

ΜΥΕ003: Ανάκτηση Πληροφορίας

Διδάσκουσα: Ευαγγελία Πιτουρά

Βαθμολόγηση. Στάθμιση όρων. Πιθανοτικό
Μοντέλο. Διανυσματικό Μοντέλο.

Ακαδημαϊκό Έτος 2024-2025

Περιεχόμενα

- Βαθμολόγηση και διάταξη εγγράφων
- Στάθμιση όρων (term weighting)
- Πιθανοτικό Μοντέλο
- 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

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

q_1	a
q_2	b
q_3	a b

Boolean Μοντέλο

- Τα Boolean ερωτήματα συχνά έχουν είτε *πολύ λίγα* 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 θα διατάξουμε-διαβαθμίσουμε τα έγγραφα μιας συλλογής με βάση ένα ερώτημα;
 - Αναθέτουμε ένα **βαθμό** (score) – ας πούμε στο $[0, 1]$ – σε κάθε έγγραφο
 - $\text{score}(d, q)$: 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

Βαθμός εγγράφου και ερώτησης

Μέτρο βαθμολογίας επικάλυψης (overlap score measure)

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


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 φορές 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
 - $df_t \leq N$
- Ορίζουμε την *αντίστροφη συχνότητα εγγράφων idf* (inverse document frequency) του t ως

$$idf_t = N/df_t$$

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

A μικρό idf

B 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

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

Διαβάθμιση με tf-idf

Ερωτήματα

q1	a
q2	b
q3	a b

νόρση

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

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

Διαβάθμιση με tf-idf

Ερωτήματα

q_1	a
q_2	b
q_3	a b

νόθεση

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

$$idf_a = \frac{6}{5}$$

$$idf_b = \frac{6}{4}$$

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

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

$$score(d_2, q_1) = 2 * \frac{6}{5}$$

$$score(d_3, q_1) = 2 * \frac{6}{5}$$

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

$$score(d_5, q_1) = 2 * \frac{6}{5}$$

$$score(d_6, q_1) = \frac{6}{5}$$

 $d_2 \quad d_5$ $d_1 \quad d_3 \quad d_6$

Το idf δεικνύει
ερωτήματα που έχουν τόσο
ελάχιστο

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

q1	a
q2	b
q3	a b

$$idf_a = 6/5$$

$$idf_b = 6/4$$

Διαβάθμιση με tf-idf

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

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

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

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

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

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

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

d_5
 d_4
 d_3
 d_1
 d_2
 d_6

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** $tf_{t,d}$ 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

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

• $0 \rightarrow 0, 1 \rightarrow 1, 2 \rightarrow 1.3, 3 \rightarrow 1.48, 4 \rightarrow 1.6 \dots 10 \rightarrow 2, 1000 \rightarrow 4, \text{ κλπ}$

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

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

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

Βάρος idf

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

$$idf_t = \log_{10} (N/df_t)$$

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

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

- 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 .

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

q1	a
q2	b
q3	a b

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

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

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

$$\begin{aligned} \text{Score}(d_1, q_1) &= w_{a,d_1} = (1 + \log_{10} tf_{a,d_1}) \cdot \log_{10}(idf_a) \\ &= (1 + \log_{10} 1) \cdot \log_{10}(idf_a) = 0.08 \quad (2) \end{aligned}$$

$$\text{score}(d_2, q_1) = (1 + \log_{10} 2) \cdot 0.08 = 1.3 \cdot 0.08 = 0.104 \quad (1)$$

$$\text{score}(d_3, q_1) = 0.104 \quad (1)$$

$$\text{score}(d_4, q_1) = 0$$

$$\text{score}(d_5, q_1) = 0.104 \quad (1)$$

$$\text{score}(d_6, q_1) = 0.08 \quad (2)$$

d_2
 d_3
 d_5
 d_1
 d_6

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

q1	a
q2	b
q3	a b

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

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

$$score(d_1, q_2) = (1 + \log_{10} \overbrace{tf_{b,d_1}}^1) \cdot \log_{10}(idf_b) = 0.18 \text{ (3)}$$

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

$$score(d_3, q_2) = 0.18 \text{ (3)}$$

$$score(d_4, q_2) = (1 + \log_{10} \underbrace{tf_{b,d_4}}_3) \cdot \log_{10}(idf_b) = 0.27 \text{ (1)}$$

$$score(d_5, q_2) = 0.234 \text{ (2)}$$

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

d_4
 d_5
 d_1
 d_3

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

q1	a
q2	b
q3	a b

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

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

d_5
 d_3
 d_4
 d_1
 d_2
 d_6

$$\begin{aligned} score(d_1, q_3) &= w_{a,d_1} + w_{b,d_1} = \\ &= (1 + \log_{10} tf_{a,d_1}) \cdot \log_{10}(idf_a) + (1 + \log_{10} tf_{b,d_1}) \cdot \log_{10}(idf_b) \\ &= (1 + \log_{10} 1) \cdot 0.08 + (1 + \log_{10} 1) \cdot 0.18 = 0.26 \text{ (4)} \\ score(d_2, q_3) &= w_{a,d_2} = (1 + \log_{10} tf_{a,d_2}) \cdot \log_{10}(idf_a) \\ &= (1 + \log_{10} 2) \cdot 0.08 = 0.104 \text{ (5)} \\ score(d_3, q_3) &= w_{a,d_3} + w_{b,d_3} = 0.104 + 0.18 = 0.284 \text{ (2)} \\ score(d_4, q_3) &= w_{b,d_4} = 0.27 \text{ (3)} \\ score(d_5, q_3) &= w_{a,d_5} + w_{b,d_5} = 0.338 \text{ (1)} \\ score(d_6, q_3) &= 0.08 \text{ (6)} \end{aligned}$$

Παραλλαγές της tf-idf στάθμισης

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

Γιατί δεν έχει σημασία η βάση του λογαρίθμου;

Κανονικοποίηση με μέγιστη συχνότητα όρου

Έστω t ο πιο 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

Περιεχόμενα

- Βαθμολόγηση και διάταξη εγγράφων
- Στάθμιση όρων (term weighting)
- Πιθανοτικό Μοντέλο
- Αναπαράσταση εγγράφων και ερωτημάτων ως διανύσματα

ΜΥΕ003: Ανάκτηση Πληροφορίας

Διδάσκουσα: Ευαγγελία Πιτουρά

Κεφάλαιο 11: Πιθανοτική ανάκτηση πληροφορίας.

Αρχή Πιθανοτικής Κατάταξης

Διάταξη των εγγράφων (d) με βάση **την πιθανότητα να είναι συναφή** με το ερώτημα (q)

Probability Ranking Principle (RPR):

decreasing probability of relevance => best overall effectiveness

«Αν η απόκριση ενός συστήματος ανάκτησης πληροφορίας σε κάθε ερώτημα είναι η *κατάταξη των εγγράφων κατά φθίνουσα πιθανότητα συνάφειας* προς τον χρήστη που υπέβαλε το ερώτημα, όπου οι πιθανότητες εκτιμώνται με τον ακριβέστερο δυνατό τρόπο βάσει των όποιων δεδομένων είναι διαθέσιμα στο σύστημα για τον σκοπό αυτό, η συλλογική *αποτελεσματικότητα είναι η βέλτιστη εφικτή* με βάση αυτά τα δεδομένα» [van Rijsbergen 1979]

Πιθανοτική Ανάκτηση Πληροφορίας

- Βασική ιδέα: Διάταξη εγγράφων με βάση την *πιθανότητα* να είναι συναφή με το ερώτημα
- Τυχαία μεταβλητή R
- Διάταξη με βάση το
$$P(R=1 | q, d)$$

Δυαδικό Μοντέλο Ανεξαρτησίας (Binary Independence Model (BIP))

Θα αναπαραστήσουμε κάθε έγγραφο d (και αντίστοιχα, το ερώτημα) ως ένα δυαδικό (Boolean) διάνυσμα (M : #όρων)

$$\vec{X} = (x_1, \dots, x_M)$$

$$x_i = 1 \quad \text{Αν το έγγραφο περιέχει τον όρο } i$$

Υπόθεση ανεξαρτησίας όρων (independence assumption):

Οι όροι εμφανίζονται στα έγγραφα ανεξάρτητα ο ένας από τον άλλο

Επίσης, η συνάφεια ενός εγγράφου είναι ανεξάρτητη από τη συνάφεια των άλλων εγγράφων (**ανεξαρτησία εγγράφων**)

Δυαδικό Μοντέλο Ανεξαρτησίας

Επειδή μας ενδιαφέρει *η διάταξη* θα χρησιμοποιήσουμε τη σχετική πιθανότητα - **odds**

$$O(R | q, \vec{x}) = \frac{p(R = 1 | q, \vec{x})}{p(R = 0 | q, \vec{x})}$$

Από τον κανόνα του Bayes

$$O(R | q, \vec{x}) = \frac{p(R = 1 | q, \vec{x})}{p(R = 0 | q, \vec{x})} = \frac{\frac{p(R = 1 | q) p(\vec{x} | R = 1, q)}{p(\vec{x} | q)}}{\frac{p(R = 0 | q) p(\vec{x} | R = 0, q)}{p(\vec{x} | q)}}$$

Δυαδικό Μοντέλο Ανεξαρτησίας

$$O(R | q, \vec{x}) = \frac{p(R = 1 | q, \vec{x})}{p(R = 0 | q, \vec{x})} = \frac{p(R = 1 | q)}{p(R = 0 | q)} \frac{p(\vec{x} | R = 1, q)}{p(\vec{x} | R = 0, q)}$$

Σταθερά για
κάθε ερώτημα

Εκ των προτέρων πιθανότητα (prior probability) να ανακτήσουμε συναφές (μη συναφές) έγγραφο

$$O(R | q)$$

Πρέπει να
εκτιμήσουμε αυτήν
την ποσότητα

Αν έχουμε ανακτήσει ένα συναφές (μη συναφές) έγγραφο, αυτό να έχει αναπαράσταση $\vec{x}$

Δυαδικό Μοντέλο Ανεξαρτησίας

Θα χρησιμοποιήσουμε την *υπόθεση ανεξαρτησίας των όρων*: η παρουσία ή απουσία ενός όρου σε ένα έγγραφο είναι ανεξάρτητη από την παρουσία ή απουσία οποιουδήποτε άλλου όρου – **Naïve Bayes conditional independent assumption**)

$$\frac{p(\vec{x} | R = 1, q)}{p(\vec{x} | R = 0, q)} = \prod_{i=1}^M \frac{p(x_i | R = 1, q)}{p(x_i | R = 0, q)}$$

Άρα:

$$O(R | q, \vec{x}) = O(R | q) \prod_{i=1}^M \frac{p(x_i | R = 1, q)}{p(x_i | R = 0, q)}$$

Δυαδικό Μοντέλο Ανεξαρτησίας

$$O(R | q, \vec{x}) = O(R | q) \prod_{i=1}^n \frac{p(x_i | R=1, q)}{p(x_i | R=0, q)}$$

Αφού τα x_i είναι είτε 1 είτε 0, χωρίζουμε το γινόμενο

$$O(R | q, \vec{x}) = O(R | q) \prod_{x_i=1} \frac{p(x_i=1 | R=1, q)}{p(x_i=1 | R=0, q)} \prod_{x_i=0} \frac{p(x_i=0 | R=1, q)}{p(x_i=0 | R=0, q)}$$

Δυαδικό Μοντέλο Ανεξαρτησίας

Για τον όρο i , ορίζουμε ως p_i την πιθανότητα να εμφανίζεται σε **συναφές** έγγραφο και ως r_i την πιθανότητα να εμφανίζεται σε **μη συναφές** έγγραφο

$$p_i = p(x_i = 1 \mid R = 1, q)$$

$$r_i = p(x_i = 1 \mid R = 0, q)$$

	Έγγραφο	Συναφές (R=1)	Μη συναφές (R=0)
Όρος i εμφανίζεται	$x_i = 1$	p_i	r_i
Όρος i δεν εμφανίζεται	$x_i = 0$	$(1 - p_i)$	$(1 - r_i)$

Δυαδικό Μοντέλο Ανεξαρτησίας

$$O(R | q, \vec{x}) = O(R | q) \prod_{x_i=1} \frac{p(x_i = 1 | R = 1, q)}{p(x_i = 1 | R = 0, q)} \prod_{x_i=0} \frac{p(x_i = 0 | R = 1, q)}{p(x_i = 0 | R = 0, q)}$$

$$O(R | q, \vec{x}) = O(R | q) \prod_{x_i=1} \frac{p_i}{r_i} \prod_{x_i=0} \frac{(1-p_i)}{(1-r_i)}$$

Δυαδικό Μοντέλο Ανεξαρτησίας

Αν υποθέσουμε ότι οι όροι που δεν εμφανίζονται στο ερώτημα είναι το ίδιο πιθανό να εμφανίζονται σε συναφή και σε μη συναφή έγγραφα, δηλαδή για αυτούς τους όρους:

$$p_i = r_i$$

$$O(R \mid q, \vec{x}) = O(R \mid q) \prod_{\substack{x_i=1 \\ q_i=1}} \frac{p_i}{r_i} \prod_{\substack{x_i=0 \\ q_i=1}} \frac{(1-p_i)}{(1-r_i)}$$

Δυαδικό Μοντέλο Ανεξαρτησίας

$$O(R \mid q, \vec{x}) = O(R \mid q) \prod_{x_i=q_i=1} \frac{p_i}{r_i} \prod_{\substack{x_i=0 \\ q_i=1}} \frac{1-p_i}{1-r_i}$$

Οι όροι του ερωτήματος που εμφανίζονται

Οι όροι του ερωτήματος που δεν εμφανίζονται

Μοιράζουμε το μεσαίο όρο (=1) στα αριστερά και δεξιά (είναι οι όροι του ερωτήματος που εμφανίζονται στο έγγραφο)

$$O(R \mid q, \vec{x}) = O(R \mid q) \prod_{\substack{x_i=1 \\ q_i=1}} \frac{p_i}{r_i} \prod_{\substack{x_i=1 \\ q_i=1}} \left(\frac{1-r_i}{1-p_i} \cdot \frac{1-p_i}{1-r_i} \right) \prod_{\substack{x_i=0 \\ q_i=1}} \frac{1-p_i}{1-r_i}$$

$$O(R \mid q, \vec{x}) = O(R \mid q) \prod_{x_i=q_i=1} \frac{p_i(1-r_i)}{r_i(1-p_i)} \prod_{q_i=1} \frac{1-p_i}{1-r_i}$$

Οι όροι του ερωτήματος που εμφανίζονται στο έγγραφο

Όλοι οι όροι

Δυναμικό Μοντέλο Ανεξαρτησίας

$$O(R | q, \vec{x}) = O(R | q) \cdot \prod_{x_i=q_i=1} \frac{p_i(1-r_i)}{r_i(1-p_i)} \cdot \prod_{q_i=1} \frac{1-p_i}{1-r_i}$$

Σταθερά για κάθε ερώτημα

Η μοναδική ποσότητα που πρέπει να εκτιμηθεί

Θα χρησιμοποιήσουμε το $\log$ αυτής της ποσότητας
Retrieval Status Value (RSV) (τιμή κατάστασης ανάκτησης)

$$RSV = \log \prod_{x_i=q_i=1} \frac{p_i(1-r_i)}{r_i(1-p_i)} = \sum_{x_i=q_i=1} \log \frac{p_i(1-r_i)}{r_i(1-p_i)}$$

Τιμή Κατάστασης Ανάκτησης (RSV)

$$RSV = \log \prod_{x_i=q_i=1} \frac{p_i(1-r_i)}{r_i(1-p_i)} = \sum_{x_i=q_i=1} \log \frac{p_i(1-r_i)}{r_i(1-p_i)}$$

$$RSV = \sum_{x_i=q_i=1} c_i$$

$$c_i = \log \frac{p_i(1-r_i)}{r_i(1-p_i)} = \log \frac{p_i}{1-p_i} + \log \frac{1-r_i}{r_i}$$

Τιμή Κατάστασης Ανάκτησης (RSV)

$$RSV = \sum_{x_i=q_i=1} c_i \quad c_i = \log \frac{p_i}{1-p_i} + \log \frac{1-r_i}{r_i}$$

	Έγγραφο	Συναφές (R=1)	Μη συναφές (R=0)
Όρος i εμφανίζεται	$x_i = 1$	p_i	r_i
Όρος i δεν εμφανίζεται	$x_i = 0$	$(1 - p_i)$	$(1 - r_i)$

Τα c_i (log odd ratios) έχουν το ρόλο των βαρών σε αυτό το μοντέλο – αθροιστικά για κάθε όρο της ερώτησης

Για έναν όρο:

- 0 αν equal odds - η πιθανότητα να εμφανίζεται και να μην εμφανίζεται σε συναφές (μη συναφές) έγγραφο είναι $\frac{1}{2}$
- + αν $p_i > 1-p_i$ (εμφανίζεται σε περισσότερα συναφή) και $r_i < 1-r_i$ (εμφανίζεται σε λιγότερα μη συναφή)

Πως θα τα υπολογίσουμε/εκτιμήσουμε;

Θεωρητική Εκτίμηση των RSV συντελεστών

Για κάθε όρο i , κατασκευάζουμε έναν πίνακα μετρητών

Έστω N έγγραφα στη συλλογή, S συναφή, και ο όρος εμφανίζεται σε συνολικά σε n έγγραφα από τα οποία τα συναφή είναι s

Έγγραφα	Συναφή (R=1)	Μη συναφή (R=0)	Συνολικά
$x_i = 1$	s	$n - s$	n
$x_i = 0$	$S - s$	$N - n - (S - s)$	$N - n$
	S	$N - S$	N

Εκτιμήσεις:

$$p_i \approx \frac{s}{S} \quad r_i \approx \frac{(n - s)}{(N - S)}$$

$$c_i \approx K(N, n, S, s) = \log \frac{s / (S - s)}{(n - s) / (N - n - S + s)}$$

Αγνοεί τους
όρους που δεν
εμφανίζονται

Εκτίμηση των RSV συντελεστών (στην πράξη)

Αν τώρα υποθέσουμε ότι τα μη συναφή έγγραφα είναι περίπου όσα όλη η συλλογή, τότε το r_i (η πιθανότητα εμφάνισης του όρου σε μη συναφές έγγραφο) είναι n/N

$$\log \frac{1-r_i}{r_i} = \log \frac{N-n-S+s}{n-s} \approx \log \frac{N-n}{n} \approx \log \frac{N}{n} = IDF!$$

Εκτίμηση των RSV συντελεστών (στην πράξη)

- Το p_i (πιθανότητα εμφάνισης σε συναφή έγγραφα) είναι πιο δύσκολο να εκτιμηθεί
- Πιθανοί τρόποι εκτίμησης του p_i :
 - Από συναφή έγγραφα, αν κάποια από αυτά είναι γνωστά
 - Με μία σταθερά – τότε απλώς χρησιμοποιούμε το idf (with $p_i=0.5$)

$$RSV = \sum_{x_i=q_i=1} \log \frac{N}{n_i}$$

- Ανάλογο των εμφανίσεων του όρου στη συλλογή
έχει προταθεί: $1/3 + 2/3 \text{ df}_i/N$

Πιθανοτική Ανάδραση Συνάφειας (Relevance Feedback)

1. Υποθέτουμε κάποιες αρχικές τιμές για τα p_i και r_i - τις οποίες χρησιμοποιούμε για να ανακτήσουμε ένα αρχικό σύνολο συναφών εγγράφων (εγγράφων με $R=1$)
2. Αλληλοεπιδρούμε με το χρήστη για να βελτιώσουμε αυτές τις τιμές: οι χρήστες χαρακτηρίζουν ένα σύνολο έγγραφων V ως συναφή ($R = 1$) και μη συναφή ($R = 0$)
3. Επανα-υπολογίζουμε τα p_i και r_i
 - Ή τα συνδυάζουμε με τις αρχικές μας εκτιμήσεις χρήση Bayesian prior):

$$p_i^{(2)} = \frac{|V_i| + \kappa p_i^{(1)}}{|V| + \kappa}$$

κ : prior weight

4. Επαναλαμβάνουμε, έως σύγκλιση

PRP και BIM

(+) Λογικές προσεγγίσεις για τις πιθανότητες

(-) Περιοριστικές υποθέσεις:

- Ανεξαρτησία όρων
- Ανεξαρτησία συνάφειας εγγράφων
- Οι όροι που δεν εμφανίζονται στο ερώτημα δεν επηρεάζουν το αποτέλεσμα
- Δυναδική αναπαράσταση (αγνοούμε tf, μήκος εγγράφου, κλπ)

Okapi BM25

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

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

$|d|$ 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 (bag of words model) – έχει σημασία ο αριθμός των εμφανίσεων αλλά όχι που εμφανίζονται στο έγγραφο
- Θα *εισάγουμε και πληροφορία θέσης αργότερα* (θυμηθείτε τα *positional index*)

ΤΕΛΟΣ (τμήματος) Κεφαλαίου 11

Ερωτήσεις?

Χρησιμοποιήθηκε κάποιο υλικό των:

- ✓ *Pandu Nayak and Prabhakar Raghavan, CS276:Information Retrieval and Web Search (Stanford)*
- ✓ *Hinrich Schütze and Christina Lioma, Stuttgart IIR class*

Περιεχόμενα

- Βαθμολόγηση και διάταξη εγγράφων
- Στάθμιση όρων (term weighting)
- Πιθανοτικό Μοντέλο
- Αναπαράσταση εγγράφων και ερωτημάτων ως διανύσματα

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 $\in \{0,1\}^{|V|}$ (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 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 $\mathbb{R}^{|V|}$

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, $\log_{10}(idf_b)$ 

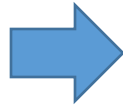
	d1	d2	d3	d4	d5
a	1	1.3	1.3	1	1.3
b	1	1	0	1.48	1
c	1	0	1.3	0	1
d	0	1	1	0	1
e	0	0	1	1	0
f	0	0	1	0	0

$$\log_{10}(idf_a) = 0 \quad \log_{10}(idf_b) = 0.097 \quad \log_{10}(idf_c) = 0.22 \quad \log_{10}(idf_d) = 0.22$$

$$\log_{10}(idf_e) = 0.398 \quad \log_{10}(idf_f) = 0.699$$

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;

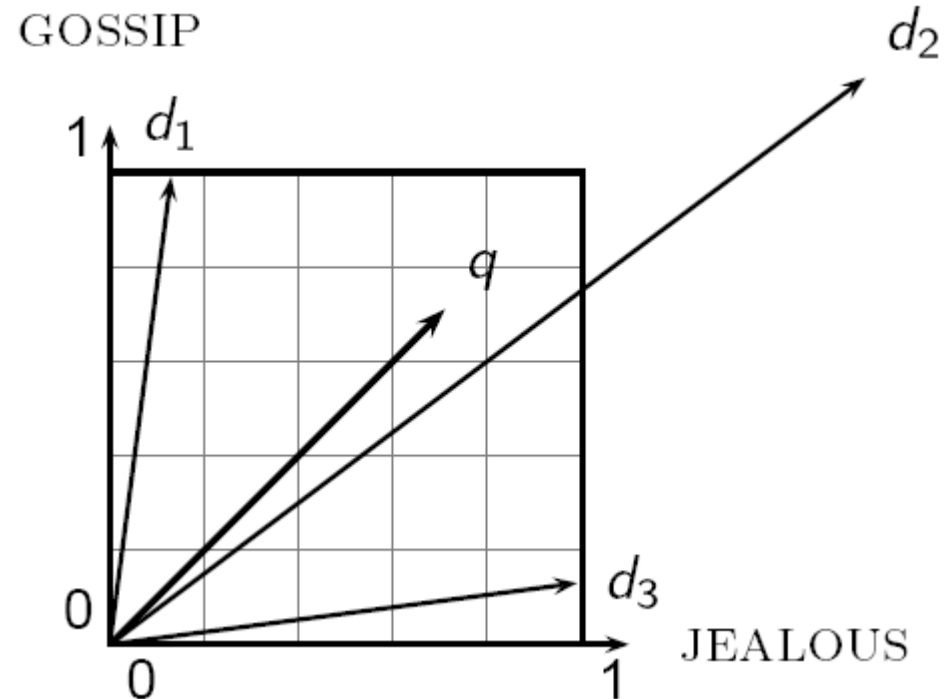
Ομοιότητα διανυσμάτων

Πρώτη προσέγγιση: απόσταση μεταξύ δυο διανυσμάτων

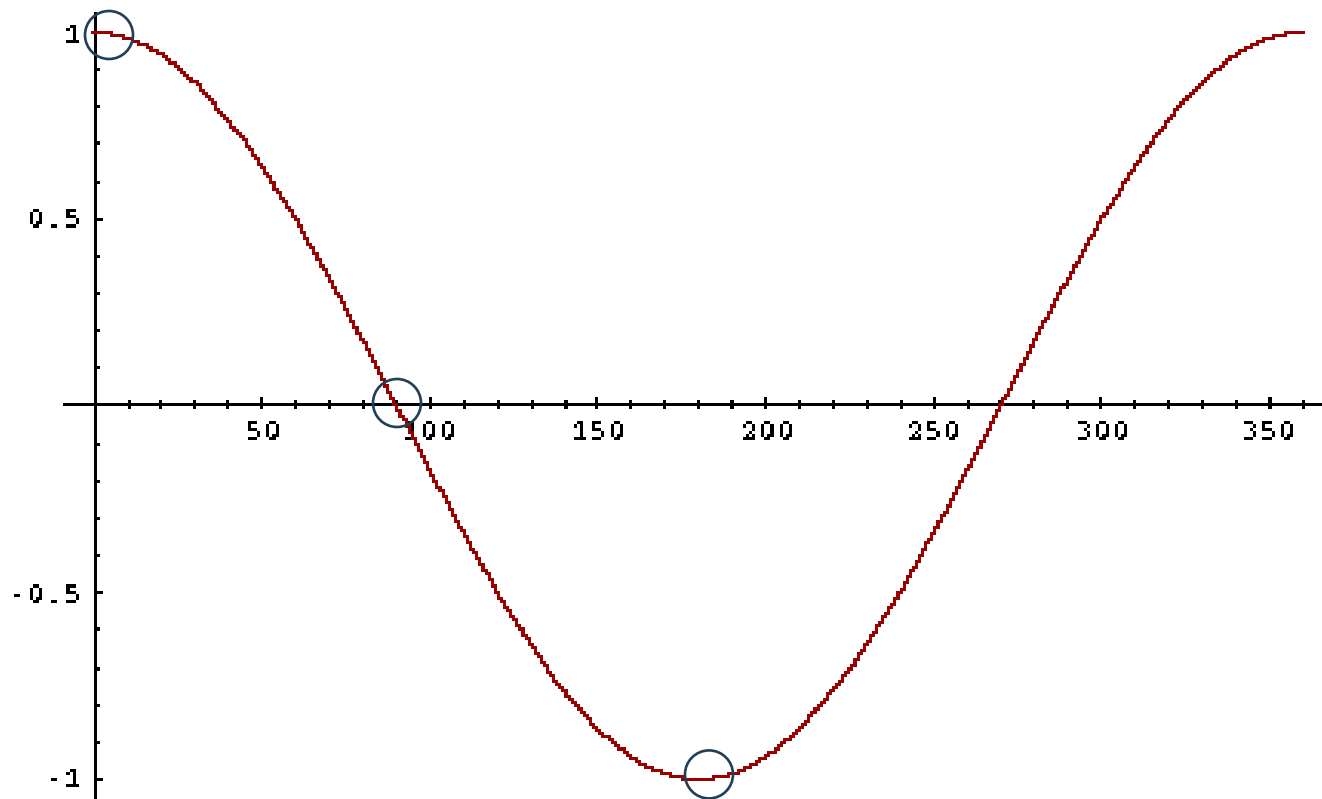
- Ευκλείδεια απόσταση;
 - Δεν είναι καλή ιδέα – είναι **μεγάλη** για διανύσματα **διαφορετικού μήκους**
- Έστω ένα έγγραφο d . Υποθέστε ότι κάνουμε append το d στον εαυτό του και έστω d' το κείμενο που προκύπτει.
- “Σημασιολογικά” το d και το d' έχουν το ίδιο περιεχόμενο
- Η Ευκλείδεια απόσταση μεταξύ 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 d και του d' είναι
μεγάλη αν και η
κατανομή των όρων
είναι παρόμοια



Από γωνίες σε συνημίτονα



Συνημίτονο μονότονα φθίνουσα συνάρτηση στο διάστημα $[0^\circ, 180^\circ]$

Ομοιότητα εγγράφων

Εσωτερικό γινόμενο

Μοναδιαία διανύσματα

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

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:

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

- 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 I

Ποιο έγγραφο είναι το πιο όμοιο με το d1;

$$\vec{d1} = (0, 0.097, 0.22, 0, 0, 0)$$

$$\vec{d2} = (0, 0.097, 0, 0.22, 0, 0)$$

$$\vec{d3} = (0, 0, 0.286, 0.22, 0.398, 0.699)$$

$$\vec{d4} = (0, 0.143, 0, 0, 0.398, 0)$$

$$\vec{d5} = (0, 0.097, 0.22, 0.22, 0, 0)$$

	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

$$sim(d1, d2) = \cos(\vec{d1}, \vec{d2}) = \frac{\vec{d1} \cdot \vec{d2}}{|\vec{d1}| |\vec{d2}|} = \frac{0.009}{\sqrt{0.058} \sqrt{0.058}} = 0.156$$

$$sim(d1, d3) = \cos(\vec{d1}, \vec{d3}) = \frac{\vec{d1} \cdot \vec{d3}}{|\vec{d1}| |\vec{d3}|} = \frac{0.062}{\sqrt{0.058} \sqrt{0.697}} = 0.308$$

$$sim(d1, d4) = \cos(\vec{d1}, \vec{d4}) = \frac{\vec{d1} \cdot \vec{d4}}{|\vec{d1}| |\vec{d4}|} = \frac{0.014}{\sqrt{0.058} \sqrt{0.179}} = 0.137$$

$$sim(d1, d5) = \cos(\vec{d1}, \vec{d5}) = \frac{\vec{d1} \cdot \vec{d5}}{|\vec{d1}| |\vec{d5}|} = \frac{0.058}{\sqrt{0.058} \sqrt{0.498}} = 0.341$$

d5
d3
d2
d4

Παράδειγμα II

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 II (συνέχεια)

Για απλοποίηση σε αυτό το παράδειγμα,
δε χρησιμοποιούμε τα idf βάρη

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

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

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

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 II (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

$$\cos(\text{SaS}, \text{PaP}) \approx$$

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

$$\approx 0.94$$

$$\cos(\text{SaS}, \text{WH}) \approx 0.79$$

$$\cos(\text{PaP}, \text{WH}) \approx 0.69$$

Γιατί $\cos(\text{SaS}, \text{PaP}) > \cos(\text{SaS}, \text{WH})$?

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 $\approx$ αντίθετο της απόστασης

cosine(query, document)

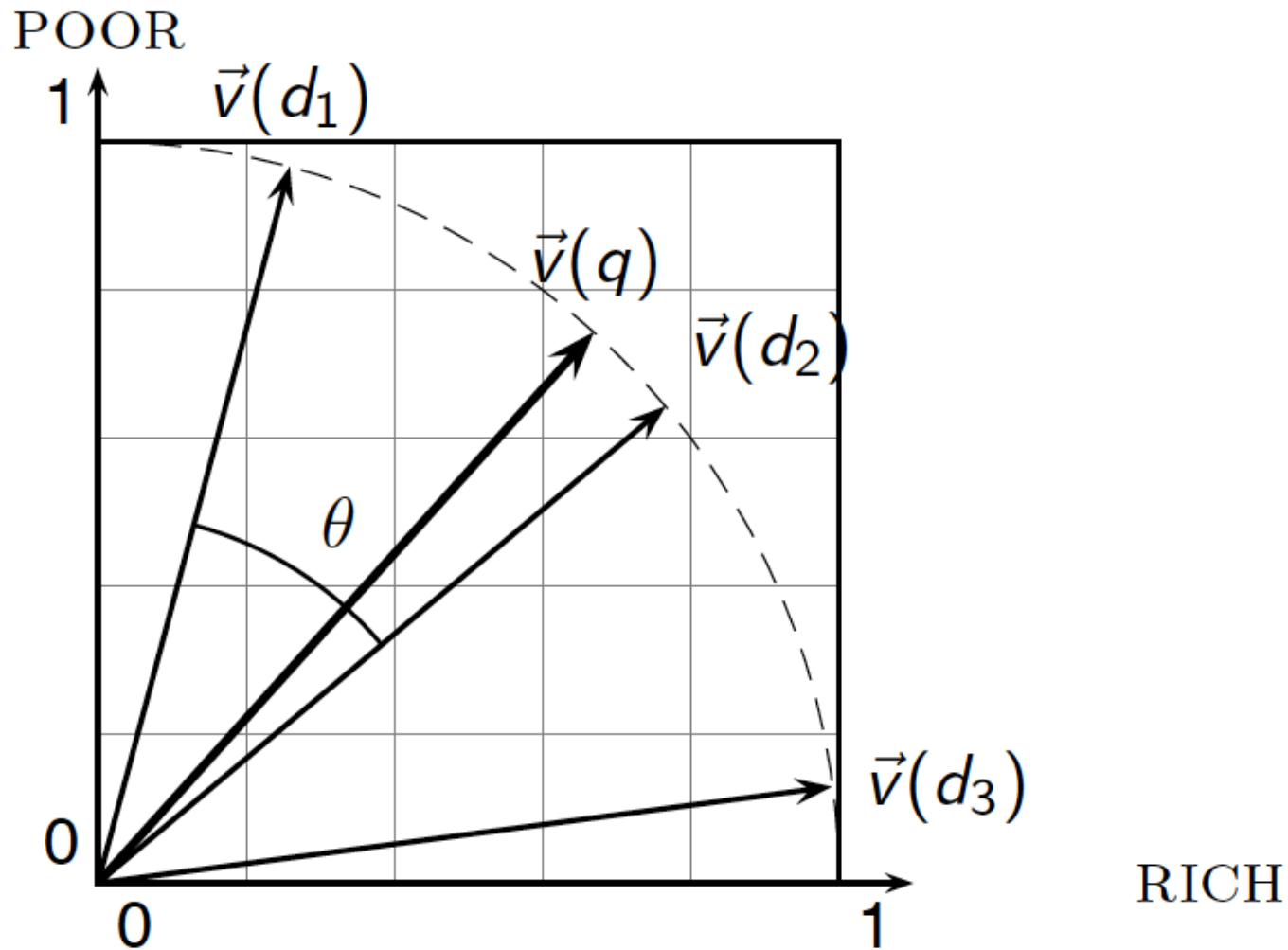
Ως $score(q, d)$ θα χρησιμοποιήσουμε το *συνημίτονο* της γωνίας της διανυσματικής αναπαράστασης του q και d

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

$\vec{q}_i$ 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

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	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	0.57	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

q1: b

d4

d1

d2

d3

	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

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

$$\text{score}(d_1, q_2) = 0.82$$

$$d_2 \quad 0.37$$

$$d_3 \quad 0.36$$

$$d_4 \quad 0.51$$

$$d_5 \quad 0.68$$

	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
- Συμβολισμός: *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 : Inc.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.Itc

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

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

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

$$\text{Score} = 0+0+(0.52*0.52=)27+(0.78*0.68=)0.53 = 0.8$$

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
<i>insurance</i>	10440	3997
<i>try</i>	10422	8760

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

Ερωτήσεις;