Latent Semantic Analysis

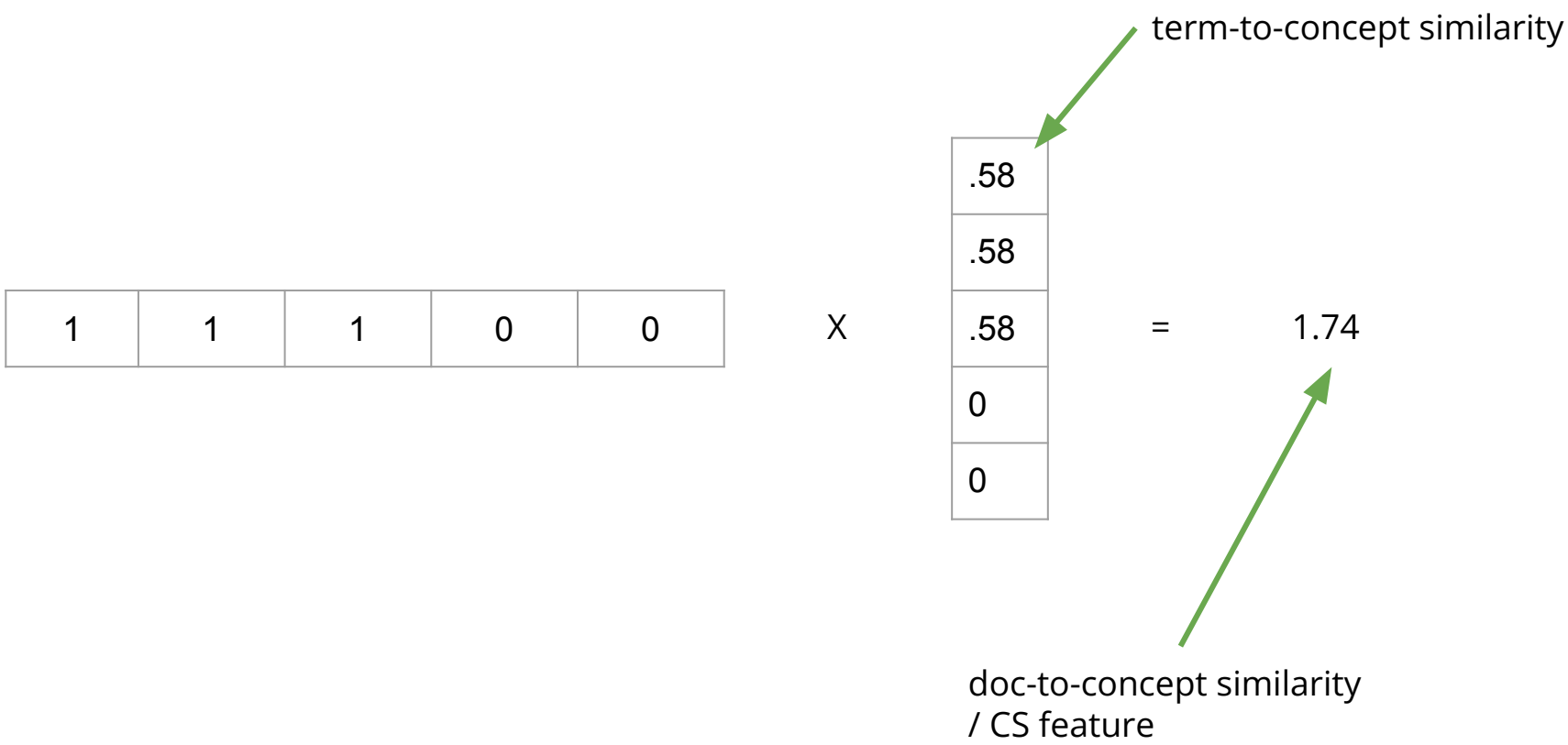
Boston University CS 506 - Lance Galletti

Latent Semantic Analysis

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

- The presence of each word (0 / 1)

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

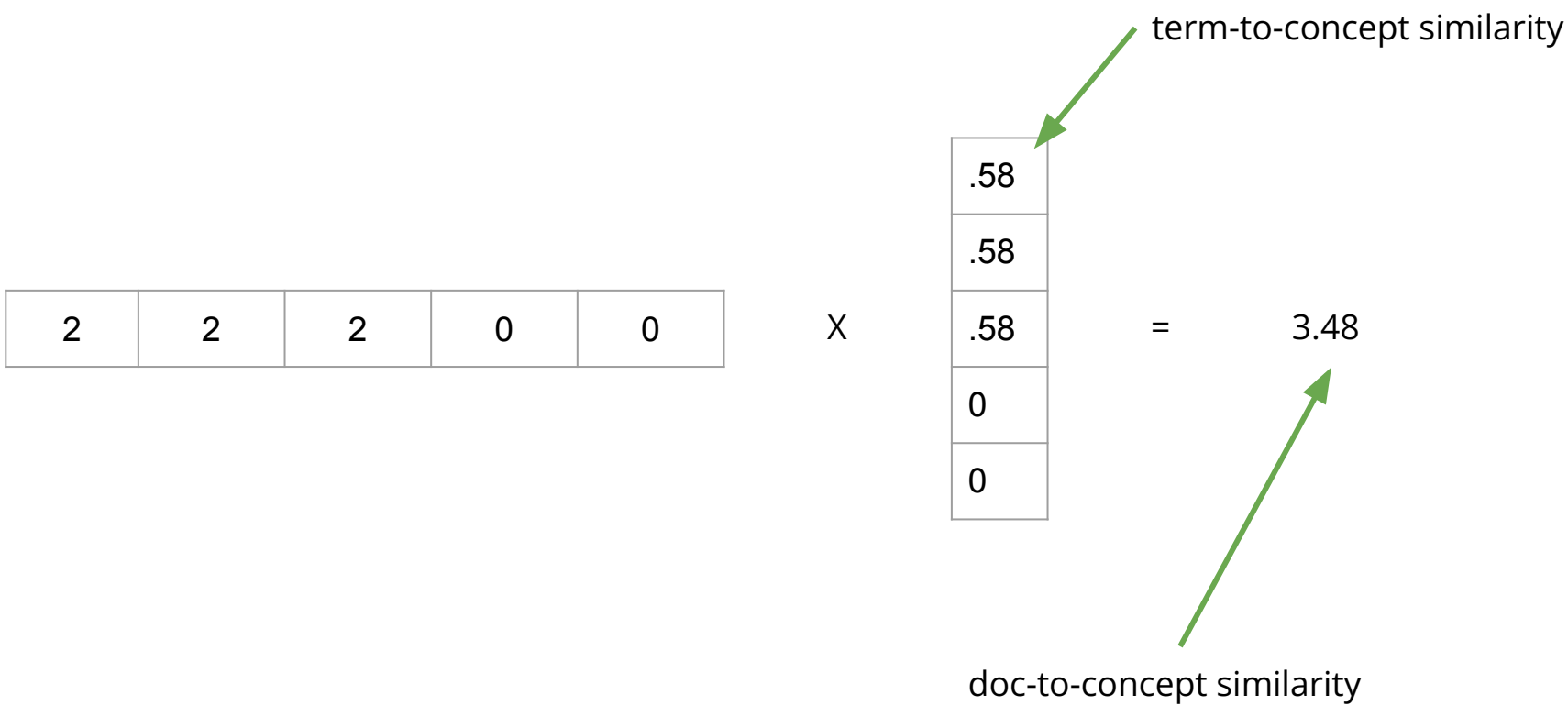

term-to-concept similarity
1
1
1
0
0
X
.58
.58
.58
0
0
=
1.74
doc-to-concept similarity
/ CS feature

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, ...)

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


term-to-concept similarity
2
2
2
0
0
X
.58
.58
.58
0
0
=
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

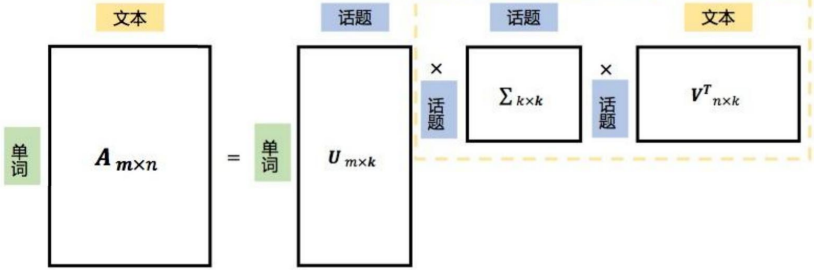
x

0.58	0.58	0.58	0	0
0	0	0	0.71	0.71

潜在语义分析（LSA）：https://www.jianshu.com/p/9fe0a7004560

Latent Semantic Analysis

把大矩阵A进行**截断奇异值分解**，分解成三个小矩阵相乘。假设主题数为k个，LSA的**物理解释**如下：



A_{ij}：单词i在文本j中出现的权值（频率）

$U_{m\times k}$：单词-话题矩阵（**每一列表示一个话题**）

$\Sigma_{k\times k}V_{n\times k}{}^T$：话题-文本矩阵

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

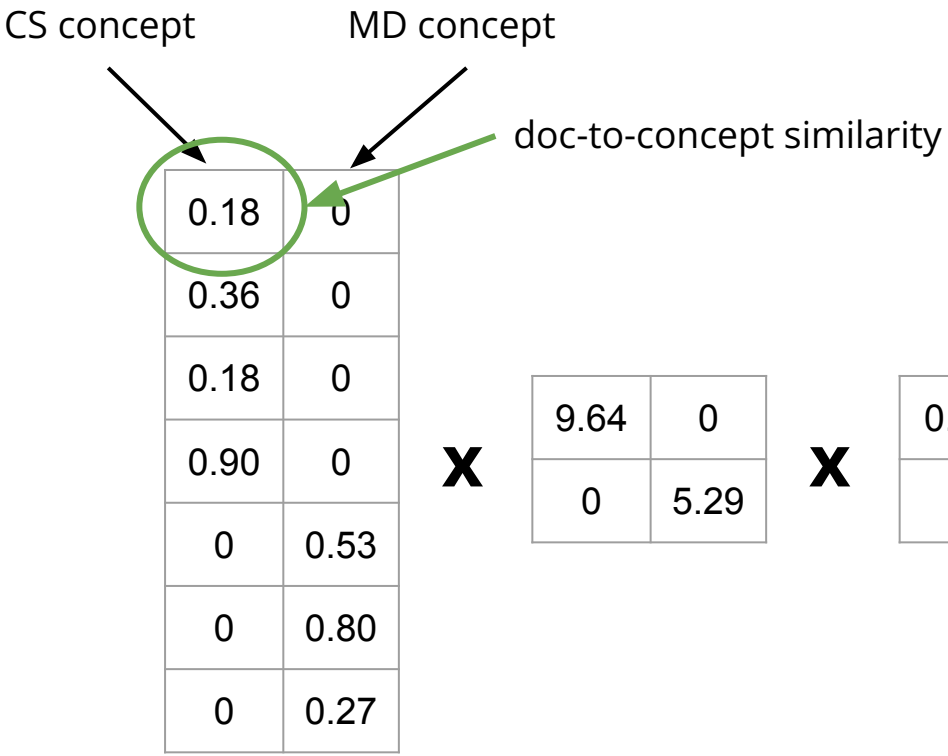
x

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



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

$$词频(TF) = \frac{某个词在文章中的出现次数}{文章的总词数}$$

$$逆文档频率(IDF) = \log\left(\frac{语料库的文档总数}{包含该词的文档数 + 1}\right)$$

$$tf \cdot idf$$

Term frequency in the document

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