

---

---

# Latent Semantic Analysis

— Boston University CS 506 - Lance Galletti —

---

---

# Latent Semantic Analysis

Inputs are documents. Each word is a feature. We can represent each document by:

- The presence of each word (0 / 1)

	<b>data</b>	<b>information</b>	<b>retrieval</b>	<b>brain</b>	<b>lung</b>
<b>CS-paper-1</b>	1	1	1	0	0

1	1	1	0	0
---	---	---	---	---

X

.58
.58
.58
0
0

term-to-concept similarity

=

1.74

doc-to-concept similarity  
/ CS feature

# Latent Semantic Analysis

Inputs are documents. Each word is a feature. We can represent each document by:

- The presence of each word (0 / 1)
- Count of the word (0, 1, ... )

	<b>data</b>	<b>information</b>	<b>retrieval</b>	<b>brain</b>	<b>lung</b>
<b>CS-paper-1</b>	2	2	2	0	0

2	2	2	0	0
---	---	---	---	---

X

.58
.58
.58
0
0

term-to-concept similarity

=

3.48

doc-to-concept similarity

# Latent Semantic Analysis

	<b>data</b>	<b>information</b>	<b>retrieval</b>	<b>brain</b>	<b>lung</b>
<b>CS-paper-1</b>	1	1	1	0	0
<b>CS-paper-2</b>	2	2	2	0	0
<b>CS-paper-3</b>	1	1	1	0	0
<b>CS-paper-4</b>	5	5	5	0	0
<b>Med-paper-1</b>	0	0	0	2	2
<b>Med-paper-2</b>	0	0	0	3	3
<b>Med-paper-3</b>	0	0	0	1	1



# Latent Semantic Analysis

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

 $=$ 

0.18	0
0.36	0
0.18	0
0.90	0
0	0.53
0	0.80
0	0.27

 $\times$ 

9.64	0
0	5.29

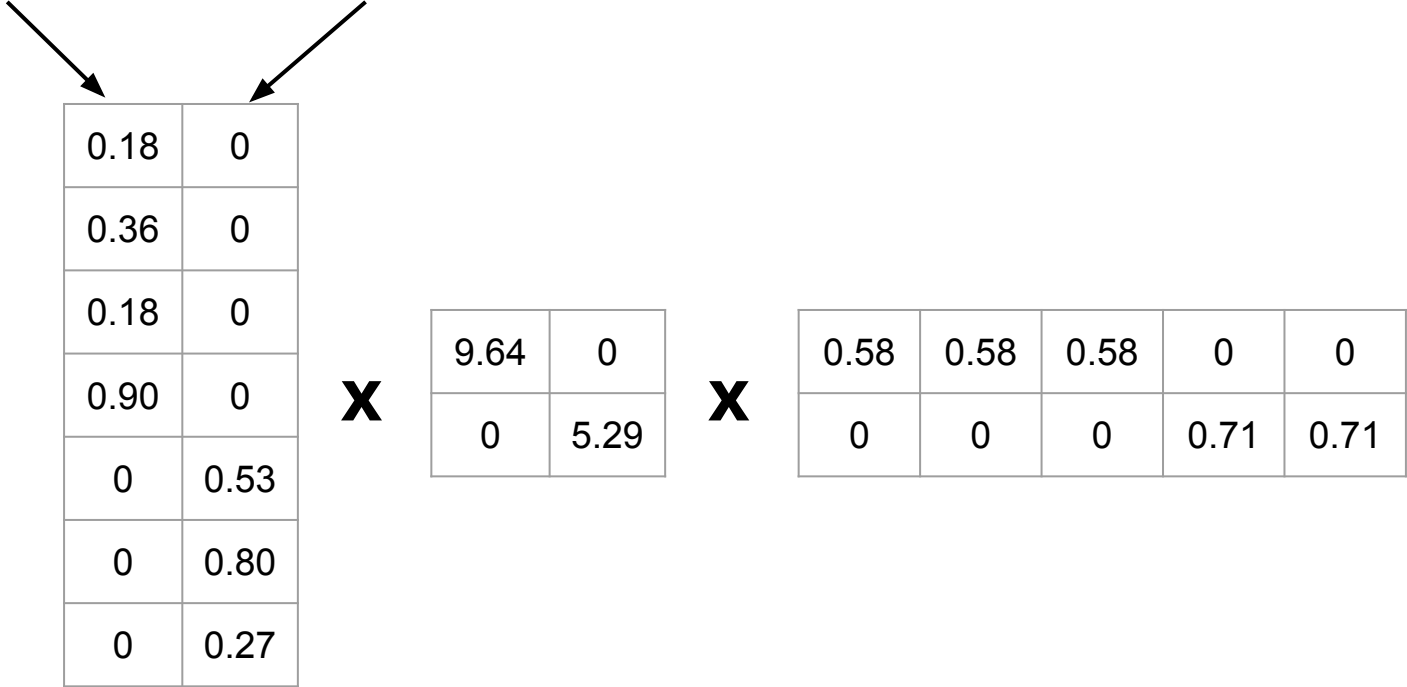
 $\times$ 

0.58	0.58	0.58	0	0
0	0	0	0.71	0.71

# Latent Semantic Analysis

CS concept

MD concept



0.18	0
0.36	0
0.18	0
0.90	0
0	0.53
0	0.80
0	0.27

**X**

9.64	0
0	5.29

**X**

0.58	0.58	0.58	0	0
0	0	0	0.71	0.71

# Latent Semantic Analysis

CS concept

MD concept

doc-to-concept similarity

0.18	0
0.36	0
0.18	0
0.90	0
0	0.53
0	0.80
0	0.27

**X**

9.64	0
0	5.29

**X**

0.58	0.58	0.58	0	0
0	0	0	0.71	0.71

# Latent Semantic Analysis

doc-to-concept  
similarity matrix

0.18	0
0.36	0
0.18	0
0.90	0
0	0.53
0	0.80
0	0.27

**X**

9.64	0
0	5.29

**X**

0.58	0.58	0.58	0	0
0	0	0	0.71	0.71

# Latent Semantic Analysis

doc-to-concept  
similarity matrix

0.18	0
0.36	0
0.18	0
0.90	0
0	0.53
0	0.80
0	0.27

**X**

9.64	0
0	5.29

**X**

0.58	0.58	0.58	0	0
0	0	0	0.71	0.71

"strength" of the CS concept



# Latent Semantic Analysis

doc-to-concept  
similarity matrix

0.18	0
0.36	0
0.18	0
0.90	0
0	0.53
0	0.80
0	0.27

**X**

"strength" of the  
each concept

9.64	0
0	5.29

**X**

0.58	0.58	0.58	0	0
0	0	0	0.71	0.71

# Latent Semantic Analysis

doc-to-concept  
similarity matrix

0.18	0
0.36	0
0.18	0
0.90	0
0	0.53
0	0.80
0	0.27

**X**

"strength" of the  
each concept

9.64	0
0	5.29

**X**

term-to-concept similarity

0.58	0.58	0.58	0	0
0	0	0	0.71	0.71

# Latent Semantic Analysis

doc-to-concept  
similarity matrix

0.18	0
0.36	0
0.18	0
0.90	0
0	0.53
0	0.80
0	0.27

**X**

"strength" of the  
each concept

9.64	0
0	5.29

**X**

term-to-concept similarity  
matrix

0.58	0.58	0.58	0	0
0	0	0	0.71	0.71



# Latent Semantic Analysis

We can better represent each document by:

- Frequency of the word ( $n_i / \sum n_i$ )
- TfiDf

