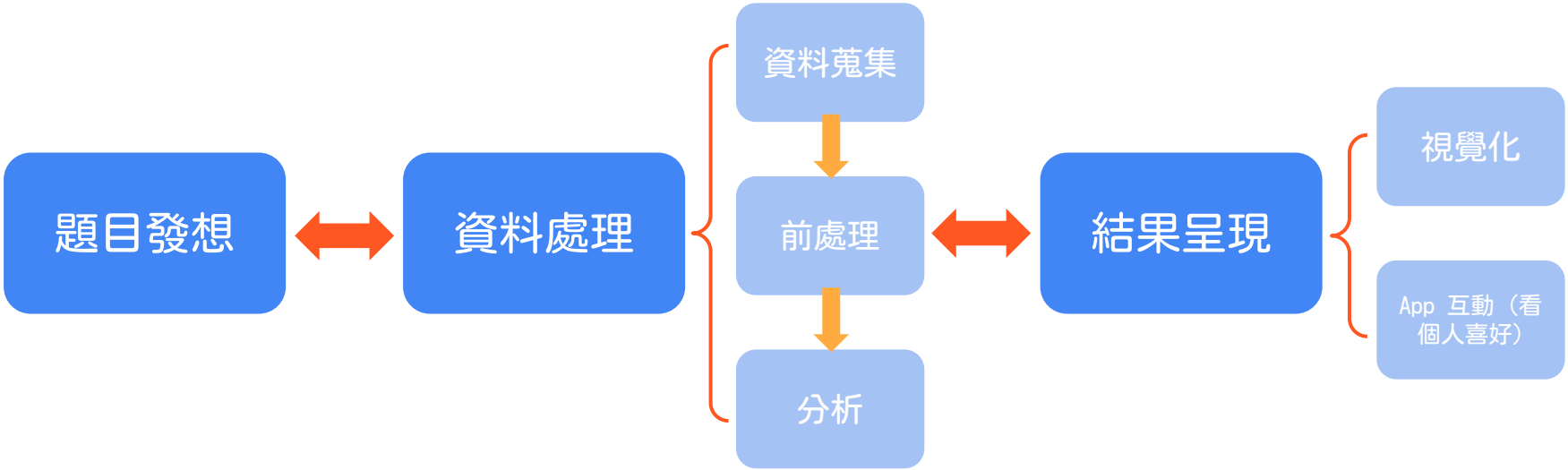
專案示範

兩岸社群媒體中的臺灣形象：
以 PTT 和微博 2020年 Q2-Q3 數據為例

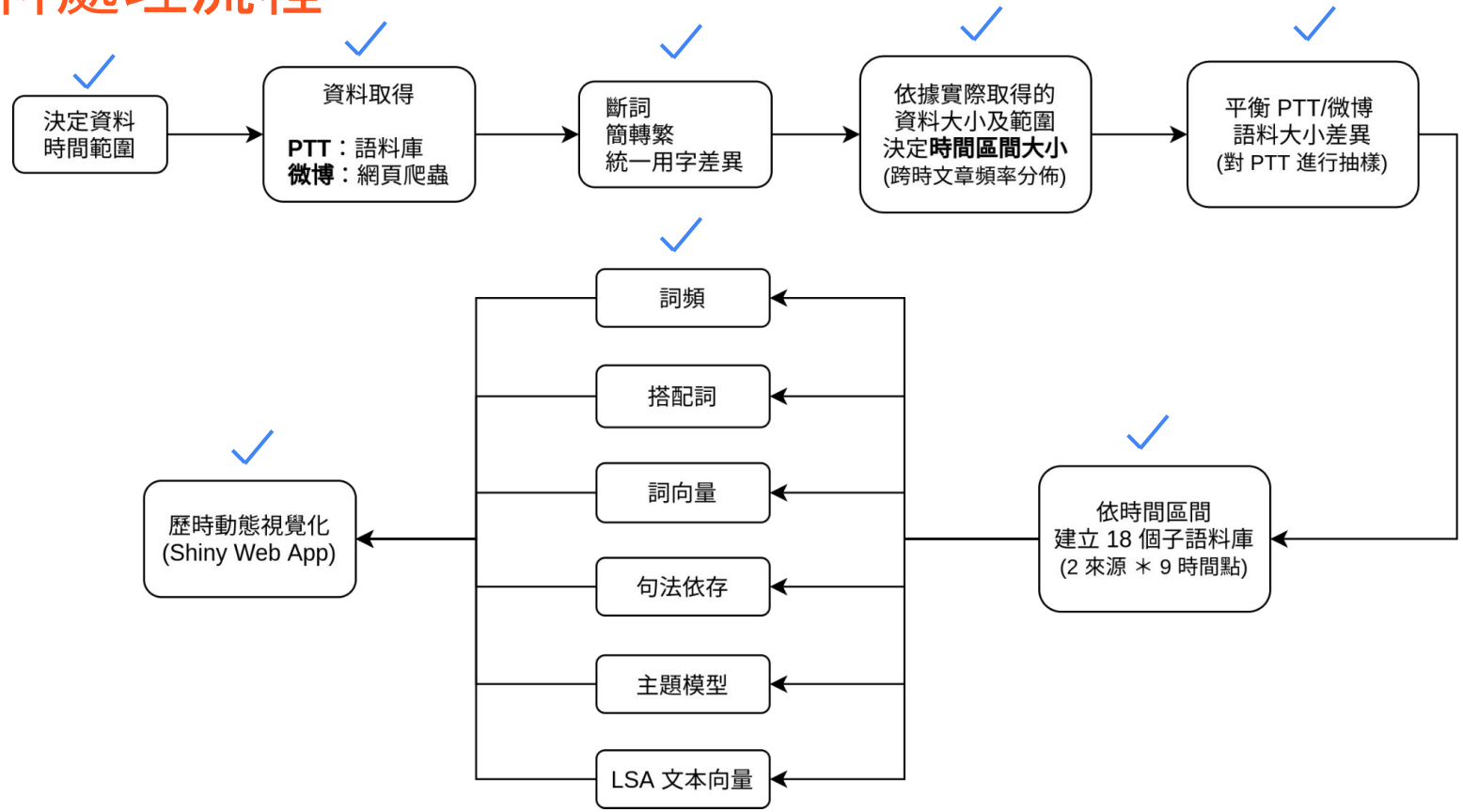
Workflow



研究動機

收集兩岸著名論壇中與臺灣有關的貼文內容，進行不同的面向的文字分析，以探測不同時段兩岸對特定詞彙的使用差異與貼文主題的變化。

資料處理流程


決定資料
時間範圍
資料取得
PTT：語料庫
微博：網頁爬蟲
斷詞
簡轉繁
統一用字差異
依據實際取得的
資料大小及範圍
決定時間區間大小
(跨時文章頻率分佈)
平衡 PTT/微博
語料大小差異
(對 PTT 進行抽樣)
詞頻
搭配詞
詞向量
句法依存
主題模型
LSA 文本向量
依時間區間
建立 18 個子語料庫
(2 來源 ✶ 9 時間點)
歷時動態視覺化
(Shiny Web App)

PDF

各分析說明

分析	處理單位	說明
詞頻 (Word frequency)	詞彙	計算詞彙標準化後的頻率
搭配詞 (Collocation)		計算詞彙前後搭配詞的 MI值
詞向量 (Word embedding)		計算詞彙語意之間的相似程度
句法依存 (Dependency parser)		計算詞彙所搭配動詞的頻率
主題模型 (Topic modeling)	文本	計算文本間的主題分群
LSA 文本向量 (Latent semantic analysis)		計算文本之間的相似程度

Demo

參考資料

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1489–1501. Berlin, Germany: Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1141

Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), Corpus linguistics. An international handbook (Vol. 2, pp. 1212–1248). Berlin: Walter de Gruyter.

Levshina, N. (2015). How to do Linguistics with R: Data exploration and statistical analysis. http://doi.org/10.1075/z.195.website

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., ... Suchomel, V. (2014). The Sketch Engine: Ten years on. Lexicography, 1(1), 7–36.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. the Journal of machine Learning research, 3, 993-1022.

Schweinberger, M. (2021, April 22). Topic Modeling with R. Language Technology and Data Analysis Laboratory. https://slcladal.github.io/topicmodels.html

Tatman, R. (2018, April 4). NLP in R: Topic Modelling. Kaggle. https://www.kaggle.com/rtatman/nlp-in-r-topic-modelling