Applied Text Analytics & Natural Language Processing

with Dr. Mahdi Roozbahani
& Wafa Louhichi

Topic Modeling
Latent Semantic Indexing

This lecture is adopted based on Polo Chau LSI slides in DVA

Learning Objectives

In this lesson, you will learn a topic-learning model named as Latent Semantic Indexing (LSI)

- SVD
- Document-term matrix
- Queries with LSI

What is Topic Modeling?

It is an unsupervised learning technique (no labels needed) to extract topics from documents and find documents that potentially share a common context.

This technique is used to query documents that may not have all the keywords, but are still related to a topic.

We may retrieve documents that DON’T have the term “system”,
but they contain almost everything else (“data”, “retrieval”)

Latent Semantic Indexing (LSI)

Main idea

- Map each **document** into some '**concepts**'
- Map each **term** into some '**concepts**'

'**Concept**': a set of terms, with weights

For example, **DBMS_concept**:

- "data" (0.8),
- "system" (0.5),
- "retrieval" (0.6)

We Need to Construct the Document-term Matrix

It is similar to the unigrams Bag of Words:

	data	system	retireval	lung	ear
doc1	1	1	1		
doc2	1	1	1		
doc3				1	1
doc4				1	1

Convert Document-term Matrix into a Concept Matrix

term-concept matrix

	database concept	medical concept
data	1	
system	1	
retrieval	1	
lung		1
ear		1

… and

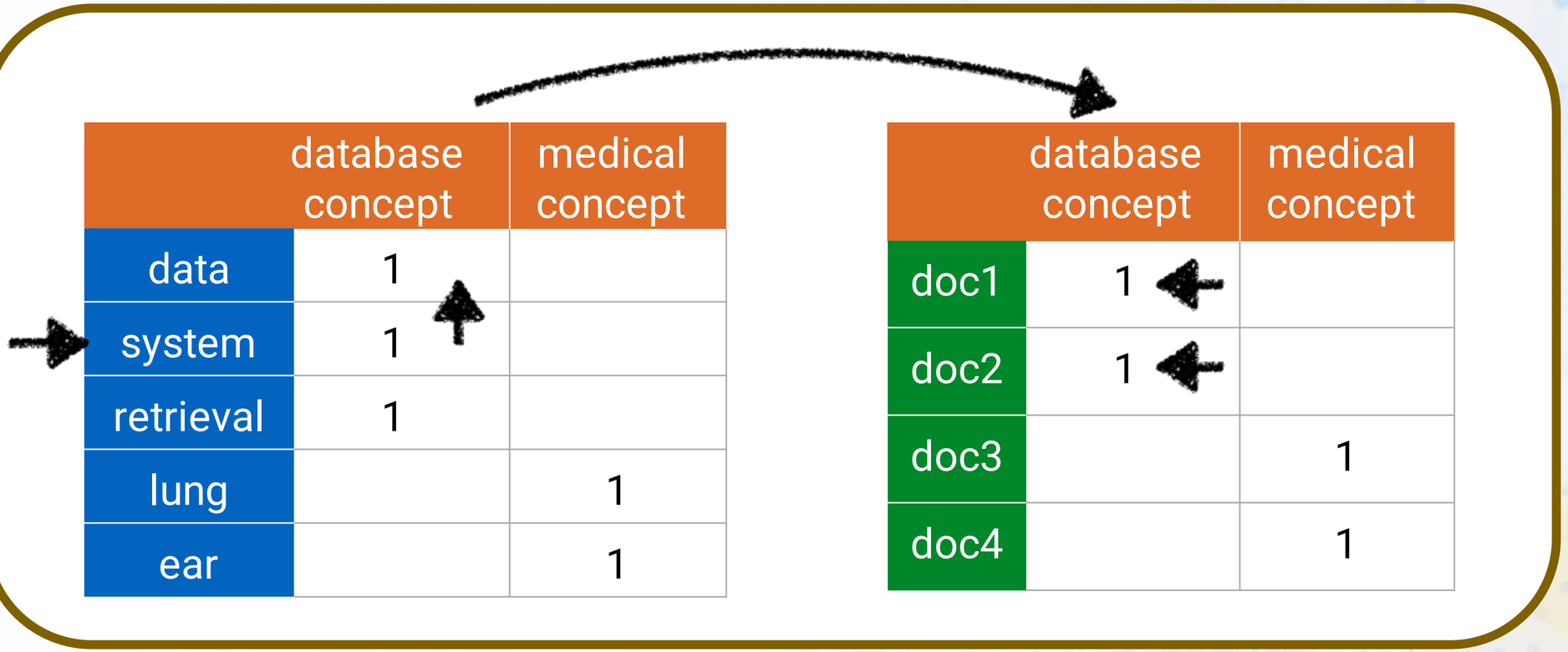
document-concept matrix

	database concept	medical concept
doc1	1	
doc2	1	
doc3		1
doc4		1

Query Using the Concepts

Q: How to search, e.g., for “system”?

A: find the corresponding concept(s); and the corresponding documents

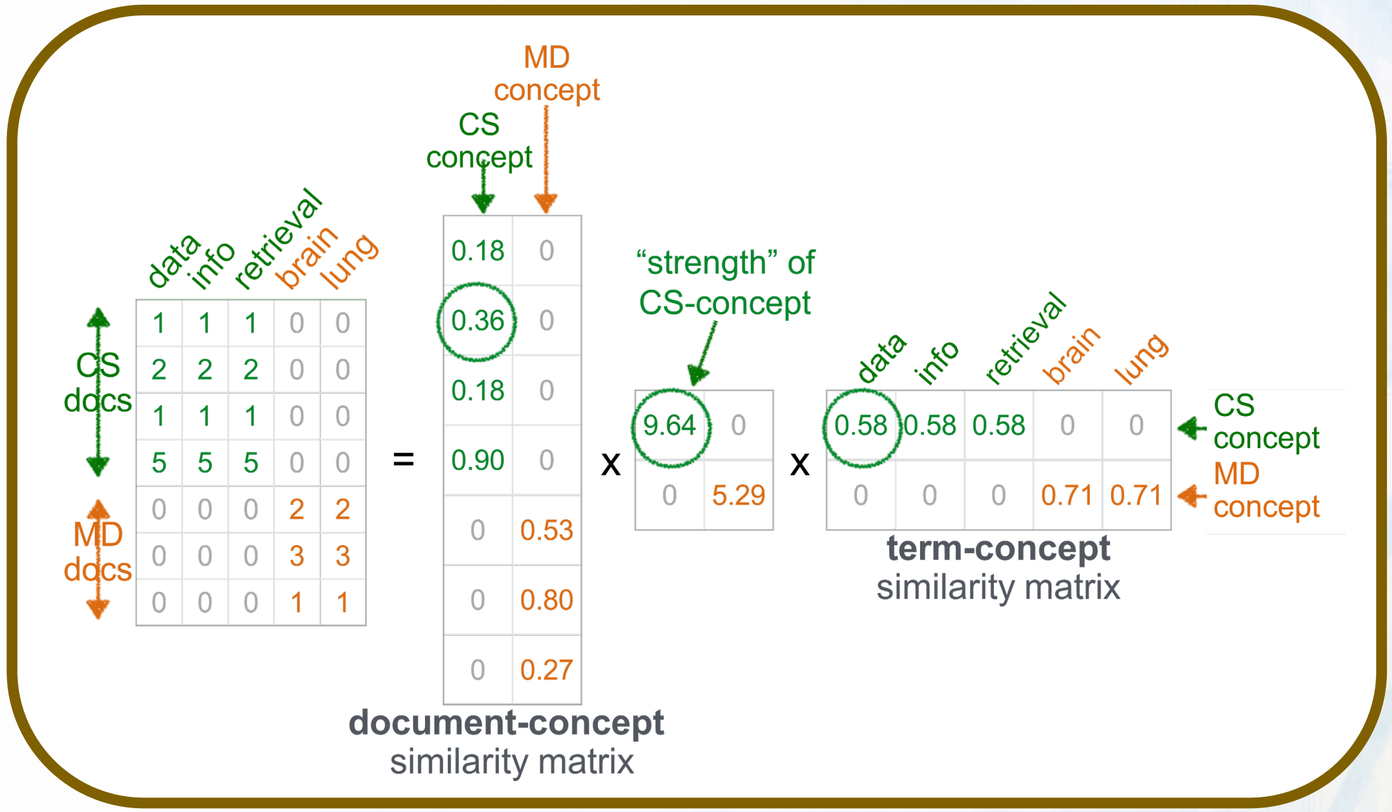


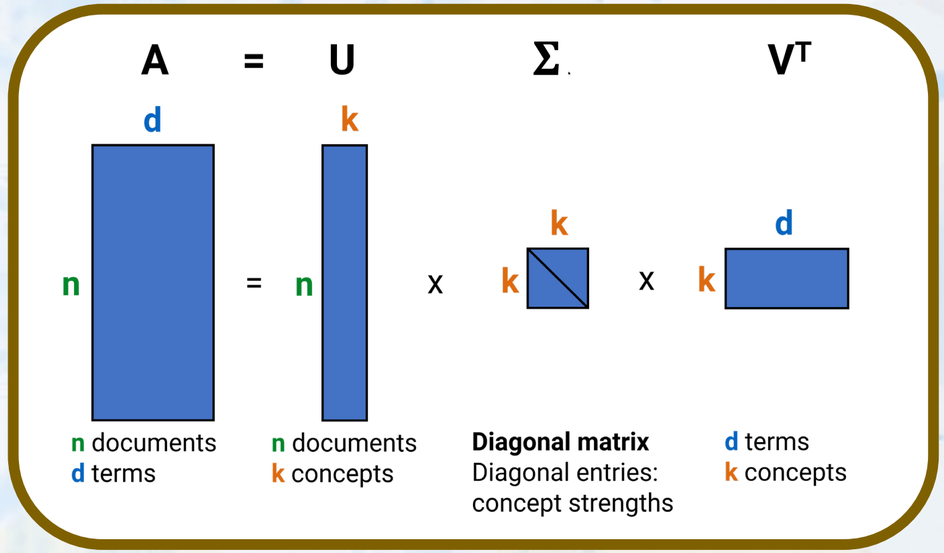
	database concept	medical concept
data	1	
system	1	
retrieval	1	
lung		1
ear		1

	database concept	medical concept
doc1	1	
doc2	1	
doc3		1
doc4		1

We Need SVD to Create the Concepts

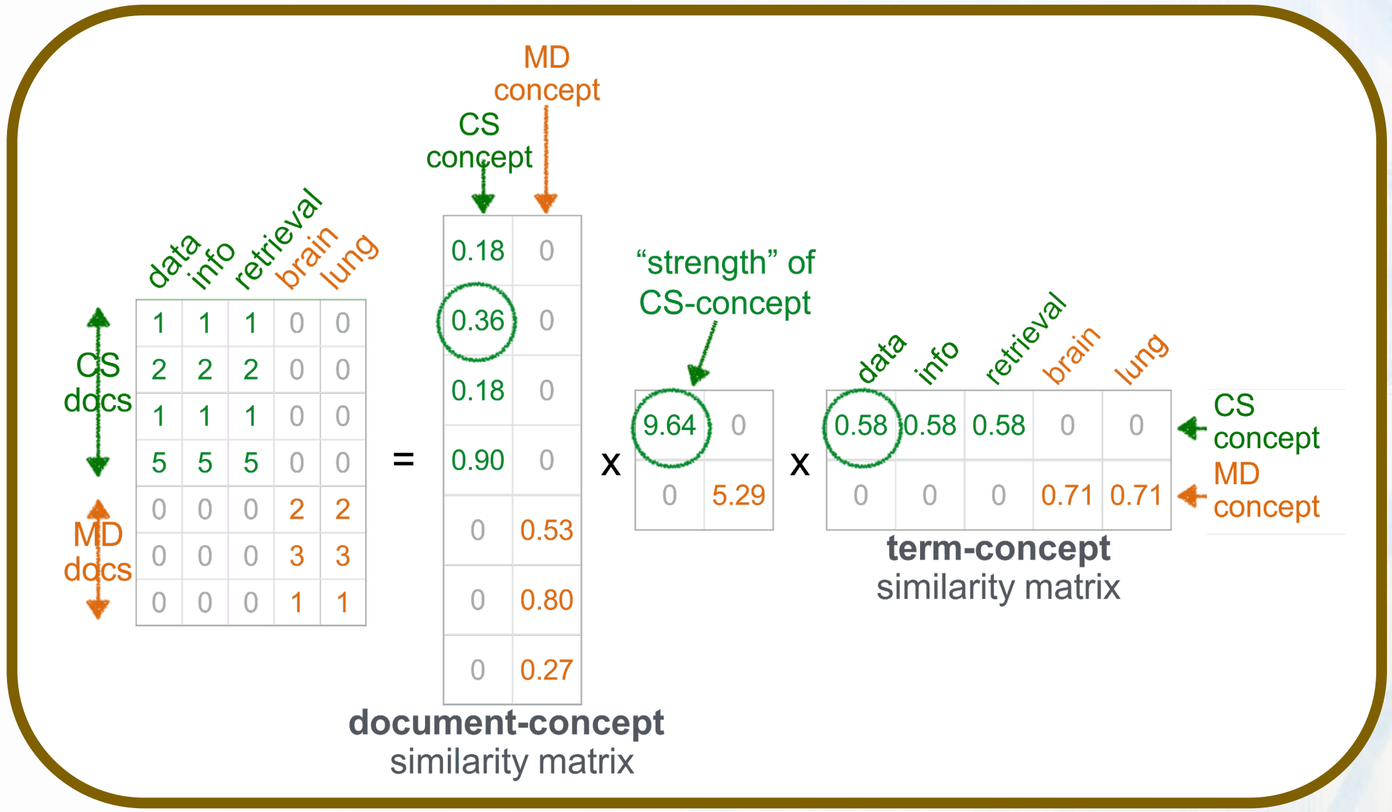
(Very similar to what we had for Dimensionality reduction lecture)

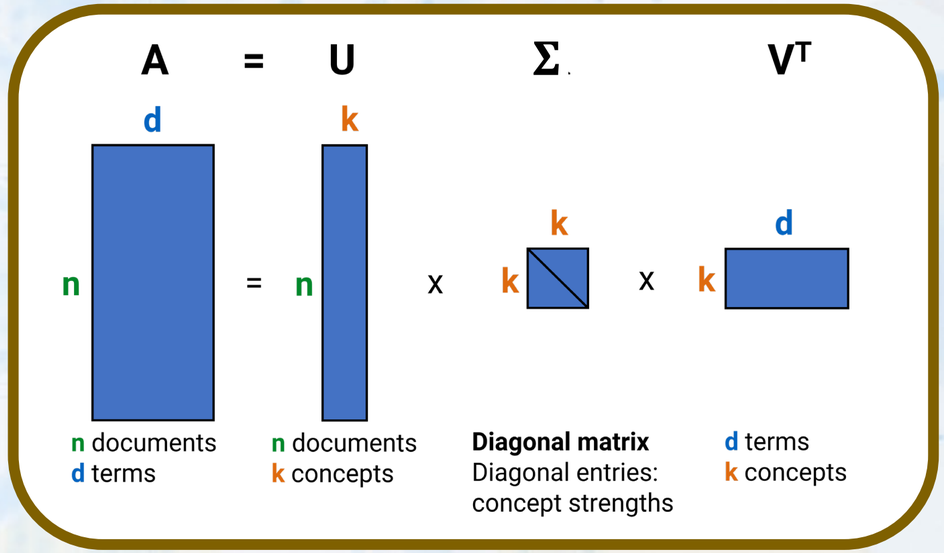




We Need SVD to Create the Concepts

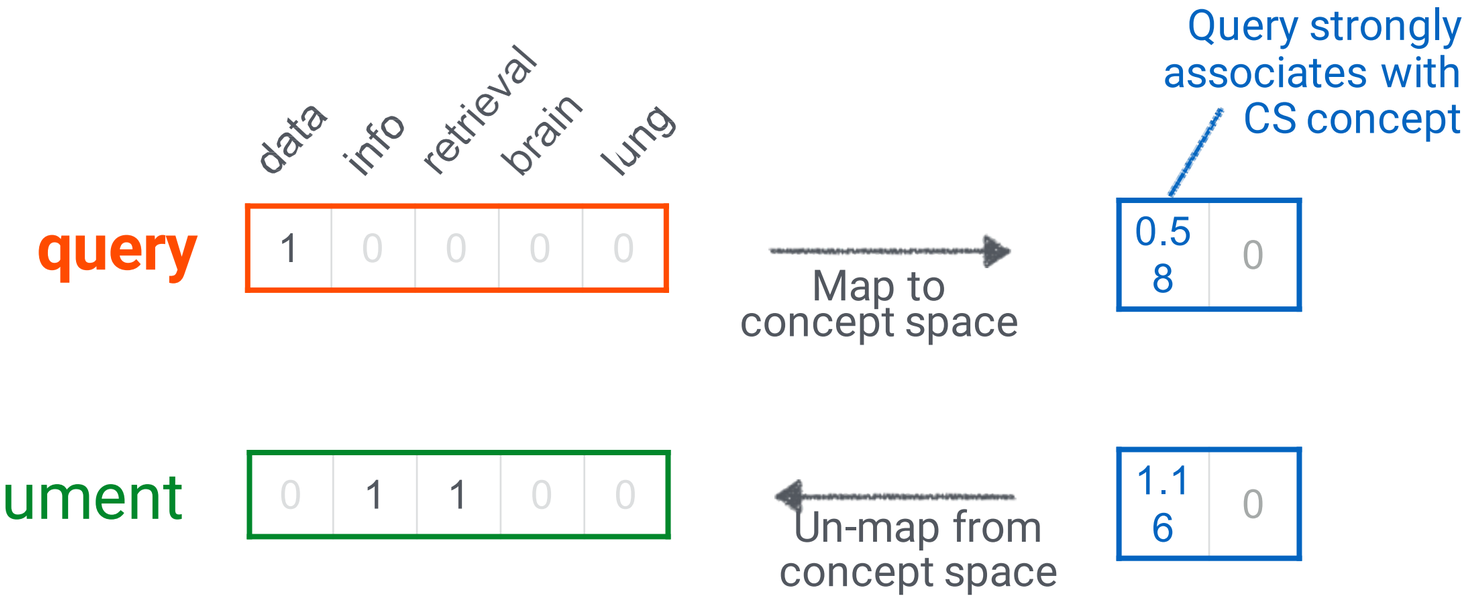
(Very similar to what we had for Dimensionality reduction lecture)





Query on Both Document and Term

Document (‘information’, ‘retrieval’) will be retrieved by **query** (‘data’), even though it does not contain ‘data’!!



Summary

- We learned an unsupervised technique (topic modeling)
- LSI and SVD
- Query