ΜΥΕ003: Ανάκτηση Πληροφορίας

Διδάσκουσα: Ευαγγελία Πιτουρά

Βαθμολόγηση. Στάθμιση όρων.

Ακαδημαϊκό Έτος 2022-2023

Τι θα δούμε σήμερα;

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

Boolean Μοντέλο

- Μέχρι τώρα, τα ερωτήματα που είδαμε ήταν Boolean.
  - Τα έγγραφα είτε ταιριάζουν στο ερώτημα, είτε όχι

Έγγραφα		Ερωτήματα	
d_1	a ….. b …	q1	a
d_2	a … a …	q2	b
d_3	a … a … b	q3	a b
d_4	b ….. b … b		
d_5	a … a … b … b		
d_6	a ….		

Boolean Μοντέλο

- Τα Boolean ερωτήματα συχνά έχουν είτε *πολύ λίγα* (=0) είτε *πάρα πολλά* (χιλιάδες) αποτελέσματα (“feast or famine”)
  - Ερώτημα 1: “*standard user dlink 650*” → 200,000 hits
  - Ερώτημα 2: “*standard user dlink 650 no card found*”: 0 hits

- Χρειάζεται επιδεξιότητα για να διατυπωθεί μια ερώτηση που έχει ως αποτέλεσμα ένα διαχειρίσιμο αριθμό ταιριασμάτων
  - AND πολύ λίγα - OR πάρα πολλά

Boolean Μοντέλο

- Κατάλληλο *για ειδικούς* με σαφή κατανόηση των αναγκών τους και γνώση της συλλογής
  - Επίσης, καλό *για εφαρμογές*: οι εφαρμογές μπορούν να επεξεργαστούν χιλιάδες αποτελεσμάτων.

- Αλλά, όχι κατάλληλο για την πλειοψηφία των χρηστών
  - Είναι δύσκολο για τους περισσότερους χρήστες να διατυπώσουν Boolean ερωτήματα
  - Οι περισσότεροι χρήστες δεν θέλουν να διαχειριστούν χιλιάδες αποτελέσματα.
    - Ιδιαίτερα στην περίπτωση των αναζητήσεων στο web

Μοντέλα διαβαθμισμένης ανάκτησης

- Αντί ενός *συνόλου* εγγράφων που ικανοποιούν το ερώτημα, η *διαβαθμισμένη ανάκτηση (ranked retrieval)* επιστρέφει μια *διάταξη* των (κορυφαίων) για την ερώτηση εγγράφων της συλλογής

- Όταν το σύστημα παράγει ένα διατεταγμένο σύνολο αποτελεσμάτων, τα μεγάλα σύνολα δεν αποτελούν πρόβλημα
  - Δείχνουμε απλώς τα *κορυφαία* (top) *k* (≈ 10) αποτελέσματα
  - Δεν παραφορτώνουμε το χρήστη

Προϋπόθεση: ο αλγόριθμος διάταξης να δουλεύει σωστά

Μοντέλα διαβαθμισμένης ανάκτησης

- Η διαβαθμισμένη ανάκτηση συνήθως με *ερωτήματα ελεύθερου κειμένου*
  - *Ερωτήματα ελεύθερου κειμένου (Free text queries)*: Μία ή περισσότερες λέξεις σε μια φυσική γλώσσα (αντί για μια γλώσσα ερωτημάτων με τελεστές και εκφράσεις)

Βαθμολόγηση ως βάση της διαβαθμισμένης ανάκτησης

- Θέλουμε να επιστρέψουμε τα αποτελέσματα διατεταγμένα με βάση *το πόσο πιθανό είναι να είναι χρήσιμα στο χρήστη* ή με βάση *τη συνάφεια τους με το ερώτημα*

- Πως διατάσουμε-διαβαθμίσουμε τα έγγραφα μιας συλλογής με βάση ένα ερώτημα;
  - Αναθέτουμε ένα βαθμό (score) – ας πούμε στο [0, 1] – σε κάθε έγγραφο
  - score(*d*, *q*): μετρά πόσο καλά το *έγγραφο d* "ταιριάζει" (match) με το *ερώτημα q*

Βαθμός ταιριάσματος ερωτήματος-εγγράφου

- Χρειαζόμαστε ένα τρόπο για να αναθέσουμε ένα βαθμό σε κάθε ζεύγος ερωτήματος (*q*), εγγράφου (*d*)

  score(d, q)

  - Αν κανένα όρος του ερωτήματος δεν εμφανίζεται στο έγγραφο, τότε ο βαθμός θα πρέπει να είναι 0

  - Όσο *πιο συχνά* εμφανίζεται ο όρος του ερωτήματος σε ένα έγγραφο, *τόσο μεγαλύτερος* θα πρέπει να είναι ο βαθμός

- Θα εξετάσουμε κάποιες εναλλακτικές για αυτό

Προσπάθεια 1: Συντελεστής Jaccard

Υπενθύμιση: συνηθισμένη μέτρηση της επικάλυψης δύο συνόλων *A* και *B*

jaccard$(A,B) = |A \cap B| \,/\, |A \cup B|$

- jaccard$(A, A) = 1$
- jaccard$(A, B) = 0$ if $A \cap B = 0$

- *Τα A* και *B* δεν έχουν απαραίτητα το ίδιο μέγεθος
- Αναθέτει πάντα έναν αριθμό μεταξύ του 0 και του 1

- Θεωρούμε το ερώτημα και το έγγραφο ως *σύνολα όρων*

Έγγραφα		Ερωτήματα	
d_1	a ….. b …	q1	a
d_2	a … a …	q2	b
d_3	a … a … b	q3	a b
d_4	b ….. b … b		
d_5	a … a … b … b		
d_6	a ….		

Διαβάθμιση με Jaccard Distance (JD)

Βαθμός εγγράφου και ερώτησης

Μέτρο βαθμολογίας επικάλυψης (overlap score measure)

$$\text{score}(q, d) = \sum_{t \in q \cap d} \text{w}(\text{t}, \text{d})$$

κοινοί όροι

Συχνότητα όρου - Term frequency (tf)

Η συχνότητα όρου $tf_{t,d}$ του όρου *t* σε ένα έγγραφο *d* ορίζεται ως ο αριθμός των φορών που το *t* εμφανίζεται στο *d* (το πλήθος των εμφανίσεων του όρου *t* στο έγγραφο *d*)

Συχνότητα εγγράφου (Document frequency)

- Οι *σπάνιοι* όροι παρέχουν περισσότερη πληροφορία από τους συχνούς όρους
  - Θυμηθείτε τα stop words (διακοπτόμενες λέξεις)
- Θεωρείστε έναν όρο σε μια ερώτηση που είναι σπάνιος στη συλλογή (π.χ., *arachnocentric*)
  - Το έγγραφο που περιέχει αυτόν τον όρο είναι πιο πιθανό να είναι περισσότερο συναφές με το ερώτημα από ένα έγγραφο που περιέχει ένα λιγότερο σπάνιο όρο του ερωτήματος
- Θέλουμε να δώσουμε *μεγαλύτερο βάρος στους σπάνιους όρους* – αλλά πως; df

Βάρος idf

- df_t είναι η συχνότητα εγγράφων του *t*: το πλήθος των εγγράφων της συλλογής που περιέχουν το *t*
  - df_t είναι η αντίστροφη μέτρηση της πληροφορίας που παρέχει ο όρος *t*
  - $\mathrm{df}_t \leq N$

- Ορίζουμε την *αντίστροφη συχνότητα εγγράφων idf* (inverse document frequency) του *t* ως

$$\mathrm{idf}_t = N/\mathrm{df}_t$$

Βαθμός εγγράφου και ερώτησης

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

- Μεγάλο για όρους που εμφανίζονται πολλές φορές σε λίγα έγγραφα (μεγάλη *διακριτική δύναμη* (discriminating power) σε αυτά τα έγγραφα)
- Μικρότερο όταν ο όρος εμφανίζεται λίγες φορές σε ένα έγγραφο ή όταν εμφανίζεται σε πολλά έγγραφα
- Το μικρότερο για όρους που εμφανίζονται σχεδόν σε όλα τα έγγραφα

- Υπάρχουν πολλές άλλες παραλλαγές
  - Πως υπολογίζεται το “tf” (με ή χωρίς log)
  - Αν δίνεται βάρος και στους όρους του ερωτήματος
  - …

Στάθμιση tf-idf

Ποιο είναι το idf ενός όρου που εμφανίζεται σε κάθε έγγραφο (ποια η σχέση με stop words);

Τα stop words έχουν
Α μικρό idf
Β μεγάλο idf

- tf-idf των παρακάτω όρων:

	Doc1	Doc2	Doc3
car	27	4	24
auto	3	33	0
insurance	0	33	29
best	14	0	17

υπόθεση

$w_{t,d} = tf_{t,d} * idf_t$

Έγγραφα		Ερωτήματα	
d_1	a ….. b …	q1	a
d_2	a … a …	q2	b
d_3	a … a … b	q3	a b
d_4	b ….. b … b		
d_5	a … a … b … b		
d_6	a ….		

Διαβάθμιση με tf-idf

Έγγραφα		Ερωτήματα	
d_1	a ….. b …	q1	a
d_2	a … a …	q2	b
d_3	a … a … b	q3	a b
d_4	b ….. b … b		
d_5	a … a … b … b		
d_6	a ….		

Διαβάθμιση με tf-idf

υπόθεση

$$w_{t,d} = tf_{t,d} * idf_t$$

$$idf_a = 6/5$$

$$idf_b = 6/4$$

$$score(d_1,q_1) = \sum_{t \in d_1 \cap q_1} tf_{t,d_1} * idf_t$$

$$= tf_{a,d_1} * idf_a = 1 * 6/5 = 6/5$$

$$score(d_2,q_1) = 2 * 6/5$$

$$score(d_3,q_1) = 2 * 6/5$$

$$score(d_4,q_1) = 0$$

$$score(d_5,q_1) = 2 * 6/5$$

$$score(d_6,q_1) = 6/5$$

d2 d5

d1 d3 d6

Το idf δεν επηρεάζει ερωτήματα που έχουν μόνο έναν όρο

Έγγραφα		Ερωτήματα	
d_1	a ….. b …	q1	a
d_2	a … a …	q2	b
d_3	a … a … b	q3	a b
d_4	b ….. b … b		
d_5	a … a … b … b		
d_6	a ….		

$idf_a = 6/5$

$idf_b = 6/4$

Διαβάθμιση με tf-idf

$$score(d_1, q_3) = \sum_{t \in d_1 \cap q_3} tf_{t,d_1} \times idf_t$$

$$= 1 \cdot 6/5 + 1 \cdot 6/4 = 2.7 \quad 4$$

$$score(d_2, q_3) = 2 \cdot 6/5 = 2.4 \quad 5$$

$$score(d_3, q_3) = 2 \cdot 6/5 + 1 \cdot 6/4 = 3.9 \quad 3$$

$$score(d_4, q_3) = 3 \cdot 6/4 = 4.5 \quad 2$$

$$score(d_5, q_3) = 2 \cdot 6/5 + 2 \cdot 6/4 = 5.4 \quad 1$$

$$score(d_6, q_3) = 1 \cdot 6/5 = 6/5 = 1.2 \quad 6$$

d5

d4

d3

d1

d2

d6

Η επίδραση του idf στη διάταξη

- Το idf δεν επηρεάζει τη διάταξη για ερωτήματα *με ένα μόνο όρο*, όπως iPhone

- Το idf επηρεάζει μόνο τη διάταξη για ερωτήματα *με τουλάχιστον δύο όρους*
  - Για το ερώτημα *capricious person*, η idf στάθμιση έχει ως αποτέλεσμα
    - οι εμφανίσεις του capricious να μετράνε *περισσότερο* στην τελική διάταξη των εγγράφων από ότι οι εμφανίσεις του person.
    - *ένα έγγραφο που περιέχει μόνο το capricious είναι πιο σημαντικό από ένα που περιέχει μόνο το person*

Στάθμιση tf-idf

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

Υπάρχουν πολλές άλλες παραλλαγές

- Πως υπολογίζεται το “tf” (με ή χωρίς log)
- Αν δίνεται βάρος και στους όρους του ερωτήματος
- ..

Συχνότητα όρου - Term frequency (tf)

Υπενθύμιση: Η συχνότητα όρου $\text{tf}_{t,d}$ του όρου *t* σε ένα έγγραφο *d* ορίζεται ως ο αριθμός των φορών που το *t* εμφανίζεται στο *d*.

Φτάνει μόνο η συχνότητα;

- Ένα έγγραφο με 10 εμφανίσεις ενός όρου είναι πιο σχετικό από ένα έγγραφο με 1 εμφάνιση του όρου. *Αλλά είναι 10 φορές πιο σχετικό*;

Η συνάφεια (relevance) δεν αυξάνει αναλογικά με τη συχνότητα εμφάνισης όρου

Στάθμιση με Log-συχνότητας

- Η 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, 3-> 1.48, 4->1.6 … 10 → 2, 1000 → 4, κλπ*

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

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

  - Ο βαθμός είναι 0 όταν κανένας από τους όρους του ερωτήματος δεν εμφανίζεται στο έγγραφο

Έγγραφα		Ερωτήματα	
d_1	a ….. b …	q1	a
d_2	a … a …	q2	b
d_3	a … a … b	q3	a b
d_4	b ….. b … b		
d_5	a … a … b … b		
d_6	a ….		

log_10(2) = 0.3
log_10(3) = 0.48

Διαβάθμιση με log tf

Βάρος idf

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

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

Παράδειγμα idf, έστω *N* = 1 εκατομμύριο

term	df_t	idf_t
calpurnia	1	6
animal	100	4
sunday	1,000	3
fly	10,000	2
under	100,000	1
the	1,000,000	0

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

- Κάθε όρος στη συλλογή έχει μια τιμή idf
- Ολική μέτρηση (επίσης, αλλάζει συνεχώς)

Στάθμιση tf-idf

Διάταξη εγγράφων με βάση:

$$score(d, q) = \sum_{t \in d \cap q} w_{t,d}$$

Το **tf-idf βάρος** ενός όρου είναι το γινόμενο του βάρους $tf_{t,d}$ και του βάρους idf_t.

- $w_{t,d} = \left(1 + \ log_{10} tf_{t,d}\right) log_{10}(idf_t)$

- Το *πιο γνωστό σχήμα διαβάθμισης* στην ανάκτηση πληροφορίας
  - Εναλλακτικά ονόματα: tf.idf, tf x idf
- Αυξάνει με τον αριθμό εμφανίσεων του όρου στο έγγραφο
- Αυξάνει με τη σπανιότητα του όρου

Έγγραφα	
d_1	a ..... b ...
d_2	a ... a ...
d_3	a ... a ... b
d_4	b ..... b ... b
d_5	a ... a ... b ... b
d_6	a ....

Ερωτήματα	
$\boxed{\textbf{q1} \quad \textbf{a}}$	
q2	b
q3	a b

$$w_{t,d} = (1 + log_{10}tf_{t,d})\, log_{10}(idf_t)$$

d_2
d_3
d_5
d_1
d_6

$$idf_a = \frac{6}{5} \qquad idf_b = \frac{6}{4}$$

$$\log_{10}(idf_a) = 0.08 \qquad \log_{10}(idf_b) = 0.18$$

$$Score(d_1, q_1) = w_{a,d_1} = (1 + \log_{10} tf_{a,d_1}) \cdot \log_{10}(idf_a)$$

$$= (1 + \log_{10} 1) \cdot \log_{10}(idf_a) = 0.08 \; (2)$$

$$score(d_2, q_1) = (\underbrace{1 + \log 2}) \cdot 0.08 = 1.3 \cdot 0.08 = 0.104 \; (1)$$

$$score(d_3, q_1) = 0.104 \; (1)$$

$$Score(d_4, q_1) = 0$$

$$Score(d_5, q_1) = 0.104 \; (1)$$

$$score(d_6, q_1) = 0.08 \; (2)$$

Έγγραφα		Ερωτήματα	
d_1	a ..... b ...	q1	a
d_2	a ... a ...	**q2**	**b**
d_3	a ... a ... b	q3	a b
d_4	b ..... b ... b		
d_5	a ... a ... b ... b		
d_6	a		

$$w_{t,d} = (1 + \log_{10} tf_{t,d}) \log_{10}(idf_t)$$

d4
d5
d1
d3

$$score(d, q) = \sum_{t \in d \cap q} w_{t,d}$$

$$score(d_1, q_2) = (1 + \log_{10} \overbrace{tf_{b,d_1}}^{1}) \cdot \log_{10}(idf_b)$$

$$= 0.18 \quad 3$$

$$score(d_2, q_2) = 0$$

$$score(d_3, q_2) = 0.18 \quad 3$$

$$score(d_4, q_2) = (1 + \log_{10} \overset{3}{tf_{b,d_4}}) \cdot \log_{10}(idf_b) \quad 1$$

$$= 0.27$$

$$score(d_5, q_2) = 0.234 \quad 2$$

$$score(d_6, q_2) = 0$$

Έγγραφα		Ερωτήματα	
d_1	a ..... b ...	q1	a
d_2	a ... a ...	q2	b
d_3	a ... a ... b	**q3**	**a b**
d_4	b ..... b ... b		
d_5	a ... a ... b ... b		
d_6	a		

$$w_{t,d} = (1 + log_{10} tf_{t,d}) \, log_{10}(idf_t)$$

$$Score(q,d) = \sum_{t \in q \cap d} w_{t,d}$$

d5
d3
d4
d1
d2
d6

$Score(d_1, q_3) = w_{a,d_1} + w_{b,d_1} =$
$(1 + log_{10} tf_{a,d_1}) \cdot log_{10}(idf_a) + (1 + log_{10} tf_{b,d_1}) \cdot log$
$= (1 + log_{10} 1) \cdot 0.08 + (1 + log_{10} 1) \cdot 0.18 = 0.26$ (4)

$score(d_2, q_3) = w_{a,d_2} = (1 + log_{10} tf_{a,d_2}) \cdot log_{10}(id$
$= (1 + log_{10} 2) \cdot 0.08 = 0.104$ (5)

$score(d_3, q_3) = w_{a,d_3} + w_{b,d_3} = 0.104 + 0.18 = 0.284$ (2)

$score(d_4, q_3) = w_{b,d_4} = 0.27$ (3)

$score(d_5, q_3) = w_{a,d_5} + w_{b,d_5} = 0.338$ (1)

$score(d_6, q_3) = 0.08$ (6)

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)	$\frac{1 + \log(\mathrm{tf}_{t,d})}{1 + \log(\mathrm{ave}_{t \in d}(\mathrm{tf}_{t,d}))}$				

Γιατί δεν έχει σημασία η βάση του λογαρίθμου;

Κανονικοποίηση με μέγιστη συχνότητα όρου

Έστω *τ* ο ποιο συχνός όρος σε *ένα* έγγραφο *d* και *tfmax(d)* η συχνότητα του

Διαιρούμε τη συχνότητα $tf_{t,d}$ κάθε όρου *t* στο *d* με αυτήν την τιμή

Γιατί;
Στα μεγάλα έγγραφα μεγάλες συχνότητες όρων απλώς γιατί υπάρχει επανάληψη

Προβλήματα:

- Ασταθής (πχ τροποποίηση stopwords)
- Ιδιαίτερη λέξη (outlier) που εμφανίζεται συχνά
- Πρέπει να υπάρχει διαφορά ανάμεσα σε έγγραφα με ομοιόμορφη και skewed κατανομή

Okapi BM25

$$score(d,q) = \sum_{t \in d \cap q} fdf(t) \frac{tf_{t,d}\,(k_1 + 1)}{tf_{t,d} + k_1(1 - b + b\frac{|d|}{avgd1}}$$

$$fdf(t) = \ln(\frac{N - df(t) + 0.5}{df(t) + 0.5} + 1)$$

$|d|$ document length in words
$avgd1$ average document length
$k_1$parameter, if no learning $k_1 \in [1.2, 2]$
b parameter, if no learning $b = 0.75$
N number of documents in the collection

Based on the probabilistic retrieval model

Bag of words model

- Η tf-idf διαβάθμιση *δεν εξετάζει τη διάταξη των λέξεων* σε ένα έγγραφο
  - *John is quicker than Mary* και
  - *Mary is quicker than John*

- Αυτό λέγεται μοντέλο σάκου λέξεων (<u>bag of words</u> model) – έχει σημασία ο αριθμός των εμφανίσεων αλλά όχι που εμφανίζονται στο έγγραφο

- Θα *εισάγουμε και πληροφορία θέσης αργότερα (θυμηθείτε τα positional index)*

ΜΥΕ003: Ανάκτηση Πληροφορίας

Διδάσκουσα: Ευαγγελία Πιτουρά

Το μοντέλο διανυσματικού χώρου

Ακαδημαϊκό Έτος 2022-2023

Δυαδική μήτρα σύμπτωσης (binary term-document incidence matrix)

	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 ∈ $\{0,1\}^{|V|}$(την αντίστοιχη στήλη)

Τα έγγραφα ως διανύσματα (vector space model)

- Έχουμε ένα |V|-διάστατο διανυσματικό χώρο
  - Οι όροι είναι οι άξονες αυτού του χώρου
  - Τα έγγραφα 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

Ο πίνακας με μετρητές

- Θεωρούμε τον tf, αριθμό (πλήθος) των εμφανίσεων ενός όρου σε ένα έγγραφο:
  - Κάθε έγγραφο είναι ένα $\boxed{\text{διάνυσμα μετρητών στο } \mathbb{N}^{|v|}}$: μια στήλη παρακάτω

	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

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 $R^{|V|}$

Παράδειγμα

Έστω μια συλλογή που περιέχει τα ακόλουθα έγγραφα:

d1: a b c

d2: a a d b

d3: a c d e c a f

d4: b e a b b

d5: a a b d c

Θα μπορούσαμε να θεωρήσουμε και το 1 + $log_{10}(tf_{t,d})$

d1: a b c
d2: a a d b
d3: a c d e c a f
d4: b e a b b
d5: a a b d c

	d1	**d2**	**d3**	**d4**	**d5**
a	1	1.3	1.3	1	1.3
b	1	1	0	1.48	1
c	1	0	1.3	0	1
d	0	1	1	0	1
e	0	0	1	1	0
f	0	0	1	0	0

Θα μπορούσαμε να χρησιμοποιήσουμε και το $w_{t,d}$

Οι τιμές στη γραμμή t πολλαπλασιάζονται με $log_{10}(idf_t)$

Για παράδειγμα, $log_{10}(idf_b)$

	d1	d2	d3	d4	d5
a	1	1.3	1.3	1	1.3
b	1	1	0	1.48	1
c	1	0	1.3	0	1
d	0	1	1	0	1
e	0	0	1	1	0
f	0	0	1	0	0

$log_{10}(idf_a) = 0$ $log_{10}(idf_b) = 0.097$ $log_{10}(idf_c) = 0.22$ $log_{10}(idf_d) = 0.22$
$log_{10}(idf_e) = 0.398$ $log_{10}\left(idf_f\right) = 0.699$

Πίνακας σύμπτωσης

	d1	d2	d3	d4	d5
a	1	1.3	1.3	1	1.3
b	1	1	0	1.48	1
c	1	0	1.3	0	1
d	0	1	1	0	1
e	0	0	1	1	0
f	0	0	1	0	0

	d1	d2	d3	d4	d5
a	0	0	0	0	0
b	0.097	0.097	0	0.143	0.097
c	0.22	0	0.286	0	0.22
d	0	0.22	0.22	0	0.22
e	0	0	0.398	0.398	0
f	0	0	0.699	0	0

- Θεωρήσαμε και την σπανιότητα των όρων στην αναπαράσταση των εγγράφων
- Διαισθητικά είναι πιο σημαντικό να έχουν δυο έγγραφα τον ίδιο σπάνιο όρο από το να έχουν τον ίδιο συχνό όρο

Αποθήκευση

- Που υπάρχει αυτή η πληροφορία στο σύστημα ανάκτησης πληροφορίας;

Ομοιότητα διανυσμάτων

Πρώτη προσέγγιση: απόσταση μεταξύ δυο διανυσμάτων

- Ευκλείδεια απόσταση;
  - Δεν είναι καλή ιδέα – είναι μεγάλη για διανύσματα διαφορετικού μήκους

- Έστω ένα έγγραφο *d*. Υποθέστε ότι κάνουμε append το *d* στον εαυτό του και έστω *d′* το κείμενο που προκύπτει.

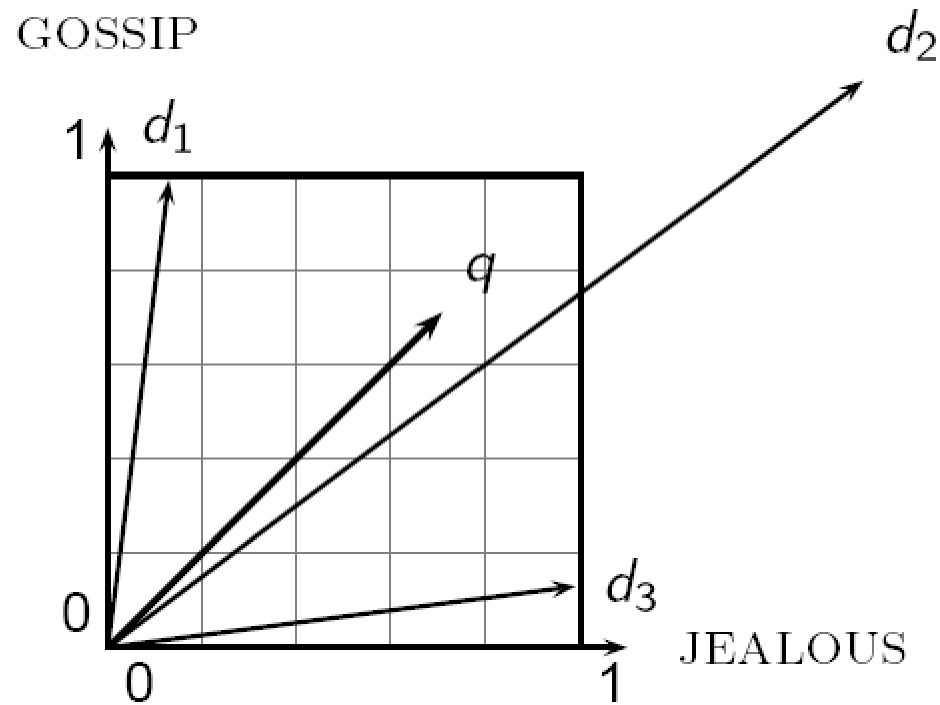
- "Σημασιολογικά" το *d* και το *d′* έχουν το ίδιο περιεχόμενο

- Η Ευκλείδεια απόσταση μεταξύ τους μπορεί να είναι πολύ μεγάλη

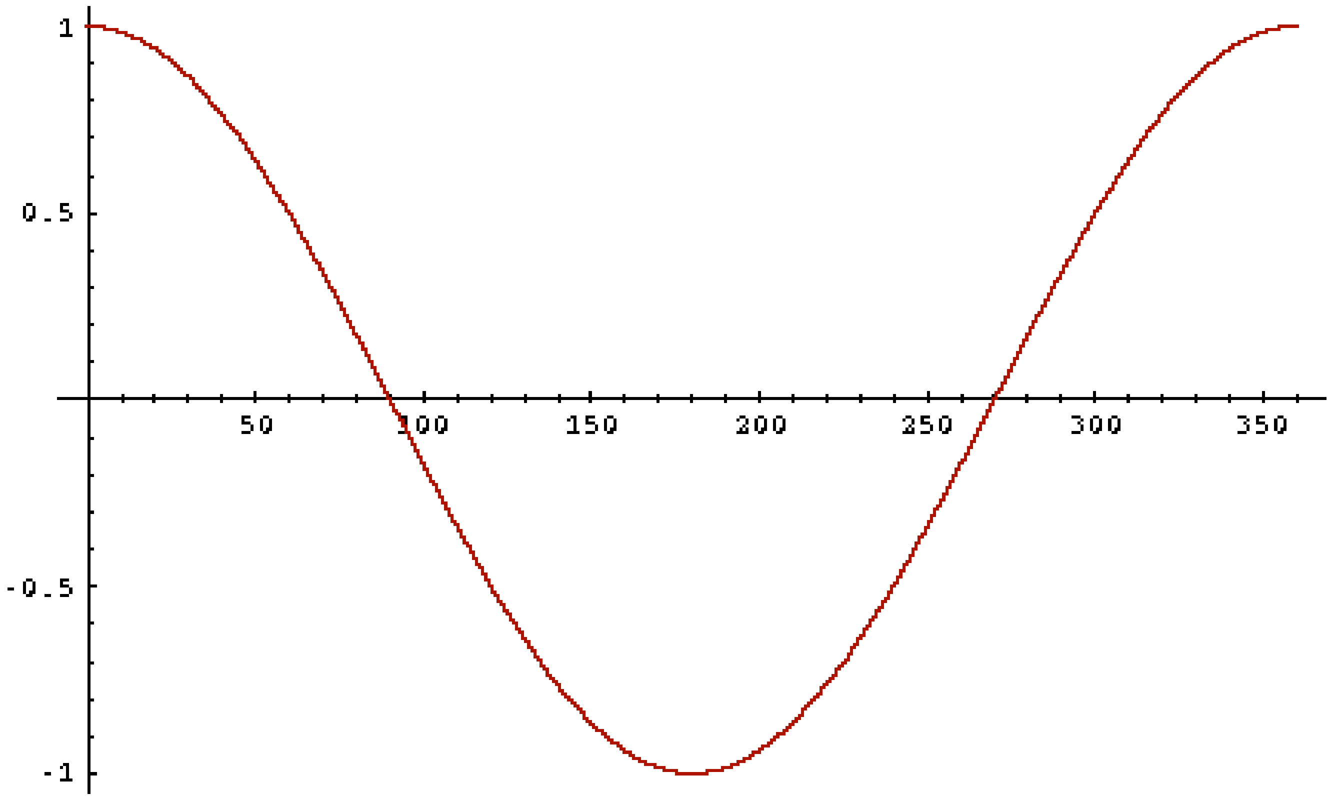
- Η γωνία όμως είναι 0 (αντιστοιχεί στη μεγαλύτερη ομοιότητα) => χρήση της γωνίας

Χρήση της γωνίας αντί της απόστασης

Η Ευκλείδεια απόσταση μεταξύ του $\vec{d_1}$ και του $\vec{d_2}$ είναι μεγάλη αν και η κατανομή των όρων είναι παρόμοια



Από γωνίες σε συνημίτονα



Συνημίτονο μονότονα φθίνουσα συνάρτηση στο διάστημα [0°, 180°]

Ομοιότητα εγγράφων

Εσωτερικό γινόμενο

Μοναδιαία διανύσματα

$$sim(\vec{d}',\vec{d}) = \cos(\vec{d}',\vec{d}) = \frac{\vec{d}'\bullet\vec{d}}{|\vec{d}'||\vec{d}|} = \frac{\vec{d}'}{|\vec{d}'|}\bullet\frac{\vec{d}}{|\vec{d}|} = \frac{\sum_{i=1}^{|V|} d'_i\, d_i}{\sqrt{\sum_{i=1}^{|V|} d'^2_i}\sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

d_i (d'_i) είναι το tf-idf βάρος του *i*-οστού όρου στο έγγραφο *d* *(d')*

$\cos(\vec{d}', \vec{d})$ η ομοιότητα συνημιτόνου μεταξύ $\vec{d}'$ and $\vec{d}$ ή,
Ισοδύναμα, το συνημίτονο της γωνίας μεταξύ των $\vec{d}'$ και $\vec{d}$.

Κανονικοποίηση του μήκους

- Ένα διάνυσμα μπορεί να κανονικοποιηθεί διαιρώντας τα στοιχεία του με το μήκος του, με χρήση της L_2 νόρμας:

$$\left\|\vec{x}\right\|_2 = \sqrt{\sum_i x_i^2}$$

- Διαιρώντας ένα διάνυσμα με την L_2 νόρμα το κάνει μοναδιαίο
  - *Ως αποτέλεσμα, μικρά και μεγάλα έγγραφα έχουν συγκρίσιμα βάρη*

- Για διανύσματα για τα οποία έχουμε κανονικοποιήσει το μήκος τους (length-normalized vectors) το συνημίτονο είναι απλώς το εσωτερικό γινόμενο (dot or scalar product):

$$\cos(\vec{d}\,', \vec{d}\,) = \vec{d}\,' \bullet \vec{d} = \sum_{i=1}^{|V|} d'_i \, d_i$$

Παράδειγμα Ι

Ποιο έγγραφο είναι το ποιο όμοιο με το d1;

$$\overrightarrow{d1} = (0, 0.097, 0.22, 0, 0, 0)$$
$$\overrightarrow{d2} = (0, 0.097, 0, 0.22, 0, 0)$$
$$\overrightarrow{d3} = (0, 0, 0.286.\ 0.22, 0.398, 0.699)$$
$$\overrightarrow{d4} = (0, 0.143, 0, 0, 0.398, 0)$$
$$\overrightarrow{d5} = (0, 0.097, 0.22, 0.22, 0, 0)$$

	d1	d2	d3	d4	d5
a	0	0	0	0	0
b	0.097	0.097	0	0.143	0.097
c	0.22	0	0.286	0	0.22
d	0	0.22	0.22	0	0.22
e	0	0	0.398	0.398	0
f	0	0	0.699	0	0

$$sim(d1, d2) = \cos\left(\overrightarrow{d1}, \overrightarrow{d2}\right) = \frac{\overrightarrow{d1}\ \overrightarrow{d2}}{\left|\overrightarrow{d1}\right|\left|\overrightarrow{d2}\right|} = \frac{0.009}{\sqrt{0.058}\sqrt{0.058}} = 0.156$$

$$sim(d1, d3) = \cos\left(\overrightarrow{d1}, \overrightarrow{d3}\right) = \frac{\overrightarrow{d1}\ \overrightarrow{d3}}{\left|\overrightarrow{d1}\right|\left|\overrightarrow{d3}\right|} = \frac{0.062}{\sqrt{0.058}\sqrt{0.697}} = 0.308$$

$$sim(d1, d4) = \cos\left(\overrightarrow{d1}, \overrightarrow{d4}\right) = \frac{\overrightarrow{d1}\ \overrightarrow{d4}}{\left|\overrightarrow{d1}\right|\left|\overrightarrow{d4}\right|} = \frac{0.014}{\sqrt{0.058}\sqrt{0.179}} = 0.137$$

$$sim(d1, d5) = \cos\left(\overrightarrow{d1}, \overrightarrow{d5}\right) = \frac{\overrightarrow{d1}\ \overrightarrow{d5}}{\left|\overrightarrow{d1}\right|\left|\overrightarrow{d2}\right|} = \frac{0.058}{\sqrt{0.058}\sqrt{0.498}} = 0.341$$

d5
d3
d2
d4

Παράδειγμα ΙΙ

Ποια είναι οι ομοιότητα μεταξύ των έργων

SaS: *Sense and Sensibility*

PaP: *Pride and Prejudice*, and

WH: *Wuthering Heights*?

Συχνότητα όρων (μετρητές)

όρος	SaS	PaP	WH
affection	115	58	20
jealous	10	7	11
gossip	2	0	6
wuthering	0	0	38

Παράδειγμα ΙΙ (συνέχεια)

Για απλοποίηση σε αυτό το παράδειγμα,
δε θα χρησιμοποιήσουμε τα idf βάρη

όρος	SaS	PaP	WH
affection	115	58	20
jealous	10	7	11
gossip	2	0	6
wuthering	0	0	38

Log frequency βάρος (log tf)

όρος	SaS	PaP	WH
affection	3.06	2.76	2.30
jealous	2.00	1.85	2.04
gossip	1.30	0	1.78
wuthering	0	0	2.58

Μήκος

SaS $= \sqrt{3.06^2 + 2.00^2 + 1.3^2 + 0^2} \approx 3.88$

Μετά την κανονικοποίηση

όρος	SaS	PaP	WH
affection	0.789	0.832	0.524
jealous	0.515	0.555	0.465
gossip	0.335	0	0.405
wuthering	0	0	0.588

Παράδειγμα ΙΙ (συνέχεια)

όρος	SaS	PaP	WH
affection	0.789	0.832	0.524
jealous	0.515	0.555	0.465
gossip	0.335	0	0.405
wuthering	0	0	0.588

όρος	SaS	PaP	WH
affection	115	58	20
jealous	10	7	11
gossip	2	0	6
wuthering	0	0	38

cos(SaS, PaP) ≈

$0.789 \times 0.832 + 0.515 \times 0.555 + 0.335 \times 0.0 + 0.0 \times 0.0$

≈ 0.94

cos(SaS,WH) ≈ 0.79

cos(PaP,WH) ≈ 0.69

Γιατί cos(SaS,PaP) > cos(SaS,WH)?

Τα ερωτήματα ως διανύσματα

Εφαρμόζουμε το ίδιο και για τα ερωτήματα, δηλαδή, *αναπαριστούμε και τα ερωτήματα ως διανύσματα στον ίδιο χώρο*

Παράδειγμα: q: b c

Διαβάθμιση των εγγράφων με βάση το πόσο κοντά είναι στην ερώτηση σε αυτό το χώρο

- Κοντινά = ομοιότητα διανυσμάτων
- Ομοιότητα ≈ αντίθετο της απόστασης

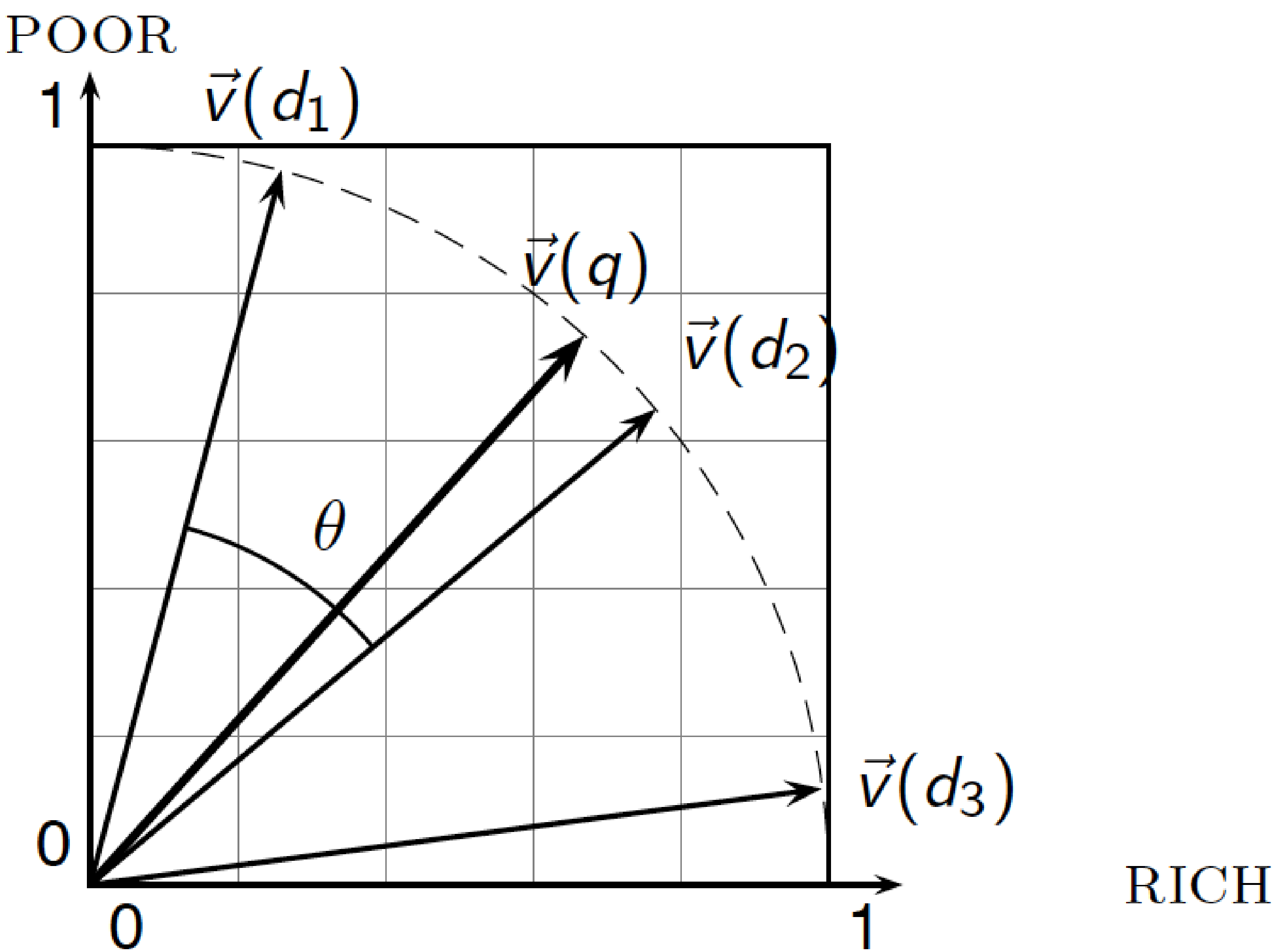
cosine(query, document)

Ως *score(q, d)* θα χρησιμοποιήσουμε το *συνημίτονο* της γωνίας της διανυσματικής αναπαράστασης του q και d

$$\cos(\vec{q},\vec{d}) = \frac{\vec{q}\bullet\vec{d}}{|\vec{q}||\vec{d}|} = \frac{\vec{q}}{|\vec{q}|}\bullet\frac{\vec{d}}{|\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2}\sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

$\vec{q_i}$ είναι το tf-idf βάρος του όρου *i* στην ερώτηση
$\vec{d_i}$ είναι το tf-idf βάρος του όρου *i* στο έγγραφο

Ομοιότητα συνημίτονου



Περίληψη βαθμολόγησης 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

Παράδειγμα

Έστω μια συλλογή που περιέχει τα ακόλουθα έγγραφα:

d1: a b c

d2: a a d b

d3: a c d e c a f

d4: b e a b b

d5: a a b d c

Απάντηση ερώτησης με χρήση διανυσματικής αναπαράστασης

q1: b

q2: b c

q3: b f

d1: a b c
d2: a a d b
d3: a c d e c a f
d4: b e a b b
d5: a a b d c

	d1	**d2**	**d3**	**d4**	**d5**
a	1	1.3	1.3	1	1.3
b	1	1	0	1.48	1
c	1	0	1.3	0	1
d	0	1	1	0	1
e	0	0	1	1	0
f	0	0	1	0	0

Θα δούμε ένα παράδειγμα *χωρίς idf* και για ερώτηση και για έγγραφο

- Κανονικοποίηση διανυσμάτων

Διαίρεση με το μήκος τους

	d1	**d2**	**d3**	**d4**	**d5**
a	0.57	0.68	0.51	0.49	0.6
b	0.57	0.52	0	0.72	0.46
c	0.57	0	0.51	0	0.46
d	0	0.52	0.39	0	0.46
e	0	0	0.39	0.49	0
f	0	0	0.39	0	0

d1: a b c
d2: a a d b
d3: a c d e c a f
d4: b e a b b
d5: a a b d c

Θα δούμε ένα παράδειγμα *χωρίς idf* και για ερώτηση και για έγγραφο

	d1	d2	d3	d4	d5
a	0.57	0.68	0.51	0.49	0.6
b	0.57	0.52	0	0.72	0.46
c	0.57	0	0.51	0	0.46
d	0	0.52	0.39	0	0.46
e	0	0	0.39	0.49	0
f	0	0	0.39	0	0

q1: b

d4
d1
d2
d3

Αποτέλεσμα

d1: a b c
d2: a a d b
d3: a c d e c a f
d4: b e a b b
d5: a a b d c

q1: b c

q_2 b c

a	0
b	1
c	1
d	0
e	0
f	0

$\|q_2\|_2 = \sqrt{2}$

0
0.71
0.71
0
0
0

$score(d_1, q_2) = 0.82$ d_2 0.37 d_3 0.36 d_4 0.51

d_5 0.68

	d1	**d2**	**d3**	**d4**	**d5**
a	0.57	0.68	0.51	0.49	0.6
b	0.57	0.52	0	0.72	0.46
c	0.57	0	0.51	0	0.46
d	0	0.52	0.39	0	0.46
e	0	0	0.39	0.49	0
f	0	0	0.39	0	0

Στάθμιση ερωτημάτων και εγγράφων

- Πολλές μηχανές αναζήτησης σταθμίζουν διαφορετικά τις ερωτήσεις από τα έγγραφα
- Συμβολισμό: *ddd.qqq*, με χρήση των ακρωνύμων του πίνακα, όπου τα πρώτα 3 γράμματα (*d*) αφορούν το έγγραφο και τα επόμενα 3 γράμματα (*q*) αφορούν το ερώτημα
  - συχνότητα όρου.συχνότητα εγγράφων.κανονικοποίηση
- Συχνό σχήμα : lnc.ltc
  - Έγγραφο: logarithmic tf (l), no idf (n), cosine normalization (c)
  - Ερώτημα: logarithmic tf (l), idf (t), cosine normalization (c)

Παράδειγμα

Ερώτημα: *best car insurance* N = 1000K

Έγγραφο: *car insurance auto insurance* Inc.ltc

Όρος	Ερώτημα (Query)						Έγγραφο				Prod
	tf-raw	tf-wt	df	idf	wt	n'lize	tf-raw	tf-wt	wt	n'lize	
auto	0	0	5000	2.3	0	0	1	1	1	0.52	0
best	1	1	50000	1.3	1.3	0.34	0	0	0	0	0
car	1	1	10000	2.0	2.0	0.52	1	1	1	0.52	0.27
insurance	1	1	1000	3.0	3.0	0.78	2	1.3	1.3	0.68	0.53

Μήκος Ερωτήματος = $\sqrt{0^2 + 1.3^2 + 2.0^2 + 3^2} \approx 3.8$

Μήκος Εγγράφου = $\sqrt{1^2 + 0^2 + 1^2 + 1.3^2} \approx 1.92$

Score = 0+0+(0.52*0.52=)27+(0.78*0.68=)0.53 = 0.8

Συχνότητα συλλογής και εγγράφου

- Η *συχνότητα συλλογής* ενός όρου *t* είναι ο αριθμός των εμφανίσεων του *t* στη συλλογή, *μετρώντας* ***και τις πολλαπλές*** **εμφανίσεις**

Παράδειγμα:

Word	Collection frequency	Document frequency
insurance	10440	3997
try	10422	8760

- Ποια λέξη είναι καλύτερος όρος αναζήτησης (και πρέπει να έχει μεγαλύτερο βάρος)?

Ερωτήσεις;