LDA 主題模型分析新冠肺炎新聞

組員: S09350204 廖書賢

摘要

這次實作的目的就是為了快速得到和理解新冠肺炎的資訊。剛開始原 本的想法是把新聞的標題和內容一個一個的複製到文檔中當作資料集，這樣發現不僅沒有效率，而且資料也收集的不多。這裡就想到要用python爬蟲去爬取需要的新聞資料，但學到現在還沒有使用過 python 爬蟲，所以去參考許多網站的爬蟲資訊，最後有成功用爬蟲爬出新聞資料，並儲存到 excel 裡面，把這個 excel 放進可以分詞的網站進行分詞，完成之後可以下載成一個分詞效果excel表，這個excel表就是這次實作的資料集。

資料集處理的時候，把用不到的部分 drop 掉。這次建立語料庫使用的是分詞效果excel表裡面的分詞資料這一行，在建立之前，先進行預處理， 這裡只去除純數字和一個字的詞。

這個實作的核心，主題建模，這裡有三個是必需有的，詞典(dictionary)、語料庫(corpus)和選擇主題數(num_topics)。選擇主題數的時後，一開始只有想說計算困惑度，就花三天去跑一千個主題的困惑度，結果發現困惑度一直下降，根本看不出來主題數。之後剛好有查到主題連貫性，發現這個可以決定主題數，後來結合這困惑度和主題連貫性就決定最終主題數。建立語料庫的時候，原本只有建立一個，但因為想看不同的語料庫訓練出來的模型會有什麼差異，最後建立詞袋語料庫和TF-IDF語料庫去進行比較。

使用詞典、語料庫(詞袋語料庫和 TF-IDF 語料庫)和主題數建立兩個LDA 模型，結果輸出的主題有點難解釋，所以把主題模型視覺化，發現同樣難以理解，但這裡結合困惑度、主題連貫性和視覺化可以比較出兩個語料庫訓練的模型差異，最終選擇詞袋語料庫訓練的 LDA 模型比較符合我的目的。理解主題想表達的意思這是後面開始實作的部分，這裡發現主題的貢 獻度可以找到跟主題最相近的新聞內容，透過新聞內容去理解主題。最後使用測試集和訓練集進行模型的分類性能評估，判斷選擇模型的好壞。

第一章 前言

1.1 背景

現在的資訊越來越多都是從網路上獲取，而這樣的情況導致你不容易找到你想知道的資訊，例如關於這次新冠肺炎新聞在網路上就數不勝數，很難有效率的尋找、整理和理解這些關於新冠肺炎新聞的資訊。

1.2 動機

假如用人工的方式把全部新冠肺炎相關的新聞全部看一次，並做內容的分析，甚至把分析出來的內容作分類，這樣不僅浪費時間，還很沒有效率。因為觀察到這樣的情況，所以我打算用 NLP 的技術去做新聞的分析。

1.3 目的

馬上獲得最近新的新冠肺炎相關資訊，不必把時間浪費在不重要和重複的資訊上，達到節省時間目的。

1.4 解決問題

1.4.1 問題

1 怎麼把新聞報導抓出來變成一個程式可以讀的檔案 ？
2 要分幾個主題 ？
3 關於新冠肺炎的專有名詞逐漸出現，分詞真的分得出來這些專有名詞 ？

1.4.2 方法

1 利用 python 爬蟲，把網路上的新聞抓出來儲存到一個excel檔裡面。
2 透過困惑度和主題連貫性，得到最佳的主題個數。
3 有些分的出來，有些分不出來，所以會用手動方式做調整。

1.5 方法概述

1.5.1 python 爬蟲

匯入爬蟲套件 → 加入使用者資訊模擬真實瀏覽網頁的情況 →抓取新聞 → 抓取新聞連結 → 爬取多個分頁的新聞連結並得到抓取新聞的標題、時間、分類、描述、內容、關鍵字、來源、網址 → 儲存資料

1.5.2 困惑度和主題連貫性

1 困惑度
計算困惑度，困惑度越低，表示模型越好。

2 主題連貫性
計算主題連貫性，主題連貫性越大，表示主題越好。

第二章 文獻探討

2.1 文件主題模型

模型使用非監督式學習，事先不做任何標籤和註解，文本就是訓練的資料，應用於規模化地整理和進行大量文本摘要。問題文件找到背後隱含的主題的結構。

2.2 潛在狄利克雷分佈(LDA)

2.2.1 基本的原則

1 每篇文件是由多個主題所組成。
2 描述每個主題可以使用多個重要用詞，不同的主題之間可以出現相同的用詞。

2.2.2 每一篇文件的用詞產生方式

1 隨機選擇一個主題分佈。
2 從第一點的主題分佈中，隨機選擇裡面的一個主題，從這主題當中，隨機選擇一個字詞。

主題一開始就會決定好，甚至沒有文件時候就已經決定好主題。主題數是固定的，由不同比例的各個主題組合成不同的文件。

2.2.3 LDA 優缺點

2.2.3.1 優點

1 快速、直覺和容易理解。
2 預測文件中沒看過的主題。

2.2.3.2 缺點

1 模型需要做不少人工的微調。
2 模型中主題難以理解想表達的意思。

第三章 研究方法

3.1 Gensim

一個強大的自然語言處理的 Python 第三方庫，自動提取文檔中的語義主題。透過一個訓練中的語料庫，檢查詞彙並統計每個詞彙出現的次數，進而發現文件語義結構。

3.1.1 演算法

1 Word2Vec
2 潛在語義分析(Latent Semantic Analysis，LSA)
3 潛在狄利克雷分佈(Latent Dirichlet Allocation，LDA)

上面的演算法都屬於非監督式學習，可以處理純文本。

3.2 生成詞袋

詞袋就是一個字典(dictionary)，裡面儲存所有經過預處理以後的單詞和索引。

3.3 doc2bow(Gensim庫)

使用 doc2bow 統計分詞資料一行的所有單詞，在分詞excel表中出現的次數。

3.4 TF-IDF

一種統計方式用來評估一個單詞對於一個文件集或語料庫中一份文件的重要程度。單詞的重要性會隨著在文件中出現的次數成正比增加，語料庫則恰好相反。

3.5 困惑度(Perplexity)

利用機率計算某個主題模型在測試集上的表現，困惑度越低，表示主題模型越好。

這裡我觀察到兩個缺點，第一個缺點，它只是測量模型的可信度，但不是準確度，所以很難用在最終評估，第二個缺點，很難在資料集(不同上下文章長度、詞彙大小)之間作比較。

3.6 主題連貫性(Coherence)

主要是評量一個主題內的詞語是否有相關性，計算方法有很多種，例如: C_UMASS、CV、CP等。

主題連貫性越大，表示主題越好。

3.7 視覺化(pyLDAvis庫)

左邊氣泡分佈表示不同的主題，右邊是主題內前三十個相關詞。淺藍色代表整個文檔中詞語出現的權重，深紅色代表這個主題中詞語所佔的權重。

3.7.1 詞語主題的相關性

調節λ參數，λ越接近1，更經常出現在主題下的詞語，跟主題更有關。λ越接近0，主題下更特殊和獨有的詞語，跟主題更有關。

3.7.2 每個主題的普遍性

氣泡的大小和編號代表主題出現的權重。

3.7.3 主題之間的關聯

兩個主題之間的位置遠近，代表主題之間的接近性。氣泡重疊則表示兩個主題想表達的意思相近。

第四章 系統設計

4.1 python 爬蟲

抓取新冠肺炎的新聞，抓取到的資料儲存到excel檔裡面(Covid-19 news.xlsx)。

4.2 分詞

用分詞對 Covid-19 news.xlsx 進行分詞，分出來的詞儲存到新的excel檔裡面(圖一)。

	A	B	C	D	E
1	原資料	分詞資料	關鍵字	序號	發佈時間
2	0	0		152830	
3	打擊新冠肺	打擊 新冠	假消息you	152831	
4	面對新冠肺	面對 新冠	疫情資訊!	152832	
5	面對新冠肺炎(COVID-19)疫情持續在全球蔓延除了政府、公共衛生主管機關全面防堵疫情散播外如何宣導傳遞正確的防疫資訊並且避免錯誤資訊、惡意資訊以及假消息傳播也是當務之急。對此社群平臺Facebook、Twitter、YouTube都公佈了	面對 新冠	肺炎相關	152833	
6	新冠肺炎!	新冠肺炎	肺炎廣告!	152834	
7	1	1		152835	
8	新增8例境	新增 8 例	境外指揮	152836	

圖一: 分詞效果excel表

4.3 資料處理

1. 第一次資料處理

   Covid-19 news.xlsx 刪除裡面不需要用到的資料(時間、分類、網址等)。

2. 第二次資料處理

   分詞效果excel表手動處理分詞資料，因為有些專有名詞被分開，需用手動方式把專有名詞進行合併。例如: 新冠 肺炎 → 新冠肺炎、COVID- 19 → COVID-19

4.4 讀檔

讀取分詞效果excel表，分詞資料一行的分詞資料使用 split() 分成詞數組，有缺失值的地方用空字串取代。

```
0                                                      NaN
1                        [打擊, 新冠肺炎, 假消息, 臉書, 推特, youtube, 全面, 防堵]
2        [面對, 新冠肺炎, covid-19, 疫情, 持續, 在, 全球蔓延, 除, 了, 政府...
3        [面對, 新冠肺炎, covid-19, 疫情, 持續, 在, 全球蔓延, 除, 了, 政府...
4                  [新冠肺炎, 廣告, facebook, twitter, youtube]
                                    ...
27260                                                  NaN
27261                          [中, 裕, 報捷, 漲, 停, 生, 醫, 族群, 紅通通]
27262    [台股, 高檔, 盤整, 之際, 中, 裕, 4147, 獲, 新冠肺炎, 最新, 單株, ...
27263    [台股, 高檔, 盤整, 之際, 中, 裕, 4147, 獲, 新冠肺炎, 最新, 單株, ...
27264                              [杏, 輝, 抗體, 疫苗, 肺炎, 生, 技]
Name: 分詞數據, Length: 27265, dtype: object
```

圖二:使用 split() 分成的詞數

4.5 預處理

對讀取分詞資料(語料庫)進行基本預處理，把純數字、一個字的詞去除。

4.6 生成詞袋

dictionary = corpora.Dictionary(processed_corpus(預處理完的語料庫))

4.6 doc2bow

corpus = [dic.doc2bow(text) for text in processed_corpus

4.8 TF-IDF

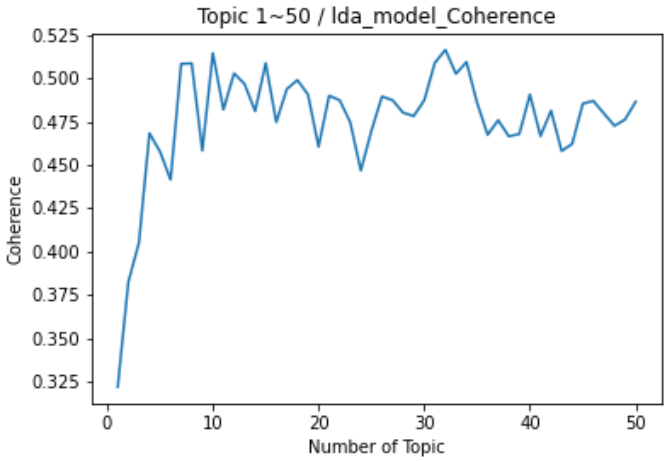
```
In [24]: # TF-IDF
         tfidf = models.TfidfModel(corpus)
         corpus_tfidf = tfidf[corpus]
```

圖三:建立 TF-IDF 語料庫

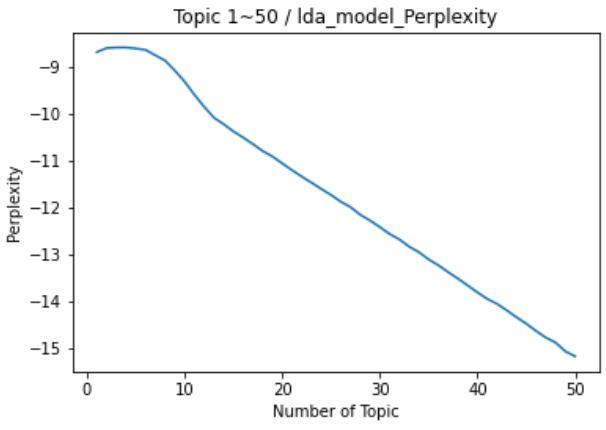
4.9 選擇主題數

模擬由兩個語料庫訓練下的 LDA 模型去計算主題數1到50的困惑度和主題連貫性。

1 詞袋語料庫

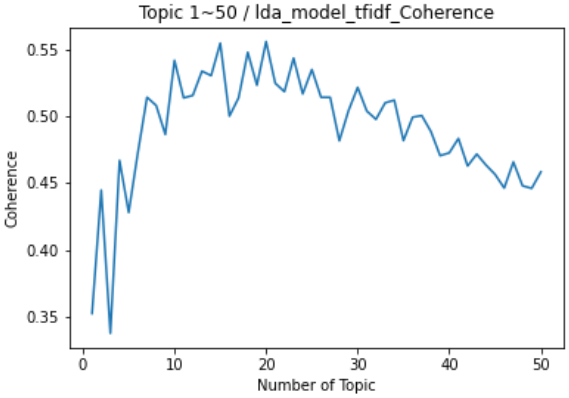


圖四:主題連貫性

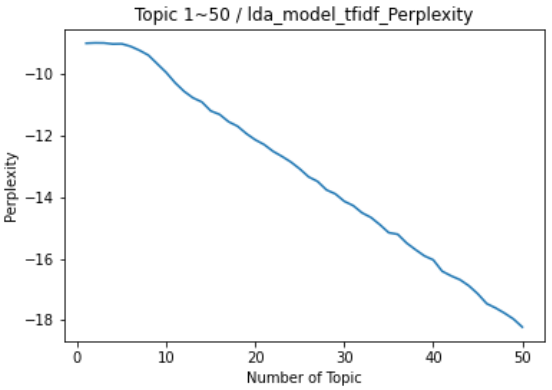


圖五:困惑度

2 TF-IDF 語料庫



圖六:主題連貫性



圖七:困惑度

由圖四、五、六、七可知，兩個語料庫訓練下來的模型困惑度一直下降，所以這樣的情況無法用困惑度去選主題數。改成看主題連貫性，發現兩個模型主題數為32和20的時候，主題連貫性最大。

3 最終結果

	詞袋語料庫	TF-IDF 語料庫
主題數	32	20

4.10 LDA 主題模型

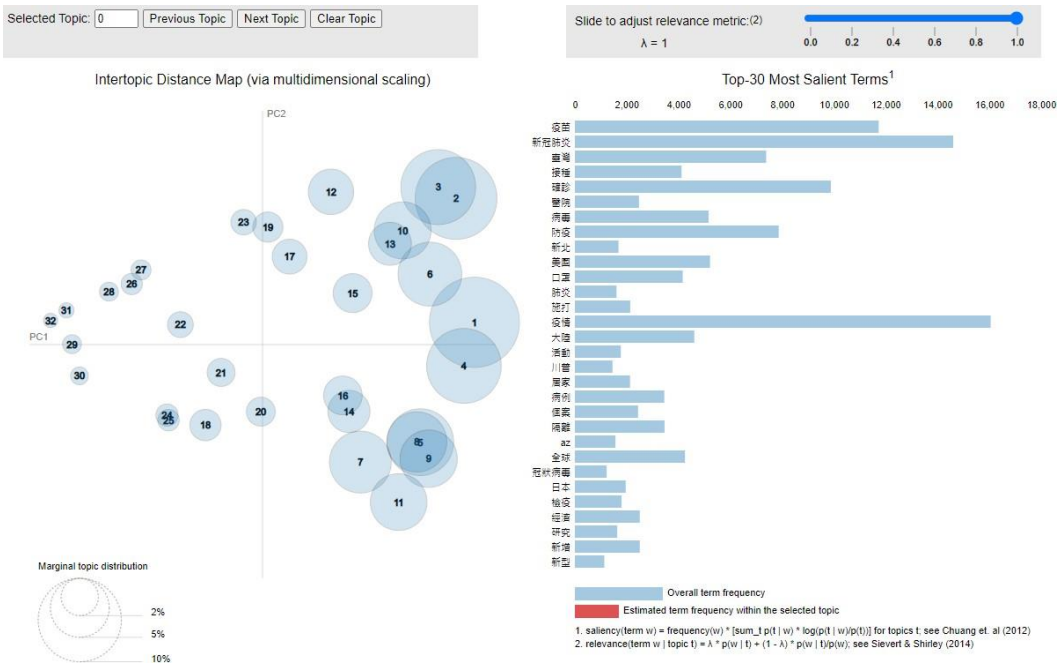
```
In [37]: # 在詞袋語料庫上訓練LDA模型
         lda_model = models.LdaModel(corpus, id2word=dic, num_topics=32) # num_topics:主題個數32
```

圖八:詞袋語料庫訓練的 LDA 模型(lda_model)

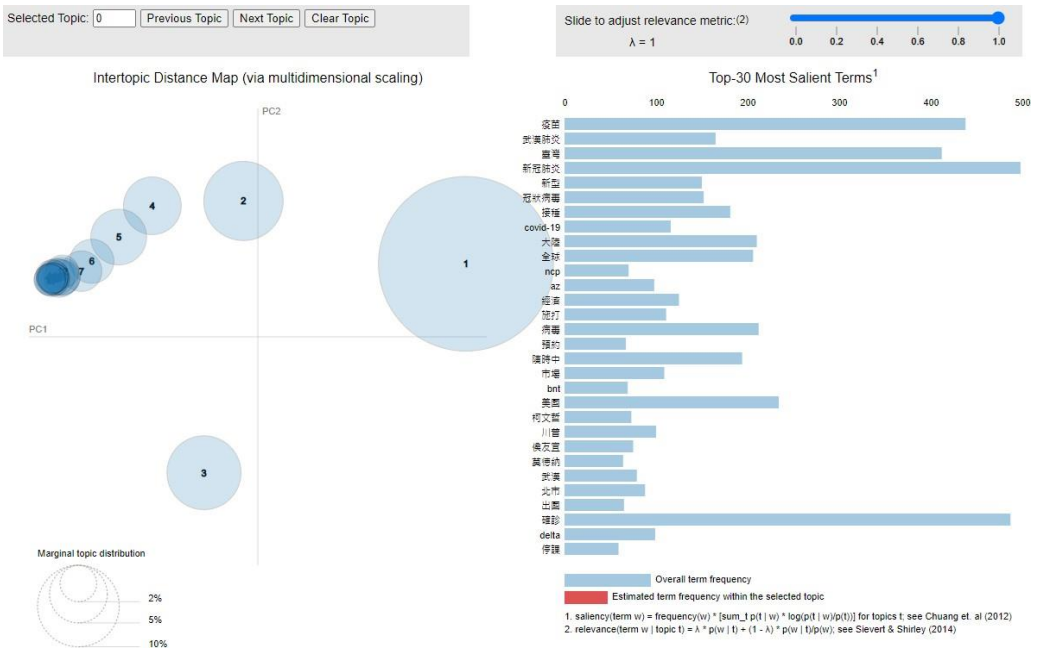
```
In [54]: # 在TF-IDF語料庫上訓練LDA模型
         lda_model_tfidf = models.LdaModel(corpus_tfidf, id2word=dic, num_topics=20) # num_topics:主題個數20
```

圖九:TF-IDF 語料庫訓練 LDA 模型(lda_model_tfidf)

4.11 視覺化



圖十:詞袋語料庫訓練的 LDA 模型



圖十一: TF-IDF 語料庫訓練的 LDA 模型

4.12 兩個模型比較

	主題連貫性最高的分數	困惑度	視覺化
lda_model	0.516533	一直下降	氣泡平均散開
lda_model_tfidf	0.555782	一直下降	氣泡重疊太多

雖然 lda_model 的主題連貫性分數比 lda_model_tfidf 低，但 lda_model_tfidf 的氣泡重疊太多，代表主題相似度太高。這會導致主題表示都差不多，無法獲取更有用的新冠肺炎資訊，所以最後選擇由詞袋語料庫訓練出來的 LDA 模型可以獲得的資訊比較多。

第五章 實驗結果

5.1 模型選擇

詞袋語料庫訓練的 LDA 模型(lda_model)

5.1.1 理解模型中每個的主題內容

	主題	主題貢獻度(百分比)	關鍵字	內容
0	0.0	0.9677	美國, 大陸, 病毒, 中國, 新冠, 臨床, 研發, 試驗, 新冠肺炎, 川普	[生華科(6492)旗下被視為治療新冠肺炎潛力新藥Silmitasertib緊急人體臨床試驗...
1	1.0	0.8616	疫情, 川普, 美國, 新冠肺炎, 宣佈, 影響, 報導, 拜登, 員工, 餐廳	[美國拜登美國民眾民眾人數]
2	2.0	0.9706	疫情, 紓困, 新冠肺炎, 政府, 表示, 業者, 申請, 活動, 影響, 產業	[新冠肺炎疫情衝擊擴大行政院會13（今）日將拍板規模逾400億的短期紓困和振興方案針對服務業...
3	3.0	0.9119	經濟, 疫情, 基金, 投資, 新興, 全球, 降息, 預測, 新冠肺炎, gdp	[投資人新興市場債券收益資金新興市場債券基金]
4	4.0	0.8616	患者, 治療, 醫院, 重症, 醫療, 病人, 住院, 症狀, 新冠肺炎, 醫師	[C肝患者染新冠, 死亡風險33倍！醫師呼籲必做9件事]
5	5.0	0.8616	北市, 柯文哲, 臺北市, 飯店, 中央, 防疫, 簡訊, 臺北, 臺灣, 新冠肺炎	[獨》驚！新冠肺炎疫情強襲, 臺北觀光飯店晚上不賣Buffet啦！]
6	6.0	0.9417	防疫, 疫情, 臺灣, 新冠肺炎, 表示, 醫療, 消毒, 透過, 工作, 安全	[面對新冠肺炎(COVID-19)疫情持續在全球蔓延除了政府、公共衛生主管機關全面防堵疫情散...
7	7.0	0.8385	新北, 侯友宜, 市長, 高雄, 新冠肺炎, 臺灣, 全台, 曝光, 雲林, 工會	[疫情新冠肺炎台灣侯友宜新北市]
8	8.0	0.9031	總統, 疫情, 病毒, 新冠肺炎, 巴西, 他們, 英國, 香港, 我們, 新冠	[病毒可停留紙鈔3小時, 香港專家呼籲注意手部衛生]
9	9.0	0.9310	確診, 結果, 新冠肺炎, 檢測, 陰性, 症狀, 感染, 陽性, 表示, 目前	[纏噗、味覺異常後國際上又出現腳指出現水泡的新冠肺炎疑似癥狀中央流行指揮中心專家諮詢小組召集...

圖十二:新冠肺炎新聞的內容和主題最相關

這裡取10筆資料去做觀察，可以發現貢獻度越高，主題就越接近內容，取裡面最高貢獻度的第三個主題作為例子。

第 2 個主題: 0.020*"疫情" + 0.013*"紓困" + 0.011*"新冠肺炎" + 0.009*"政府" + 0.009*"表示" + 0.009*"業者" + 0.008*"申請" + 0.008*"活動" + 0.008*"影響" + 0.007*"產業"

↕

內容: 疫情衝擊擴大行政院會13（今）日將拍板規模逾400億的短期紓困和振興方案針對服務業、觀光運輸業和農業三大產業紓困。其 中政院擬擴大餐飲、零售、會展、商圈、夜市及傳統市場五大內需服務業適 用範圍提供逾百億元紓困措施另外還要發「振興抵用券」以利疫情紓緩時刺激消費、提振經濟動能。

由上面的例子得知，這樣的方式可以讓我們快速的理解主題。

5.1.2 瞭解模型中主題的數量和分佈

	主題	主題的關鍵字	文件的數量	佔文件多少%
0.0	0.0	美國, 大陸, 病毒, 中國, 新冠, 臨床, 研發, 試驗, 新冠肺炎, 川普	6076.0	0.2228
1.0	6.0	防疫, 疫情, 臺灣, 新冠肺炎, 表示, 醫療, 消毒, 透過, 工作, 安全	770.0	0.0282
2.0	6.0	防疫, 疫情, 臺灣, 新冠肺炎, 表示, 醫療, 消毒, 透過, 工作, 安全	1248.0	0.0458
3.0	6.0	防疫, 疫情, 臺灣, 新冠肺炎, 表示, 醫療, 消毒, 透過, 工作, 安全	240.0	0.0088
4.0	25.0	肺炎, 新冠肺炎, 南韓, 訊息, 垃圾, 疫情, 特殊, 傳染性, 澳洲, 轉機	237.0	0.0087
5.0	0.0	美國, 大陸, 病毒, 中國, 新冠, 臨床, 研發, 試驗, 新冠肺炎, 川普	277.0	0.0102
6.0	13.0	確診, 病例, 新增, 本土, 新冠肺炎, 境外移入, 個案, 疫情, 死亡, 累計	1218.0	0.0447
7.0	13.0	確診, 病例, 新增, 本土, 新冠肺炎, 境外移入, 個案, 疫情, 死亡, 累計	224.0	0.0082
8.0	20.0	確診, 隔離, 居家, 個案, 檢疫, 接觸, 防疫, 表示, 目前, 症狀	513.0	0.0188
9.0	13.0	確診, 病例, 新增, 本土, 新冠肺炎, 境外移入, 個案, 疫情, 死亡, 累計	519.0	0.0190

圖十三

更加了解分詞效果excel表(資料集)偏向哪一個主題，並可以得知資料集最主要說明的內容。縮小資料範圍，幫助自己快速理解內容，篩掉重複的內容。

5.2 模型對訓練集和測試集進行分類的性能評估

5.2.1 分詞效果excel表(訓練集)

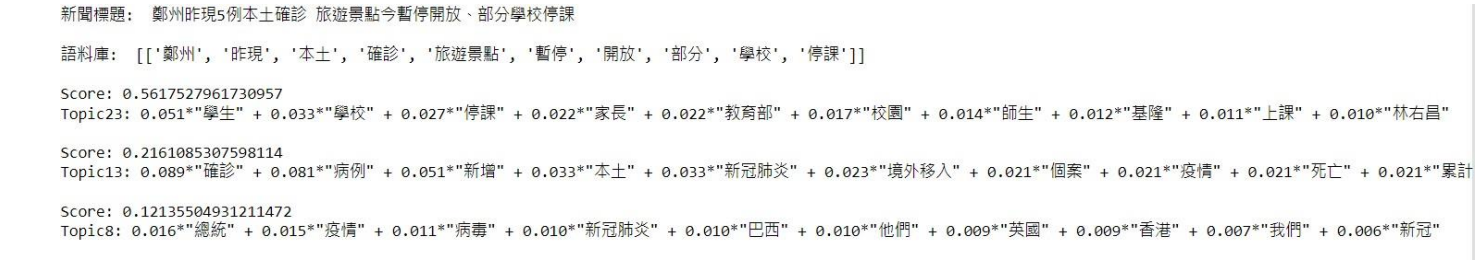
分詞效果表總計有27266筆資料。

```
原資料:  健檢病情洪子仁新光疫情

詞袋語料庫:  ['健檢', '病情', '洪子仁', '新光', '疫情']

Score: 0.3819490969181061
主題2: 0.158*"新北" + 0.094*"侯友宜" + 0.054*"市長" + 0.041*"高雄" + 0.028*"新冠肺炎" + 0.015*"臺灣" + 0.014*"全台" + 0.014*"曝光" + 0.013*"雲林" + 0.013*"工會"

Score: 0.2247644066810608
主題31: 0.158*"新北" + 0.094*"侯友宜" + 0.054*"市長" + 0.041*"高雄" + 0.028*"新冠肺炎" + 0.015*"臺灣" + 0.014*"全台" + 0.014*"曝光" + 0.013*"雲林" + 0.013*"工會"

Score: 0.2118476927280426
主題4: 0.158*"新北" + 0.094*"侯友宜" + 0.054*"市長" + 0.041*"高雄" + 0.028*"新冠肺炎" + 0.015*"臺灣" + 0.014*"全台" + 0.014*"曝光" + 0.013*"雲林" + 0.013*"工會"
```

圖十四: 隨機找一筆資料做測試

5.2.2 測試集

1. 輸入有關新冠肺炎的新聞標題



圖十五: 輸入一筆資料(新聞的標題)進行測試

2. 隨機找10筆資料(新冠肺炎的新聞標題)

圖十六:隨機找一筆資料(新聞的標題)進行測試

第六章 結論

6.1 結論與研究困境

6.1.1 結論

困惑度和主題連貫性可以幫助找到適合模組的主題數，只不過很花費時間。

這次實作比較兩種語料庫訓練的 LDA 模型，經過困惑度、主題連貫性、視覺化的比較，觀察下來，詞袋語料庫比較適合達到我的目的。

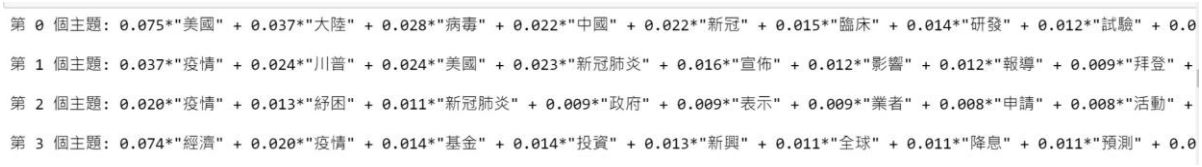
隨機找訓練集和測試集中的一筆資料進行模型的測試，發現裡面分數都不高，表示模型中的主題跟隨機找的資料的相關性很低。這樣測試下來，知道不是每筆資料都會跟模型中的主題有相關。

6.1.2 研究困境

剛開始原本的想法是把新聞的標題和內容一個一個的複製到txt檔中當作資料集，這樣發現不僅沒有效率，而且資料也收集的不多。所以之後 就使用 python 爬蟲去爬取需要的資料，但目前為止還沒有學習任何有關python 爬蟲的技術，只能從零開始學習。

計算困惑度的時候，花三天去測試一到一千個主題的困惑度，結果發現困惑度一直下降，根本看不出來主題數。

理解主題想要表達的意思，因為每個主題都顯示如圖十七，這導致理解主題變得相當困難。



圖十七: 主題顯示

隨機選一篇有關新冠肺炎的新聞標題儲存到txt檔中測試模型，發現文檔對模型每個主題的佔比都相同。

6.1.3 未來展望

抓取大量且不同種類的新聞，這次的題目是新冠肺炎，所以只使用有關疫情的新聞，但其實可以結合不同的新聞，去做更寬廣的分析，建立更龐大的語料庫。

這次的分詞處理我用的是手動，把被分開的專有名詞進行合併，這之後會自己寫code來代替手動。

使用不同主題模型的演算法來分析文章，並優化這次實作的模型。

參考文獻

[1] 文件主題模型和 LDA
https://tengyuanchang.medium.com/%E7%9B%B4%E8%A7%80%E7%90%86%E8%A7%A3-lda-latent-dirichlet-allocation-%E8%88%87%E6%96%87%E4%BB%B6%E4%B8%BB%E9%A1%8C%E6%A8%A1%E5%9E%8B-ab4f26c27184

[2] Gensim
https://blog.csdn.net/weixin_42608414/article/details/87559437
https://codertw.com/%E7%A8%8B%E5%BC%8F%E8%AA%9E%E8%A8%80/43743/

[3] 主題連貫性和困惑度
https://www.codenong.com/7d49959c74f3c1e0cf63/
https://zhuanlan.zhihu.com/p/449745831

[4] pyLDAvis
https://blog.csdn.net/qq_39496504/article/details/107125284

[5] 分詞工具
https://www.gooseeker.com/res/softdetail_13.html

[6] 分詞效果 excel 表(分詞效果_202112171711425680.xlsx)