期末專案

資料科學系 -文字探勘與自然語言處理

指導教授: 吳政隆教授/呂明潁教授
報告人: 黃士倫

問題

Problem Setting

True (News)

□ Use the sentiment analysis method to label the sentence as positive or negative.

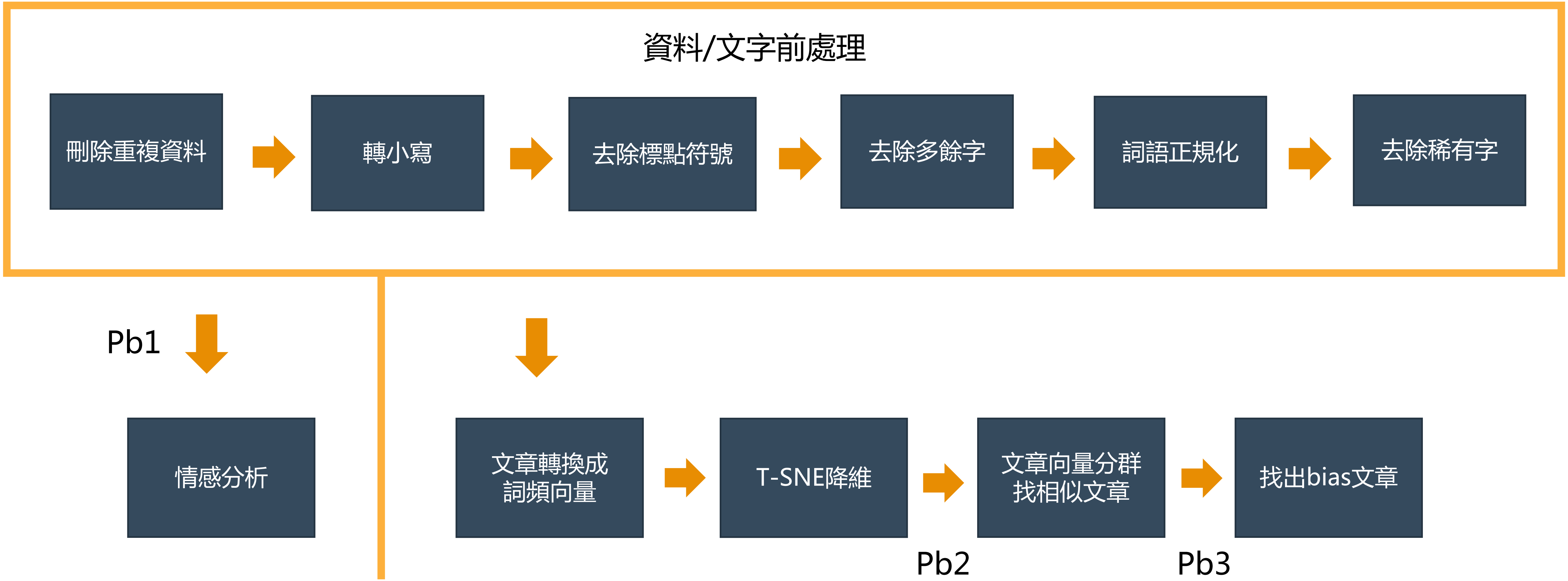
不確定

□ Measured by correlation between rows in term-document matrix. Decide how many groups (similar News) topic in this file?

不確定

□There are many different discourses in news. Which news discourses are biased? Try to define and find out those news..

流程圖



資料/文字前處理

刪除重複資料 → 轉小寫 → 去除標點符號 → 去除多餘字 → 詞語正規化 → 去除稀有字

Pb1

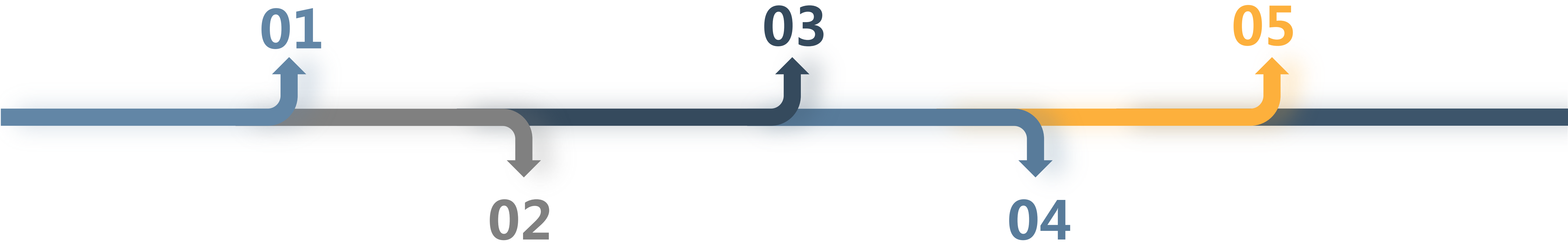
情感分析

文章轉換成詞頻向量 → T-SNE降維 → 文章向量分群找相似文章 → 找出bias文章

Pb2

Pb3

文字前處理



01

文字轉小寫

What -> what…

02

去除標點符號

去除：！“#$%&\ ‘()*+,-
./:;<=>?@[\\]^_`{|}~ “”

03

去除多餘字

(1) 去除：\xa0
(2) 去除 Stopwords : am、and…

04

文字正規化

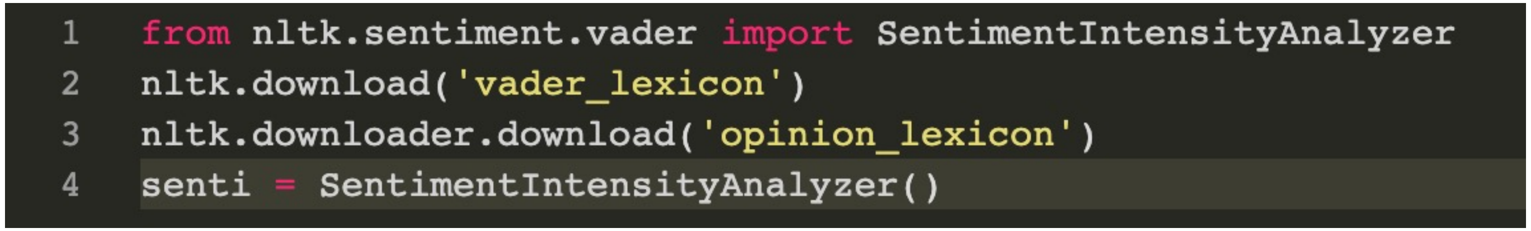
(1) 去除字根 （ stemming) :
Ex. started/starting -> start
(2) 單複數轉換: （ lemmatizing) :
Ex. papers -> paper

05

去除稀有字

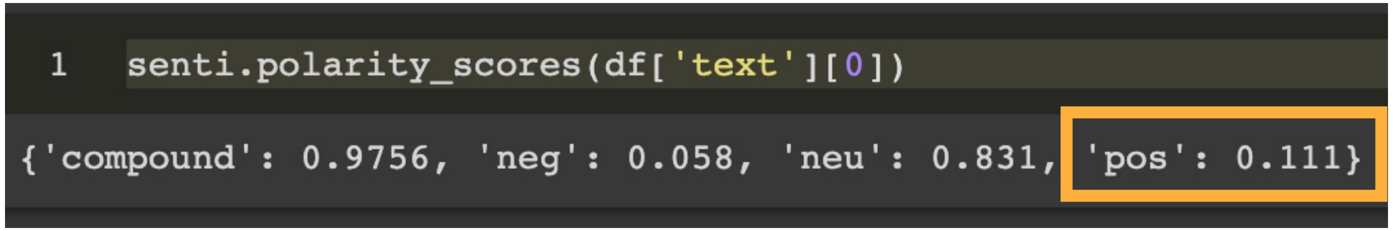
所有文章詞頻為 **1** 的詞去除，
目的：減少維度、拼字錯誤詞

情感分析結果



```python
from nltk.sentiment.vader import SentimentIntensityAnalyzer
nltk.download('vader_lexicon')
nltk.downloader.download('opinion_lexicon')
senti = SentimentIntensityAnalyzer()
```

使用 nltk 情感分析模組

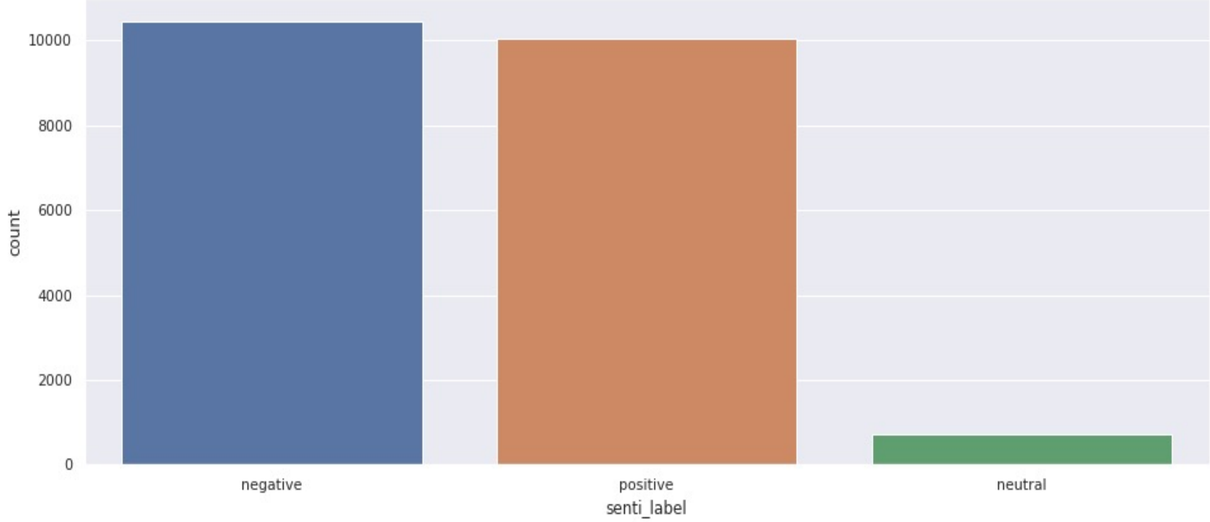


```python
senti.polarity_scores(df['text'][0])
```

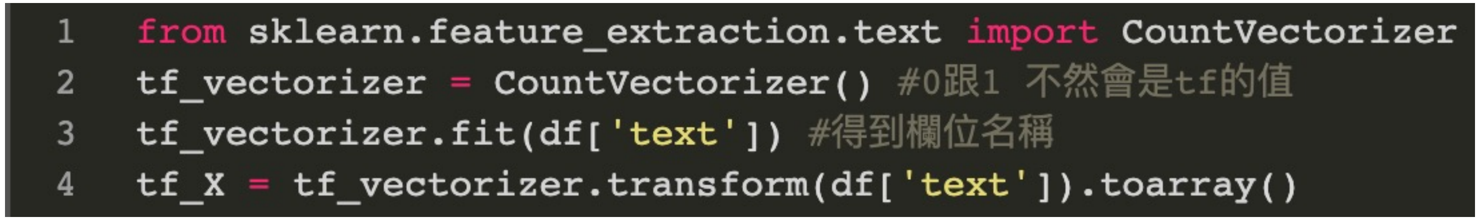
```
{'compound': 0.9756, 'neg': 0.058, 'neu': 0.831, 'pos': 0.111}
```

比較 positive、negative 決定正負向情感標籤：
- pos > neg：正面
- neg > pos：負面
- neg = pos：中立

Pb1: 情感分析結果

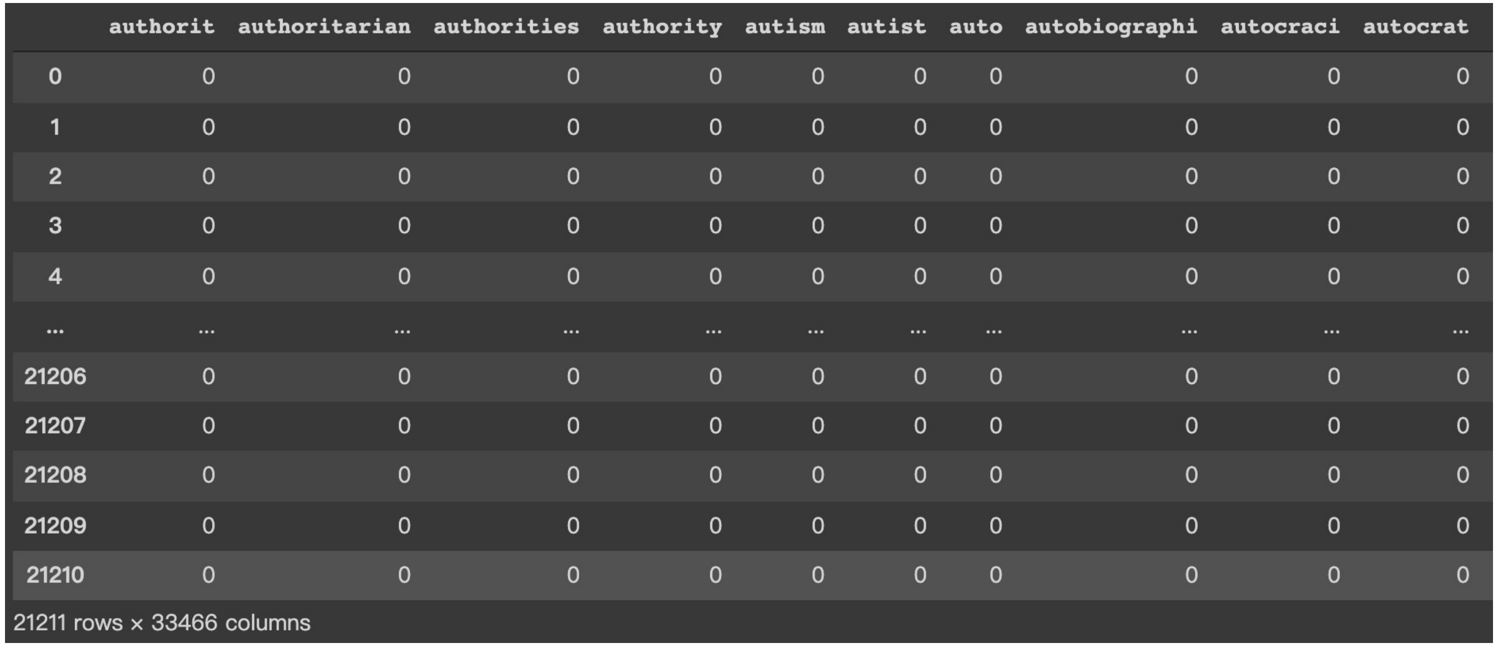


文章向量



```
from sklearn.feature_extraction.text import CountVectorizer
tf_vectorizer = CountVectorizer() #0跟1 不然會是tf的值
tf_vectorizer.fit(df['text']) #得到欄位名稱
tf_X = tf_vectorizer.transform(df['text']).toarray()
```

使用 nltk 詞頻向量模組

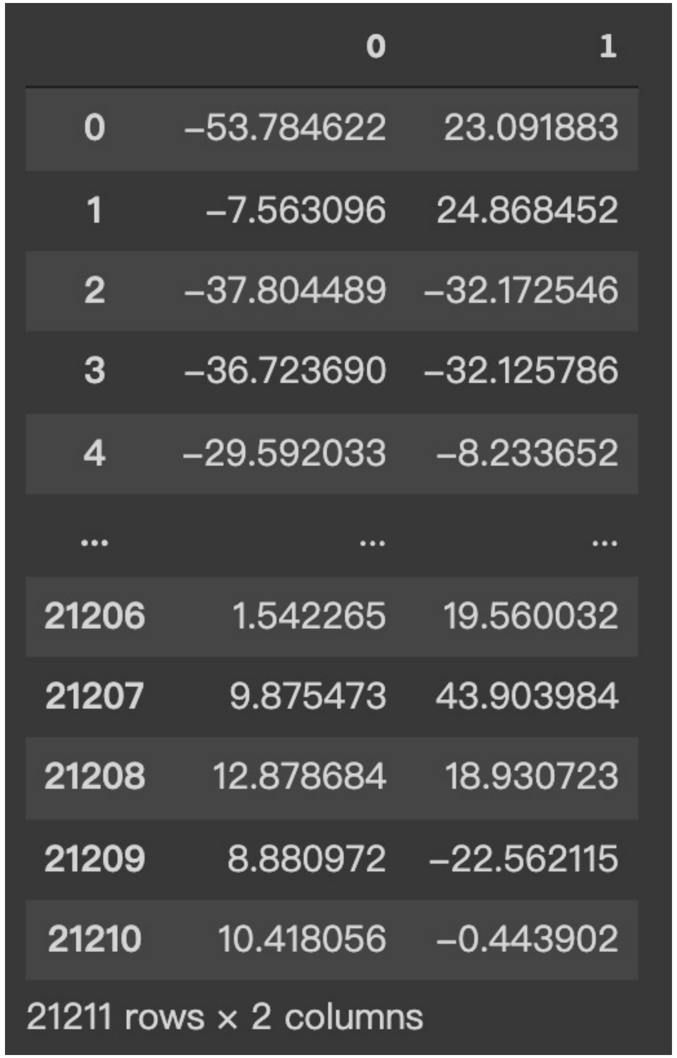


	authorit	authoritarian	authorities	authority	autism	autist	auto	autobiographi	autocraci	autocrat
0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...
21206	0	0	0	0	0	0	0	0	0	0
21207	0	0	0	0	0	0	0	0	0	0
21208	0	0	0	0	0	0	0	0	0	0
21209	0	0	0	0	0	0	0	0	0	0
21210	0	0	0	0	0	0	0	0	0	0

21211 rows × 33466 columns

欄位為該詞在此文章的詞頻(tf)

文章向量降維



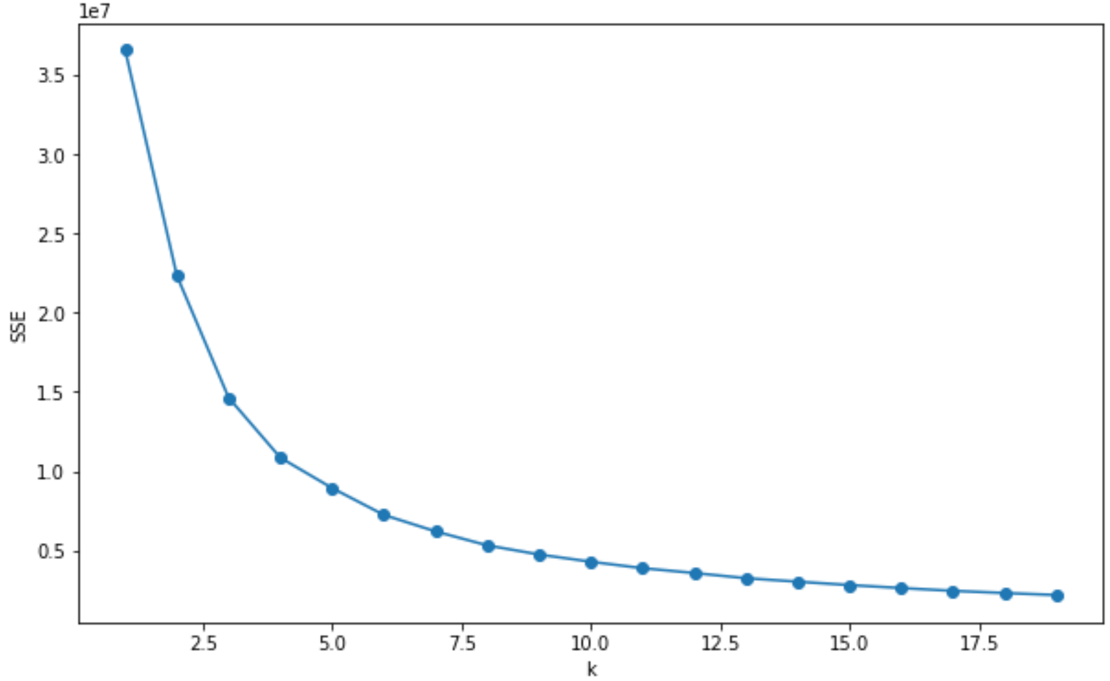
	0	1
0	–53.784622	23.091883
1	–7.563096	24.868452
2	–37.804489	–32.172546
3	–36.723690	–32.125786
4	–29.592033	–8.233652
...	...	...
21206	1.542265	19.560032
21207	9.875473	43.903984
21208	12.878684	18.930723
21209	8.880972	–22.562115
21210	10.418056	–0.443902

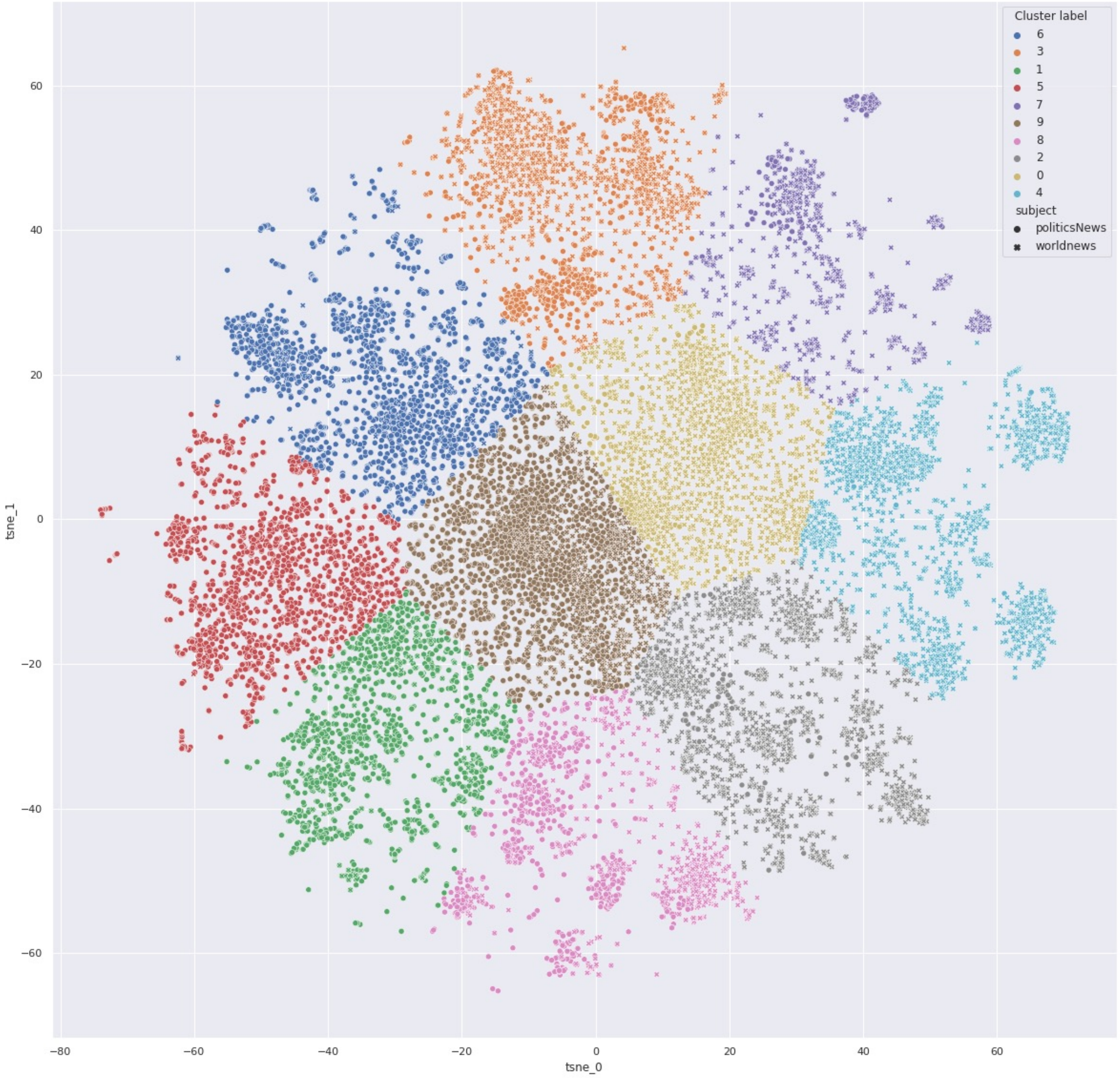
21211 rows × 2 columns

透過t-SNE 將 **33466** 降維成 **2** 維

視覺化結果-1

K-means 手肘法 +
K-means（10群）視覺化結果



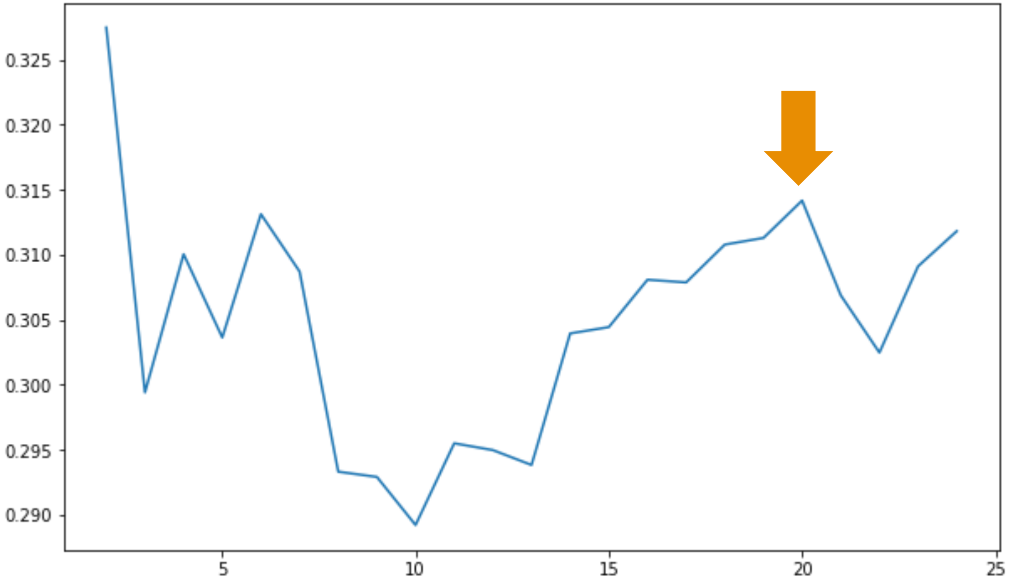


困難：較難找手肘且
分群比較不理想（明顯邊界）

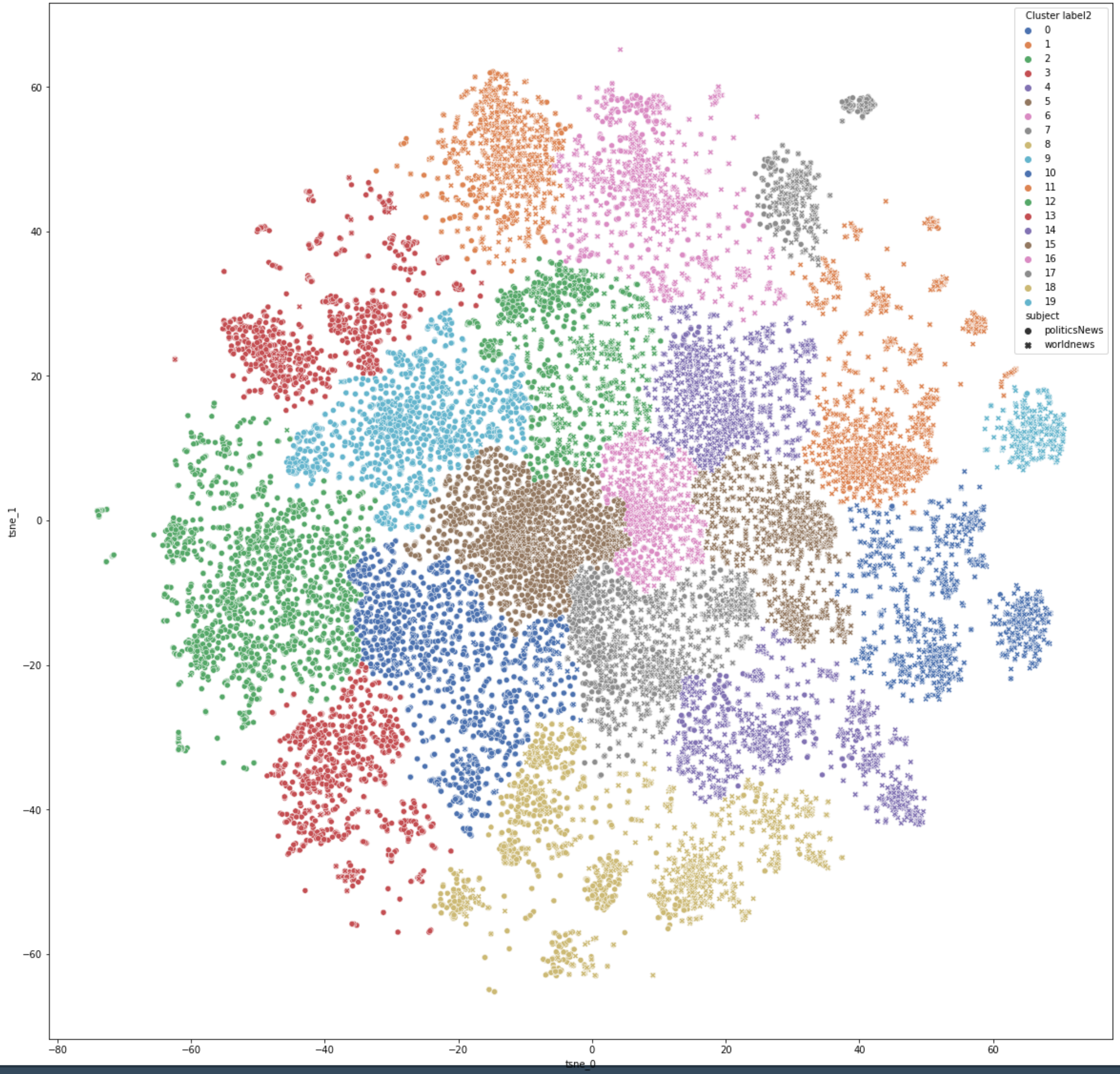
視覺化結果-2

Pb2:
文章的大至分成 20 個主題，
但是未進一步訂定每群主題

輪廓分析法（取大））



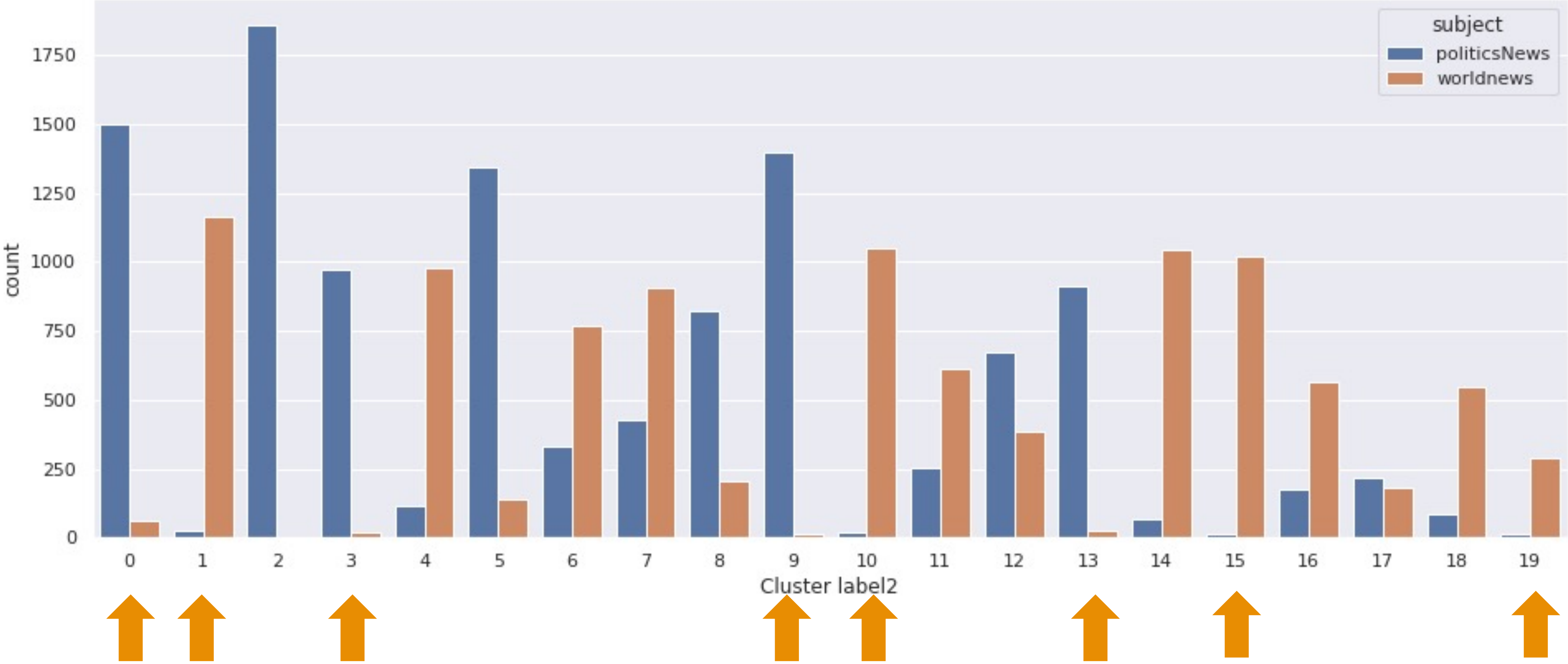
階層式聚類分群結果，群數=20



Bias news

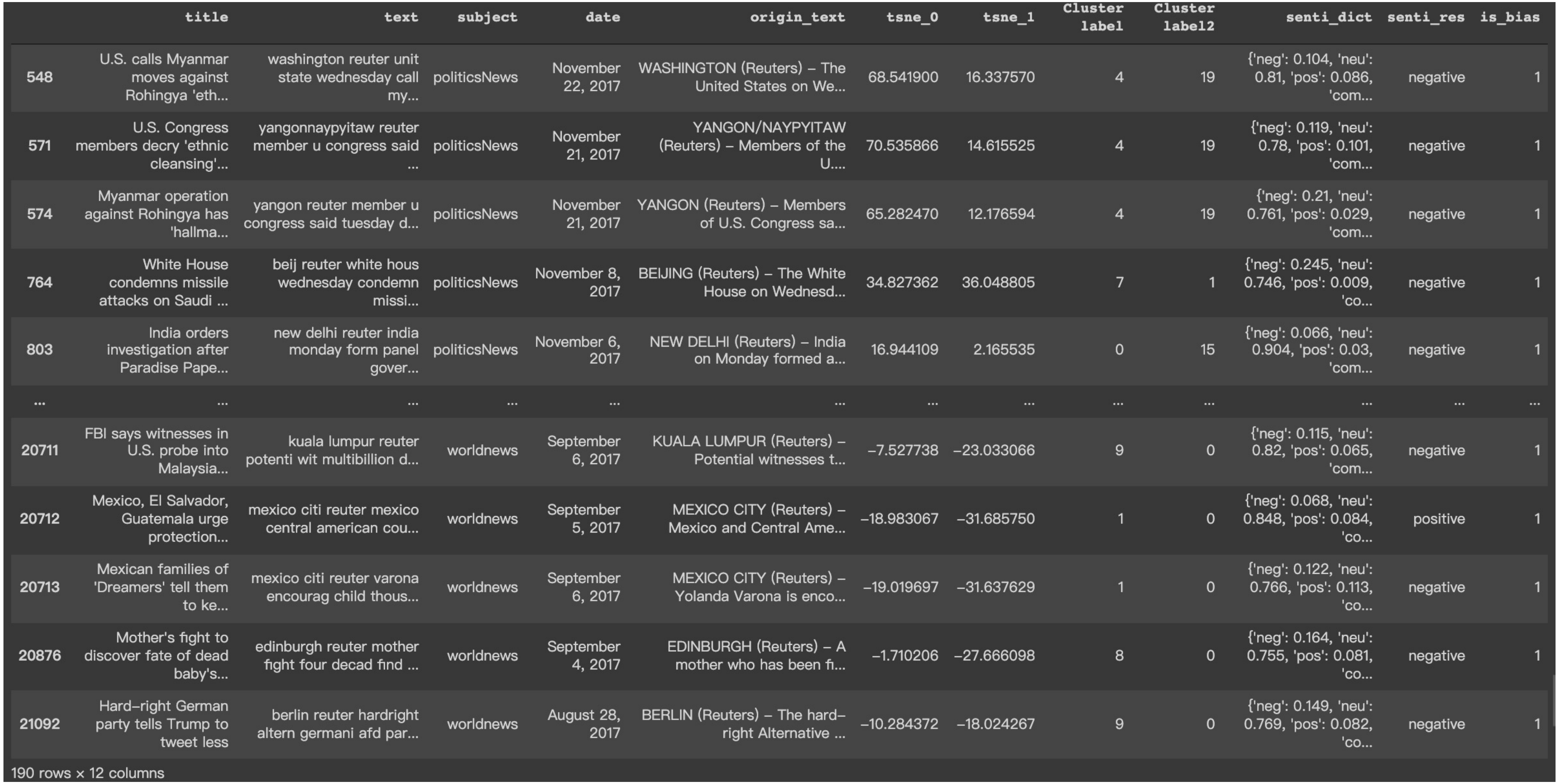
Bias rule:

subject是 Worldnews 且在第 0,3,9,13 群
subject是 PoliticsNews 且在第 1,10,15,19 群



問題

Pb3: 有 190 篇新聞視為偏誤的文章，其文章內容可能與subject較不符合



	title	text	subject	date	origin_text	tsne_0	tsne_1	Cluster label	Cluster label2	senti_dict	senti_res	is_bias
548	U.S. calls Myanmar moves against Rohingya 'eth...	washington reuter unit state wednesday call my...	politicsNews	November 22, 2017	WASHINGTON (Reuters) - The United States on We...	68.541900	16.337570	4	19	{'neg': 0.104, 'neu': 0.81, 'pos': 0.086, 'com...	negative	1
571	U.S. Congress members decry 'ethnic cleansing'...	yangonnaypyitaw reuter member u congress said ...	politicsNews	November 21, 2017	YANGON/NAYPYITAW (Reuters) - Members of the U....	70.535866	14.615525	4	19	{'neg': 0.119, 'neu': 0.78, 'pos': 0.101, 'com...	negative	1
574	Myanmar operation against Rohingya has 'hallma...	yangon reuter member u congress said tuesday d...	politicsNews	November 21, 2017	YANGON (Reuters) - Members of U.S. Congress sa...	65.282470	12.176594	4	19	{'neg': 0.21, 'neu': 0.761, 'pos': 0.029, 'com...	negative	1
764	White House condemns missile attacks on Saudi ...	beij reuter white hous wednesday condemn missi...	politicsNews	November 8, 2017	BEIJING (Reuters) - The White House on Wednesd...	34.827362	36.048805	7	1	{'neg': 0.245, 'neu': 0.746, 'pos': 0.009, 'co...	negative	1
803	India orders investigation after Paradise Pape...	new delhi reuter india monday form panel gover...	politicsNews	November 6, 2017	NEW DELHI (Reuters) - India on Monday formed a...	16.944109	2.165535	0	15	{'neg': 0.066, 'neu': 0.904, 'pos': 0.03, 'com...	negative	1
...	...	...	...	...	...	...	...	...	...	...	...	...
20711	FBI says witnesses in U.S. probe into Malaysia...	kuala lumpur reuter potenti wit multibillion d...	worldnews	September 6, 2017	KUALA LUMPUR (Reuters) - Potential witnesses t...	-7.527738	-23.033066	9	0	{'neg': 0.115, 'neu': 0.82, 'pos': 0.065, 'com...	negative	1
20712	Mexico, El Salvador, Guatemala urge protection...	mexico citi reuter mexico central american cou...	worldnews	September 5, 2017	MEXICO CITY (Reuters) - Mexico and Central Ame...	-18.983067	-31.685750	1	0	{'neg': 0.068, 'neu': 0.848, 'pos': 0.084, 'co...	positive	1
20713	Mexican families of 'Dreamers' tell them to ke...	mexico citi reuter varona encourag child thous...	worldnews	September 6, 2017	MEXICO CITY (Reuters) - Yolanda Varona is enco...	-19.019697	-31.637629	1	0	{'neg': 0.122, 'neu': 0.766, 'pos': 0.113, 'co...	negative	1
20876	Mother's fight to discover fate of dead baby's...	edinburgh reuter mother fight four decad find ...	worldnews	September 4, 2017	EDINBURGH (Reuters) - A mother who has been fi...	-1.710206	-27.666098	8	0	{'neg': 0.164, 'neu': 0.755, 'pos': 0.081, 'co...	negative	1
21092	Hard-right German party tells Trump to tweet less	berlin reuter hardright altern germani afd par...	worldnews	August 28, 2017	BERLIN (Reuters) - The hard-right Alternative ...	-10.284372	-18.024267	9	0	{'neg': 0.149, 'neu': 0.769, 'pos': 0.082, 'co...	negative	1

190 rows × 12 columns

CASE 5

THANKS