Lab 09

文本與詞彙的向量表徵

Document-Term Matrix, Latent Semantic Analysis & Word Vectors

表徵文本
(Representing documents)

How to represent documents numerically?

doc1: “I baked the cake and the muffin”
doc2: “I loved the cake”
doc3: “I wrote the book”

How to represent documents numerically?

doc1: “I baked the **cake** and the **muffin**”
doc2: “I loved the **cake**”
doc3: “I wrote the **book**”

Document-Term Matrix

doc1: “I baked the **cake** and the **muffin**”
doc2: “I loved the **cake**”
doc3: “I wrote the **book**”

	I	baked	loved	wrote	the	and	cake	muffin	book
doc1	1	1	0	0	2	1	1	1	0
doc2	1	0	1	0	1	0	1	0	0
doc3	1	0	0	1	1	0	0	0	1

Document-Term Matrix

doc1: “I baked the **cake** and the **muffin**”
doc2: “I loved the **cake**”
doc3: “I wrote the **book**”

	I	baked	loved	wrote	the	and	cake	muffin	book
doc1	1	1	0	0	2	1	**1**	**1**	0
doc2	1	0	1	0	1	0	**1**	0	0
doc3	1	0	0	1	1	0	0	0	**1**

Representing documents as vectors

doc1: “I baked the **cake** and the **muffin**”
doc2: “I loved the **cake**”
doc3: “I wrote the **book**”

	I	baked	loved	wrote	the	and	cake	muffin	book	
doc1	1	1	0	0	2	1	1	1	0	vector 1
doc2	1	0	1	0	1	0	1	0	0	vector 2
doc3	1	0	0	1	1	0	0	0	1	vector 3

Vectors quantify similarity b/t documents

Euclidean Distance $d(\overrightarrow{p}, \overrightarrow{q}) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_n - q_n)^2}$

Cosine Similarity $cos(\theta) = \frac{\overrightarrow{p} \cdot \overrightarrow{q}}{\| \overrightarrow{p} \| \| \overrightarrow{q} \|}$

	I	baked	loved	wrote	the	and	cake	muffin	book	
doc1	1	1	0	0	2	1	1	1	0	vector 1
doc2	1	0	1	0	1	0	1	0	0	vector 2
doc3	1	0	0	1	1	0	0	0	1	vector 3

Vectors quantify similarity b/t documents

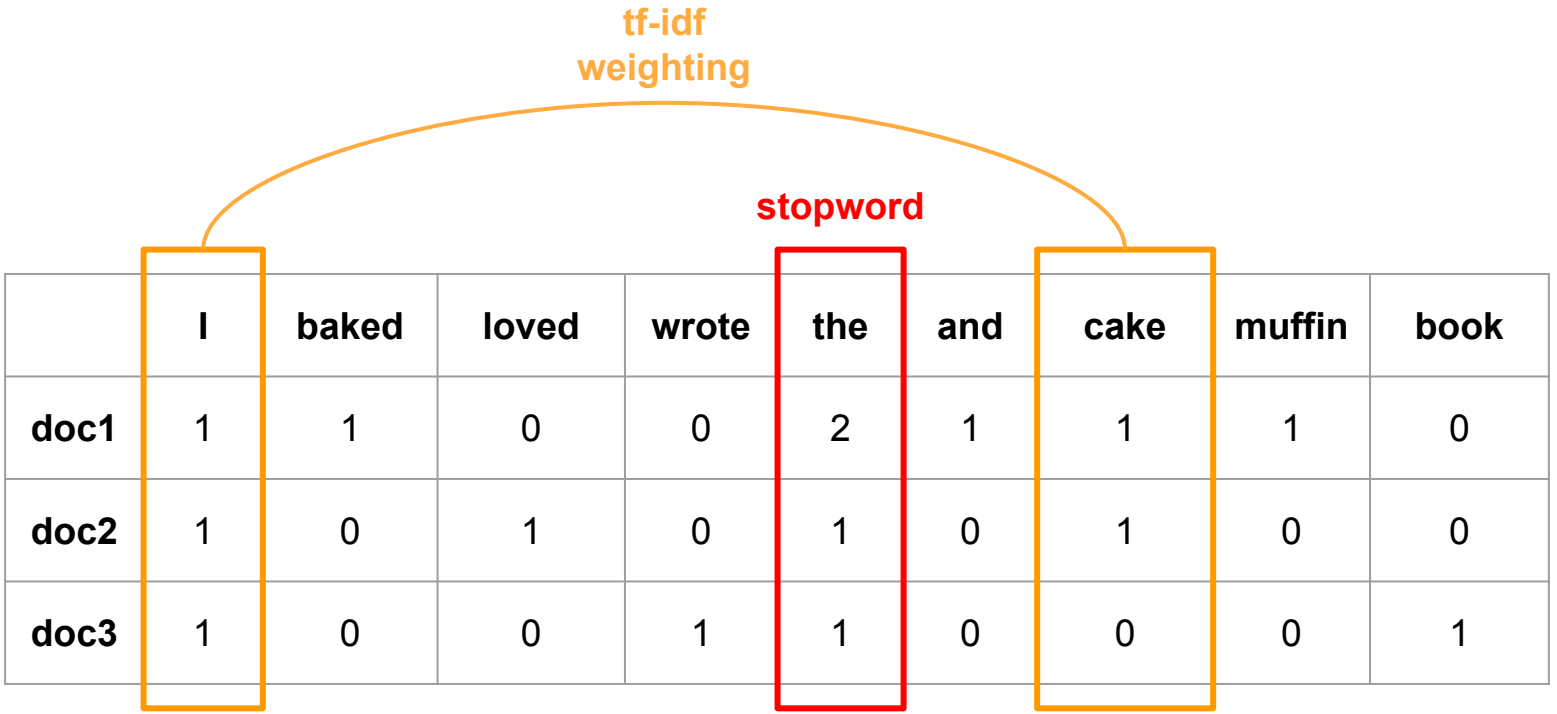
Euclidean Distance $d(\overrightarrow{p}, \overrightarrow{q}) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_n - q_n)^2}$

Cosine Similarity $cos(\theta) = \frac{\overrightarrow{p} \cdot \overrightarrow{q}}{\| \overrightarrow{p} \| \| \overrightarrow{q} \|}$

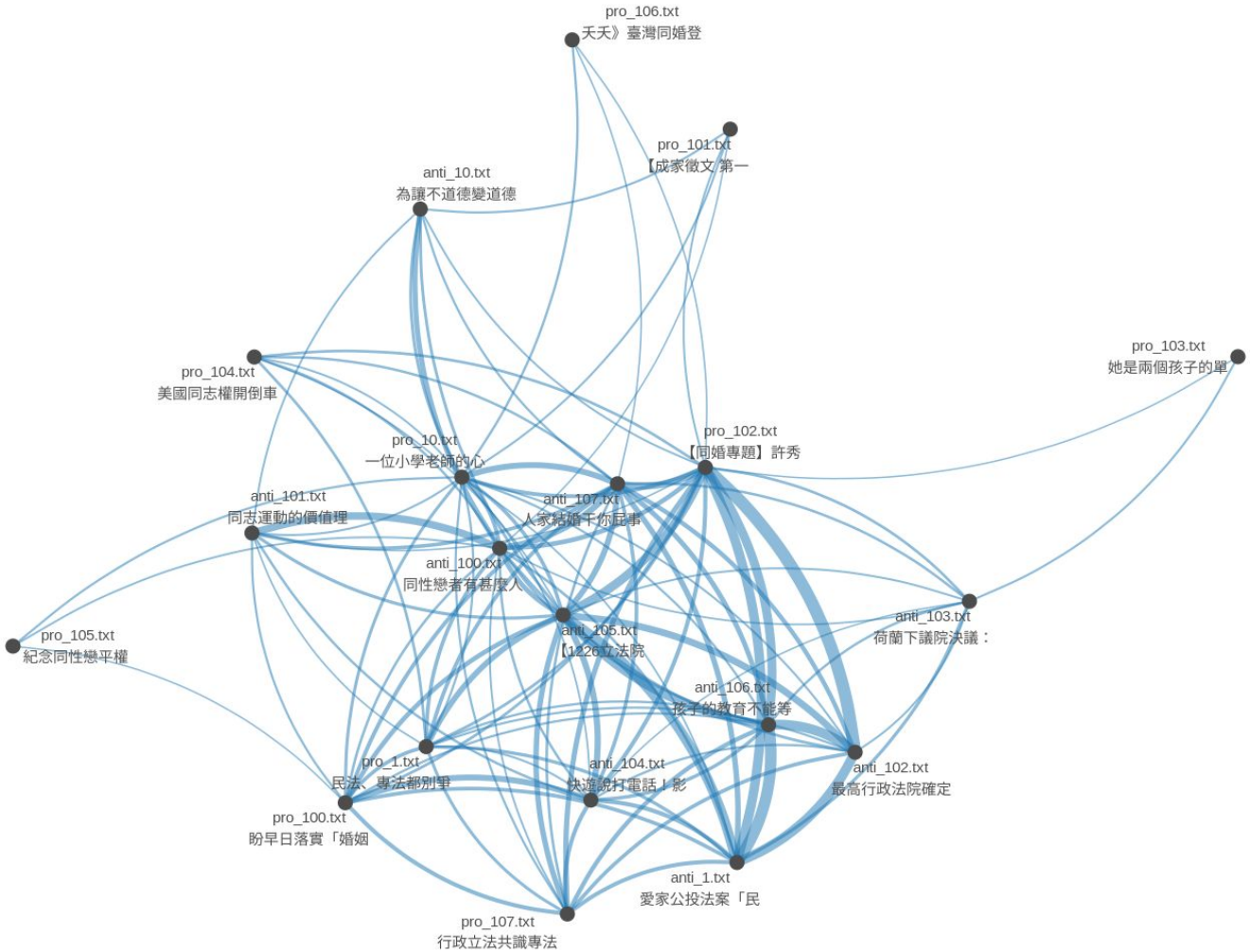
	I	cake	
doc1	1	1	vector 1
doc2	1	1	vector 2
doc3	1	0	vector 3

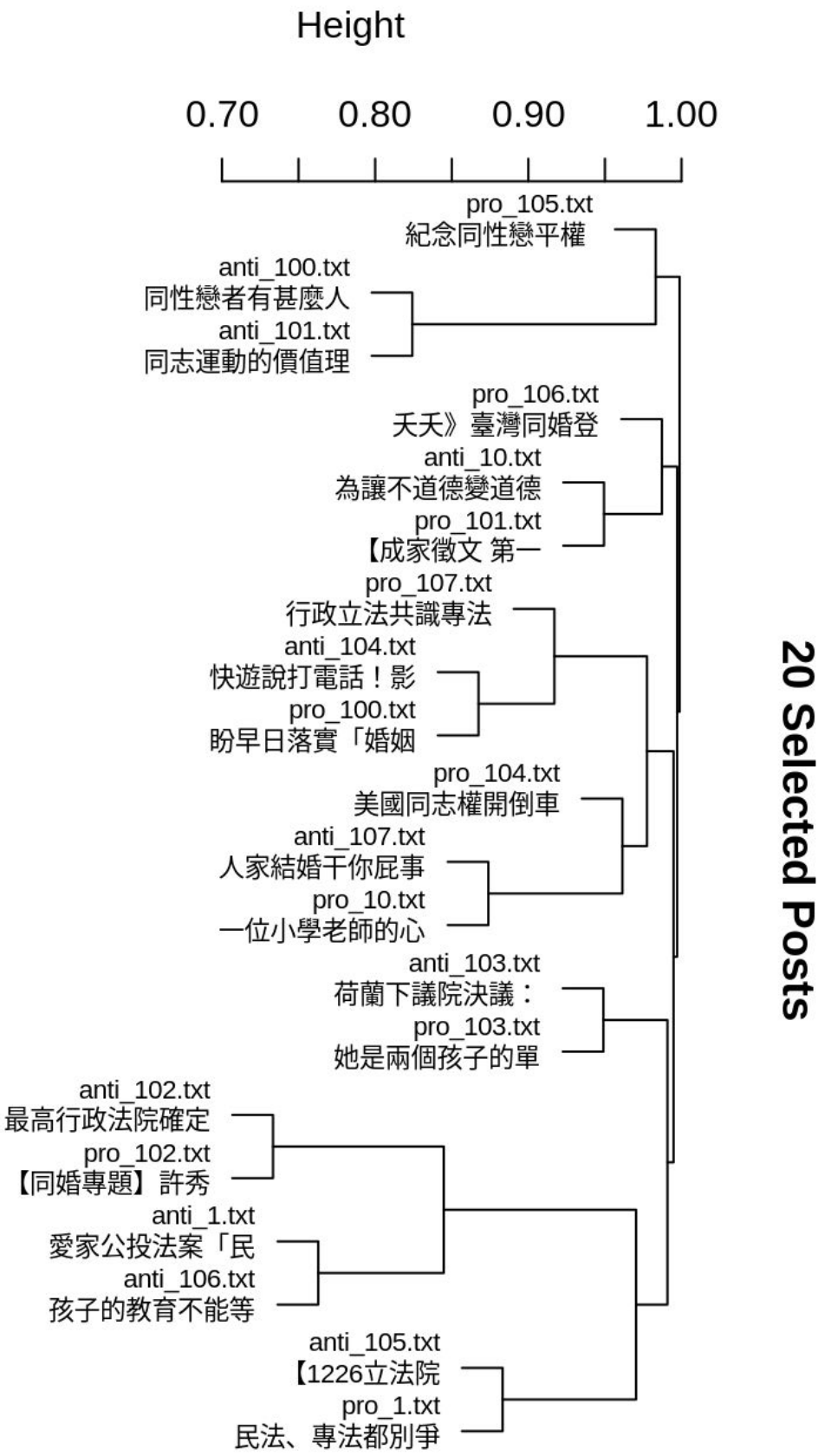
	$cos(\theta)$	$d(\overrightarrow{p}, \overrightarrow{q})$
doc1, doc2	1	0
doc1, doc3	1/√2	1
doc2, doc3	1/√2	1

Selecting & Weighting Terms



	I	baked	loved	wrote	the	and	cake	muffin	book
doc1	1	1	0	0	2	1	1	1	0
doc2	1	0	1	0	1	0	1	0	0
doc3	1	0	0	1	1	0	0	0	1


pro_106.txt
天夭》臺灣同婚登
pro_101.txt
【成家徵文 第一
anti_10.txt
為讓不道德變道德
pro_104.txt
美國同志權開倒車
pro_103.txt
她是兩個孩子的單
pro_10.txt
一位小學老師的心
pro_102.txt
【同婚專題】許秀
anti_101.txt
同志運動的價值理
anti_107.txt
人家結婚干你屁事
anti_100.txt
同性戀者有甚麼人
anti_103.txt
荷蘭下議院決議：
pro_105.txt
紀念同性戀平權
anti_105.txt
【1226立法院
anti_106.txt
孩子的教育不能等
pro_1.txt
民法、專法都別爭
anti_104.txt
快遞說打電話！影
anti_102.txt
最高行政法院確定
pro_100.txt
盼早日落實「婚姻
anti_1.txt
愛家公投法案「民
pro_107.txt
行政立法共識專法

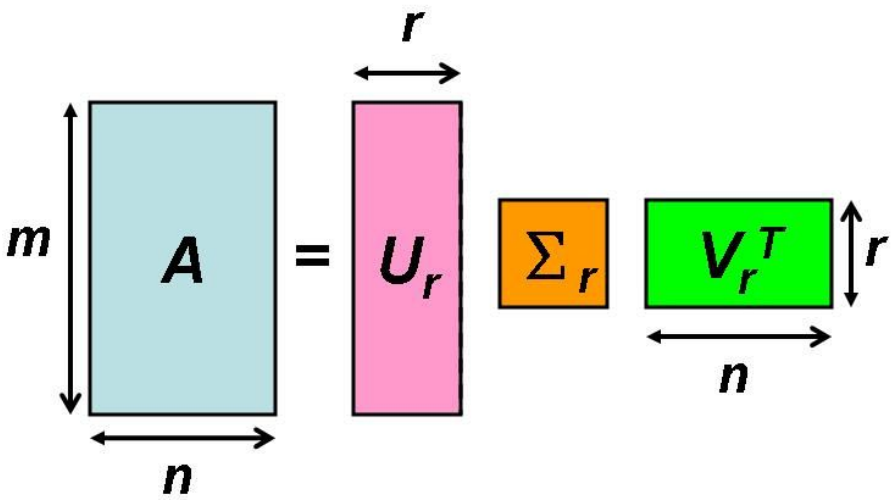

Height
0.70
0.80
0.90
1.00
pro_105.txt
紀念同性戀平權
anti_100.txt
同性戀者有甚麼人
anti_101.txt
同志運動的價值理
pro_106.txt
夭夭》臺灣同婚登
anti_10.txt
為讓不道德變道德
pro_101.txt
【成家徵文 第一
pro_107.txt
行政立法共識專法
anti_104.txt
快遊說打電話！影
pro_100.txt
盼早日落實「婚姻
pro_104.txt
美國同志權開倒車
anti_107.txt
人家結婚干你屁事
pro_10.txt
一位小學老師的心
anti_103.txt
荷蘭下議院決議：
pro_103.txt
她是兩個孩子的單
anti_102.txt
最高行政法院確定
pro_102.txt
【同婚專題】許秀
anti_1.txt
愛家公投法案「民
anti_106.txt
孩子的教育不能等
anti_105.txt
【1226立法院
pro_1.txt
民法、專法都別爭
20 Selected Posts

Dimensionality Reduction (SVD / LSA)

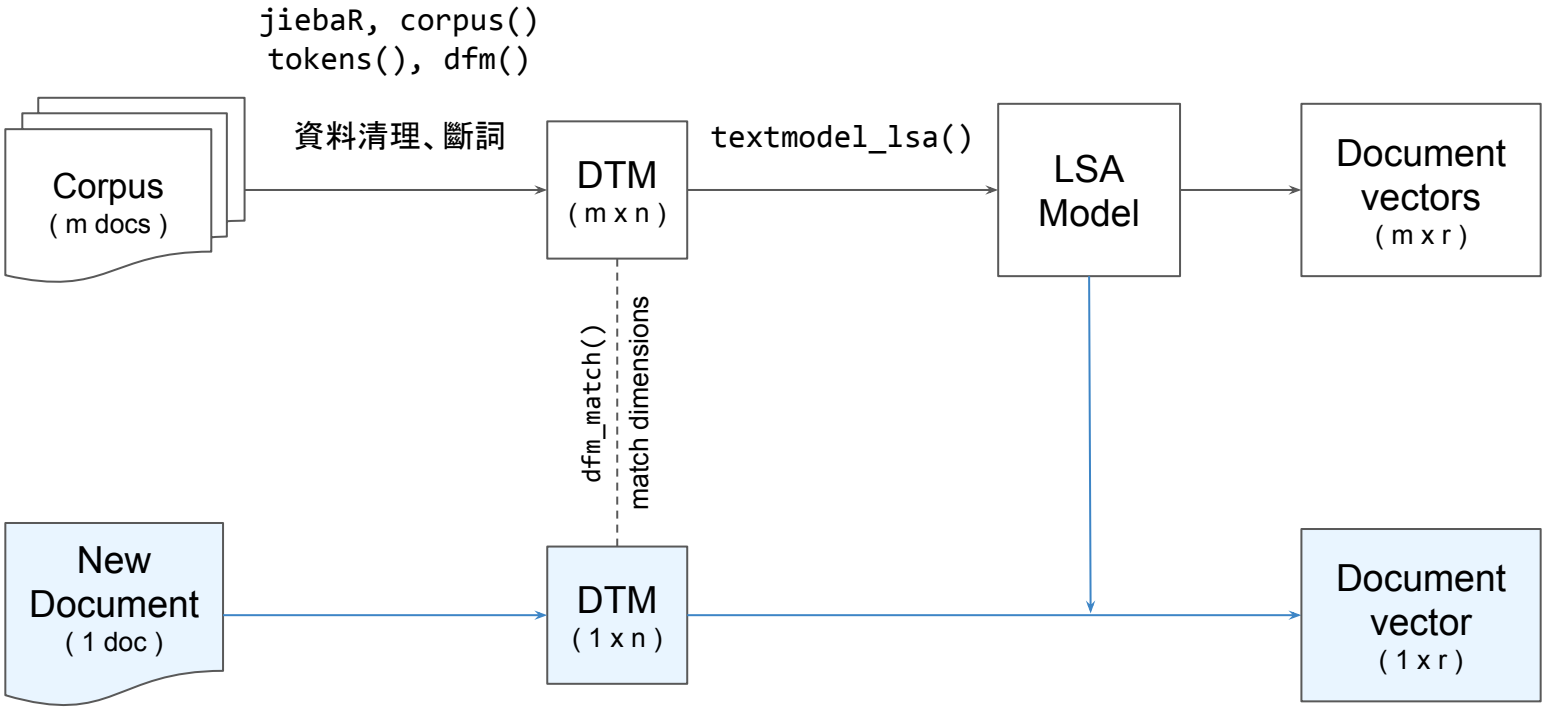
doc2: “I loved the **cake**”
doc4: “He liked my **muffin**”

	I	**He**	**liked**	**loved**	**the**	**my**	**cake**	**muffin**
doc2	1	0	0	1	1	0	1	0
doc4	0	1	1	0	0	1	0	1

Dimensionality Reduction (SVD / LSA)



圖片來源: https://ccjou.wordpress.com/2009/09/01/奇異值分解-svd/

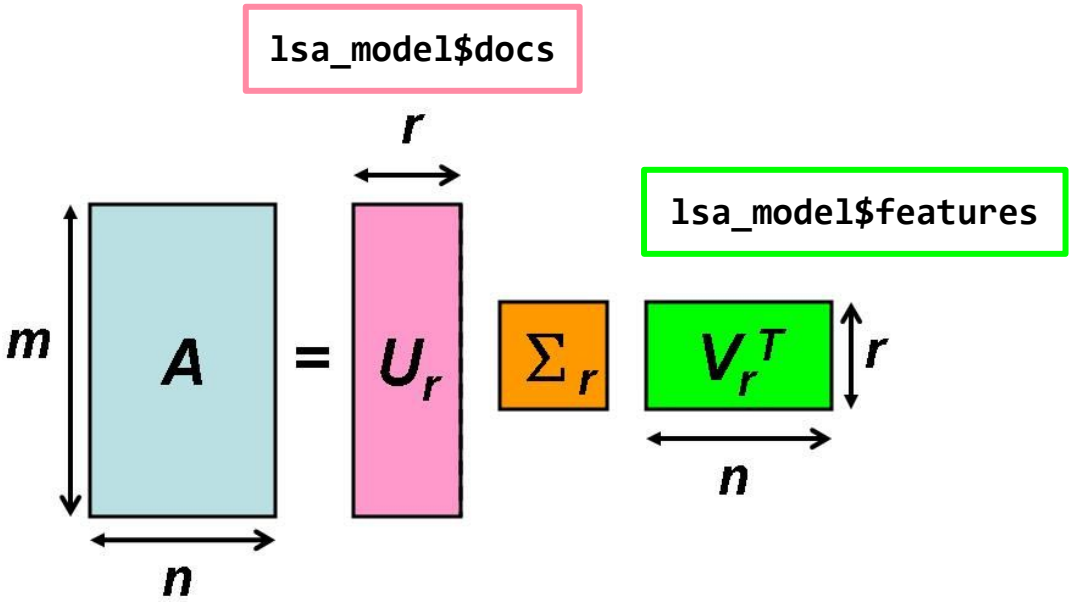

jiebaR, corpus()
tokens(), dfm()
資料清理、斷詞
Corpus
(m docs)
DTM
(m x n)
textmodel_lsa()
LSA
Model
Document
vectors
(m x r)
dfm_match()
match dimensions
New
Document
(1 doc)
DTM
(1 x n)
Document
vector
(1 x r)

表徴詞集
(Representing words)

Document-Term Matrix

	I	baked	loved	wrote	the	and	cake	muffin	book
doc1	1	1	0	0	2	1	1	1	0
doc2	1	0	1	0	1	0	1	0	0
doc3	1	0	0	1	1	0	0	0	1

LSA Word Vectors



圖片來源: https://ccjou.wordpress.com/2009/09/01/奇異值分解-svd/

Word Vectors Trained From Large Raw Text

- PPMI (count-based)
- GloVe (count-based)
- fasttext (neural network)
- word2vec (neural network)

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. [Link]