

# Homework : TF-IDF and beyond

楊立偉教授

wyang@ntu.edu.tw

© Copyright

# 常見的詞彙處理

---

- 取特徵詞 **feature selection**
  - 選取有效鑑別用的詞 (以利後續分析處理)
- 取關鍵詞 **keyword extraction**
  - 選取有代表意義的詞 (多半是為了人類閱讀)
- 切詞 **tokenization**
  - 將內容切割成多個單元 (以利後續分析處理)
- 斷詞 **word segmentation**
  - (多半依語意及文法) 將內容做 (正確且唯一的) 切詞

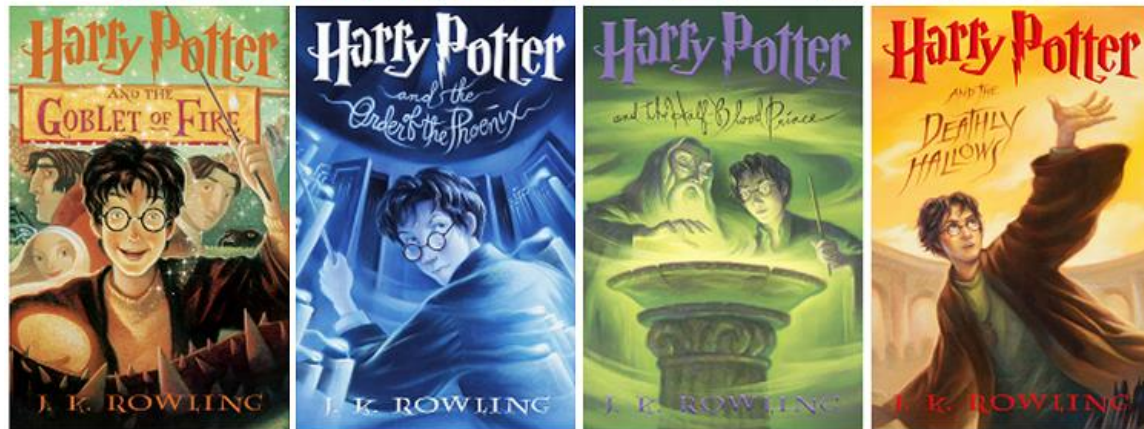
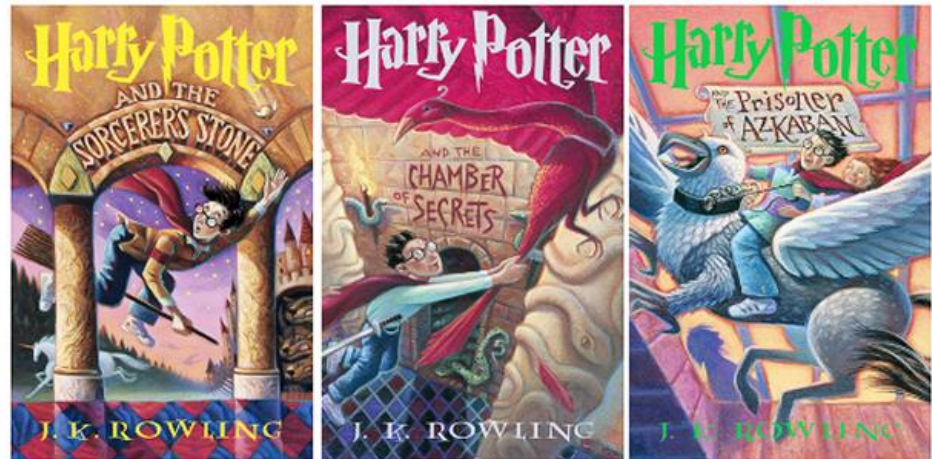
"Harry Potter" may be a good keyword  
but not a good feature for further processing (無鑑別力)

Ex. What is the tf-idf value of "Harry Potter" among the books?

hint: common terms

→ what does 'common' mean?

→ should eliminate or maintain?



- 當年Gerard Salton教授發展tf-idf方法時，面對的是類似圖書館語料：量大、具一般性、跨多主題，因此在大量詞彙的分布下，使用idf加權是很有用的，可以辨識、分離出特定的詞彙。
- 但是在小量、或是單一主題的語料時，idf加權仍可以做為語料中辨識各文件特徵之用。但若是取出該語料的代表詞，例如七本哈利波特的代表詞，應該要包含哈利波特本身才對。針對不同情境，應做適當的調整。

# Chinese Keyword Extraction

---

- Chinese keyword extraction is fundamental for many applications.
- There are two major approaches
  - Need word segmentation 需先斷詞
  - No word segmentation 不需先斷詞

# N-gram approach

---

- No word segmentation 不需先斷詞
- The keywords are in the subset of n-grams
- How to select the proper n-grams for keywords ?
  - tf-idf
  - chi-square (for variance)
  - mutual information
  - information gain, maximum entropy, and others

# N-gram approach with tf-idf

---

- Enumerate n-grams, for example, 2 to 6
- Compute tf and idf
- Sort by tf-idf descendingly
- Post-processing
  - Remove non-words Ex. 去除純英數字或特殊字元
  - Merge sub-keywords Ex. 蔡英、英文、蔡英文，應該只保留英文及蔡英文

# Demonstration

---

- 小語料集練習
  - 2016年1~11月「全部」財經新聞 90507篇
  - 其中與產業相關的「產業」新聞 5896篇
  - 其中內容有鴻海或郭台銘字樣的「鴻海」新聞 2081篇
- 預處理表格
  - 2 gram及3 gram (去除純英數字或含特殊符號者)
  - 依 tf 排序，保留出現50次以上者



	A	B	C	D	E	F	G	H	
1	90507	doc							
2	1192367	gram							
3	編號	詞	TF	DF					
4	1	表示	100945	54184					
5	2	台灣	81884	27124					
6	3	今天	58701	38822					
7	4	公司	44799	18932					
8	5	市場	43295	18346					
9	6	總統	41582	14326					
10	7	美國	41515	16231					
11	8	政府	40217	19108					
12	9	今年	39867	21477					
13	10	大陸	38863	13641					
◀ ▶	全部_2gram	全部_3gram	產業_2gram	產業_3gram	鴻海_2gram	鴻海_3gram			

編號	詞	TF	DF	TF-IDF	全部TF	全部DF	全部TF-IDF	TF期望值	DF期望值	TF卡方值(保留正負號)	DF卡方值(保留正負號)	MI(用DF)	Lift(用DF)
1	台灣	8276	2510	1.82	81884	27124	3.09	5334	1767	1622	312	-4.80	1.42
2	市場	7829	2951	1.47	43295	18346	3.91	2820	1195	8894	2580	-4.56	2.47
3	公司	7770	2735	1.63	44799	18932	3.84	2918	1233	8065	1828	-4.61	2.22
4	今年	6437	2822	1.54	39867	21477	3.50	2597	1399	5677	1447	-4.65	2.02
5	表示	6376	3527	1.07	100945	54184	1.34	6576	3530	-6	0	-4.96	1.00
6	億元	5977	2005	2.24	31050	11605	4.90	2023	756	7730	2064	-4.53	2.65
7	銀行	5049	1154	3.33	21216	6674	6.03	1382	435	9729	1190	-4.53	2.65
8	去年	4562	2084	2.10	26571	14724	4.28	1731	959	4630	1319	-4.62	2.17
9	投資	4542	1598	2.64	27548	10981	4.98	1795	715	4206	1089	-4.61	2.23
10	董事	4472	1840	2.35	19173	9227	5.24	1249	601	8317	2554	-4.47	3.06

- 以Excel進行各項實驗

# Beyond TF-IDF

---

- 1.依領域修正 Mutual Information (MI)
- 2.依分布做修正 chi square ( $\chi^2$ )
- 3.依分布做修正 lift

以下分別做介紹

# 1.依領域修正 using Mutual Information

---

- In addition to term weighting, need to consider the relevance between terms and the topic
- use Mutual Information
$$\text{tf-idf} * \text{MI}$$

- Mutual Information

$$MI = \log \frac{P(x, y)}{P(x)P(y)} = \log \frac{\frac{f(x, y)}{N}}{\frac{f(x)}{N} \frac{f(y)}{N}} = \log \frac{f(x, y)}{f(x)f(y)}$$

P : probability

N : size of the corpus

f(x) : the occurrences of term x in the corpus

f(y) : the occurrences of term x in the corpus

f(x,y) : the co-occurrences of term x and y in the corpus

- Mutual Information

- larger MI means the tendency of co-occurrences of term  $x$  and  $y$  值越大其共現率越高
- using  $\text{tf-idf} * \text{MI}$  may extract topic-related keywords more precisely

- Example

- 以政治類的前  $n$  個關鍵詞 ( $w_1 \dots w_n$ ) 重新計算
- 假設
  - $f(\text{政治}, w_i)$ : 同時出現 政治 與  $w_i$  的篇數
  - $f(\text{政治})$ : 出現 政治 的篇數
  - $f(w_i)$ : 出現  $w_i$  的篇數

- x 該類別之事件
- y 該詞之事件
- x,y 該詞在該類別之事件

							$\frac{f(x,y)}{f(x)f(y)}$		
		$f(x,y)$		$f(x)$	$f(y)$				
no	term	tf	df	N	tf-idf	df in all corpus	MI	tf-idf * MI	Rank
1	歐晉德	93	12	421	15.8	114	0.0003	0.0039	5
2	以色列	28	3	421	15.5	62	0.0001	0.0018	9
3	小白兔	26	3	421	15.2	16	0.0004	0.0068	3
4	能源	39	6	421	15.0	285	0.0001	0.0008	10
5	金門	53	9	421	14.8	150	0.0001	0.0021	8
6	不分區立委	27	4	421	14.6	15	0.0006	0.0093	1
7	假釋	15	1	421	14.5	6	0.0004	0.0057	4
8	募兵捐	34	6	421	14.4	29	0.0005	0.0071	2
9	李登輝	26	4	421	14.4	36	0.0003	0.0038	6
10	禁閉	26	4	421	14.4	60	0.0002	0.0023	7



## 2.依分布做修正 using chi square

- 對照更大的語料統計資料(背景知識)，判斷哪些是突出的
- 解決IDF的問題；更貼近人工閱讀用的關鍵詞
- Use  $\chi^2$  (chi square)

IDF會受到極端稀有詞影響，例如某單篇才有的用法，其IDF值大

$$\chi^2 = \frac{\sum(O-E)^2}{E}$$

O: observed value

E: expected value

- $\chi^2$  常用在類別檢定。假設每一篇文章都當成是「一類」，若「馬英九」一詞總共出現1000次，出現在10篇中，每篇的期望次數是 $1000/10=100$ 次。此時若有一篇出現「馬英九」200次，代表這個詞對這篇有特別意義，其 $\chi^2$  為 $(200-100)^2/100$
- 如果以類來看，有政治類60篇，運動類40篇，共100篇文章，「馬英九」總共出現1000次，期望在政治類應該出現 $1000*60/100=600$ 次，結果實際發現「馬英九」在政治類出現800次，代表這個詞對這類有特別意義，則 $\chi^2$  為 $(800-600)^2/600$

- 所以 $\chi^2$ 可用來挑出與「類別」更相關的詞，其中「類別」的單位可以是「篇」或是任何一種區分用的「類」都可以
- 換句話說， $\chi^2$ 可用來去掉那些與「類別」不相關的，例如說每一類都平均出現的詞
- $\chi^2$ 在概念上與IDF異曲同工 (挑出分布特別的)，因為有平方項放大，所以更明顯，概念上更通用
- $\chi^2$ 有平方項，特別顯著或不顯著的都會被放大，故應設法保留其正負號

### 3.依分布做修正 using lift

---

- 對照更大的語料統計資料(背景知識) ，判斷哪些是突出的

$$\text{lift} = \frac{\frac{\text{該詞出現在該類別之篇數}}{\text{該類別篇數}}}{\frac{\text{該詞出現之篇數}}{\text{總篇數}}}$$

# 合併詞處理

- n-gram可能切出多餘的子字串

原詞	DF	gram	DF	gram	DF
蔡英文	5573	蔡英	5576	英文	6283
希拉蕊	1219	希拉	1237	拉蕊	1219

- 除非有別種意義 (用法)，不然子字串之出現必包含於完整字串中
  - 以DF判斷較TF佳 (why?)
  - 若彼此gram的DF相等或在誤差內 (ex. 1%)，則合併gram
    - 蔡英、希拉、拉蕊可合併；"英文"不可合併




假設各篇來源不同，TF易受高頻次的單篇(單來源)所影響；相對的，DF代表更多的來源是這樣使用。

# Requirement (1) 個人作業

---

- 將Excel的各項指標完成
  - 全部、產業、鴻海新聞的2gram及3gram (共6張表)，  
進行觀察 (此項為練習，不需繳交)
- 挑出最具代表的keyword
  - 需為合併2gram及3gram的結果
  - 全部、產業、鴻海新聞各挑選20個keyword
  - 說明你挑選的方法和原因，連同結果寫在1頁內的  
Word，進行繳交

## 名稱

-  **exercise\_more\_news (2019).xlsx** 額外練習的語料集(不需繳交)
-  **hw1\_table.xlsx** 個人練習檔案，觀察結果，將心得寫在Word繳交
-  **hw1\_text.xlsx** 分組作業語料集 (需繳交Excel、尾附影片的簡報檔)

將預設公式向下填滿

hw1\_table.xlsx - Excel

Willie Yang

檔案 常用 插入 版面配置 公式 資料 校閱 檢視 說明 告訴我您想做什么

E4

$$=(1+\text{LOG}(C4))*\text{LOG}(\$A\$1/D4)$$

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	5896	doc												
2	327219	gram												
3	編號	詞	TF	DF	TF-IDF	全部TF	全部DF	全部TF-IDF	TF期望值	DF期望值	TF卡方值(保留正負號)	DF卡方值(保留正負號)	MI(用DF)	Lift(用DF)
4	1	台灣	8276	2510	1.82	81884	27124	3.09	5334	1767	1622	312	-4.80	1.42
5	2	市場	7829	2951	1.47	43295	18346	3.91	2820	1195	8894	2580	-4.56	2.47
6	3	公司	7770	2735	1.63	44799	18932	3.84	2918	1233	8065	1828	-4.61	2.22
7	4	今年	6437	2822	1.54	39867	21477	3.50	2597	1399	5677	1447	-4.65	2.02
8	5	表示	6376	3527	1.07	100945	54184	1.34	6576	3530	-6	0	-4.96	1.00
9	6	億元	5977	2005	2.24	31050	11605	4.90	2023	756	7730	2064	-4.53	2.65
10	7	銀行	5049	1154	3.33	21216	6674	6.03	1382	435	9729	1190	-4.53	2.65
11	8	去年	4562	2084	2.10	26571	14724	4.28	1731	959	4630	1319	-4.62	2.17
12	9	投資	4542	1598	2.64	27548	10981	4.98	1795	715	4206	1089	-4.61	2.23
13	10	董事	4472	1840	2.35	19173	9227	5.24	1249	601	8317	2554	-4.47	3.06
14	11	產品	4413	1649										
15	12	營收	4332	1307										
16	13	成長	4305	1838										
17	14	產業	4139	1563										
18	15	企業	3451	1294										
19	16	目前	3413	2134										
20	17	指出	3369	2305										
21	18	全球	3286	1553										
22	19	金融	3240	1010										

全部\_2gram 全部\_3gram 產業\_2gram 產業\_3gram 鴻海\_2gram 鴻海\_3gram

就緒

平均值: 7439.24 項目個數: 100 加總: 743923.56

115%



使用篩選功能，依照TF遞減排序，觀察前幾名的詞

hw1\_table.xlsx - Excel

Willie Yang

檔案 常用 插入 版面配置 公式 資料 校閱 檢視 說明 告訴我您想做什么 共用

C3			TF												
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	
1	5896	doc													
2	327219	gram													
3	編號	詞	TF	DF	TF-IDF	全部TF	全部DF	全部TF-IDF	TF期望值	DF期望值	TF卡方值(保留正負號)	DF卡方值(保留正負號)	MI(用DF)	Lift(用DF)	
4	1	台灣	8276	2510	1.82	81884	27124	3.09	5334	1767	1622	312	-4.80	1.42	
5	2	市場	7829	2951	1.47	43295	18346	3.91	2820	1195	8894	2580	-4.56	2.47	
6	3	公司	7770	2735	1.63	44799	18932	3.84	2918	1233	8065	1828	-4.61	2.22	
7	4	今年	6437	2822	1.54	39867	21477	3.50	2597	1399	5677	1447	-4.65	2.02	
8	5	表示	6376	3527	1.07	100945	54184	1.34	6576	3530	-6	0	-4.96	1.00	
9	6	億元	5977	2005	2.24	31050	11605	4.90	2023	756	7730	2064	-4.53	2.65	
10	7	銀行	5049	1154	3.33	21216	6674	6.03	1382	435	9729	1190	-4.53	2.65	
11	8	去年	4562	2084	2.10	26571	14724	4.28	1731	959	4630	1319	-4.62	2.17	
12	9	投資	4542	1598	2.64	27548	10981	4.98	1795	715	4206	1089	-4.61	2.23	
13	10	董事	4472	1840	2.35	19173	9227	5.24	1249	601	8317	2554	-4.47	3.06	
14	11	產品	4413	1649	2.57	19224	8523	5.42	1252	555	7977	2155	-4.48	2.97	
15	12	營收	4332	1307	3.03	16743	5871	6.21	1091	382	9632	2235	-4.42	3.42	
16	13	成長	4305	1838	2.35	24950	11053	4.93	1625	720	4418	1736	-4.55	2.55	
17	14	產業	4139	1563	2.66	20906	8943	5.35	1362	583	5663	1650	-4.53	2.68	
18	15	企業	3451	1294	2.99	19331	8381	5.46	1259	546	3814	1025	-4.58	2.37	
19	16	目前	3413	2134	2.00	28748	20342	3.54	1873	1325	1267	494	-4.75	1.61	
20	17	指出	3369	2305	1.85	38517	27718	2.87	2509	1806	295	138	-4.85	1.28	
21	18	全球	3286	1553	2.62	21551	11557	4.77	1404	753	2523	850	-4.64	2.06	
22	19	金融	3240	1010	3.46	16292	6863	5.84	1061	447	4472	709	-4.60	2.26	

全部\_2gram 全部\_3gram 產業\_2gram 產業\_3gram 鴻海\_2gram 鴻海\_3gram

就緒

平均值: 5012.526316 項目個數: 20 加總: 95238 115%

依照TF-IDF遞減排序，觀察前幾名的詞

hw1\_table.xlsx - Excel

Willie Yang

檔案 常用 插入 版面配置 公式 資料 校閱 檢視 說明 告訴我您想做什么

共用

E3															
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	
1	5896	doc													
2	327219	gram													
3	編號	詞	TF	DF	TF-IDF	全部TF	全部DF	全部TF-IDF	TF期望值	DF期望值	TF卡方值(保留正負號)	DF卡方值(保留正負號)	MI(用DF)	Lift(用DF)	
4	6805	寬庭	51	1	10.21	59	3	12.41	4	0	579	3	-4.25	5.12	
5	7093	捷流	50	1	10.18	50	1	13.38	3	0	671	13	-3.77	15.35	
6	6262	志聖	56	4	8.71	62	8	11.32	4	1	668	23	-4.07	7.68	
7	5692	神準	61	6	8.33	109	33	10.44	7	2	409	7	-4.51	2.79	
8	3472	永冠	97	11	8.15	245	56	10.87	16	4	411	15	-4.48	3.02	
9	5710	朝陽	61	7	8.15	318	86	10.58	21	6	78	0	-4.86	1.25	
10	6731	致茂	52	6	8.13	79	22	10.47	5	1	427	15	-4.33	4.19	
11	1876	越鋼	169	18	8.12	306	57	11.16	20	4	1115	55	-4.27	4.85	
12	5310	精測	65	8	8.07	301	64	10.96	20	4	105	4	-4.67	1.92	
13	6060	光洋	57	7	8.06	202	30	11.50	13	2	146	13	-4.40	3.58	
14	3856	味全	88	11	8.04	440	88	10.97	29	6	123	5	-4.67	1.92	
15	1639	外勞	191	22	7.97	569	153	10.41	37	10	639	15	-4.61	2.21	
16	2535	王品	130	17	7.91	547	107	10.94	36	7	250	14	-4.57	2.44	
17	3636	紀珠	93	13	7.89	202	58	10.55	13	4	484	23	-4.42	3.44	
18	3642	李紀	93	13	7.89	208	61	10.52	14	4	466	21	-4.44	3.27	
19	6531	塵器	53	8	7.81	141	70	9.80	9	5	209	3	-4.71	1.75	
20	6541	吸塵	53	8	7.81	145	69	9.86	9	4	201	3	-4.71	1.78	
21	6587	臺銀	53	8	7.81	128	25	11.06	8	2	239	25	-4.27	4.91	
22	3932	虎航	87	13	7.81	671	143	10.72	44	9	43	1	-4.81	1.40	

就緒

平均值: 8.264826969 項目個數: 20 加總: 157.0317124

115%

wyang@ntu.edu.tw

依照MI(用DF)遞減排序，觀察前幾名的詞

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
	編號	詞	TF	DF	TF-IDF	全部TF	全部DF	全部TF-IDF	TF期望值	DF期望值	TF卡方值(保留正負號)	DF卡方值(保留正負號)	MI(用DF)	Lift(用DF)
1	5896	doc												
2	327219	gram												
3	7093	捷流		50	10.18	50	1	13.38	3	0	671	13	-3.77	15.35
5	2206	北報		147	5.14	171	163	8.87	11	11	1657	1601	-3.83	13.28
6	5311	科昨		65	6.16	76	47	9.46	5	3	728	399	-3.86	12.41
7	6003	海昨		58	5.99	69	51	9.22	4	3	637	405	-3.88	12.04
8	4119	光昨		83	5.80	103	81	9.18	7	5	867	588	-3.89	11.56
9	5069	金昨		68	5.73	89	75	9.09	6	5	667	535	-3.90	11.46
10	4718	板債		73	6.70	112	45	10.07	7	3	592	198	-3.99	9.21
11	5893	紐行		59	6.73	86	37	9.94	6	2	509	159	-4.00	9.13
12	4507	際板		76	6.74	123	47	10.15	8	3	577	187	-4.01	8.82
13	3976	行昨		86	5.58	148	130	9.01	10	8	605	507	-4.02	8.74
14	2404	華車		137	6.50	212	88	10.02	14	6	1099	342	-4.02	8.72
15	7094	G技		50	6.40	76	44	9.54	5	3	410	171	-4.02	8.72
16	5364	今周		64	6.00	122	76	9.49	8	5	395	292	-4.02	8.69
17	5766	電昨		60	5.86	102	82	9.15	7	5	428	309	-4.02	8.61
18	4541	恩平		76	6.89	114	43	10.16	7	3	633	160	-4.02	8.57
19	6322	三席		55	6.10	96	64	9.40	6	4	380	228	-4.03	8.39
20	6727	融圈		52	5.89	100	77	9.21	7	5	318	244	-4.05	7.97
21	3778	管指		90	5.49	186	161	8.99	12	10	501	488	-4.06	7.82
22	7095	投片		50	6.11	99	63	9.46	6	4	294	190	-4.06	7.80

wyang@ntu.edu.tw

# Requirement (2) 分組作業

---

- 實作六項主題
  - 銀行、信用卡、匯率、台積電、台灣、日本
- 列出每一主題的前100名keyword
  - 列舉2~6字詞
  - 去除純英數字或含特殊符號者
  - 合併可能多餘的子字串
  - 列出排名、keyword、tf、df、tf-idf、或其它你們用到的指標，依序貼在 Excel 的工作表，進行繳交

# Deadline

---

- 公布2周後繳交個人及分組作業
  - 繳交期限及方式由助教通知
  - 分組作業除Excel外，需繳簡報檔(尾附影片連結)，另錄製八分鐘內的說明影片，解說成果及過程。影片中若能以程式化處理並實際執行 (live demo) 者加分。
- 分組作業將開放彼此觀摩。

# 程式化處理技巧參考 (1)

---

- 對每一主題
  - 篩選出包含主題關鍵字詞的文章
  - 對每篇文章
    - 使用雙層迴圈及移動指標切割2~6gram，排除純英數字或含特殊符號者
    - 累計每個gram的TF及DF
    - 為避免記憶體不足，可定期對所有gram進行排序，清除TF或DF過低的gram
  - 對每個留下的gram
    - 篩除滿足合併條件 (被包含且次數相近) 之gram
    - 得到一候選之gram list



# 程式化處理技巧參考 (2)

---

- 對全部每篇文章
  - 使用雙層迴圈及移動指標切割2~6gram，排除純英數字或含特殊符號者
  - 累計每個gram在全部文章中的TF'及DF' (有在候選gram list中的才需要做)
- 對候選gram list依分數排序，印出前幾名之詞
  - 可以是TF-IDF、 $\chi^2$ 、MI、Lift，或自行設計之指標

# 補充練習 1 (不需繳交)

---

- 另提供2019全年新聞語料，請練習
  - 取出前50個代表詞
  - 若範圍限縮至財經新聞，取出50個代表詞
  - 若範圍限縮至鴻海新聞，取出50個代表詞
- 說說看你覺得效果好嗎？怎麼樣做可以更好？

## 補充練習 2 (不需繳交)

- 嘗試使用中文斷詞結果取代n-gram
  - 正體中文可參考中研院[CkipTagger](#)或[MONPA](#)
  - 斷詞正確率\* 參考如下

Tool	(WS) prec	(WS) rec	(WS) f1	(POS) acc
CkipTagger	97.49%	97.17%	97.33%	94.59%
CKIPWS(classic)	95.85%	95.96%	95.91%	90.62%
Jieba-zh_TW	90.51%	89.10%	89.80%	--

- 其他加權排序方法不變，一樣試著取出代表詞
- 你覺得斷詞取代n-gram做的效果好嗎？優缺點為何？