ΑΡΙΣΤΟΤΕΛΕΙΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ

ΑΝΟΙΚΤΑ
ΑΚΑΔΗΜΑΪΚΑ
ΜΑΘΗΜΑΤΑ

Ανάκτηση Πληροφορίας

Ενότητα 8: Λανθάνουσα Σημασιολογική Ανάλυση
(Latent Semantic Analysis)

Απόστολος Παπαδόπουλος
Τμήμα Πληροφορικής

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «Ανοικτά Ακαδημαϊκά Μαθήματα στο Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης» έχει χρηματοδοτήσει μόνο την αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

ΑΡΙΣΤΟΤΕΛΕΙΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ

ΑΝΟΙΚΤΑ
ΑΚΑΔΗΜΑΪΚΑ
ΜΑΘΗΜΑΤΑ

Λανθάνουσα Σημασιολογική Ανάλυση (Latent Semantic Analysis)

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο

ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ

ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Περιεχόμενα ενότητας

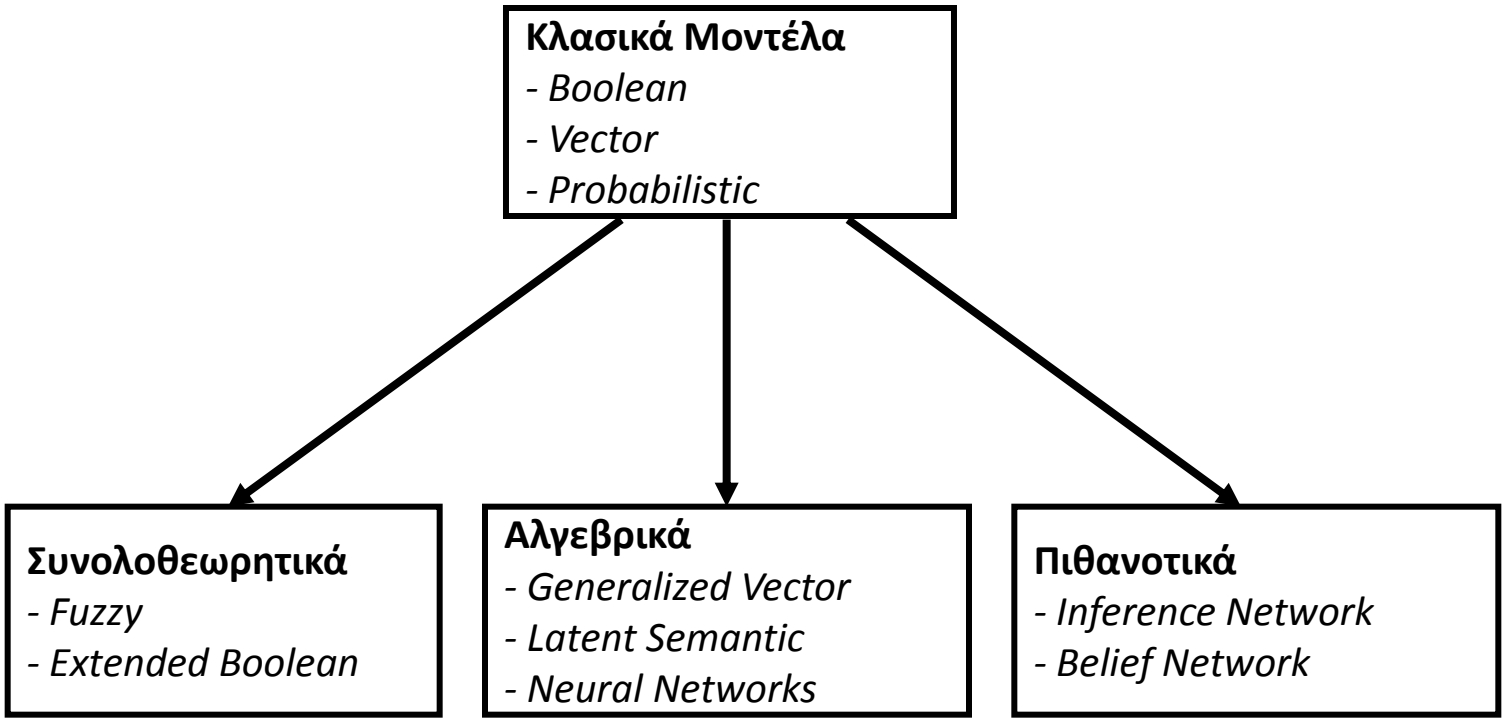
1. Λανθάνουσα Σημασιολογική Ανάλυση

ΑΡΙΣΤΟΤΕΛΕΙΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ

Latent Semantic Analysis

Λανθάνουσα Σημασιολογική Ανάλυση

Μοντέλα IR


Κλασικά Μοντέλα
- Boolean
- Vector
- Probabilistic
Συνολοθεωρητικά
- Fuzzy
- Extended Boolean
Αλγεβρικά
- Generalized Vector
- Latent Semantic
- Neural Networks
Πιθανοτικά
- Inference Network
- Belief Network

Κίνητρο-1

Βασικό μειονέκτημα των μέχρι τώρα μοντέλων είναι ότι υποθέτουν την ανεξαρτησία των όρων (η πιθανότητα ένας όρος να ανήκει σε ένα έγγραφο δεν εξαρτάται από την ύπαρξη κάποιου άλλου όρου στο έγγραφο).

Επίσης, έχουμε δύο βασικά προβλήματα με τους όρους: **πολυσημία** (ένας όρος έχει διαφορετικό νόημα σε διαφορετικές περιοχές, π.χ. ποντίκι) και **συνωνυμία** (δύο όροι έχουν ακριβώς το ίδιο ή πολύ κοντινό νόημα, άστρο, αστέρας).

Κίνητρο-2

Ποια η επίδραση της **πολυσημίας** και της **συνωνυμίας** στην ανάκληση και την ακρίβεια;

Η πολυσημία μειώνει την ακρίβεια και
η συνωνυμία μειώνει την ανάκληση

Γιατί;

Στόχος

- Να μπορέσουμε να βρούμε συσχετίσεις μεταξύ των όρων έτσι ώστε να μεταβούμε από τους όρους στις θεματικές περιοχές (concepts).
- Να μειώσουμε την επίδραση της πολυσημίας και της συνωνυμίας.
- Να μειώσουμε τον αριθμό των διαστάσεων.

Singular Value Decomposition-1

- Η τεχνική LSI (Latent Semantic Indexing) βασίζεται στο μαθηματικό εργαλείο SVD (Singular Value Decomposition).

Singular Value Decomposition-2

- Problem:
  - #1: Find concepts in text
  - #2: Reduce dimensionality

term / document	data	information	retrieval	brain	lung
CS-TR1	1	1	1	0	0
CS-TR2	2	2	2	0	0
CS-TR3	1	1	1	0	0
CS-TR4	5	5	5	0	0
MED-TR1	0	0	0	2	2
MED-TR2	0	0	0	3	3
MED-TR3	0	0	0	1	1

SVD - Definition

$$\mathbf{A}_{[n \times m]} = \mathbf{U}_{[n \times r]} \Lambda_{[r \times r]} (\mathbf{V}_{[m \times r]})^T$$

- **A**: *n x m* matrix (e.g., n documents, m terms)
- **U**: *n x r* matrix (n documents, r concepts)
- Λ: *r x r* diagonal matrix (strength of each 'concept') (r: rank of the matrix)
- **V**: m x r matrix (m terms, r concepts)

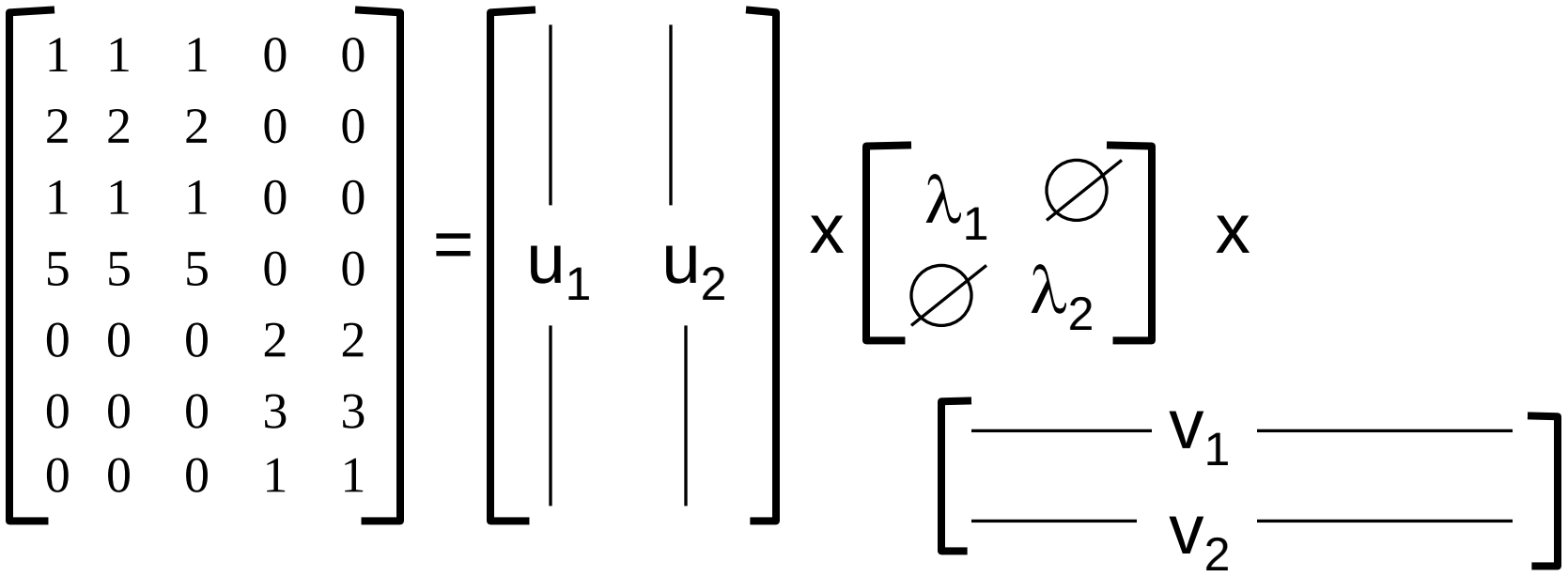
SVD – Properties-1

THEOREM [Press+92]: always possible to decompose matrix **A** into **A** = **U** Λ **V**^{T} , where

- **U, Λ, V**: unique (*)
- **U**, **V**: column orthonormal (ie., columns are unit vectors, orthogonal to each other)
  - **U**^{T}**U** = **I**; **V**^{T}**V** = **I (I:** identity matrix**)**
- **Λ:** singular value are positive, and sorted in decreasing order

SVD – Properties-2

'spectral decomposition' of the matrix:



SVD - Interpretation

'documents', 'terms' and 'concepts':

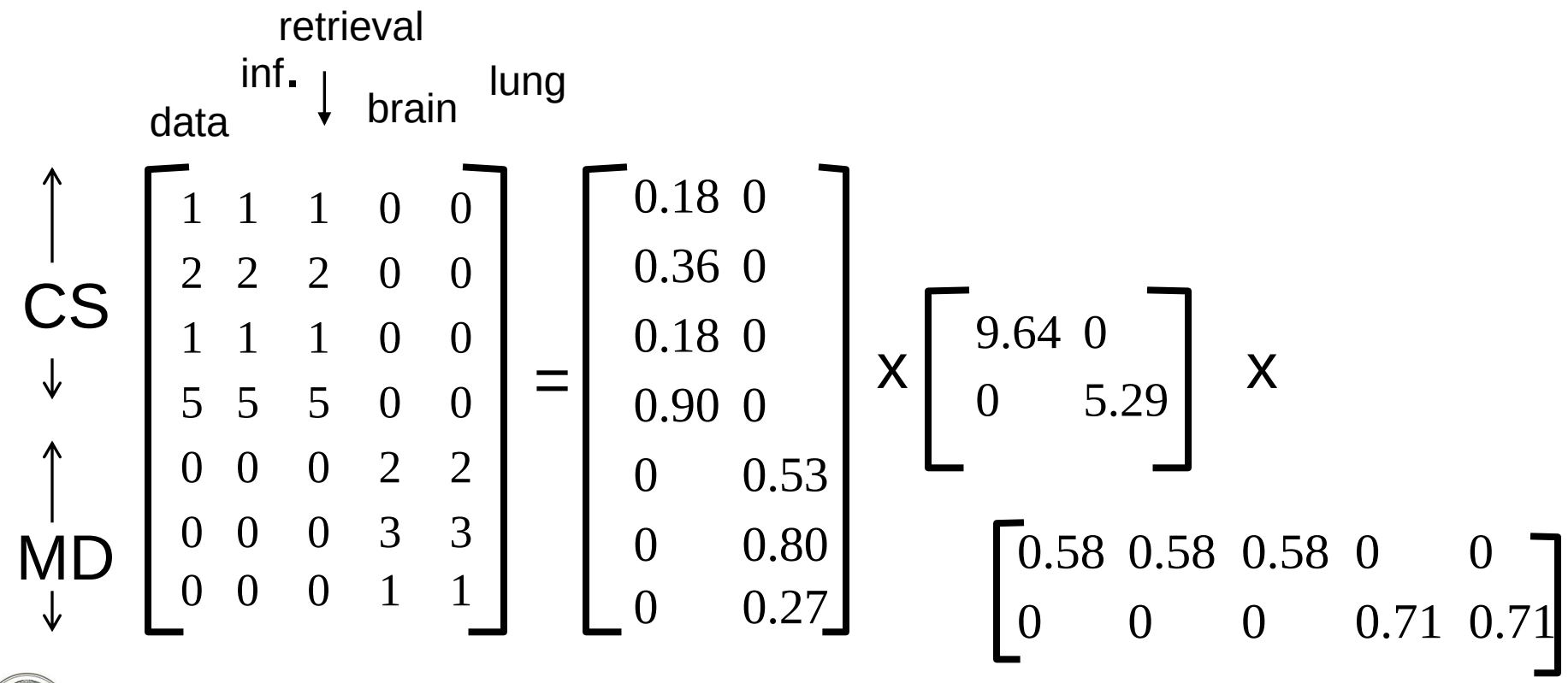
- **U**: document-to-concept similarity matrix
- **V**: term-to-concept similarity matrix
- Λ: its diagonal elements: 'strength' of each concept

Projection:

- best axis to project on: ('best' = min sum of squares of projection errors)

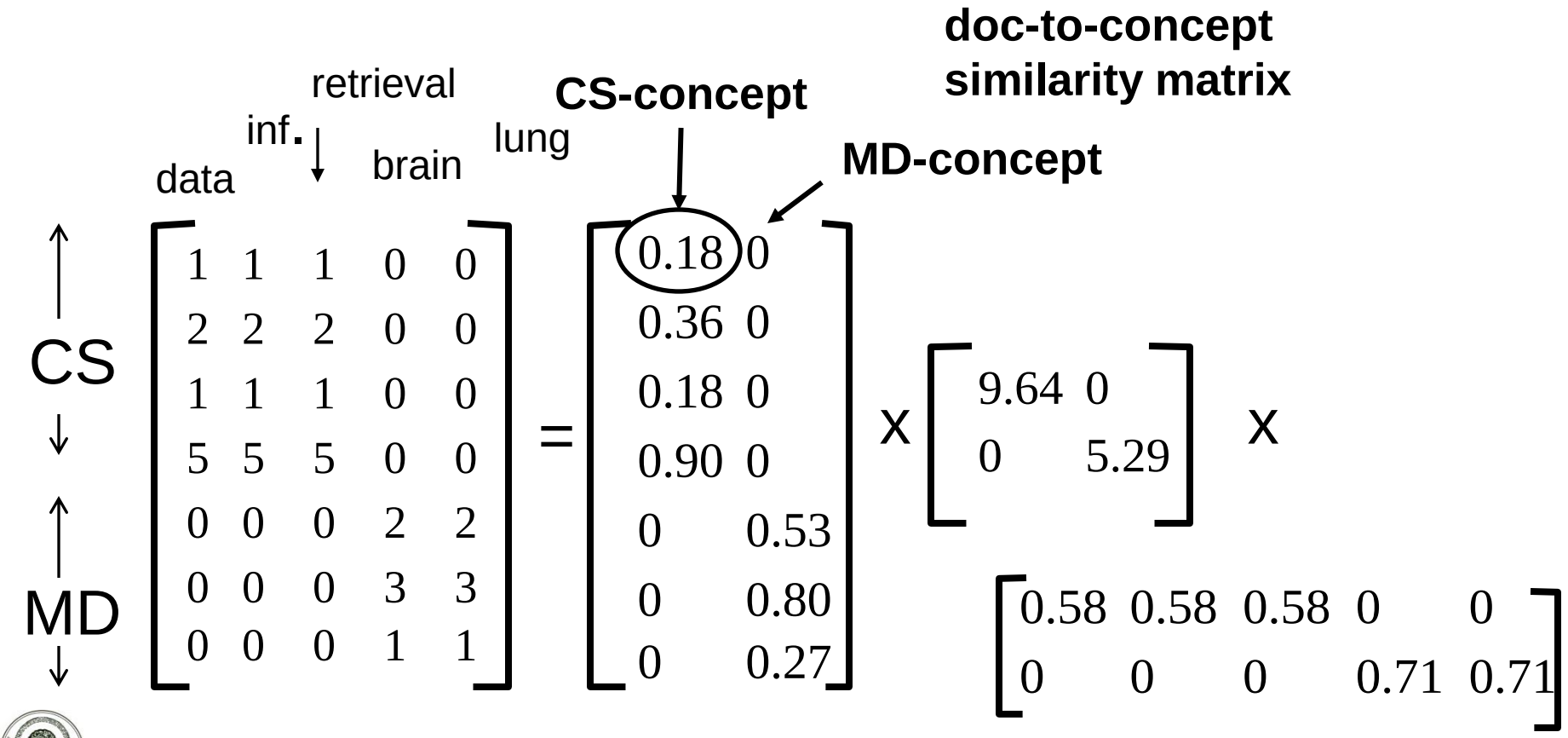
SVD – Example-1

- $\mathbf{A} = \mathbf{U}\,\Lambda\,\mathbf{V}^T$ - example:



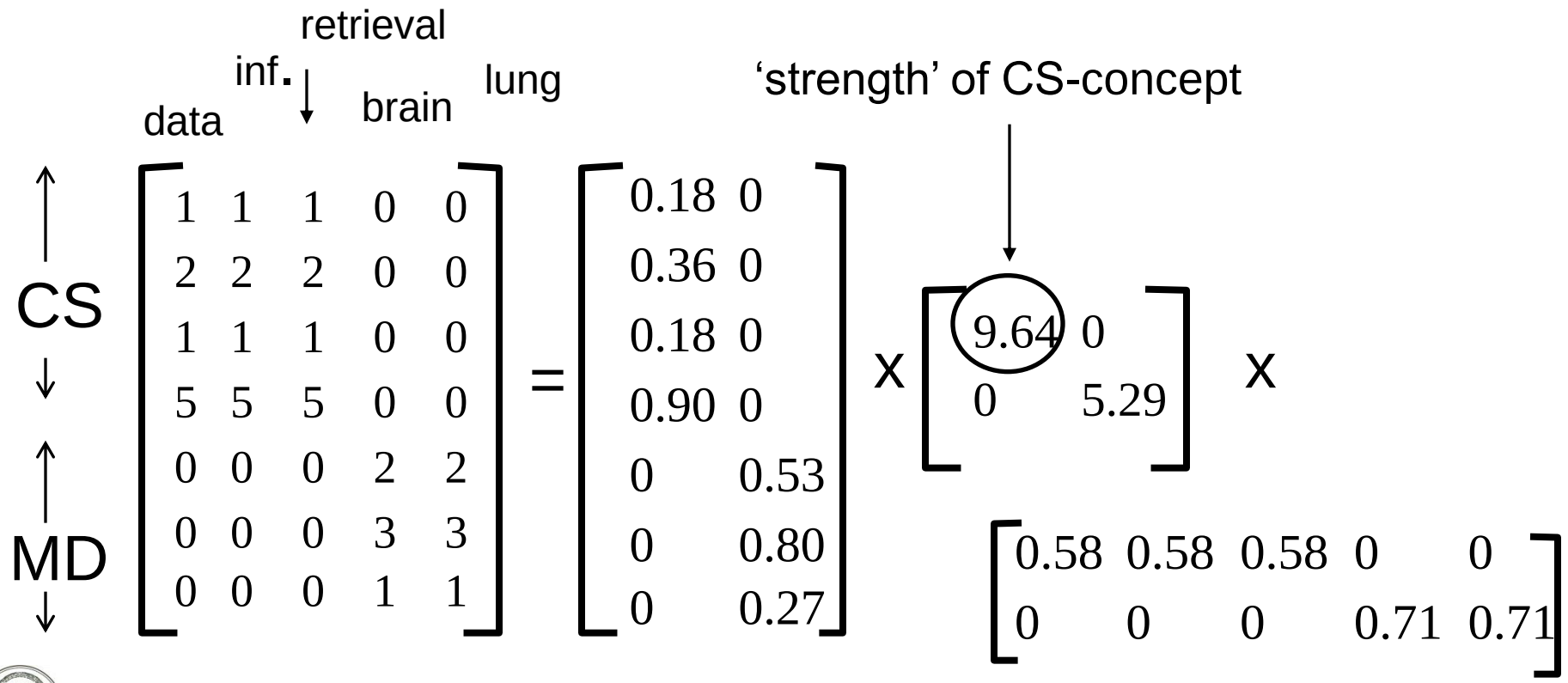
SVD – Example-2

- $\mathbf{A} = \mathbf{U}\,\Lambda\,\mathbf{V}^T$ - example:



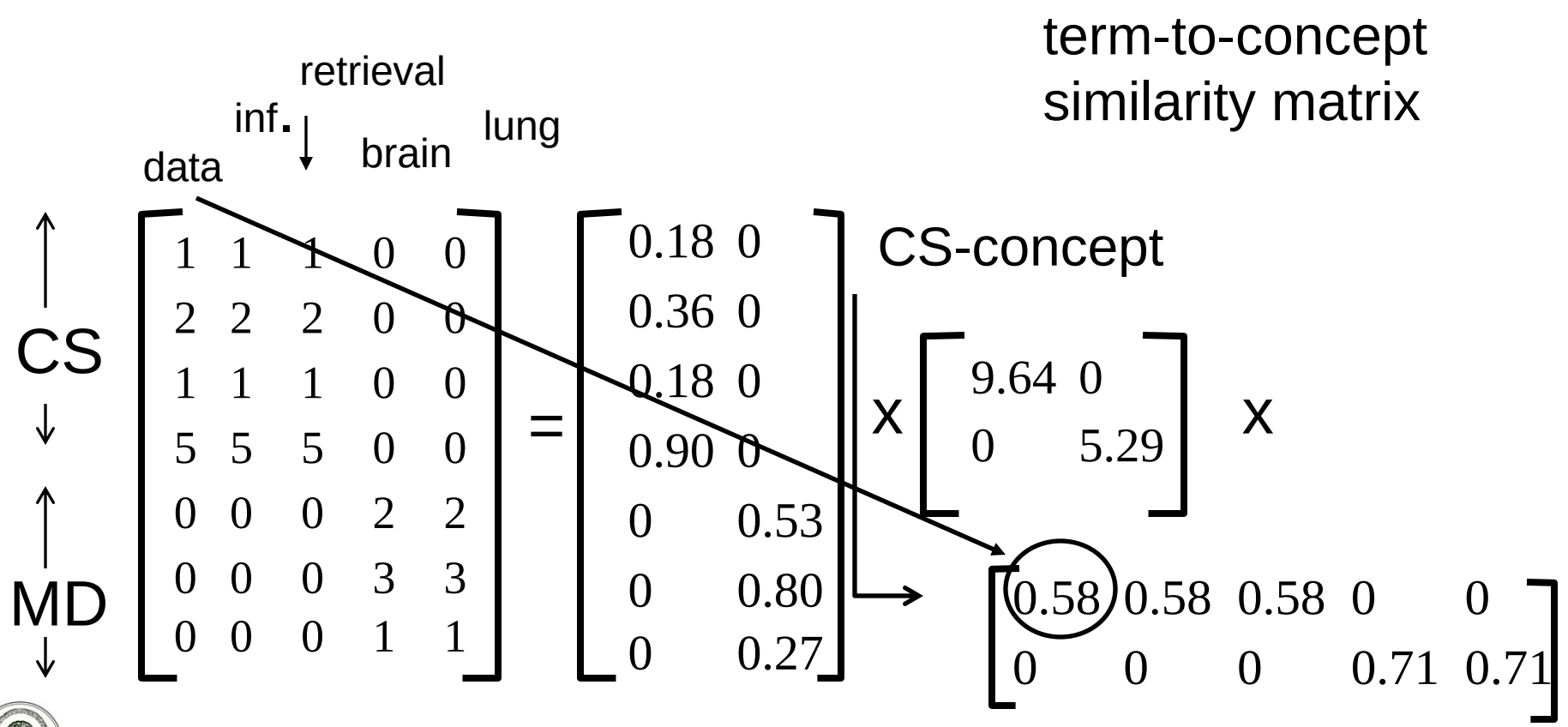
SVD – Example-3

- **A** = **U** Λ $\mathbf{V}^T$ - example:



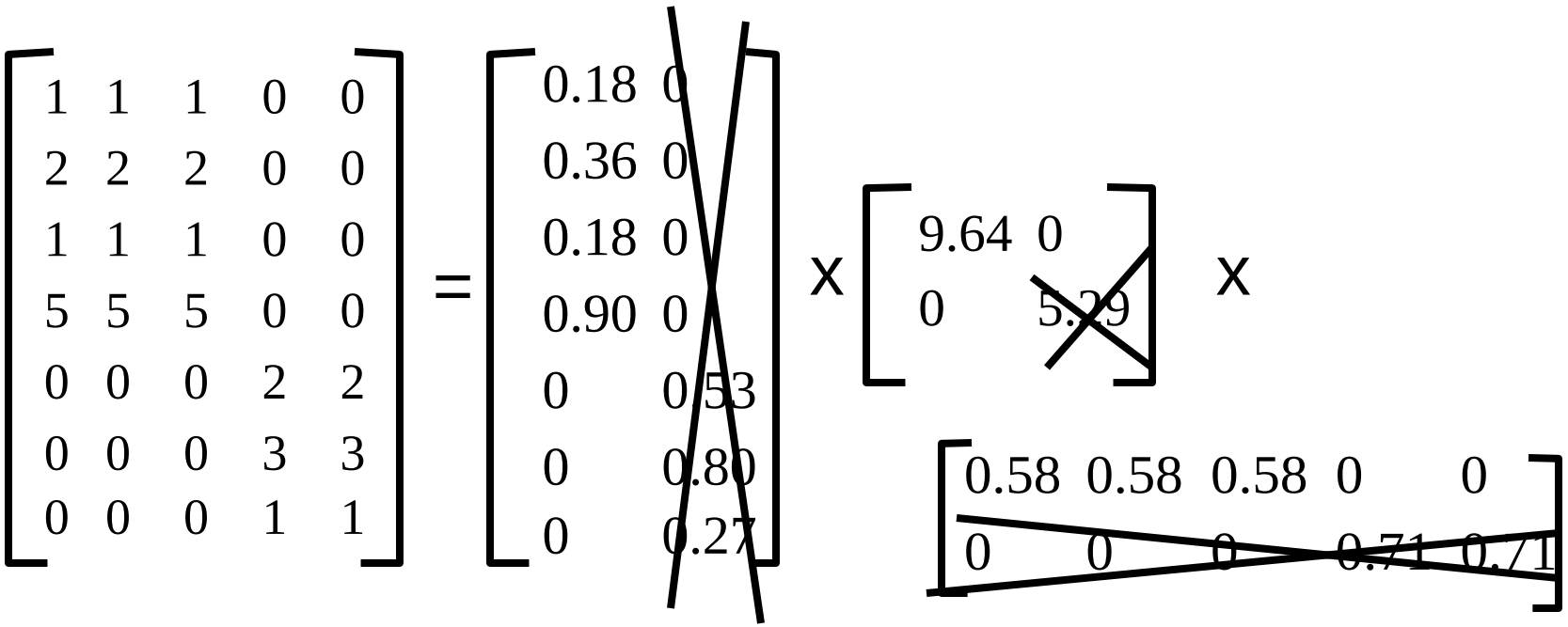
SVD – Example-4

- **A** = **U** Λ **V**^{T} - example:



SVD – Dimensionality reduction-1

- Q: how exactly is dim. reduction done?
- A: set the smallest singular values to zero:



SVD - Dimensionality reduction-2

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \sim \begin{bmatrix} 0.18 \\ 0.36 \\ 0.18 \\ 0.90 \\ 0 \\ 0 \\ 0 \end{bmatrix} \times \begin{bmatrix} 9.64 \end{bmatrix} \times \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \end{bmatrix}$$

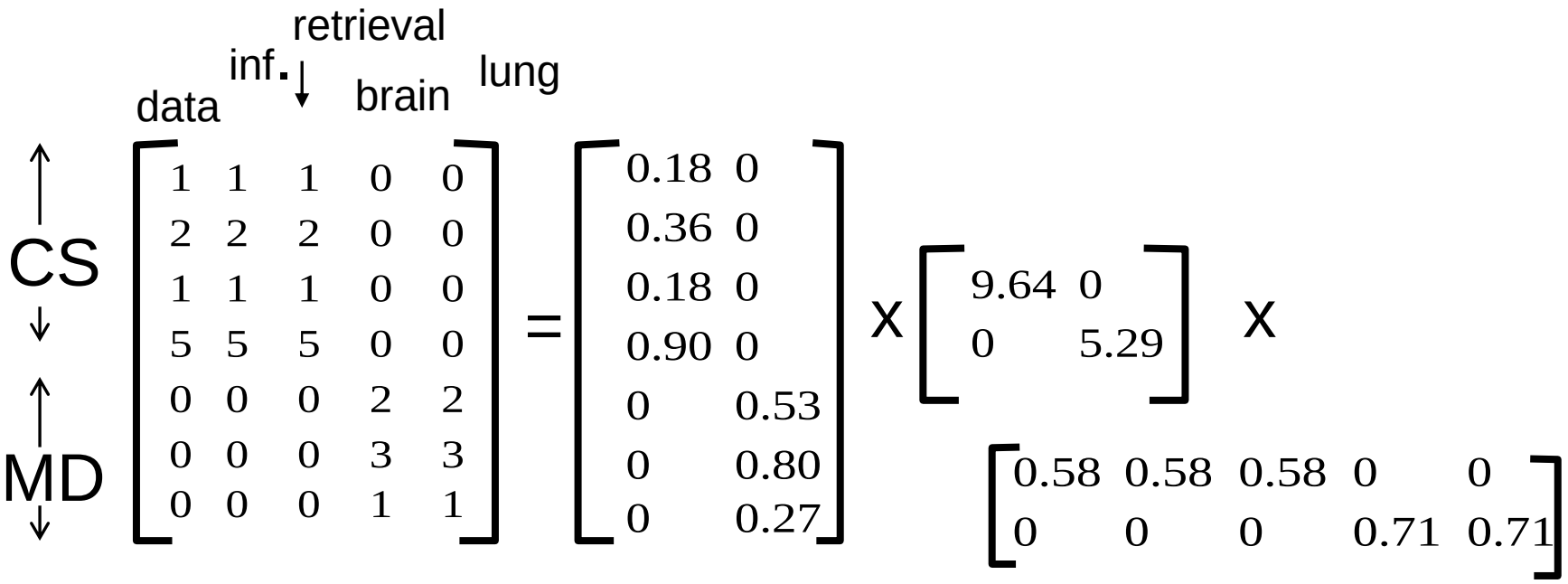
SVD - Dimensionality reduction-3

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \sim \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

LSI-1

Q1: How to do queries with LSI?

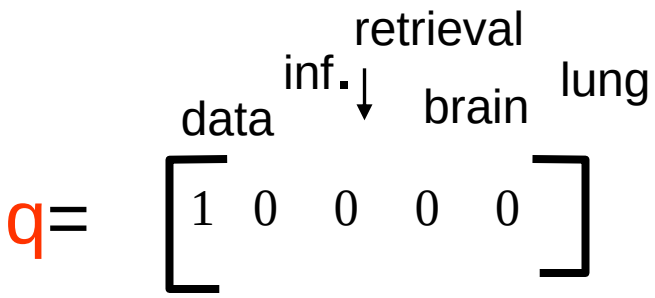
A: map query vectors into 'concept space' – how?

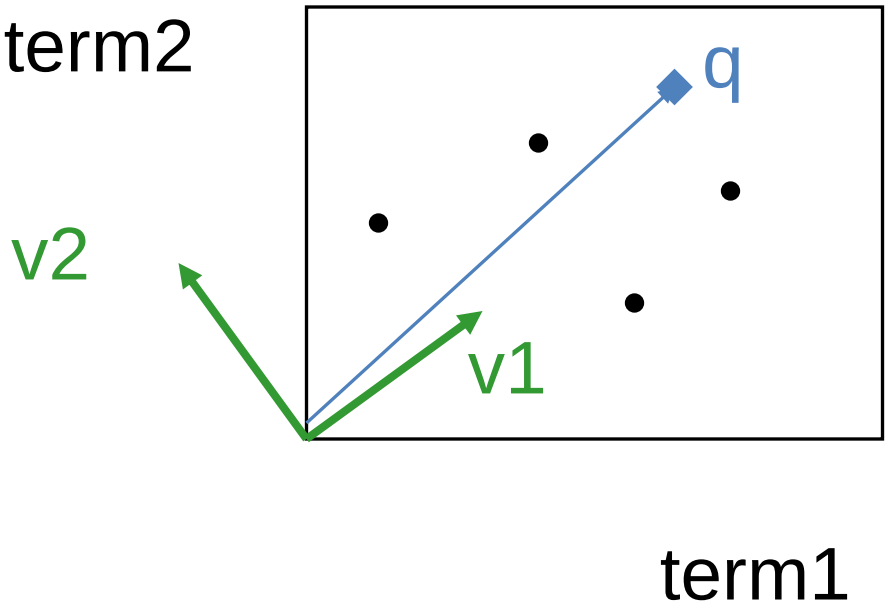


LSI-2

Q: How to do queries with LSI?

A: map query vectors into 'concept space' – how?





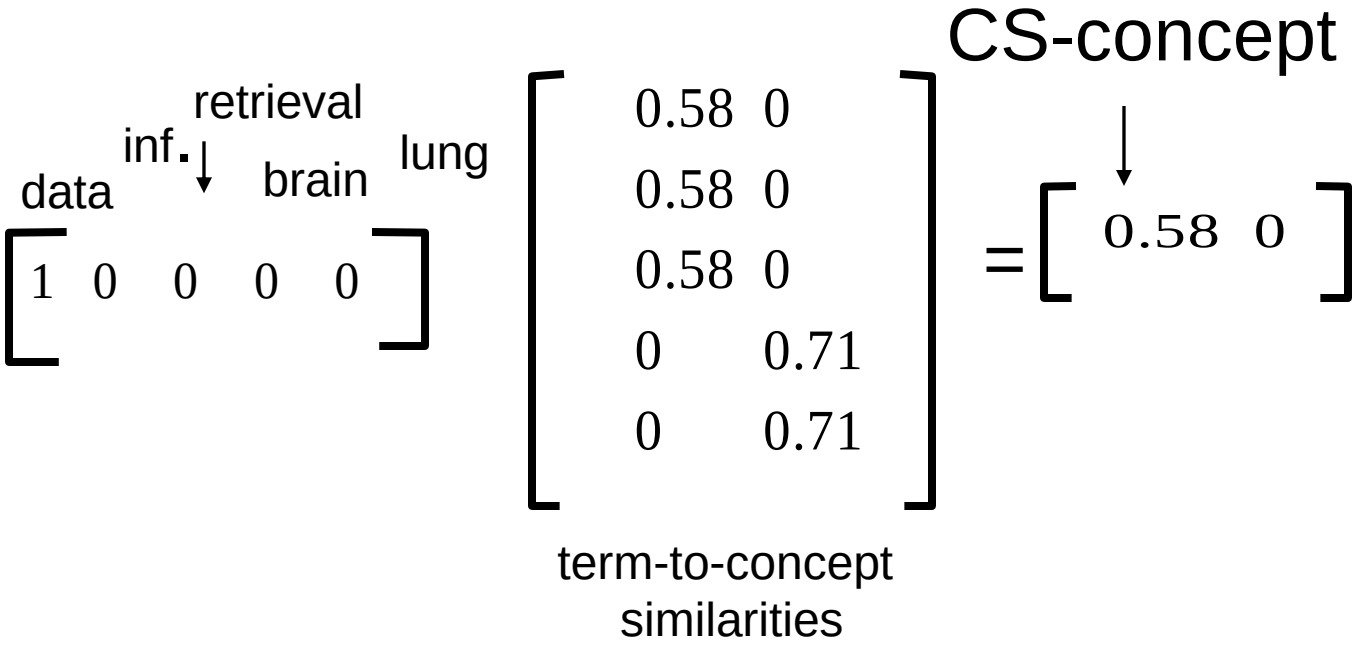
A: inner product
(cosine similarity)
with each 'concept' vector v_i

LSI-3

compactly, we have:

$q_{concept} = q\ \mathbf{V}$

e.g.:

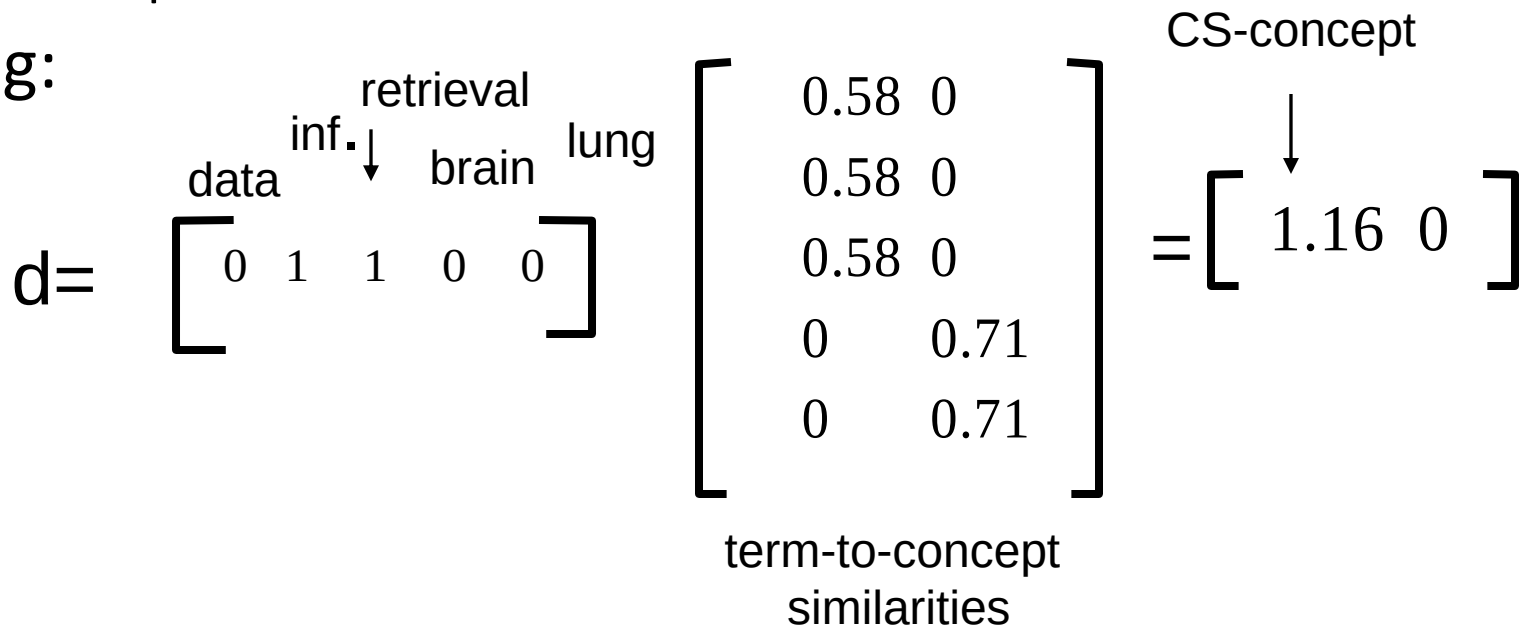


Little example

How would the document ('information', 'retrieval') handled by LSI? A: SAME:

$d_{concept} = d\,\mathbf{V}$

Eg:



Προσοχή

- Εάν στον αρχικό πίνακα Α τα διανύσματα των εγγράφων είναι στις γραμμές του πίνακα τότε για τη μείωση των διαστάσεων χρησιμοποιείται ο μετασχηματισμός:

$$A' = A * Vk$$

- Εάν στον αρχικό πίνακα Α τα διανύσματα των εγγράφων είναι στις στήλες του Α τότε:

$$A' = Uk^T * A$$

Σημείωμα Αναφοράς

Σημείωμα Αδειοδότησης

Το παρόν υλικό διατίθεται με τους όρους της άδειας χρήσης Creative Commons Αναφορά - Μη Εμπορική Χρήση - Όχι Παράγωγα Έργα 4.0 [1] ή μεταγενέστερη, Διεθνής Έκδοση. Εξαιρούνται τα αυτοτελή έργα τρίτων π.χ. φωτογραφίες, διαγράμματα κ.λ.π., τα οποία εμπεριέχονται σε αυτό και τα οποία αναφέρονται μαζί με τους όρους χρήσης τους στο «Σημείωμα Χρήσης Έργων Τρίτων».

Ο δικαιούχος μπορεί να παρέχει στον αδειοδόχο ξεχωριστή άδεια να χρησιμοποιεί το έργο για εμπορική χρήση, εφόσον αυτό του ζητηθεί.

Ως **Μη Εμπορική** ορίζεται η χρήση:

- που δεν περιλαμβάνει άμεσο ή έμμεσο οικονομικό όφελος από την χρήση του έργου, για το διανομέα του έργου και αδειοδόχο
- που δεν περιλαμβάνει οικονομική συναλλαγή ως προϋπόθεση για τη χρήση ή πρόσβαση στο έργο
- που δεν προσπορίζει στο διανομέα του έργου και αδειοδόχο έμμεσο οικονομικό όφελος (π.χ. διαφημίσεις) από την προβολή του έργου σε διαδικτυακό τόπο

[1] http://creativecommons.org/licenses/by-nc-nd/4.0/

Τέλος ενότητας

Επεξεργασία: <Μαυρίδης Απόστολος>
Θεσσαλονίκη, <Εαρινό εξάμηνο 2013-2014>

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Σημειώματα

Διατήρηση Σημειωμάτων

Οποιαδήποτε αναπαραγωγή ή διασκευή του υλικού θα πρέπει να συμπεριλαμβάνει:

- το Σημείωμα Αναφοράς
- το Σημείωμα Αδειοδότησης
- τη δήλωση Διατήρησης Σημειωμάτων
- το Σημείωμα Χρήσης Έργων Τρίτων (εφόσον υπάρχει)

μαζί με τους συνοδευόμενους υπερσυνδέσμους.