

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

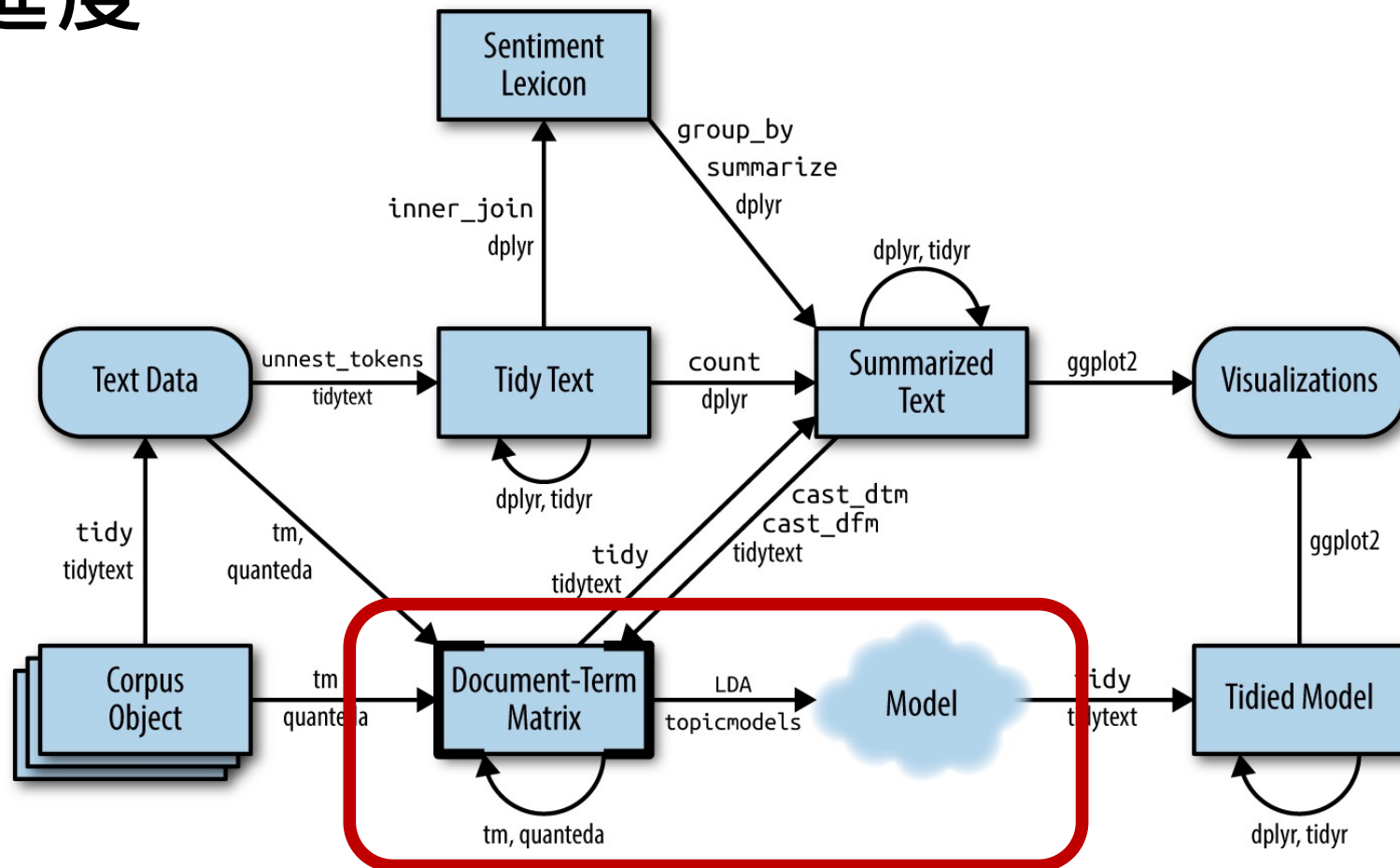
國立臺灣大學政治學系博士、兼任講師

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」

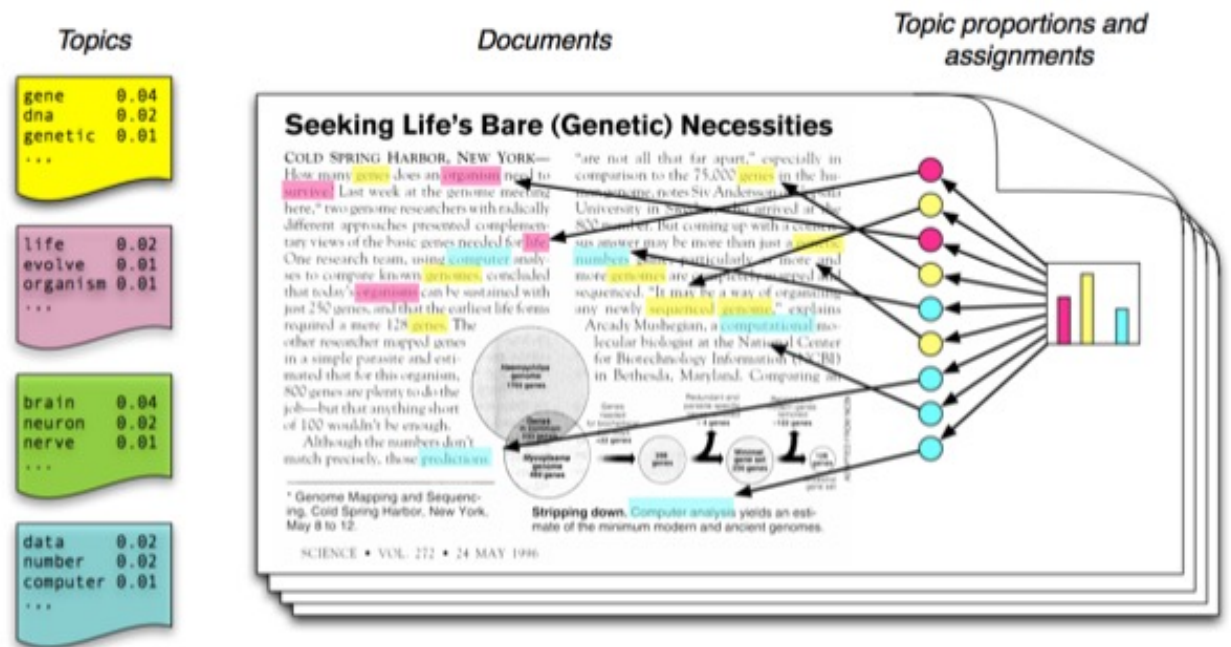


Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77-84.

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

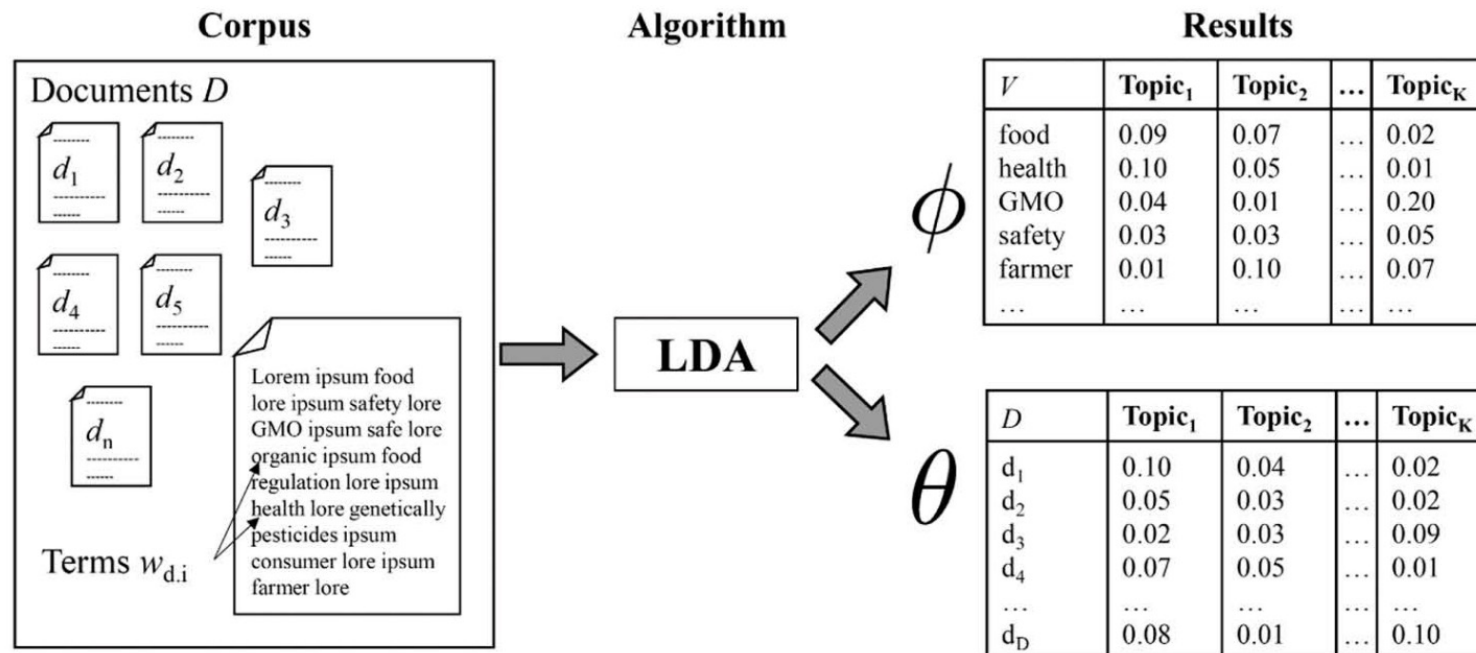


Figure 1. Application of LDA to a Corpus. *Note.* LDA = latent Dirichlet allocation.

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)

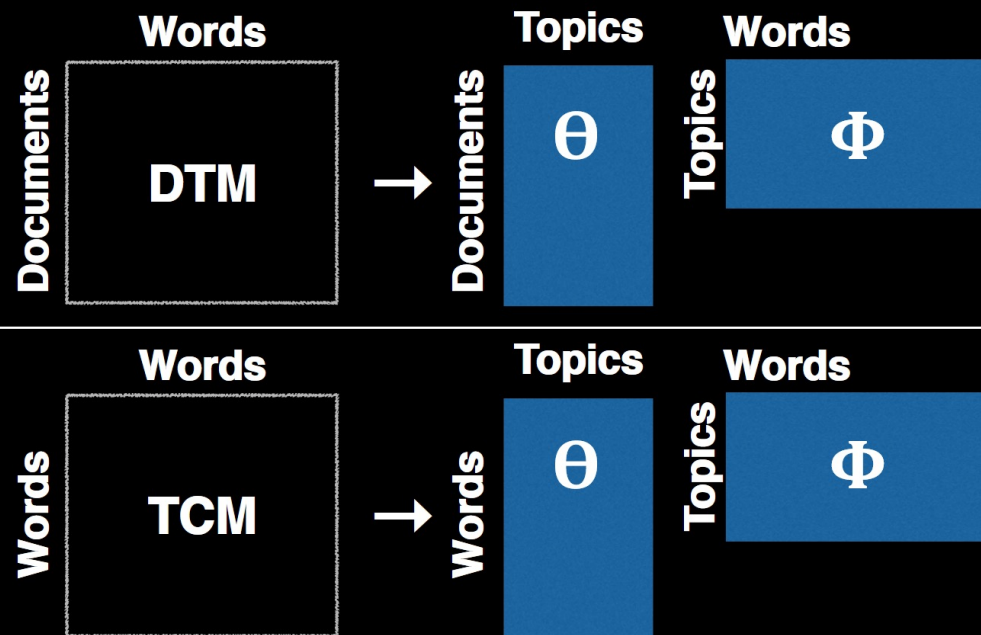
	reduce health	policy	food choice	study sodium	social	...
Document 1	1		1			
Document 2		1				
Document 3			2			
Document 4		2			1	
Document 5	1		1		3	
...						

Term Co-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		
...						

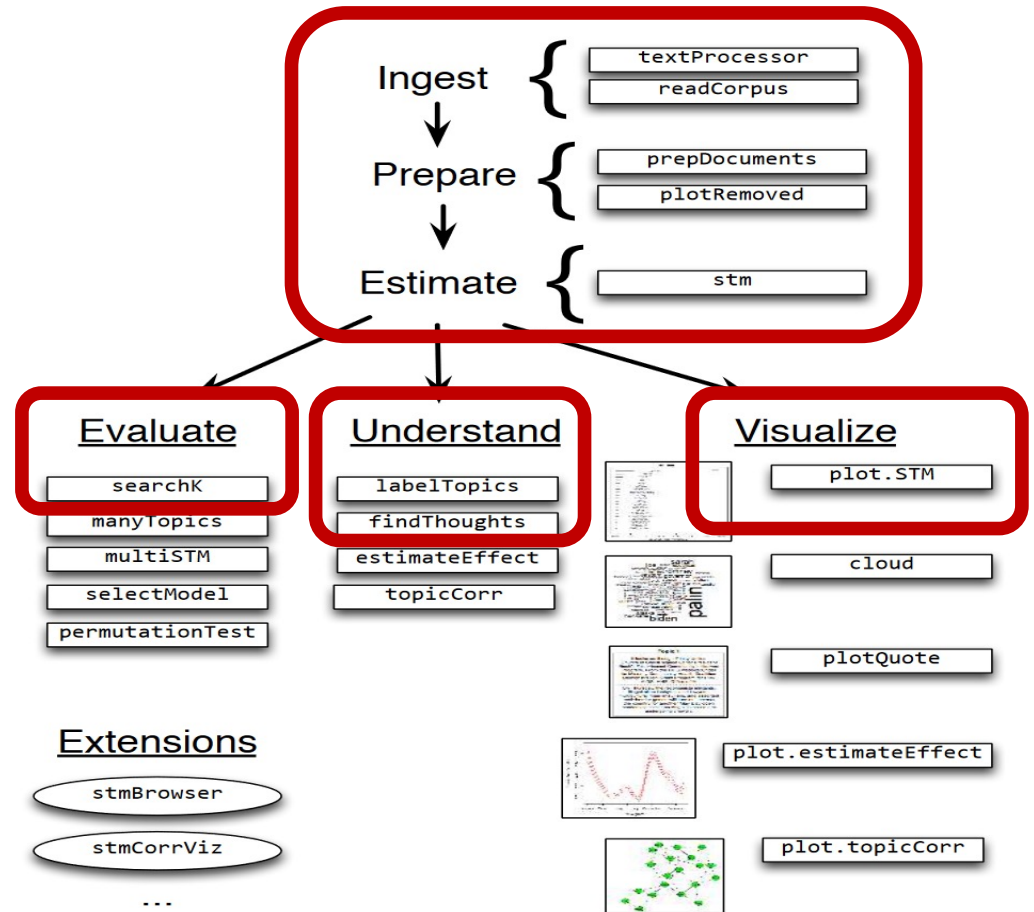
文字矩陣怎麼抽取主題

Topic Models are Functions Mapping
a DTM (TCM) to 2 or More Matrices



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/>

Access provided by 國立臺灣大學

JSTOR

Search

Advanced Search Browse Tools

About Support

Text Analyzer **BETA**

BROUGHT TO YOU BY JSTOR LABS

Help us make this better | About Text Analyzer

Uploading your document

Analyzing text and identifying topics

Generating recommendations

Explore JSTOR

By Subject
By Title
By Publisher
Advanced Search
Data for Research

Get Access
Support
Libguides
Research
Basics

About JSTOR
Mission and History
What's in JSTOR
Get JSTOR
News

JSTOR Labs
JSTOR Daily
Careers
Contact Us

For Librarians
For Publishers

JSTOR is part of ITHAKA, a not-for-profit organization helping the academic community use digital technologies to preserve the scholarly record and to advance research and teaching in sustainable ways.

JSTOR

Search

Advanced Search Browse Tools

About Support

Text Analyzer **BETA**

BROUGHT TO YOU BY JSTOR LABS

Analyze Another Document

Help us make this better | About Text Analyzer

ANALYSIS

Prioritized terms
Adjust results by changing the weights for each term.

✕ Fairy tales
✕ Victorians
✕ Fantasy fiction
✕ History of science
✕ Philosophy of religion

RESULTS

Results with the prioritized terms: Fairy tales, Victorians, Fantasy fiction, History of science, Philosophy of religion

Search Filters: content I can access from 1900 - 2019

FREE ARTICLE

Nature's Invisibilia: The Victorian Microscope and the Miniature Fairy

Laura Forsberg
Victorian Studies, Vol. 57, No. 4 (Summer 2015), pp. 638-666

Download PDF

Save

Cite This Item

Identified terms
Click to add to Prioritized Terms.

TOPICS
American literature
Antiscience
Art history
Behavioral sciences
Childrens literature
Decision analysis
Ethics
Evolutionary psychology
Fairy tales