Latent Semantic Analysis

Boston University CS 506 - Lance Galletti

doc-to-term similarity X term-to-concept similarity = doc-to-concept similarity

doc-to-term similarity X term-to-concept similarity = doc-to-concept similarity

Latent Semantic Analysis

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

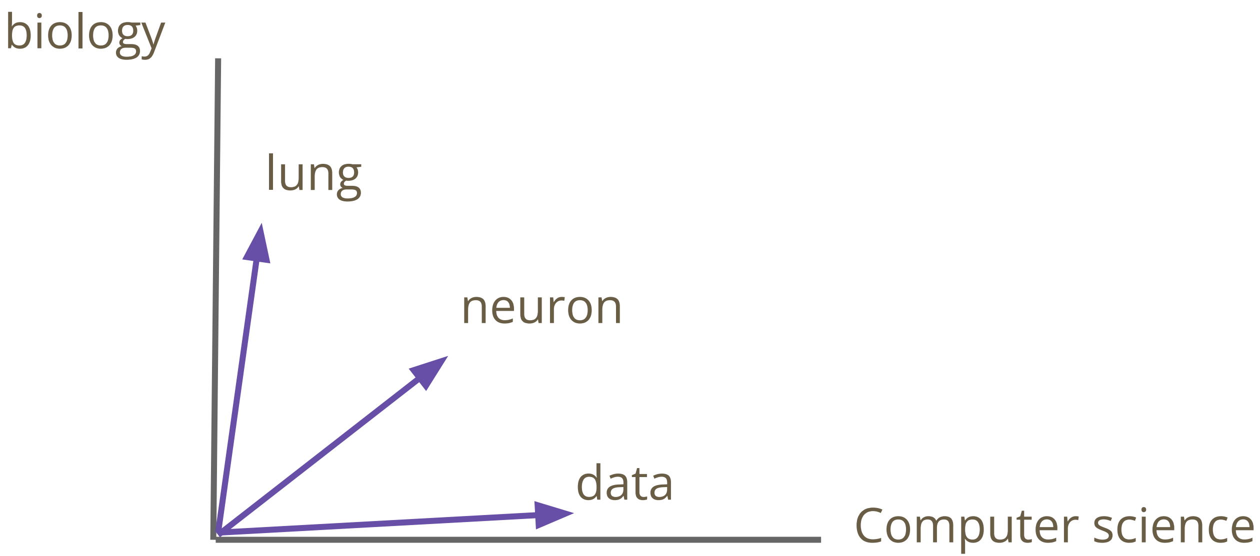
- The presence of each word (0 / 1)

	data	information	retrieval	brain	lung
CS-paper-1	1	1	1	0	0

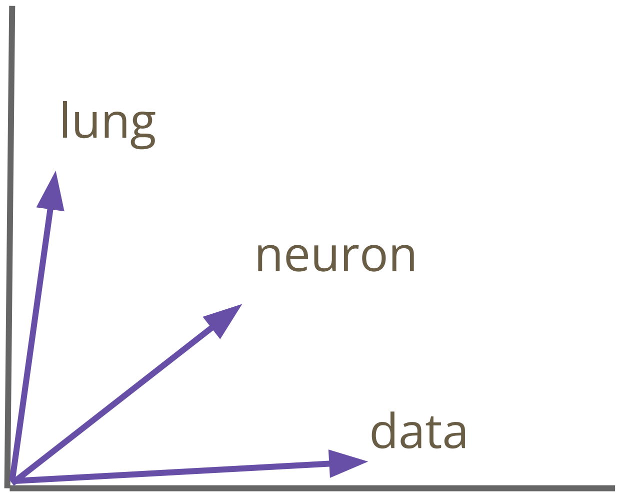
doc-to-term similarity X term-to-concept similarity = doc-to-concept similarity

doc-to-term similarity X term-to-concept similarity = doc-to-concept similarity

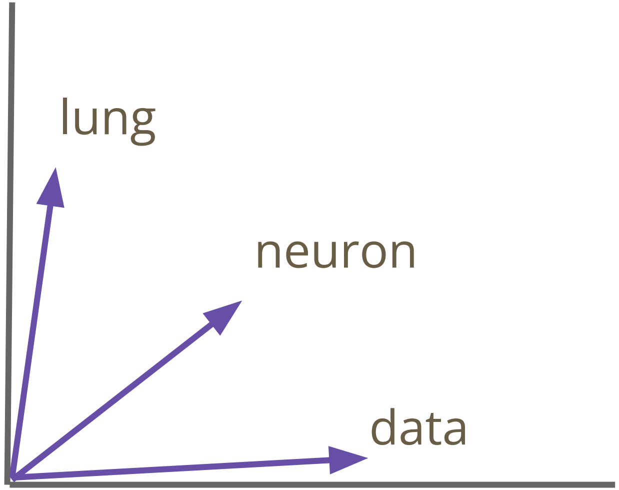
In theory



In practice

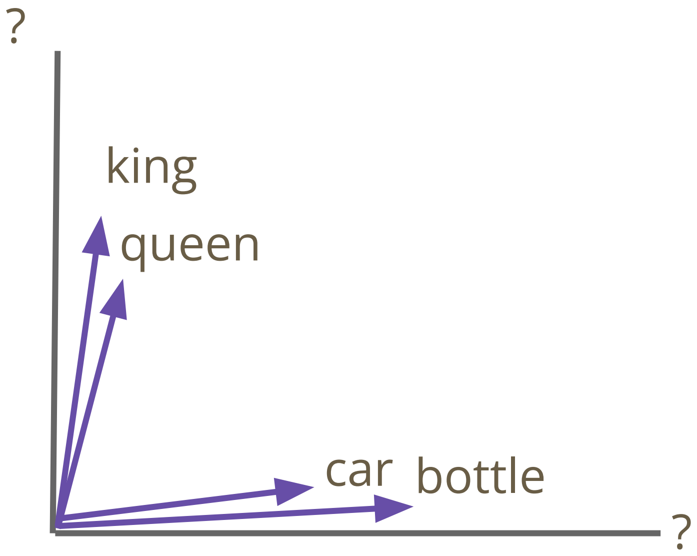


In practice



Words with similar semantic meanings should be close

In actual practice



Words with similar semantic meanings should be close

Lots of ways to generate embeddings. SVD is one of them term-to-concept similarity

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} .58 \\ .58 \\ .58 \\ 0 \\ 0 \end{bmatrix} = 1.74$$

embedding

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, ...)

	data	information	retrieval	brain	lung
CS-paper-1	2	2	2	0	0

term-to-concept similarity

$$\begin{bmatrix} 2 & 2 & 2 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} .58 \\ .58 \\ .58 \\ 0 \\ 0 \end{bmatrix} = 3.48$$

doc-to-concept similarity

Latent Semantic Analysis

	data	information	retrieval	brain	lung
CS-paper-1	1	1	1	0	0
CS-paper-2	2	2	2	0	0
CS-paper-3	1	1	1	0	0
CS-paper-4	5	5	5	0	0
Med-paper-1	0	0	0	2	2
Med-paper-2	0	0	0	3	3
Med-paper-3	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

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

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

Doc to term similarity

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

“strength” of the CS 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

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 / Σn_i)
- TfiDf

$$tf \cdot idf$$

Term frequency in the document

$$\log\left(\frac{\text{number of documents}}{\text{number of documents that contain the term}}\right)$$