

R 語言 詞嵌入模型

王貿

國立臺灣大學行為與資料科學研究中心助理研究員

國立臺灣大學政治學系博士、兼任講師

maowang01@gmail.com

課程主題重點

- 詞嵌入 (word embedding) 的基本概念
(又稱word to vector, word2vec)
- 詞共現矩陣 (term co-occurrence matrix, tcm)
- 詞向量空間 (vector spaces)
- 餘弦相似性 (cosine similarity)

- You shall know **a word by the company it keeps.**

(John R. Firth, 1957)



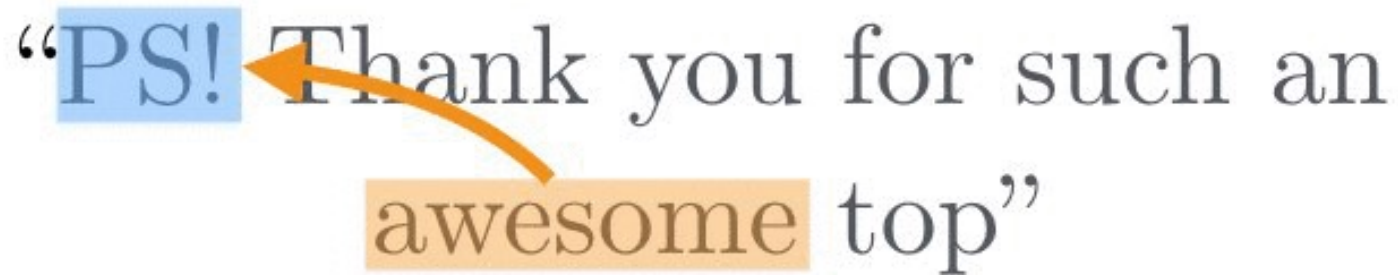
DTM的邏輯



詞嵌入的邏輯 (TCM | FCM)

word2vec

“PS! Thank you for such an
awesome top”



Window size

 : Center Word

 : Context Word

c=0 The cute  jumps over the lazy dog.

c=1 The    over the lazy dog.

c=2      the lazy dog.

文字矩陣 (DTM, TCM)

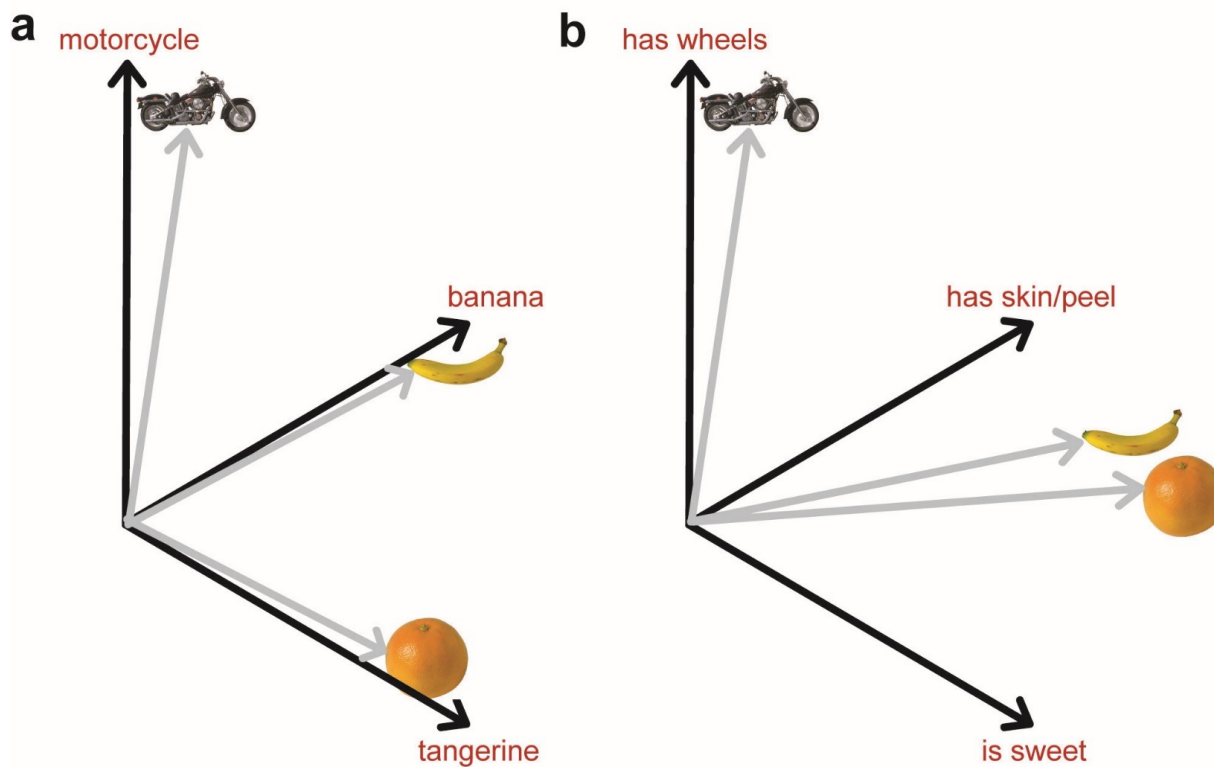
Document Term Matrix (DTM)

	reduce health	policy	food choice	study sodium	social	...
Document 1	1		1			
Document 2		1				
Document 3			2			
Document 4		2			1	
Document 5	1		1		3	
...						

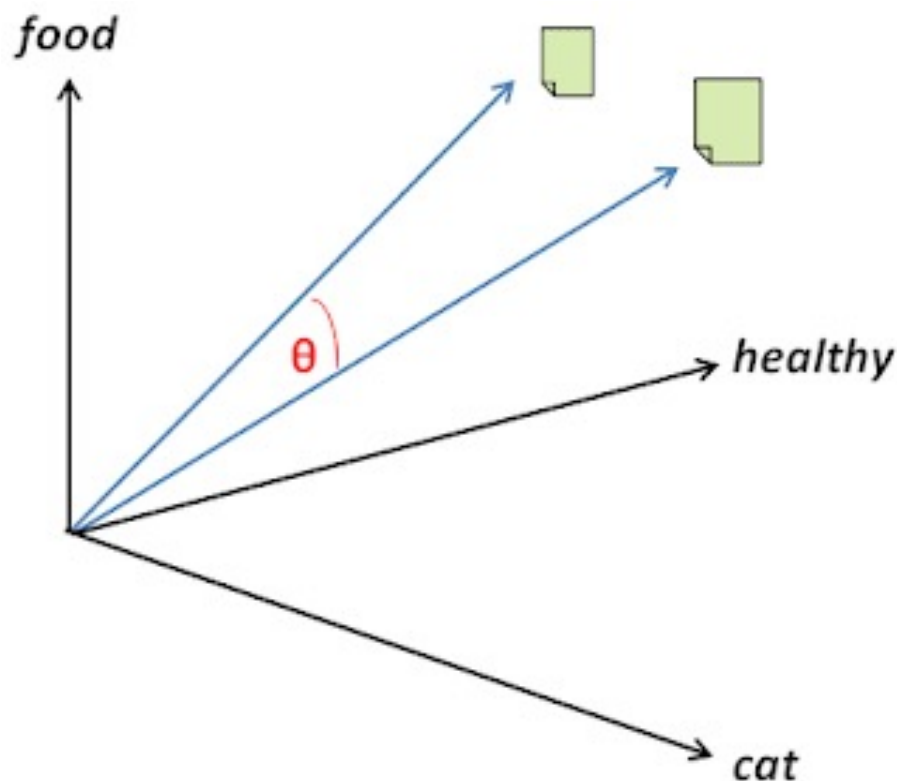
Term Co-occurrence Matrix (TCM)

	reduce health	policy	food choice	study sodium	social	...
reduce health		1	1			
policy	2				1	
food choice		2		1		
study sodium		2			1	
social	1			3		
...						

詞向量空間 (vector spaces)



餘弦相似性 (cosine similarity)



這張圖是顯示DTM的餘弦相似性，代表文件的相似性。如果是使用TCM，就變成是字詞的餘弦相似性。