R 語言
主題模型

王貿

國立臺灣大學行為與資料科學研究中心助理研究員

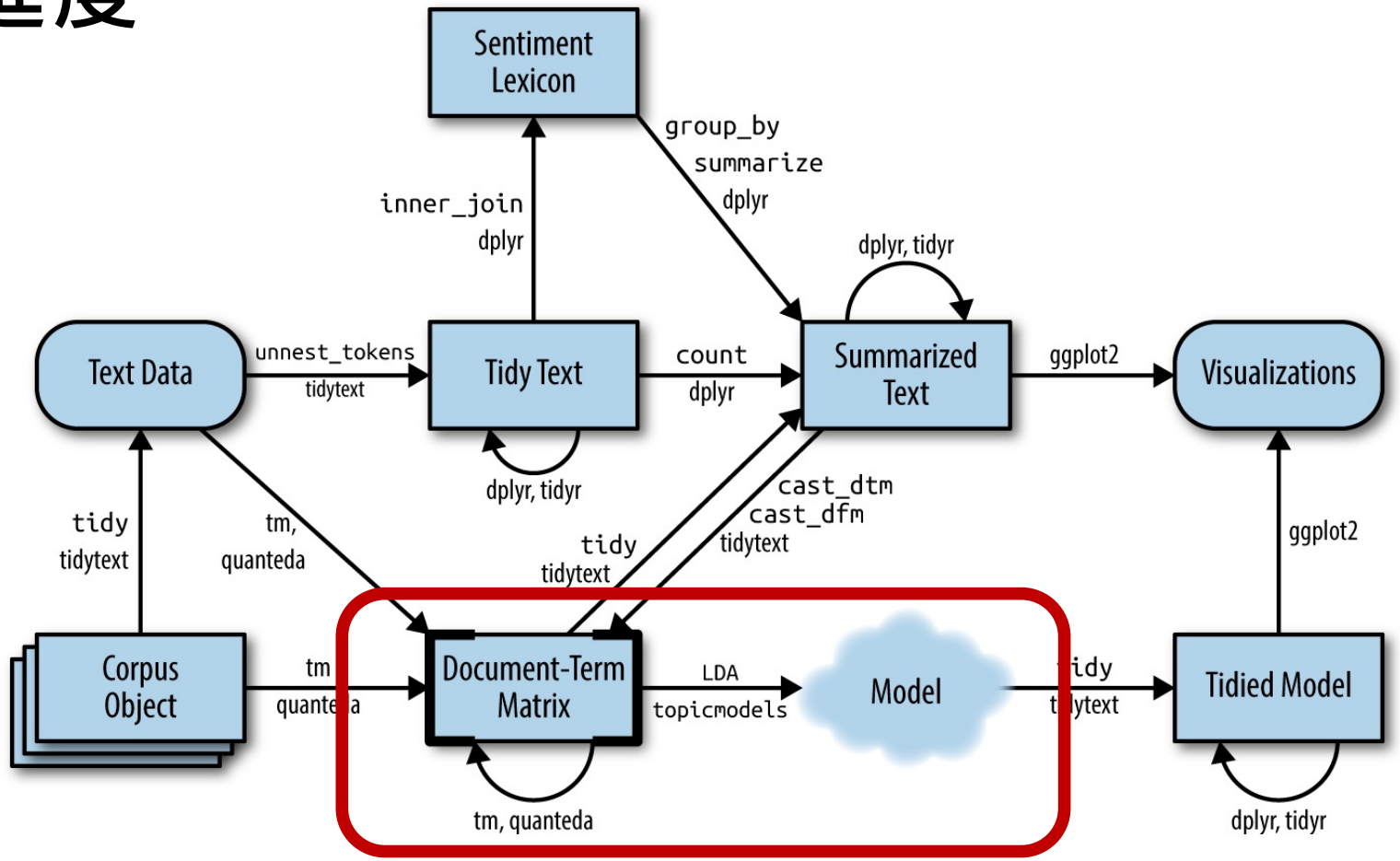
國立臺灣大學政治學系博士、兼任講師

maowang01@gmail.com

課程主題重點

- 常用的主題模型（topic modeling）
- 改良版的主題模型
- 怎麼決定主題數？

目前進度



https://www.tidytextmining.com/images/tmwr_0601.png

主題模型概覽

- 機器學習：
  監督式學習vs. 非監督式學習

- 從大量的文字資料中，分析出其中討論的主題。
- 主題模型就屬於「非監督式學習」

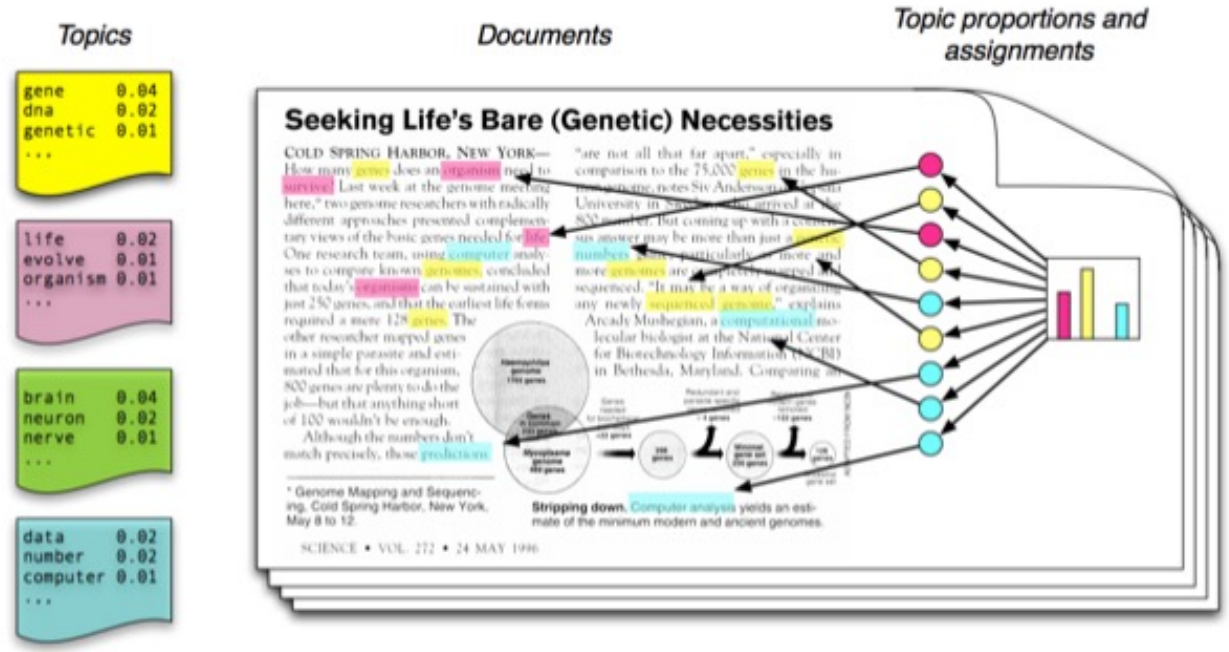
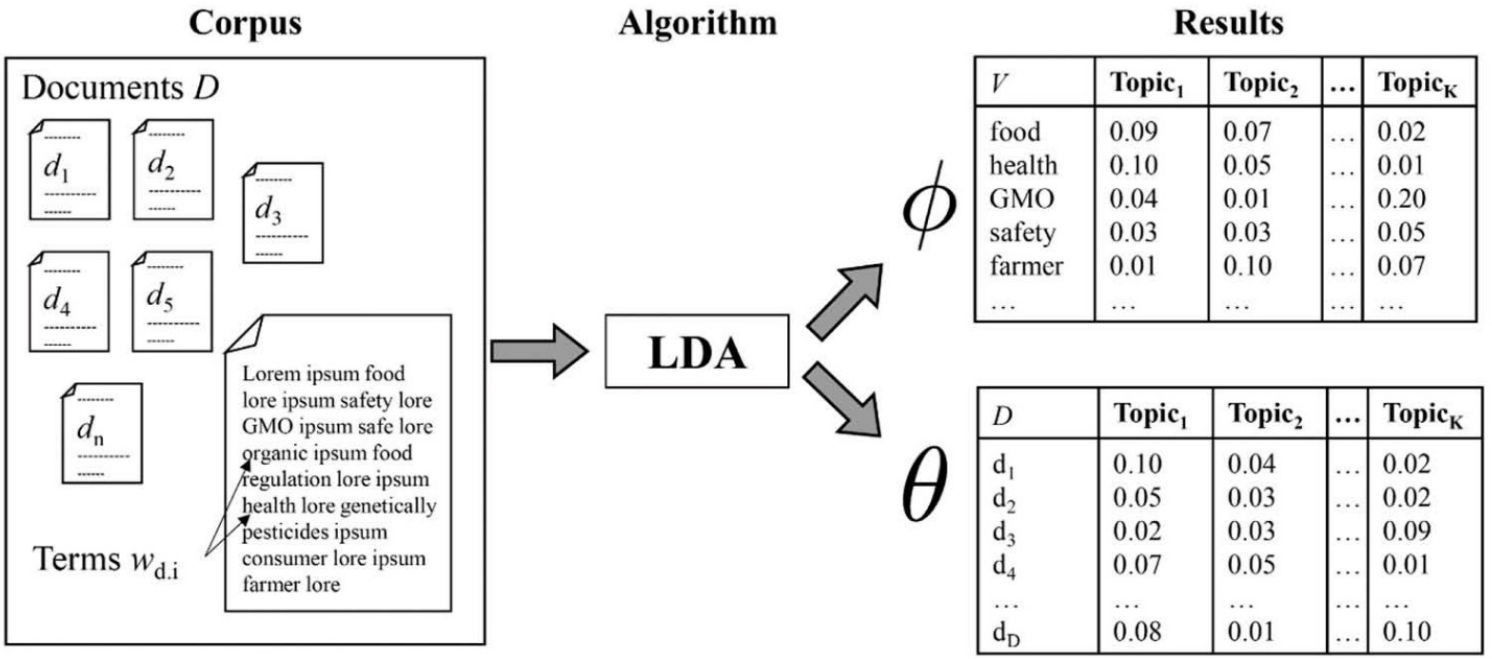


Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

主題模型是什麼？

- 應用「潛在狄利克雷分布」（Latent Dirichlet allocation, LDA）
- 主題模型是一個「**分群**」（clustering）演算法。
- 主題模型與一般的分群演算法（K-means, KNN）不同，一個文件內有多個主題，所以是一種軟式（soft clustering）分群。

主題模型分析的流程



V	Topic$_1$	Topic$_2$	…	Topic$_K$
food	0.09	0.07	…	0.02
health	0.10	0.05	…	0.01
GMO	0.04	0.01	…	0.20
safety	0.03	0.03	…	0.05
farmer	0.01	0.10	…	0.07
…	…	…	…	…

D	Topic$_1$	Topic$_2$	…	Topic$_K$
d_1	0.10	0.04	…	0.02
d_2	0.05	0.03	…	0.02
d_3	0.02	0.03	…	0.09
d_4	0.07	0.05	…	0.01
…	…	…	…	…
d_D	0.08	0.01	…	0.10

Figure 1. Application of LDA to a Corpus. *Note.* LDA = latent Dirichlet allocation.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., & Adam, S. (2018). Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology. *Communication Methods and Measures*, *12*(2-3), 93-118. https://doi.org/10.1080/19312458.2018.1430754

主題模型的假設

- 一份文件有**多個主題**。
- 基於詞袋模型的假設，字詞的順序、文件的順序對於分析沒有影響。
- 必須**先知道主題的數量**。
- （最原始的主題模型）主題之間沒有關聯。

文字矩陣（DTM, TCM）

Document Term Matrix (DTM)

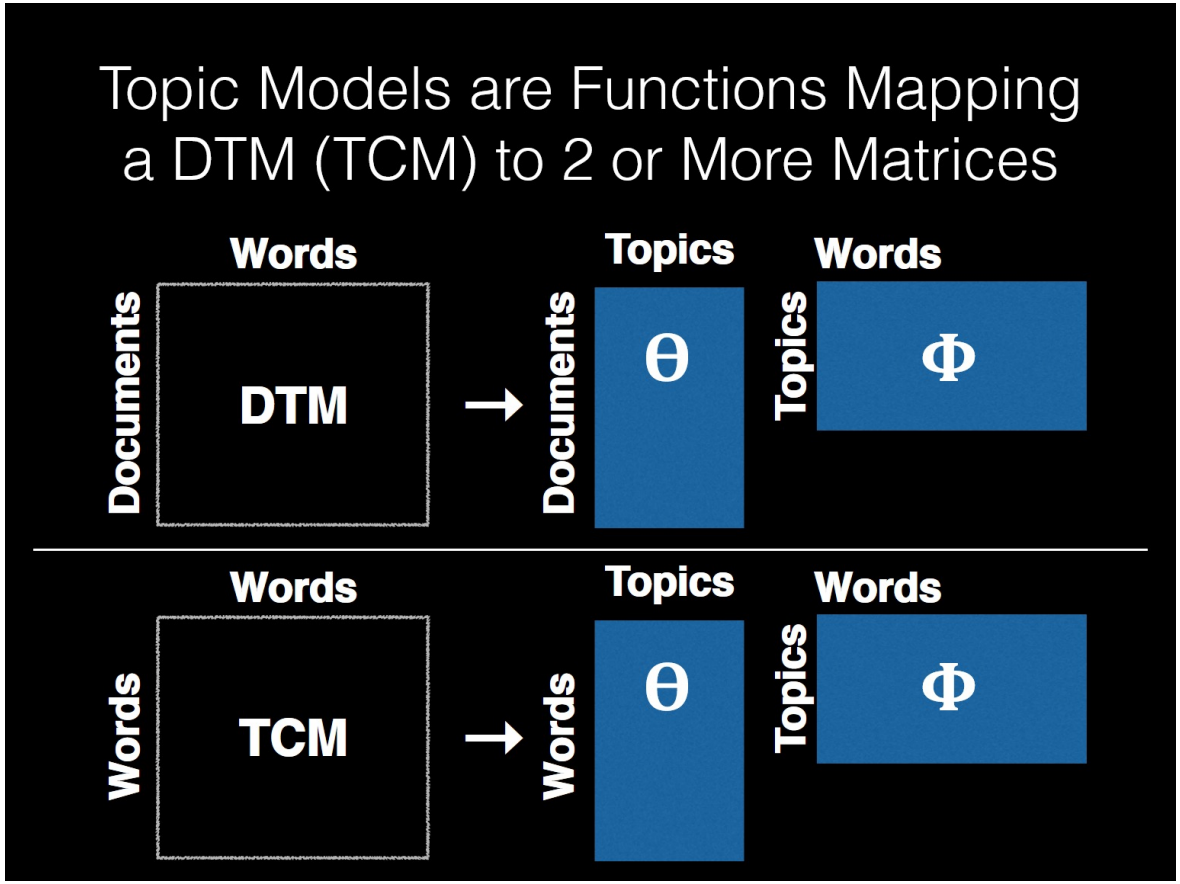
	reduce health	policy	food choice	study sodium	social	…
Document 1	1		1			
Document 2		1				
Document 3			2			
Document 4		2			1	
Document 5	1		1		3	
…						

Term Co-occurence 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		
…						

http://tommyjones.github.io/StatProgDC_2014_04_27/assets/player/KeynoteDHTMLPlayer.html#8

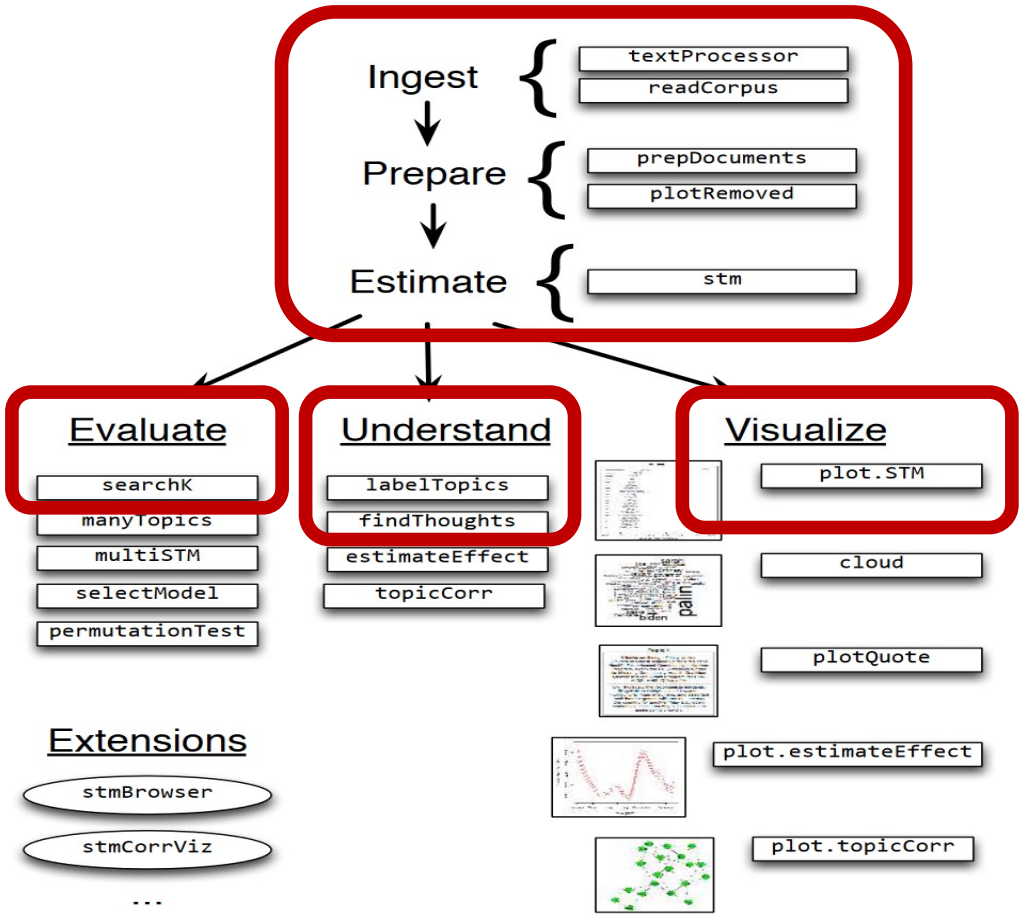
文字矩陣怎麼抽取主題



http://tommyjones.github.io/StatProgDC_2014_04_27/assets/player/KeynoteDHTMLPlayer.html#38

Structural Topic Model (stm)

- 優勢：納入文件的後設資料（metadata），主題之間**可以有關連**。
- 功能完整且強大。



其他學習資源

- Topic Modeling in R
  (https://www.datacamp.com/courses/topic-modeling-in-r)

實際應用範例

- https://www.jstor.org/analyze/