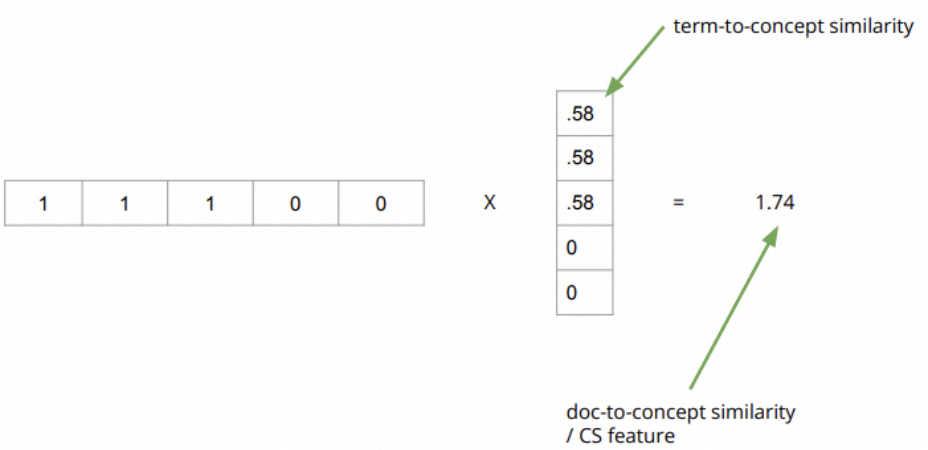
CAS CS 506
Lec 10

Latent Semantic Analysis

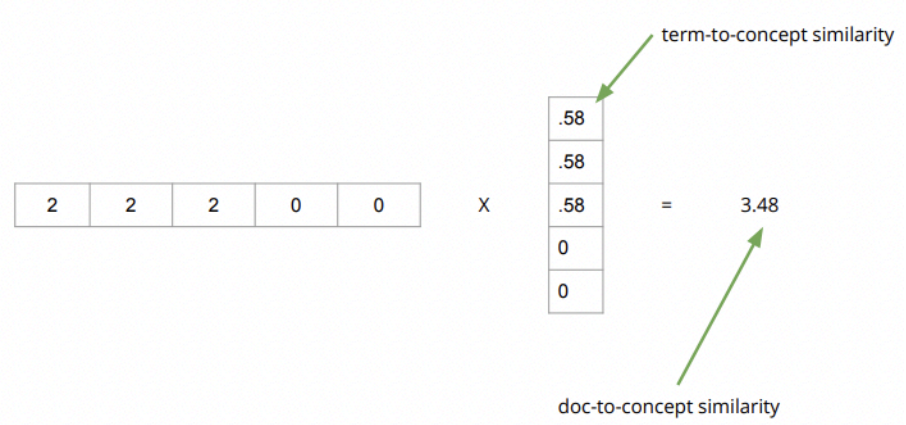
1. Latent Semantic Analysis
    a. Inputs are documents. Each word is a feature. We can represent each document by
        i. The presence of each word (0 / 1)

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



        ii. Count of the word (0, 1, …)

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



2. Latent Semantic Analysis

a.

	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

b.

1	1	1	0	0		0.18	0		9.64	0		0.58	0.58	0.58	0	0
2	2	2	0	0		0.36	0		0	5.29		0	0	0	0.71	0.71
1	1	1	0	0		0.18	0									
5	5	5	0	0	=	0.90	0	x			x					
0	0	0	2	2		0	0.53									
0	0	0	3	3		0	0.80									
0	0	0	1	1		0	0.27									

c.

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

d. Each document can be better represented by
   i. Frequency of the word (n_i / / Σn_i)
   ii. TfiDf
      1. tf: term frequency in the document
      2. idf: log(number of documents / number of documents that contain the term)