Εναλλακτικές βαθμονομήσεις και αναπαραστάσεις

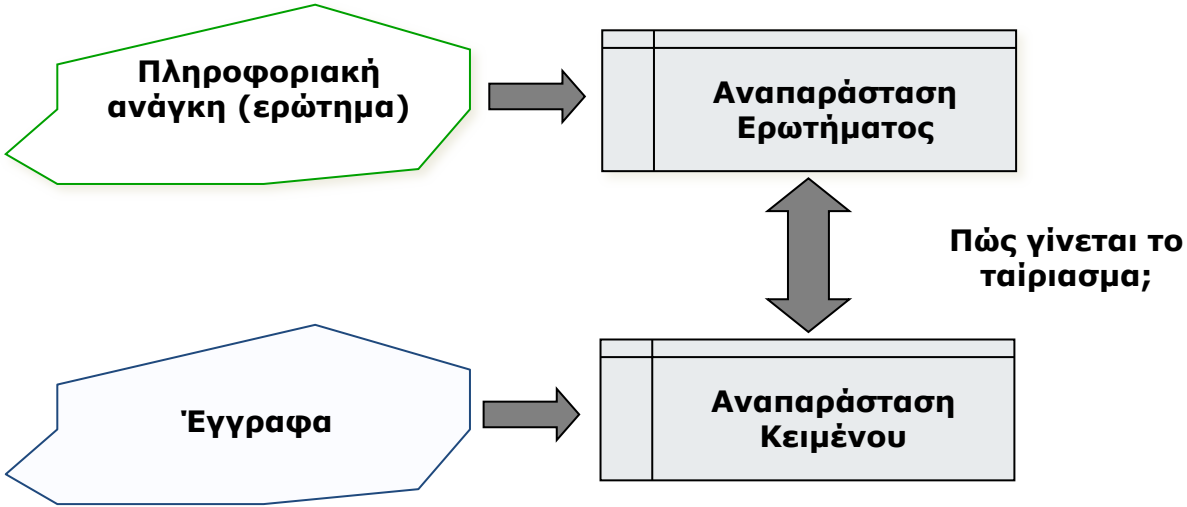
Ηρακλής Βαρλάμης

Περιεχόμενα

- Probabilistic models
- LSI (Latent Semantic Indexing)
- Singular Value Decomposition

Πιθανοτικά μοντέλα

- Βασίζονται στην υπόθεση ότι δεν είμαστε πάντοτε βέβαιοι για το πόσο σχετικό ή όχι είναι το περιεχόμενο ενός κειμένου προς ένα ερώτημα.
- Στο διανυσματικό μοντέλο (VSM) το ταίριασμα μεταξύ εγγράφου και ερωτηματος βασίζεται σε έναν σημασιολογικά μη-ακριβή χώρο των λέξεων του ευρετηρίου, λόγω και της πολυσημίας και συνωνυμίας των λέξεων
- Οι πιθανότητες παρέχουν ένα πλαίσιο διαχείρισης της αβεβαιότητας.



Θεώρημα του Bayes

- Για δύο events *A* και *B:*
- Bayes' Rule

$$p(A,B) = p(A \cap B) = p(A \mid B)p(B) = p(B \mid A)p(A)$$

$$p(A \mid B) = \frac{p(B \mid A)p(A)}{p(B)} = \frac{p(B \mid A)p(A)}{\sum_{X=A,\bar{A}} p(B \mid X)p(X)}$$

Prior

Posterior

- Odds ratio:

$$O(A) = \frac{p(A)}{p(\bar{A})} = \frac{p(A)}{1-p(A)}$$

Βαθμονόμηση κειμένων (ως προς τη σχετικότητα με την ερώτηση)

- Οι μέθοδοι βαθμονόμησης είναι βασικές στο IR
- Θέλουμε τα πιο σχετικά κείμενα να εμφανίζονται πιο ψηλά
- Ιδέα: Κατάταξη με βάση την πιθανότητα ένα κείμενο x να είναι σχετικό με το ερώτημα p(R=1|x, query) - R=1⇒σχετικό, R=0⇒μη-σχετικό
- Για ένα δεδομένο query:

$$p(R=1 \mid x) = \frac{p(x \mid R=1)p(R=1)}{p(x)}$$

$$p(R=0 \mid x) = \frac{p(x \mid R=0)p(R=0)}{p(x)}$$

$$p(R=0 \mid x) + p(R=1 \mid x) = 1$$

p(*R=1)*, p(*R=0*) - η πιθανότητα να βρω ένα σχετική (ή μη-σχετικό) κείμενο στη συλλογή

p(*x*|*R*=1), (p(*x*|*R*=0))- η πιθανότητα αν ανακτήσω ένα σχετικό κείμενο (ή μη σχετικό) αυτό να είναι το *x*.

Παραδοχές απλούστευσης

- Δε γνωρίζουμε τις ακριβείς πιθανότητες, άρα πρέπει να χρησιμοποιήσουμε εκτιμήσεις
- Το πιο απλό μοντέλο εκτίμησης είναι το Binary Independence Model (BIM)
- Παραδοχές (συζητήσιμες)
    - Η σχετικότητα ενός κειμένου είναι ανεξάρτητη από αυτή των άλλων κειμένων (σε πολλές θεματικές συλλογές υπάρχουν πολύ σχετικά μεταξύ τους κείμενα)
    - Η συνεισφορά κάθε όρου στη σχετικότητα είναι ανεξάρτητη από των υπολοίπων όρων (συχνά η εμφάνιση μιας λέξης σε ένα κείμενο είναι άμεσα συνδεδεμένη με την εμφάνιση και μιας άλλης λέξης - π.χ. Όνομα και Επώνυμο)

Binary Independence Model

- Queries: binary vectors που αφορούν την ύπαρξη ή όχι ενός όρου
- Για ένα δεδομένο ερώτημα q,
  - Για κάθε κείμενο d πρέπει να υπολογίσουμε το p(R|d,q).
  - Εναλλακτικά να υπολογίσουμε το p(R|x,q) όπου x είναι το binary term incidence vector του d (BoW).
  - Μας ενδιαφέρει η βαθμονόμηση μόνο (ranking) και όχι τα απόλυτα νούμερα
- Χρησιμοποιούμε τα odds από τον κανόνα του Bayes

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

Binary Independence Model

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

Σταθερό για το ίδιο ερώτημα = O(R|q)

Πρέπει να υπολογιστεί

- Υπόθεση ανεξαρτησίας

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

Ακριβώς τα ίδια που κάνουμε στον Naive Bayes

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

Binary Independence Model

Ο κάθε όρος x_i μπορεί να εμφανίζεται (x_i=1) ή να μην εμφανίζεται (x_i=0) στο ερώτημα

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

Έστω $p_i = p(x_i=1|R=1,q); r_i = p(x_i=1|R=0,q)$

η πιθανότητα εμφάνισης του όρου x_i σε ένα σχετικό ή μη σχετικό κείμενο.

Και υποθέτουμε ότι για όσους όρους δεν υπάρχουν στο ερώτημα (q_i=0) τα p_i=r_i

$$O(R|\overrightarrow{x},q) = O(R|,q)\cdot\prod_{x_i=1,q_i=1}\frac{p_i}{r_i}\cdot\prod_{x_i=0,q_i=1}^{n}\frac{(1-p_i)}{(1-r_i)}$$

Οι όροι του ερωτήματος που υπάρχουν και στα κείμενα

Οι όροι του ερωτήματος που δεν υπάρχουν στα κείμενα

Binary Independence Model

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

Οι όροι του ερωτήματος που υπάρχουν και στα κείμενα

Οι όροι του ερωτήματος που δεν υπάρχουν στα κείμενα

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

=1

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

Σταθερό για κάθε query

Οι όροι που ταιριάζουν

Όλοι οι όροι του ερωτήματος (ανεξάρτητα από το αν υπάρχουν ή όχι στα κείμενα)

Binary Independence Model

Ranking function: $$\prod_{x_i=q_i=1} \frac{p_i(1-r_i)}{r_i(1-p_i)}$$

Τα βάρη των όρων

Αν λογαριθμήσουμε: $$log \prod_{x_i=q_i=1} \frac{p_i(1-r_i)}{r_i(1-p_i)} = \sum_{x_i=q_i=1} log \frac{p_i(1-r_i)}{r_i(1-p_i)} = \sum_{x_i=q_i=1} c_i$$

$$\sum_{x_i=q_i=1} log \left(\frac{1-r_i}{r_i} \cdot \frac{p_i}{1-p_i} \right)$$

Binary Independence Model

Για κάθε όρο i κοιτάμε στον ακόλουθο πίνακα

Documents	Relevant	Non-Relevant	Total
$x_i=1$	s	$n-s$	n
$x_i=0$	$S-s$	$N-n-S+s$	$N-n$
Total	S	$N-S$	N

- Estimates: $p_i \approx \frac{s}{S}$ $r_i \approx \frac{(n-s)}{(N-S)}$ $c_i = log(\frac{1-r_i}{r_i} \cdot \frac{p_i}{1-p_i})$

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

Υπολογισμός σχετικότητας - Παραδοχή

- Τα μη-σχετικά κείμενα είναι σχεδόν όσα τα κείμενα της συλλογής (για μεγάλες συλλογές και συγκεκριμένα ερωτήματα), οπότε r_i=n/N

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

$$\sum_{x_i=q_i=1} log\left(\frac{1-r_i}{r_i}\cdot\frac{p_i}{1-p_i}\right)$$

Υπολογισμός σχετικότητας

Μένει να υπολογίσουμε το p_i (πιθανότητα εμφάνισης μιας λέξης του ερωτήματος στα σχετικά κείμενα.

- Αν γνωρίζουμε κάποια σχετικά έγγραφα μπορούμε να το υπολογίσουμε (relevance feedback)
- Αν το θεωρήσουμε σταθερό (π.χ. 0.5) τότε βασιζόμαστε στο IDF $\sum_{x_i=q_i=1} \log \frac{N}{n_i}$
- Αν το θεωρήσουμε αναλογικό της πιθανότητας εμφάνισης του όρου στη συλλογή: $1/3 + 2/3\ df_i/N$

Εναλλακτικές του BIM

- Okapi, BM25 θεωρούν non-binary μοντέλα στο κατά πόσο ένα κείμενο σχετίζεται ή όχι με την ερώτηση.
- Λαμβάνει υπόψη τη συχνότητα εμφάνισης κάθε όρου (tf) και το μήκος κάθε κειμένου.

Μια πολύ ενδιαφέρουσα παρουσίαση από τον Victor Lavrenko (Univ. of Edinburgh): https://rb.gy/gqp6e0

Περιεχόμενα

- Probabilistic models
- LSI (Latent Semantic Indexing)
- Singular Value Decomposition

Λέξεις ή Έννοιες

- Το ακριβές λεκτικό ταίριασμα (exact keyword/term matching) δεν δουλεύει πάντοτε καλά
  - Κυρίως όταν είναι μικρά τα κείμενα ή τα ερωτήματα
  - Οι άνθρωποι χρησιμοποιούν διαφορετικούς όρους για να περιγράψουν τα ίδια πράγματα (τις ίδιες έννοιες) π.χ. τρένο, αμαξοστοιχία
- Μια λύση είναι να έχουμε ένα θησαυρό που θα μας λέει ποιες λέξεις συνδέονται σημασιολογικά με ποιες άλλες λέξεις (π.χ. ότι τρένο και αμαξοστοιχία είναι συνώνυμα, ή ότι αμάξι και αυτοκίνητο είναι συνώνυμα και ότι το όχημα είναι υπέρνυμο του αυτοκινήτου)

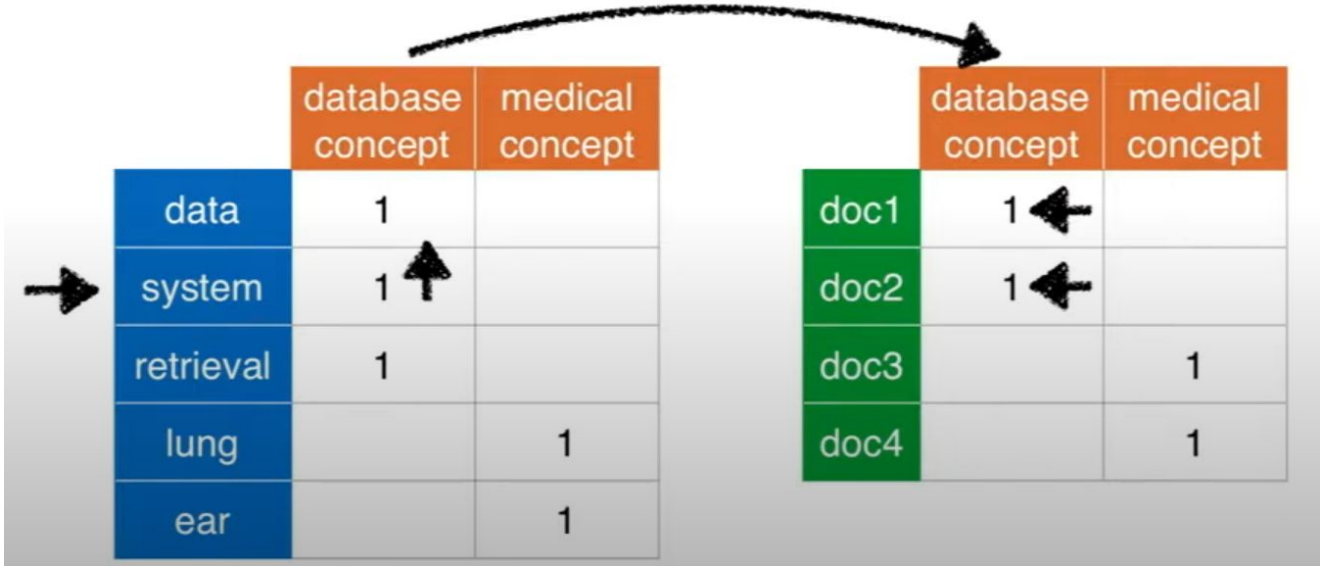
Latent Semantic Indexing (LSI)

- Η μέθοδος ευρετηρίασης της υποκρυπτόμενης σημασιολογικής πληροφορίας (indexing of latent semantics - LSI) χρησιμοποιεί τις στατιστικές συσχετίσεις μεταξύ λέξεων
- Σε πολλά κείμενα που μιλάνε για τρένα εμφανίζεται η λέξη αμαξοστοιχία, άρα οι δύο λέξεις έχουν μια συσχέτιση
- Εισάγει την έννοια του “concept”: a set of terms, with weights
- Το LSI χρησιμοποιεί μια στατιστική τεχνική, το Singular Value Decomposition (SVD) για να υπολογίσει την κρυμμένη (λανθάνουσα/latent) σημασιολογική δομή των λέξεων
- **Κρυμμένος χώρος «εννοιών»:** Συσχετίζει συντακτικά διαφορετικούς αλλά σημασιολογικά όμοιους όρους και έγγραφα

Οι όροι αντιστοιχίζονται σε έννοιες με κάποιο βάρος

Η αναζήτηση για έναν όρο ⇒ την αναζήτηση για μια έννοια

Πρακτικά προσπαθούμε να φτιάξουμε έναν θησαυρό όρων με αυτόματο τρόπο



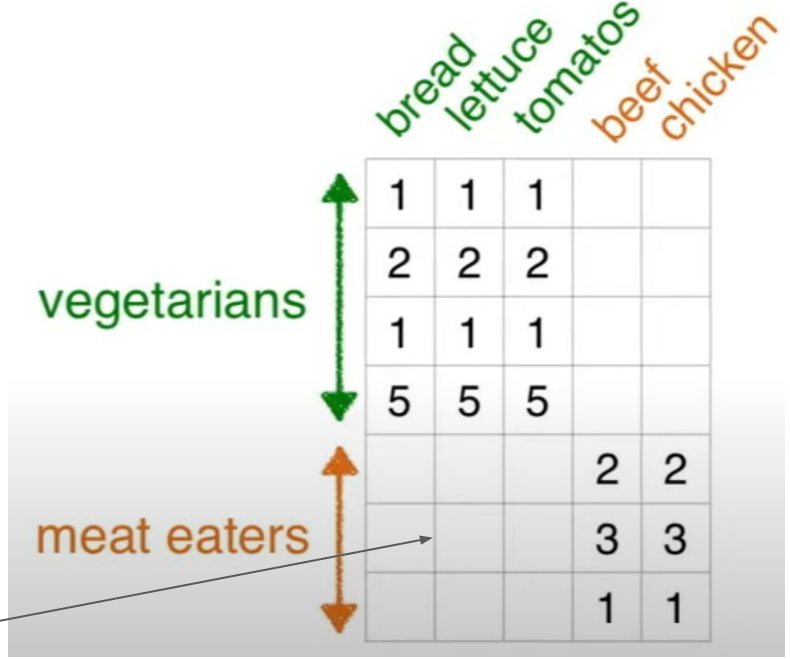
Περιεχόμενα

- Probabilistic models
- LSI (Latent Semantic Indexing)
- Singular Value Decomposition

Singular Value Decomposition (SVD)

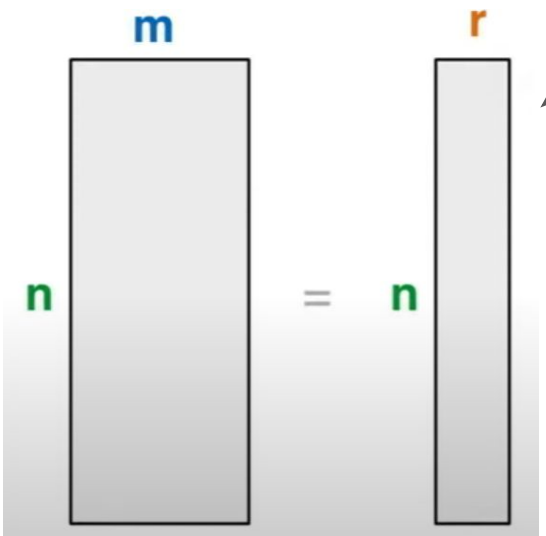
Θέλουμε να μπορούμε

- Να βρίσκουμε τις έννοιες μέσα από πίνακες
- Να συμπιέσουμε τις διαστάσεις του χώρου αναπαράστασης στις έννοιες (αντί για τους όρους)



Πώς ορίζεται το SVD

$A_{[nxm]}=U_{[nxr]}\,\Lambda_{[rxr]}\,(V_{[mxr]})^T$



dense

n κείμενα
m όροι

n κείμενα
r έννοιες

Διαγώνιος πίνακας:
Οι τιμές δηλώνουν
την ισχύ κάθε έννοιας
(πάνω αριστερά οι
πιο ισχυρές)

m όροι
r έννοιες

Ιδιότητες του SVD

- Υπάρχει πάντα μια τέτοια αποσύνθεση του A σε $A=U\Lambda V^T$
- Τα U, Λ, V είναι μοναδικά συνήθως
- Οι στήλες των U, V είναι **ορθοκανονικές**
  - $U^TU=I$ και $V^TV=I$
- Ο Λ είναι διαγώνιος πίνακας με μη-αρνητικές τιμές στη διαγώνιο, με φθίνουσα σειρά.

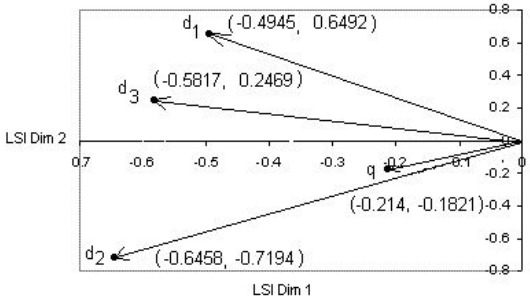
Παράδειγμα

- Βαθμονόμησε τα έγγραφα d1,d2,d3 ως προς το ερώτημα q (gold silver truck)

Αποσύνθεσε τον πίνακα A και βρες τους πίνακες U, S και V^T

Terms	d1	d2	d3		q
a	1	1	1		0
arrived	0	1	1		0
damaged	1	0	0		0
delivery	0	1	0		0
fire	1	0	0		0
gold	1	0	1		1
in	1	1	1		0
of	1	1	1		0
shipment	1	0	1		0
silver	0	2	0		1
truck	0	1	1		1

A = q =



- Δουλεύει καλύτερα από τους παραδοσιακούς αλγορίθμους που βασίζονται σε λέξεις
- Αλλά έχει μεγάλη πολυπλοκότητα (κακό για το Web)