Tokenization via nltk: *103208062*

爬蟲資料來源:

https://www.nbcnews.com/health/health-news/covid-booster-shots-are-rolling-mean-rcna2254

採取步驟:

(1) nltk.word_tokenize: 單字斷詞

(2) Multi-Word Tokenizer: (合併專有名詞)

收穫: 我發現無法批次寫入，因此附上改寫程式在 ipynb。

(3) Tokenization of Sentences:

依句斷詞，發現可以用 split(".")取代

(4)Ngram:

詞彙每 n 個間的關聯性 > nltk.FreqDist 繪製圖表，中文未必適用，因為不確定中文該切幾個字，這次資料直接換成 nbc 的英文報導(Covid booster shots are rolling out. What does that mean for you?)

(5)Stopwords 停用字

(5-1). mw_tokenizer.add_mwe

> 我發現可以透過匯入更齊全的停用字典，因此附上改寫程式，上週已經完成中文停用字寫入，這週用 collab drive 匯入再與 nltk 的 import stopword concate，使停用字更齊全

(5-2) *Stemming 詞幹提取* vs. Lemmatization 詞形還原

收穫:

我發現將單詞的複雜形態轉變成最基礎的形態 依據 aggressive 程度輕到重，分為 Porter<-Snowball(porter2) <- Lancaster，因此未來該用 Snowball(porter2)來操作，不過其實 Porter 跟 Lancaster 儘管效果部分相同，但是背後演算法與意義並不相同，可以依據文字處理的目的採用。