

分析目標

| 目標 A |

從不同**電子媒體**(中央社、聯合新聞網、自由時報電子報、中國時報電子報、蘋果電子報)，看其對不同**候選人**(蔡英文、賴清德、柯文哲、韓國瑜、王金平、朱立倫)的文章**用詞**，是否有差異。

◆ 比較方式

- 從詞頻
- 從textrank
- 從TF-IDF (sklearn)
- 從針對同一個人的相似詞 (genism / Word2vec)

| 目標 B |

看不同電子媒體的文章間，是否有敘述上的差異。

◆ 比較方式：從各媒體文章間的文章向量看相似度 (genism / Doc2vec)



資料處理

| 資料來源 |

<https://www.largitdata.com/>

| 資料型態 | (下載時已設定為包含任 候選人姓名之文章)

	id	url	title	content	region	type	source	date	time
150	a60832d8-5b6b-5fd4-879d-4006334c4589	https://udn.com/news/story/11311/3795329	賴清德簽書會 潘孟安意外現身	民進黨總統初選協調陷入僵局，未來是否納入手機民調須經雙方協調。為拚民調支持度，行政院前院長賴...	台灣	新聞	聯合新聞網	2019-05-05	23:59:00
151	7c853bc6-81eb-58f9-9b6f-92ef15132270	https://udn.com/news/story/11311/3795327	扁新書質疑黨產處理「歷史問題 能一刀切嗎」	爭取民進黨提名的總統蔡英文，走訪地方行程曾多次為推動年改及處理國民黨不當黨產的政策辯護，強調...	台灣	新聞	聯合新聞網	2019-05-05	23:58:00
188	6c096bb6-7db4-5697-82da-fb468ea4f5f4	https://www.chinatimes.com/realtimenews/201905...	時力黃捷刁難自經區翻白眼 洛杉基怒批太陽花	高雄市長韓國瑜日前到議會備詢，時代力量議員黃捷提問自由經濟示範區，韓市長多次以「高雄要發大財...	台灣	新聞	中國時報	2019-05-05	23:56:00

資料處理

| 資料分類 |

◆ 判斷報導「屬於」何候選人，並幫文章標籤。

- 用jieba斷詞後計算候選人名稱出現次數。以報導中提到次數最多的人物為判斷標準。
- 標籤可重複。
- 名稱採計標準：
 - 蔡英文：蔡英文、蔡總統、小英
 - 賴清德：賴清德、賴神
 - 柯文哲：柯文哲、柯P、柯市長、台北市長、臺北市長
 - 韓國瑜：韓國瑜、韓總、韓市長、高雄市長、新市長
 - 郭台銘：郭台銘、郭董
 - 王金平：王金平、王前院長
 - 朱立倫：朱立倫

url	title	content	region	type	source	date	time	...	pusstime	pushauthor	韓	蔡	柯	賴	郭	王	朱	標籤
https://tw.appledaily.com/iplay/article/201905...	打臉韓粉「愛河乾淨可以喝」 瀟如黃河；黑鳥屍噁爆	高雄市長韓國瑜推動愛情產業鏈，曾在愛河河岸舉辦多項活動，近日卻屢次被踢爆愛河鮮如黃河，表面更...	台灣	新聞	蘋果新聞網	2019-05-05	06:00:00	...	NaN	NaN	3	0	0	0	0	0	0	韓國瑜
https://tw.appledaily.com/iplay/article/201905...	空降華視月薪真沒17萬 視網膜：但現在超過了	24歲網紅視網膜（陳子見），因網路新聞「眼球中央電視台」成名，常於鏡頭前自稱「中華民國唯一正...	台灣	新聞	蘋果新聞網	2019-05-08	06:00:00	...	NaN	NaN	1	0	0	0	1	0	0	韓國瑜郭台銘
https://tw.appledaily.com/iplay/article/201905...	博恩開脫口秀DISS初衷全因政治人物太好笑	以「大奶微微」脫口秀出名的YouTuber博恩，以噁練的吐槽風格，狂吸44萬粉絲訂閱，螢幕上...	台灣	新聞	蘋果新聞網	2019-05-07	06:00:00	...	NaN	NaN	1	1	0	1	0	0	0	韓國瑜蔡英文賴清德
https://tw.appledaily.com/new/realtime/2019050...	郭台銘被拱捐財產30年老友：交付信託退居大股東 標率高	鴻海集團董事長郭台銘有意角逐總統大位，財經大老紛紛跳出來給建議，財信傳媒董事長謝金河今天再...	台灣	新聞	蘋果日報	2019-05-05	22:36:00	...	NaN	NaN	0	1	0	0	16	0	0	郭台銘

資料處理

資料分類

- ◆ 將資料依不同媒體、候選人分為35類。
- ◆ 總計9425篇。

	自由時報電子報	蘋果電子報	中央社	中國時報電子報	聯合新聞網	sum
韓國瑜	1197	518	138	986	597	3436
蔡英文	468	279	183	662	436	2028
柯文哲	303	159	57	283	246	1048
賴清德	245	135	94	284	257	1015
郭台銘	335	224	95	165	307	1126
王金平	99	45	24	108	87	363
朱立倫	123	41	22	118	105	409
sum	2770	1401	613	2606	2035	9425



A 從詞頻、textrank、TFIDF與w2v看不同媒體的差異

目標A 執行流程



$(O_Q)(Q_Q)(Q_O)(Q_Q)$

嗚嗚嗚嗚一團混亂



A 從詞頻、textrank、TFIDF與w2v看不同媒體的差異

| 詞頻 | ※by sklearn

	蘋果電子報_詞	詞頻	中國時報電子報_詞	詞頻.1	自由時報電子報_詞	詞頻.2	中央社_詞	詞頻.3	聯合電子報_詞	詞頻.4
0	韓國瑜	3209	韓國瑜	4816	韓國瑜	6402	台灣	1838	台灣	3450
1	台灣	3208	台灣	4528	台灣	5342	總統	1315	韓國瑜	2897
2	總統	2129	總統	3731	總統	3807	今天	919	總統	2499
3	蔡英文	1438	民進黨	3423	中國	2667	韓國瑜	847	郭台銘	1758
4	民調	1425	蔡英文	2938	國民黨	2316	賴清德	781	賴清德	1704
5	郭台銘	1367	國民黨	2563	蔡英文	2243	中央社	765	蔡英文	1628
6	中國	1356	民調	2373	郭台銘	2174	郭台銘	744	民進黨	1624
7	民進黨	1306	賴清德	2182	賴清德	1923	民進黨	593	國民黨	1528
8	國民黨	1292	支持	1712	高雄	1921	編輯	578	柯文哲	1164
9	賴清德	1194	高雄	1666	臉書	1781	禁總統	531	經濟	1064
10	支持	1182	初選	1522	民進黨	1755	台北	527	初選	1044
11	市長	925	經濟	1506	市長	1730	中國	526	民調	1025
12	政治	916	政府	1500	議員	1634	民調	478	問題	1013
13	問題	904	問題	1424	自經區	1610	國民黨	460	政府	1011
14	經濟	868	大陸	1397	台北	1566	媒體	439	自經區	985
15	自經區	853	柯文哲	1395	今天	1566	經濟	427	今天	977



A 從詞頻、textrank、TFIDF與w2v看不同媒體的差異

| textrank |

※by jieba
※利用詞在句子中的前後關係
建立權重，據以判斷每個詞
的重要性。

	中央社	textrank.1	中時	textrank.2	自由	textrank.3	聯合	textrank.4	蘋果	textrank.5
0	台灣	1.000000	台灣	1.000000	韓國瑜	1.000000	台灣	1.000000	台灣	1.000000
1	表示	0.698498	韓國瑜	0.924648	台灣	0.977189	韓國瑜	0.808831	韓國瑜	0.878488
2	韓國瑜	0.464703	表示	0.550603	表示	0.550396	表示	0.719115	表示	0.405091
3	賴清德	0.373799	禁英文	0.503972	禁英文	0.305859	賴清德	0.476880	禁英文	0.367251
4	郭台銘	0.337116	賴清德	0.398928	高雄市長	0.294929	郭台銘	0.422702	賴清德	0.357729
5	禁總統	0.269011	政府	0.302176	郭台銘	0.292856	禁英文	0.416968	郭台銘	0.349855
6	政府	0.245478	支持	0.301129	賴清德	0.279601	柯文哲	0.331333	支持	0.326595
7	希望	0.212830	柯文哲	0.281040	柯文哲	0.247306	政府	0.311637	政治	0.303351
8	指出	0.204725	政治	0.278990	政府	0.216371	政治	0.262089	柯文哲	0.255045
9	日本	0.183793	高雄市長	0.263435	認為	0.211039	希望	0.254751	政府	0.234434
10	台北	0.164448	希望	0.221847	臉書	0.210009	認為	0.233358	認為	0.234311
11	柯文哲	0.162836	認為	0.219979	台北	0.194776	支持	0.221793	高雄市長	0.219316
12	禁英文	0.161896	郭台銘	0.215840	指出	0.182673	高雄市長	0.216242	可能	0.210922
13	支持	0.155995	可能	0.211756	希望	0.181148	大家	0.203224	希望	0.184016
14	認為	0.151465	指出	0.196146	政治	0.169045	朱立倫	0.181213	市府	0.163074
15	大家	0.151061	大家	0.178717	高雄市	0.166429	禁總統	0.168652	民眾	0.152988



A 從詞頻、textrank、TFIDF與w2v看不同媒體的差異

| TFIDF |
※by sklearn

	中央社_詞	TFIDF	中國時報電子報_詞	TFIDF.1	自由時報電子報_詞	TFIDF.2	蘋果電子報_詞	TFIDF.3	聯合電子報_詞	TFIDF.4
0	台灣	0.411216	韓國瑜	0.336840	韓國瑜	0.408729	韓國瑜	0.356473	台灣	0.357698
1	總統	0.294205	台灣	0.316697	台灣	0.341054	台灣	0.356362	韓國瑜	0.300363
2	今天	0.205608	總統	0.260953	總統	0.243054	總統	0.236501	總統	0.259098
3	韓國瑜	0.189499	民進黨	0.239411	中國	0.170272	禁英文	0.159741	郭台銘	0.182270
4	賴清德	0.174733	禁英文	0.205489	國民黨	0.147863	民調	0.158297	賴清德	0.176672
5	中央社	0.171154	國民黨	0.179261	禁英文	0.143202	郭台銘	0.151854	禁英文	0.168792
6	郭台銘	0.166455	民調	0.165972	郭台銘	0.138797	中國	0.150632	民進黨	0.168377
7	民進黨	0.132672	賴清德	0.152613	賴清德	0.122772	民進黨	0.145078	國民黨	0.158424
8	編輯	0.129316	支持	0.119741	高雄	0.122644	國民黨	0.143522	柯文哲	0.120684
9	禁總統	0.118801	高雄	0.116523	臉書	0.113706	賴清德	0.132636	經濟	0.110316
10	台北	0.117906	初選	0.106452	民進黨	0.112046	支持	0.131303	初選	0.108243
11	中國	0.117682	經濟	0.105333	市長	0.110450	市長	0.102754	民調	0.106273
12	民調	0.106943	政府	0.104913	議員	0.104321	政治	0.101754	問題	0.105028
13	國民黨	0.102916	問題	0.099597	自經區	0.102789	問題	0.100421	政府	0.104821
14	媒體	0.098218	大陸	0.097709	台北	0.099980	經濟	0.096422	自經區	0.102125
15	經濟	0.095533	柯文哲	0.097569	今天	0.099980	自經區	0.094756	今天	0.101296

現在
大選
無效
處理
進行
阿扁
2020
柯P
大
杯
活動
台北
這樣
應該

民進黨
台北市長
媒體
台灣
評估
支持
指出
市議員

A 從詞頻、textrank、TFIDF與w2v看不同媒體的差異

| 詞向量 |

※by word2vec

●查詢字：蔡英文

	所有報導	蘋果	中央社	中時	自由	聯合
0	(禁總統, 0.7485520839691162)	(若選上, 0.8385324478149414)	(行政院長, 0.9654462337493896)	(梯子, 0.7191623449325562)	(川普, 0.8103059530258179)	(馬英九, 0.8672090172767639)
1	(小英, 0.7258952260017395)	(柯文哲, 0.8207341432571411)	(前新北市長, 0.961142897605896)	(馬英九, 0.7031769752502441)	(馬英九, 0.7626925706863403)	(川普, 0.8545684814453125)
2	(馬英九, 0.6094121932983398)	(若當, 0.8149627447128296)	(今天, 0.960478663444519)	(民以, 0.693260908126831)	(朱立倫, 0.7456996440887451)	(當, 0.8277769088745117)
3	(呂副, 0.6043545603752136)	(賴清德, 0.8035781383514404)	(前, 0.9462913274765015)	(川普, 0.6838102340698242)	(小英, 0.7446953058242798)	(陳水扁, 0.8277614116668701)
4	(選台灣, 0.5669070482254028)	(川普, 0.8001359701156616)	(高雄市長, 0.9462077021598816)	(選投, 0.6700340509414673)	(馬, 0.7440183758735657)	(複製, 0.8228482604026794)
5	(聽令, 0.5564980506896973)	(參選, 0.7955875396728516)	(和鴻海, 0.9282220602035522)	(郭, 0.6695970296859741)	(蘇貞昌, 0.7363768815994263)	(當選, 0.813261866569519)
6	(禁, 0.5412042140960693)	(小英, 0.7899320125579834)	(總統, 0.9136713147163391)	(她當, 0.6656322479248047)	(前, 0.6961289048194885)	(參選, 0.7825443744659424)
7	(她當, 0.5239473581314087)	(王金平, 0.7838467359542847)	(賴清德, 0.907250702381134)	(小英, 0.6654617786407471)	(有意思, 0.6930927038192749)	(小英, 0.7734874486923218)
8	(川普, 0.5062028169631958)	(蘇貞昌, 0.7756749391555786)	(郭台銘, 0.8898956775665283)	(呂副, 0.6576769351959229)	(想當, 0.6906386613845825)	(面前, 0.7556087970733643)
9	(馬, 0.49379169940948486)	(朱立倫, 0.772159218788147)	(吳敦義, 0.8891596794128418)	(馬, 0.6521287560462952)	(王金平, 0.6851978302001953)	(陳建仁, 0.7541725635528564)
10	(直選, 0.47619009017944336)	(馬英九, 0.770453929901123)	(廣播節目, 0.8835947513580322)	(當選, 0.6491608023643494)	(參選, 0.6701527833938599)	(一任, 0.7508127689361572)
11	(曖昧不明, 0.47391247749328613)	(一夕, 0.7597014904022217)	(董事長, 0.8832857608795166)	(參選, 0.6472944617271423)	(郭, 0.6679664850234985)	(蘇貞昌, 0.7413616180419922)
12	(選投, 0.4659371078014374)	(阮昭雄, 0.759059488773346)	(主席, 0.8815983533859253)	(下屆, 0.6429166793823242)	(禁總統, 0.6672786474227905)	(她當, 0.7387921810150146)
13	(陳水扁, 0.4478181004524231)	(我當, 0.7552016377449036)	(馬英九, 0.8771017789840698)	(直轄, 0.6403439044952393)	(人當, 0.6660230755805969)	(行政院長, 0.7327979803085327)
14	(蘇貞昌, 0.4441763162612915)	(分得, 0.7510144710540771)	(院長, 0.8730279207229614)	(身分, 0.6381335258483887)	(他當, 0.6560335159301758)	(防盜, 0.7313763499259949)



A 從詞頻、textrank、TFIDF與w2v看不同媒體的差異

| 詞向量 |

※by word2vec

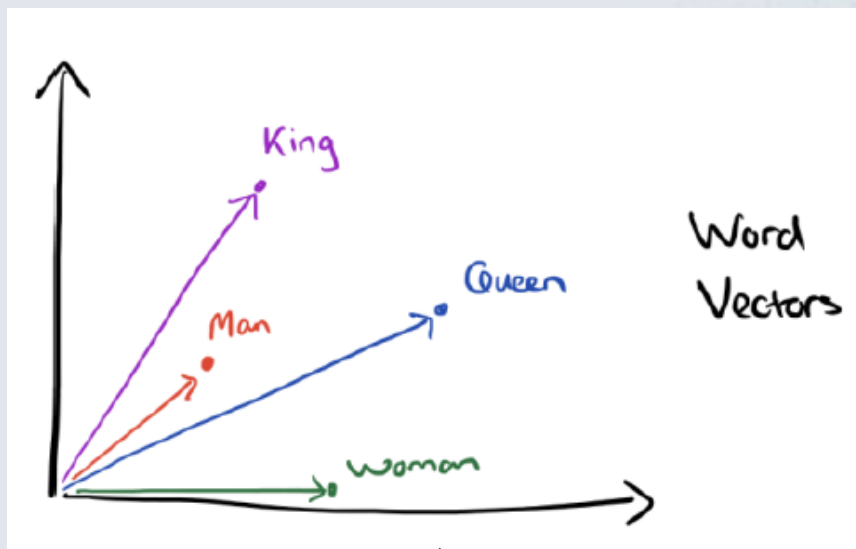
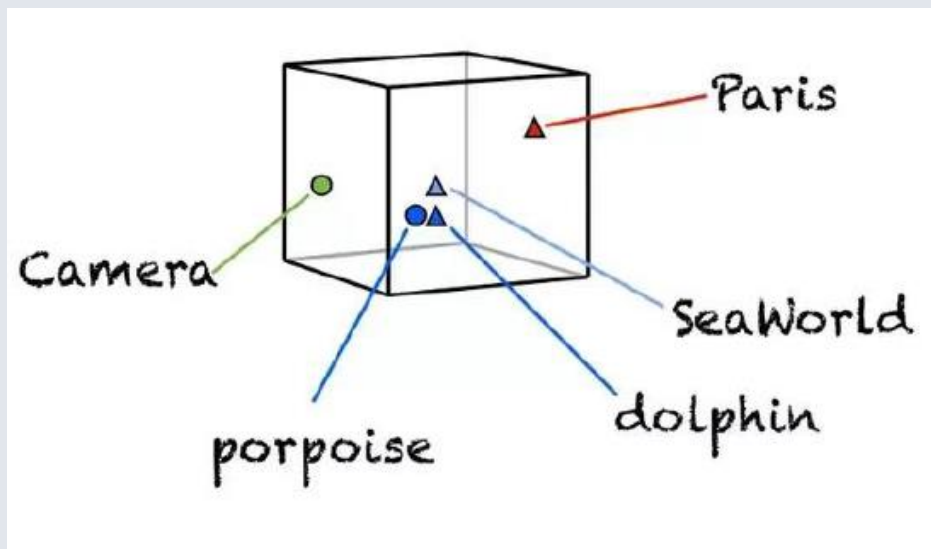
●查詢字：韓國瑜

	所有報導	蘋果	中央社	中時	自由	聯合
0	(韓, 0.6878882646560669)	(阮昭雄, 0.8668632507324219)	(朱立倫, 0.9780074954032898)	(柯P, 0.6536548733711243)	(柯文哲, 0.7631950378417969)	(柯文哲, 0.8663908839225769)
1	(韓市長, 0.6598403453826904)	(柯, 0.8381662368774414)	(蔡總統, 0.9747902154922485)	(蔡總統, 0.6513549089431763)	(韓, 0.7490894198417664)	(林, 0.8610551357269287)
2	(韓的, 0.6064039468765259)	(萬安, 0.8331362009048462)	(接受, 0.9647842645645142)	(韓, 0.6432051062583923)	(郭, 0.7461196184158325)	(柯, 0.8556434512138367)
3	(陳其邁, 0.532960832118988)	(韓, 0.8329923152923584)	(媒體, 0.962814450263977)	(綠營, 0.6355359554290771)	(郭董, 0.722244381904602)	(蘇貞昌, 0.8447281122207642)
4	(他, 0.49731600284576416)	(蔣, 0.8262654542922974)	(柯文哲, 0.9604077935218811)	(柯文哲, 0.6343829035758972)	(朱立倫, 0.6728173494338989)	(韓, 0.8378952741622925)
5	(韓流, 0.4787111282348633)	(游錫堃, 0.8215013146400452)	(賴清德, 0.9466328620910645)	(韓市長, 0.6335500478744507)	(蘇貞昌, 0.6706538200378418)	(朱立倫, 0.8324960470199585)
6	(柯P, 0.4525686800479889)	(郭董, 0.808106541633606)	(表示, 0.9435243606567383)	(外界, 0.6331431865692139)	(蔡總統, 0.6705432534217834)	(他, 0.8313823938369751)
7	(新市長, 0.45170050859451294)	(蘇貞昌, 0.8078925609588623)	(而來, 0.9412826299667358)	(震驚, 0.6284695863723755)	(康裕成, 0.6690094470977783)	(蔡總統, 0.8297350406646729)
8	(對方, 0.45065921545028687)	(他, 0.8067440986633301)	(受訪, 0.9405523538589478)	(開車, 0.6270347833633423)	(王, 0.6616251468658447)	(陳其邁, 0.8260768055915833)
9	(黃捷, 0.4505678415298462)	(聖哲, 0.8055979013442993)	(李孟諺, 0.9383920431137085)	(韓的, 0.625450849533081)	(陳其邁, 0.6564271450042725)	(賴清德, 0.8254470825195312)
10	(林智鴻, 0.44796043634414673)	(賴清德, 0.8022460341453552)	(王金平, 0.93697190284729)	(一路, 0.6247544884681702)	(侯友宜, 0.6526901125907898)	(媒體, 0.8193050622940063)
11	(康裕成, 0.4457937180995941)	(她, 0.798859179019928)	(問及此事, 0.9346781373023987)	(更連, 0.6231253743171692)	(在會談中, 0.6458081603050232)	(卡韓, 0.8157269358634949)
12	(選總統, 0.4410300552845001)	(郭, 0.7954713702201843)	(郭台銘, 0.9334653615951538)	(韓流, 0.6182922124862671)	(潘恒旭, 0.6457281112670898)	(王金平, 0.814621090888977)
13	(朱立倫, 0.44067028164863586)	(一夕, 0.7868301272392273)	(游錫堃, 0.9303030371665955)	(李佳芬, 0.6126223802566528)	(兩人, 0.6397140026092529)	(陳水扁, 0.8142418265342712)
14	(跳針, 0.43899229168891907)	(王鴻薇, 0.7856839895248413)	(李晏榕, 0.9217962622642517)	(孫大千, 0.6122015714645386)	(砸, 0.6372174620628357)	(外界, 0.8071917295455933)

B 電子媒體文章間的相似性

| Vector Space Model 空間向量模型 |

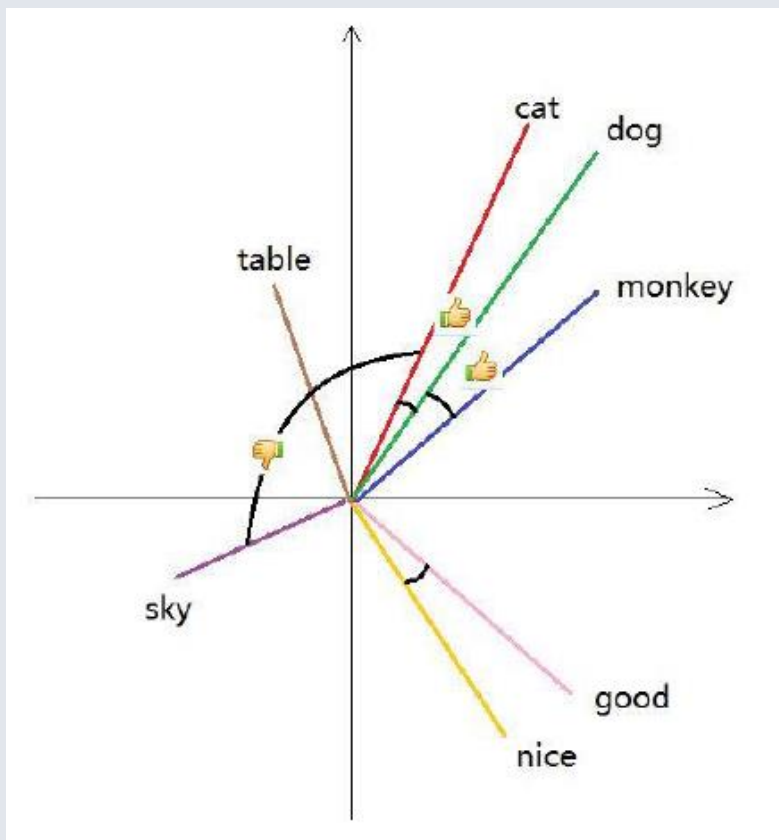
◆ Word Embedding 建立字詞向量 (Word Vector)



夾角越小，字詞(文章)相似度越高。

B 電子媒體文章間的相似性

| Vector Space Model 空間向量模型 |



◆ 文章相似度之計算

- Cosine similarity
- 值介於0 ~ 1之間
 - 1:兩篇文章相同
 - 0:兩篇文章完全不同
- $\text{score} = \text{sim}(d, q)$
- rank and select top-scoring



B 電子媒體文章間的相似性

| CKIP中文詞彙向量 |

◆ w2v_CNA_ASBC_300d.vec

已訓練好的、共517,015個字詞 X 300維向量的資料集。

517015 300↓

</s> 0.001334 0.001473 -0.001277 -0.001093 0.000456 0.001007 0.000314 0.000070 -0.001201 0.000739 -0.001452 0.000417 -0.000250 -
, 0.109873 0.305187 0.055786 -0.204820 -0.052799 -0.087339 0.018882 0.107194 -0.103589 0.020845 -0.032038 -0.050904 -0.023751 -0.0
的 0.065502 0.410419 0.171388 -0.250337 -0.161988 -0.131177 -0.026453 0.018281 -0.097931 0.088129 -0.016080 -0.020586 0.038033 0.
。 0.030542 0.295328 0.285679 -0.077693 -0.222915 -0.052738 0.154245 0.098286 -0.066998 0.027795 0.027546 0.085325 0.049827 0.06
、 0.220428 0.251800 0.171489 -0.245964 0.063861 -0.142520 0.016984 -0.209704 0.013219 -0.035599 -0.162719 -0.125440 -0.001546 0.
在 0.150613 0.373346 0.141946 -0.140266 -0.087243 -0.281908 -0.017116 0.105825 -0.061439 0.010109 0.089515 -0.075474 0.010003 0.0
「 0.011989 0.306780 0.154640 -0.333304 0.004869 -0.056118 -0.222169 0.109915 -0.076349 0.073621 0.192804 -0.007725 0.148064 -0.0
」 0.012244 0.316489 0.173132 -0.321783 -0.000873 -0.019910 -0.184563 0.104150 -0.059957 0.072824 0.182865 0.031786 0.138528 -0.0
是 0.095198 0.271638 0.254880 -0.324430 -0.162894 -0.086183 -0.000879 -0.004193 -0.088286 -0.013450 -0.017316 0.005516 0.137432 0.
一 0.051683 0.280593 0.213069 -0.299573 -0.221149 -0.131377 -0.125912 0.100719 0.084247 -0.021310 0.091330 -0.050152 0.034088 0.0
將 0.149498 0.399703 0.114655 -0.066493 0.008630 -0.195956 -0.006352 0.076218 -0.060790 0.187824 0.014881 -0.040152 0.053350 -0.1
有 0.018656 0.236593 0.248120 -0.199100 -0.088172 0.103156 0.035791 0.086370 -0.128381 0.009284 -0.068267 -0.023028 0.033590 0.0

B 電子媒體文章間的相似性

| CKIP中文詞彙向量 |

◆ 使用方法

```
In [1]: 1 import gensim
        2 from gensim.models import KeyedVectors
        3 from gensim.models import Word2Vec
        4
```

```
In [3]: 1 model=KeyedVectors.load_word2vec_format('w2v_CNA_ASBC_300d.vec',binary=False,encoding = "utf8",unicode_errors='ignore')
```

```
In [5]: 1 model.similarity("蔡英文","賴清德")
```

```
Out[5]: 0.31934118
```

```
In [8]: 1 model.similarity("國民黨","民進黨")
```

```
Out[8]: 0.79556143
```

同黨候選人大不同，藍綠政黨較相似。



B 電子媒體文章間的相似性

| Word2vec 和 Doc2vec |

- ◆ 都是gensim 套件，訓練方式類似，原理不同。

- ◆ Word2vec

詞袋模型，每次訓練只截取句子中一小部分詞訓練，忽略除了本次訓練詞以外該句子中的其他詞。向量表達只是每個詞的向量加在一起。缺點是忽略了文本詞序問題。

- ◆ Doc2vec

非監督式演算法，在輸入層添加句子向量「Paragraph vector」，以預測出的向量來表示各個的文檔。此結構克服了詞袋模型的缺點。



B 電子媒體文章間的相似性

| 以Doc2Vec計算不同媒體報導的相似度 |

- ◆ 以108年5月1日五大媒體之新聞資料做訓練。

```
In [1]: 1 import jieba
        2 import jieba.posseg as pos
        3 import codecs
        4 import pandas as pd
        5 import numpy as np
        6 import gensim
        7 from gensim import corpora, models, similarities
        8 from gensim.models.doc2vec import Doc2Vec, TaggedDocument
```

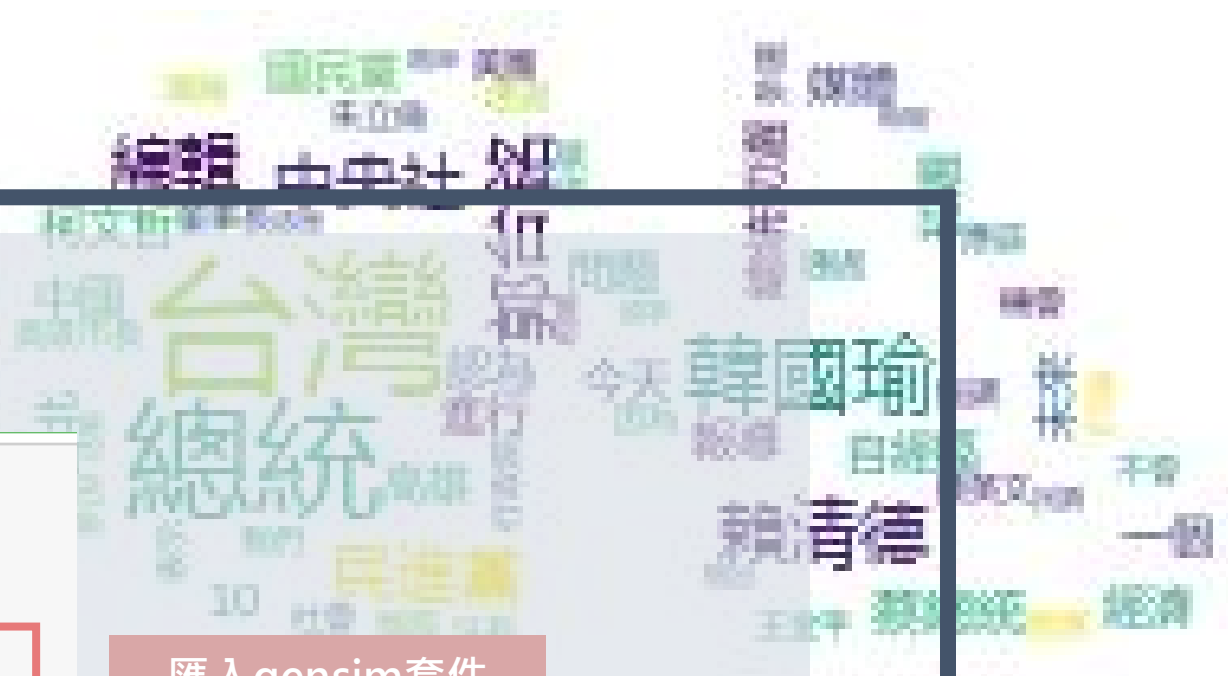
匯入gensim套件

```
In [4]: 1 jieba.set_dictionary("dict.txt.big")
        2 jieba.load_userdict("userdict.txt")

Building prefix dict from C:\Users\USER\NLP\dict.txt.big ...
Loading model from cache C:\Users\USER\AppData\Local\Temp\jieba.u8be97d5cc3017653aa149
Loading model cost 2.618 seconds.
Prefix dict has been built successfully.
```

```
In [5]: 1 def cut_sentence(text):
        2     result=[]
        3     for i in range(len(text)):
        4         each_cut=jieba.cut(text[i],cut_all=False)
        5         result.append(" ".join(filter(lambda x: x not in stopwords, each_cut)))
        6     return result
```

jieba斷字



B 電子媒體文章間的相似性

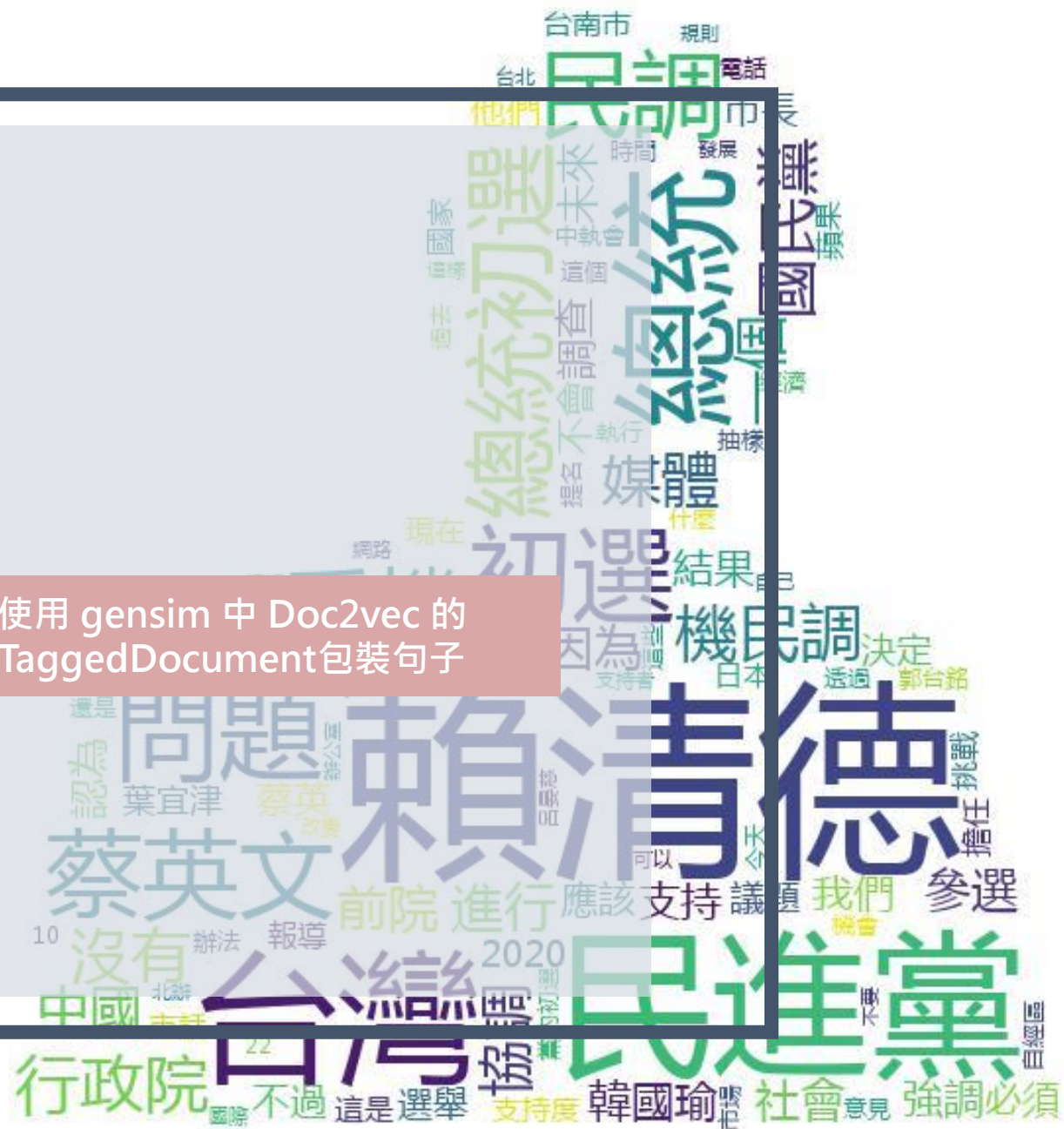
| 以Doc2Vec計算不同媒體報導的相似度 |

◆ 設定Doc2Vec輸入層

```
In [7]: 1 TaggedDocument = gensim.models.doc2vec.TaggedDocument

In [24]: 1 def X_train(cut_sentence):
2     x_train=[]
3     cut_store=[]
4     for i,text in enumerate(cut_sentence):
5         word_list=text.split(' ')
6         cut_store.append(word_list)
7         l=len(word_list)
8         word_list[l-1]=word_list[l-1].strip()
9         document=TaggedDocument(word_list,tags=[i])
10        print(document)
11        x_train.append(document)
12        df['cutwords']=cut_store
13        return x_train
14
```

使用 gensim 中 Doc2vec 的
TaggedDocument 包裝句子



B 電子媒體文章間的相似性

以Doc2Vec計算不同媒體報導的相似度

◆ Doc2vec輸入data格式

```
c=X_train(b)
```

C

`TaggedDocument(words=['感觉','不错','','','一日游','不错','选择'],tags=[12]),`

`TaggedDocument(words=['有趣','hai','xing'],tags=[13]),`

`TaggedDocument(words=['荡气回肠','', '', '10','年','', , '留下来','照片','', , '必然','再','!', 'br','n'],tags=[14]),`

`TaggedDocument(words=['景色','超级','棒','', , '美丽','故事','', , '乘船','游览','', , '沿湖','浏览','', , '累','乘坐','观光车','!','关键','门票','!','!', '!'],tags=[15]),`

`TaggedDocument(words=['南锣鼓巷','北京市','中心','一条','胡同','', , '地理位置','靠近','什刹海','', , '成为','北京','休闲','娱乐','好去处','', , '特别','外国','游人','特别','', , '许多','中国','符号','文化','商品','买','', , '餐饮','商家','各具','鲜明','特点','', , '使','整个','地区','一种','中西合璧','感觉','。','每次','世界杯','期间','', , '气氛','无比','热烈','', , '来到','一定','感染','', , '之前'银西'准备'充足','', , ''],tags=[16])`

`TaggedDocument(words=['个人感觉','卖','小商品','地方','', , '便宜','', , '晚上','夜景','挺','好看']tags=[17]),`

`TaggedDocument(words=['性价比','超高'],tags=[18]),`

`TaggedDocument(words=['挺','普通','', , '楼下','拍','几张','图片','', , '反正','进不去'],tags=[19]),`

`TaggedDocument(words=['太大','', , '里面','走','好长时间','不到','五分之一','。','周围'],ags=[20]),`

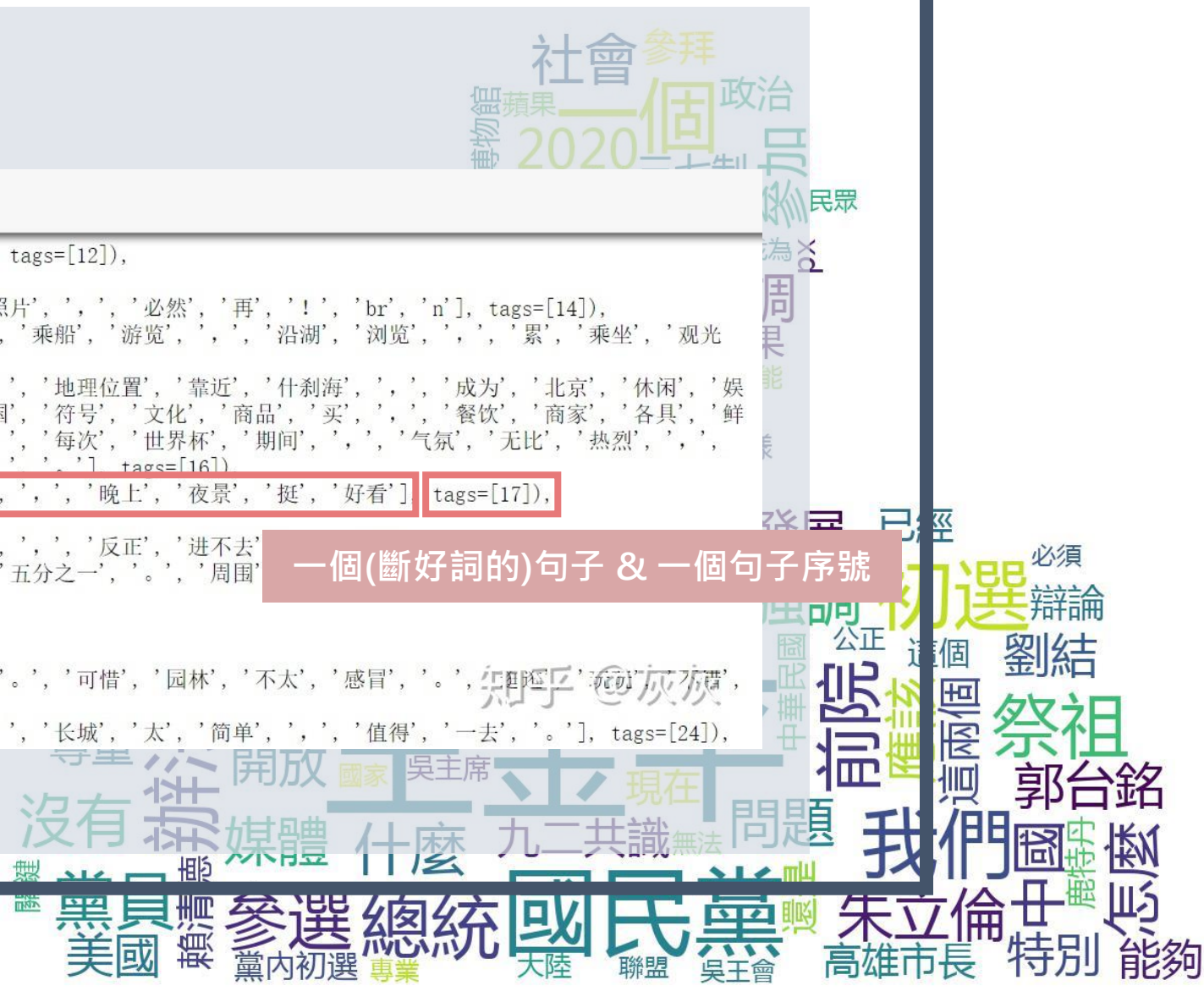
`TaggedDocument(words=['迪士尼'],tags=[21]),`

`TaggedDocument(words=['亲子','游'],tags=[22]),`

`TaggedDocument(words=['苏州','总是','欣赏','一下','古典','园林','。','可惜','园林','不太','感冒','。','通过'游玩'为精','。'],tags=[23]),`

`TaggedDocument(words=['不到长城非好汉','', , '爬','华山','来说','', , '长城','太','简单','', , '值得','一去','。'],tags=[24]),`

一個(斷好詞的)句子 & 一個句子序號



B 電子媒體文章間的相似性

| 以Doc2Vec計算不同媒體報導的相似度 |

◆ 訓練模型

```
1 model_dm=Doc2Vec(c,min_count=1, window = 5, size = 300, sample=1e-3, negative=5, workers=4)
2 model_dm.train(c, total_examples=model_dm.corpus_count, epochs=10)
3 model_dm.save("doc2vecmodel")
```

```
1 model_dm = Doc2Vec.load("doc2vecmodel")
```

- size：特徵向量的維度，預設為100，這裡設為300。
- min_count：詞頻少於min_count次數的單字會被丟棄。
- window：表示當前詞與預測詞在一個句子中的最大距離(即在當前詞往前、往後幾個字以內)。
- sample：高頻詞彙的隨機降採樣的配置閾值。
- workers：用於控制訓練的並行數。



B 電子媒體文章間的相似性

| 以Doc2Vec計算不同媒體報導的相似度 |

◆ 預測文本向量

```
1 model_dm = Doc2Vec.load("doc2vecmodel")
2 str1='郭台銘 見 川普 ? 白宮 : 周二 沒 安排 但 不 排除 可能性'
3 test_text=str1.split(' ')
4
5 #推測文本的向量
6 inferred_vector_dm = model_dm.infer_vector(doc_words=test_text,alpha=0.025,steps=500)
7 print(inferred_vector_dm)
8
```

```
[-0.16190535  0.2539202  1.0946412 -1.1268462  0.5784756  0.5092622
 0.418299  0.23291637 -0.4645703  0.22751336 -0.04762065 -0.15540494
 1.0412499 -1.1477922 -1.2879146  0.12593672  0.55762494  0.38904643
-0.09067855  0.6635338 -0.7640902  0.388205  0.9108293  0.5166687
 0.5620671 -0.03177515 -0.21766117  0.35372043  0.15297693 -0.22031955
 0.22581752 -0.3626845 -0.58652323 -0.9841081 -0.4768838  0.3480377
 0.53794163  0.12346935 -0.62255263 -0.15973766 -0.08609062  0.00937343
 0.03289044 -0.33033484 -0.40064624 -0.2831377 -0.80798036  0.82742846
-1.1849666  0.4446896  0.15719397  0.03834464 -0.10774498  0.36945826
-0.5210649  0.2538464  0.47869563  0.706772  0.81741637 -0.38555747
 0.1478742  0.61294144  0.27758154  0.50616115  0.5791702  1.0690265
 0.2507171 -0.24921432 -0.18642041  0.0042947  0.55943763  0.37207264]
```

媒體 董事長 這是 新北市 還是 說明 兩岸 希望 台灣 侯友宜 蔡英文 台北 民進黨 認為 一個 問題 勞工 王金平 臉書 提出 2020 總統 韓國瑜 自由貿易 發展 活動 前新北市長 中國 我們

不會 桃園 什麼 必須 機場 針對 市長 決定 中國 我們

B 電子媒體文章間的相似性

| 以Doc2Vec計算不同媒體報導的相似度 |

◆ 計算各家新聞文本向量，取最高相似者。

➤ 以108年5月1日五大媒體關於總統候選人之文本做分析。

```
1 vector=[]
2 simi_press=[]
3 for i in range(len(df)):
4     inferred_vector_dm = model_dm.infer_vector(doc_words=df['cutwords'][i],alpha=0.025,steps=100)
5     sims = model_dm.docvecs.most_similar([inferred_vector_dm], topn=1)
6     vector.append(sims[0][1])
7     most_simi=df.loc[sims[0][0],"source"]
8     simi_press.append(most_simi)
9
10 df['vec']=vector
11 df['simi_press']=simi_press
```



B 電子媒體文章間的相似性

| 以Doc2Vec計算不同媒體報導的相似度

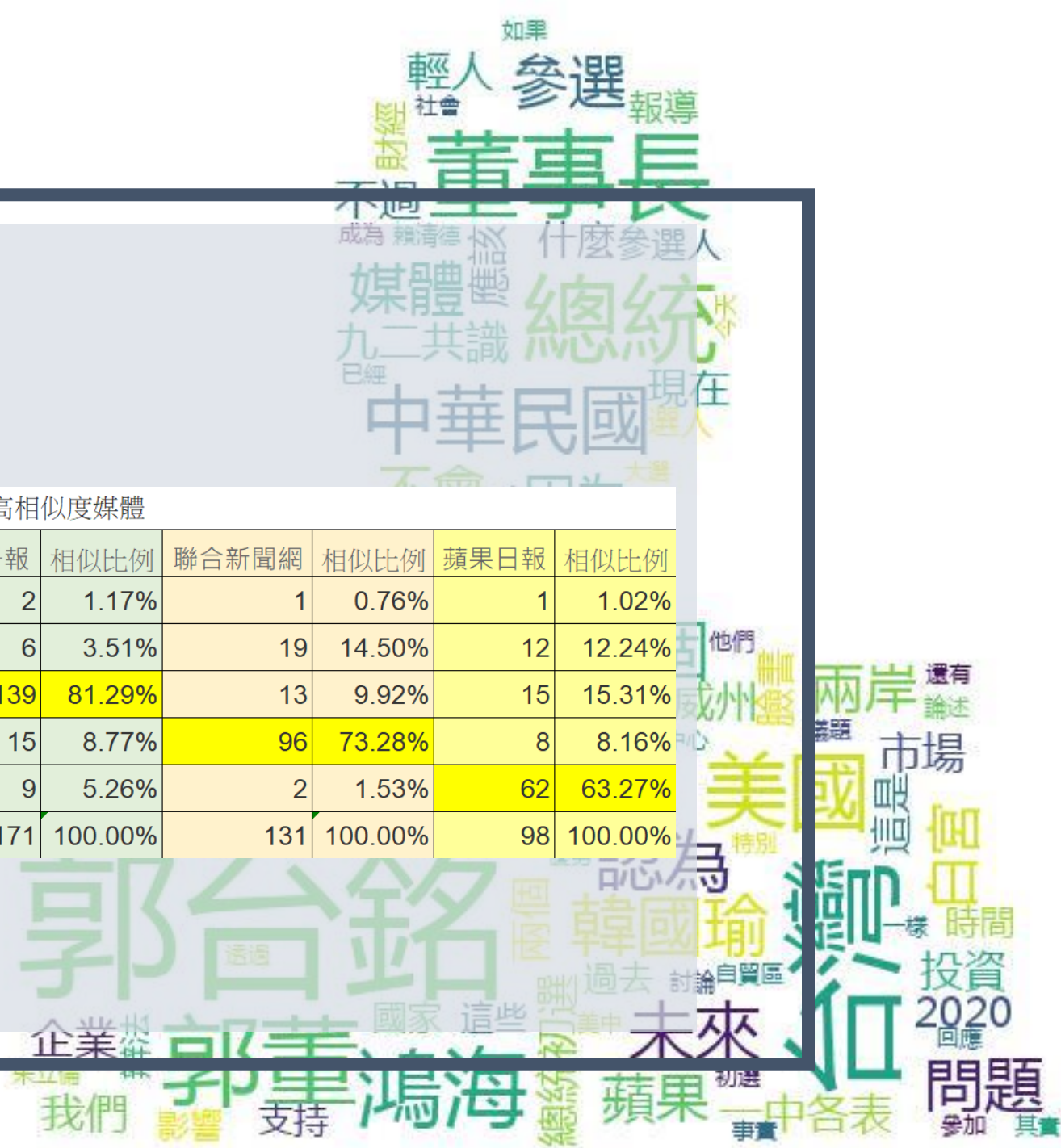
◆ 計算各家新聞文本向量，取最高相似者。

➤ 分析結果

A	B	C	D	E	F
	content	source	cutwords	vector	simi_press
0	稍早外電報導引述美	蘋果日報	['稍早', '外電報導', '引述', '美國	0.869073272	蘋果日報
1	高雄市長韓國瑜昨天	蘋果日報	['高雄市長', '韓國瑜', '昨天', '參	0.875359654	自由時報電子報
2	表態爭取國民黨總統	蘋果日報	['表態', '爭取', '國民黨', '總統', ']	0.828693509	蘋果日報
3	鴻海董事長郭台銘昨	蘋果日報	['鴻海', '董事長', '郭台銘', '飛往	0.767674685	蘋果日報
4	高雄市長韓國瑜今天	蘋果日報	['高雄市長', '韓國瑜', '今天', '世	0.712661743	中央社
5	民進黨今天中常會討	蘋果日報	['民進黨', '今天', '中常會', '討論	0.830900311	蘋果日報
6	總統蔡英文與行政院	蘋果日報	['總統', '蔡英文', '行政院', '前', ']	0.899855971	自由時報電子報
7	高雄市長韓國瑜昨天	蘋果日報	['高雄市長', '韓國瑜', '昨天', '國	0.835886657	蘋果日報
8	總統蔡英文今天下午	蘋果日報	['總統', '蔡英文', '今天下午', '前	0.824768662	蘋果日報
9	伍斯研／國民黨員、	蘋果日報	['伍斯研', '／', '國民黨員', '營業	0.792444289	蘋果日報
10	立法院長前院長王金	蘋果日報	['立法', '院長', '前', '院長', '王金	0.678807616	自由時報電子報
11	高雄市長韓國瑜昨日	蘋果日報	['高雄市長', '韓國瑜', '昨日', '30'	0.843437016	蘋果日報
12	台北市長柯文哲將於	蘋果日報	['台北市長', '柯文哲', '將於', '本	0.652706385	聯合新聞網
13	前總統陳水扁家族涉	蘋果日報	['前', '總統', '陳水扁', '家族', '涉	0.566683769	蘋果日報
14	針對美國通過《台灣	蘋果日報	['美國', '台灣', '關係法', '承諾', ']	0.914840221	蘋果日報
15	高雄市長韓國瑜在吳	蘋果日報	['高雄市長', '韓國瑜', '吳韓', '會	0.840412736	自由時報電子報
16	高雄市長韓國瑜今天	蘋果日報	['高雄市長', '韓國瑜', '今天', '世	0.809948862	蘋果日報
17	親民黨主席宋楚瑜日	蘋果日報	['親民黨', '主席', '宋楚瑜', '日前	0.905854702	蘋果日報
18	人稱「台南阿姐」的	蘋果日報	['人稱', '台南', '阿姐', '民進黨', ']	0.728755474	蘋果日報
19	今天是五一勞動節，	蘋果日報	['今天', '五一勞動節', '國民黨', ']	0.852750599	蘋果日報
20	高雄市長韓國瑜今天	蘋果日報	['高雄市長', '韓國瑜', '今天', '世	0.793950379	蘋果日報

B 電子媒體文章間的相似性

- ◆ 計算各家新聞文本向量，取最高相似者。
 - 分析結果
 - 5月1日媒體文章最高相似度統計



倒閉 國庫 收場
關心 機制
經濟
什麼 關閉 變成
度 | 初選
九二共識

什麼 關閉 變成 初選 九二共識 藍綠 唱快歌 這樣 問題 加參 選舉

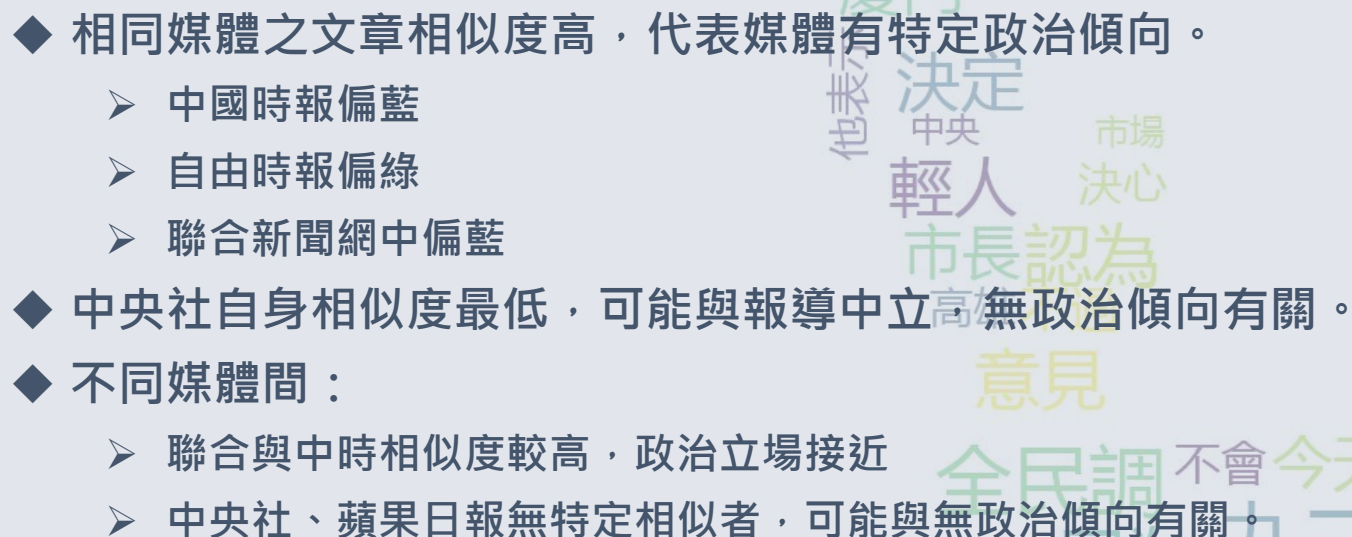
- | | |
|----|--------|
| 日報 | 相似比例 |
| 10 | 5.41% |
| 37 | 20.00% |
| 79 | 42.70% |

開放 總統 台灣 參選
 時間 表示 觀光 開始 意識 成長 回應
 大家 一個 過程 國民黨
 現在 桃園 蔡政府 節目 這些 新北

活動
執政 進行
機場
參選
認為 我們
說明 自由貿易
市長 提名
辯論 爭取 建設

B 電子媒體文章間的相似性

結論

- 
- ◆ 相同媒體之文章相似度高，代表媒體有特定政治傾向。
 - 中國時報偏藍
 - 自由時報偏綠
 - 聯合新聞網中偏藍
 - ◆ 中央社自身相似度最低，可能與報導中立，無政治傾向有關。
 - ◆ 不同媒體間：
 - 聯合與中時相似度較高，政治立場接近
 - 中央社、蘋果日報無特定相似者，可能與無政治傾向有關。



THANK YOU~