pays very close attention to “dwell time”: how long people spend on your page when coming from a Google search. This is also sometimes referred to as “long clicks vs short clicks”.

# Special Algorithm Rules

1. Query **Freshness**
2. Query Deserves **Diversity**: Google may add diversity to a SERP for ambiguous keywords, such as “Ted”, “WWF” or “ruby”.
3. User **Browsing History**: Sites that you frequently visit while signed into Google
4. User **Search History**: Search chain influence search results for later searches. For example, if you search for “reviews” then search for “toasters”, Google is more likely to show toaster review sites higher in the SERPs.
5. **Geo Targeting**: Google gives preference to sites with a local server IP and country-specific domain name extension.
6. Safe Search: Search results with curse words or adult content not appear for people with Safe Search turned on.
7. Google+ Circles: Google shows higher results for authors and sites added to your Google Plus Circles
8. DMCA Complaints: downranks pages with DMCA complaints (copyright act).

# Special Algorithm Rules (II)

9. Domain Diversity: The "*Bigfoot Update*" supposedly added more domains to each SERP page.
10. Transactional Searches: Google sometimes displays different results for shopping-related keywords, like flight searches.
11. Local Searches: Google often places Google+ Local results above the "normal" organic SERPs.
12. Google News Box: Certain keywords trigger a Google News box:
13. Big Brand Preference: After the Vince Update, Google began giving big brands a boost for certain short-tail searches.
14. Shopping Results: Google sometimes displays Google Shopping results in organic SERPs
15. Image Results: Google elbows our organic listings for image results for searches commonly used on Google Image Search.

# Special Algorithm Rules (II)

16. Easter Egg Results: Google has a dozen or so Easter Egg results. For example, when you search for ”Atari Breakout” in Google image search, the search results turn into a playable game (!).
17. Single Site Results for Brands: Domain or brand-oriented keywords bring up several results from the same site.

# Social Signals

1. Number of Tweets:
2. Authority of Twitter Users Accounts:

3. Number of Facebook Likes: Although Google can not see most Facebook accounts, likely they consider the number of Facebook likes a page receives as a weak ranking signal.
4. Facebook Shares: Facebook shares — because more similar to a backlink — may have a stronger influence than Facebook likes.
5. Authority of Facebook User Accounts: As with Twitter, Facebook shares and likes coming from popular Facebook pages may pass more weight.

6. Pinterest Pins: popular and lots of public data.

7. Votes on Social Sharing Sites: possible that Google uses shares at sites like Reddit, Stumbleupon and Digg as another type of social signal.

# Social Signals (II)

8. Number of Google+1's:
9. Authority of Google+ User Accounts:
10. Verified Google+ Authorship: may already be a trust signal.

11. Social Signal Relevancy: Google probably uses relevancy information from the account sharing the content and the text surrounding the link.
12. Site Level Social Signals: Site-wide social signals may increase a site overall authority, which will increase search visibility for all of its pages.

# Brand Signals

1. Brand Name Anchor Text: Branded anchor text is a simple — but strong — brand signal.
2. Branded Searches: people search for brands. If people search for your site in Google (ie. “Backlinko twitter”, Backlinko + “ranking factors”), Google likely takes this into consideration when determining a brand.
3. Site Has Facebook Page and Likes: Brands tend to have Facebook pages with lots of likes.
4. Site has Twitter Profile with Followers: Twitter profiles with a lot of followers signals a popular brand.
5. Official Linkedin Company Page: Most real businesses have company Linkedin pages.
6. Employees Listed at Linkedin:
7. Legitimacy of Social Media Accounts: A social media account with 10,000 followers and 2 posts is probably interpreted a lot differently than another 10,000-follower strong account with lots of interaction.

# Brand Signals (II)

8. Brand Mentions on News Sites
9. Co-Citations: Brands get mentioned without getting linked to. Google likely looks at non-hyperlinked brand mentions as a brand signal.
10. Number of RSS Subscribers
11. Brick and Mortar Location With Google+ Local Listing:
12. Website is Tax Paying Business

# On Site Webspam Factors

1. Panda Penalty: Sites with low-quality content (particularly content farms) are less visible in search after getting hit by a Panda penalty.
2. Links to Bad Neighborhoods: Linking out to “bad neighborhoods” — like pharmacy or payday loan sites — may hurt your search visibility.
3. Redirects: not just penalized, but de-indexed.
4. Popups or Distracting Ads:
5. Site Over-Optimization: Includes on-page factors like keyword stuffing, header tag stuffing, excessive keyword decoration.
6. Page Over-Optimization: Many people report that — unlike Panda — Penguin targets individual page (and even then just for certain keywords).
7. Ads Above the Fold: The “Page Layout Algorithm” penalizes sites with lots of ads (and not much content) above the fold.
8. Hiding Affiliate Links: Going too far when trying to hide affiliate links (especially with cloaking) can bring on a penalty.
9. Affiliate Sites: sites that monetize with affiliate links under extra scrutiny.

# On Site Webspam Factors

10. Autogenerated Content:
11. Excess PageRank Sculpting:
12. IP Address Flagged as Spam: .
13. Meta Tag Spamming: Keyword stuffing can also happen in meta tags.

# Off Site Webspam Factors

1. Unnatural Influx of Links: A sudden (and unnatural) influx of links
2. Penguin Penalty: Sites that were hit by Google Penguin are significantly less visible in search.
3. Link Profile with High % of Low Quality Links: Lots of links from sources commonly used by black hat SEOs (like blog comments and forum profiles) may be a sign of gaming the system.
4. Linking Domain Relevancy: sites with an unnaturally high amount of links from unrelated sites were more susceptible to Penguin.
5. Unnatural Links Warning: Google sent out thousands of "Google Webmaster Tools notice of detected unnatural links" messages. This usually precedes a ranking drop
6. Links from the Same Class C IP: Getting an unnatural amount of links from sites on the same server IP may be a sign of blog network link building.

# Off Site Webspam Factors

7. “Poison” Anchor Text: Having “poison” anchor text (especially pharmacy keywords) pointed to your site may be a sign of spam or a hacked site.
8. Manual Penalty:
9. Selling Links: Selling links can definitely impact toolbar PageRank and may hurt your search visibility.
10. Google Sandbox: New sites that get a sudden influx of links are sometimes put in the Google Sandbox, which temporarily limits search visibility.
11. Google Dance: The Google Dance can temporarily shake up rankings. According to a Google Patent, this may be a way for them to determine whether or not a site is trying to game the algorithm.
12. Disavow Tool: Use of the Disavow Tool may remove a manual or algorithmic penalty for sites that were the victims of negative SEO.
13. Reconsideration Request: A successful reconsideration request can lift a penalty.