
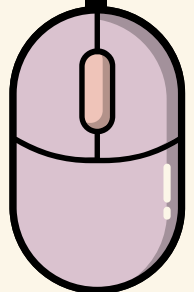


# Lab3

## 自然語言處理套件 與分詞比較



第四組 | 蕭名妍 劉貞莉 黃詩涵

>>>>>

- [illegible]

# 讀取資料與篩選

- 統計"CHANNEL NAME"每個店家的總留言數量
- 後續以「夢工廠」的留言進行分析

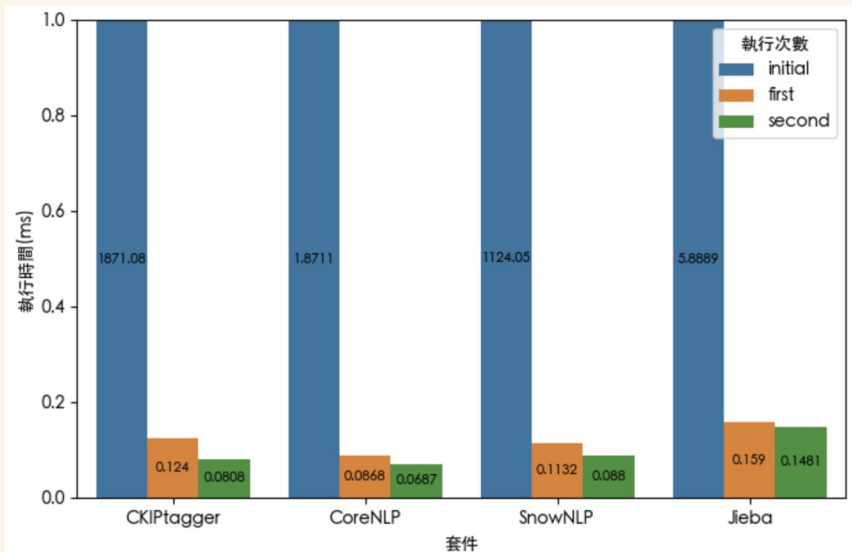
從10,774筆留言中隨機挑選10,000筆

MESSAGE	
CHANNEL NAME	
251.TC(貳伍壹潮流店)	4618
279-現貨直播車	3462
279-現貨直播車(怡君服飾)	192
TR Box寶藏屋：傘的專家、居的職人	10651
夢工場	10774
工讀生寵物	12818
拼鮮水產 足度男直播買賣	12890
蔥媽媽直播	1768
蔥媽媽直播(蔥媽媽食材有限公司)	5621
邦成-自倉(抓單)	44892
阿清服裝	16196

# 比較：套件載入時間

斷詞套件選擇：CKIPtagger、CoreNLP、SnowNLP、Jieba

- 初次載入執行時間：CKIPtagger > SnowNLP > Jieba > CoreNLP
- 後續載入執行時間：Jieba > CKIPtagger > SnowNLP > CoreNLP



- \* initial：重啟 kernel 後，載入單一套件
- \* first、second：以 reload() 重新載入套件兩次

執行次數	套件	執行時間 (ms)
0	initial CKIPtagger	1871.0761
1	initial CoreNLP	1.8711
2	initial SnowNLP	1124.0542
3	initial Jieba	5.8889
4	first CKIPtagger	0.1240
5	first CoreNLP	0.0868
6	first SnowNLP	0.1132
7	first Jieba	0.1590
8	second CKIPtagger	0.0808
9	second CoreNLP	0.0687
10	second SnowNLP	0.0880
11	second Jieba	0.1481

# CKIPtagger 斷詞結果

套件	id	原始留言	斷詞結果	執行時間 (ms)	累計執行時間 (ms)
0	CKIPtagger	0	幫我改8118舖四雙人、8202舖四加大	16.2849	16.2849
1	CKIPtagger	1	我兒子👉超愛60細棉夏被的，都早早去睡了	13.4110	29.6960
2	CKIPtagger	2	昨天下的單子打錯了. 請幫我消掉. 謝謝	11.7331	41.4290
3	CKIPtagger	3	請問這個跟剛才2組3980的有何不同呢？	8.2331	49.6621
4	CKIPtagger	4	阿公說女神剪短髮好俏麗，像女大學生👩	10.4289	60.0910
...	...	...	...	...	...
9995	CKIPtagger	9995	7S -77 珍珠天絲 醋酸纖維(乙酸)\n\n生產工藝\n醋酸纖維取材於可再生的木漿與...	262.9828	318266.2921
9996	CKIPtagger	9996	△停~止~關~鍵~字~△\n持續分享+按讚👍👍👍\n△停~止~關~鍵~字~△\n持續分享...	278.9309	318545.2230
9997	CKIPtagger	9997	多美丽的玫瑰花\n多可爱的玫瑰花\n我就这样深深的爱上她\n多美丽的玫瑰花\n多可爱的玫瑰花...	280.7581	318825.9811
9998	CKIPtagger	9998	△停~止~關~鍵~字~△\n持續分享+按讚👍👍👍\n△停~止~關~鍵~字~△\n持續分享...	318.4311	319144.4123
9999	CKIPtagger	9999	△停~止~關~鍵~字~△\n持續分享+按讚👍👍👍\n△停~止~關~鍵~字~△\n持續分享...	375.3932	319519.8054
10000 rows x 6 columns					

# CoreNLP 斷詞結果

套件	id	原始留言	斷詞結果	執行時間 (ms)	累計執行時間 (ms)
0	CoreNLP	0	幫我改8118舖四雙人、8202舖四加大	13.5398	13.5398
1	CoreNLP	1	我兒子👉超愛60細棉夏被的，都早早去睡了	31.8511	45.3908
2	CoreNLP	2	昨天下的單子打錯了。請幫我消掉。謝謝	21.0023	66.3931
3	CoreNLP	3	請問這個跟剛才2組3980的有何不同呢？	19.8679	86.2610
4	CoreNLP	4	阿公說女神剪短髮好俏麗，像女大學生👩	10.7379	96.9989
...	...	...	...	...	...
9995	CoreNLP	9995	7S -77 珍珠天絲 醋酸纖維(乙酸)\n\n生產工藝\n醋酸纖維取材於可再生的木漿與...	11.4338	44294.0545
9996	CoreNLP	9996	△停~止~關~鍵~字~△\n持續分享+按讚👍👍👍\n△停~止~關~鍵~字~△\n持續分享...	9.6440	44303.6985
9997	CoreNLP	9997	多美丽的玫瑰花\n多可爱的玫瑰花\n我就这样深深的爱上她\n多美丽的玫瑰花\n多可爱的玫瑰花...	11.6048	44315.3033
9998	CoreNLP	9998	△停~止~關~鍵~字~△\n持續分享+按讚👍👍👍\n△停~止~關~鍵~字~△\n持續分享...	11.0898	44326.3931
9999	CoreNLP	9999	△停~止~關~鍵~字~△\n持續分享+按讚👍👍👍\n△停~止~關~鍵~字~△\n持續分享...	33.3230	44359.7162
10000 rows x 6 columns					

# SnowNLP 斷詞結果

	套件	id	原始留言	斷詞結果	執行時間 (ms)	累計執行時間 (ms)
0	SnowNLP	0	幫我改8118舖四雙人、8202舖四加大	幫/我/改/8118/舖/四/雙/人/、8202/舖/四/加大	0.9739	0.9739
1	SnowNLP	1	我兒子😊超愛60細棉夏被的，都早早去睡了	我/兒子/😊/超/愛/60/細/棉夏/被/的/，/都/早早/去/睡/了	1.9121	2.8861
2	SnowNLP	2	昨天下的單子打錯了。請幫我消掉。謝謝	昨/天/下/的/單/子/打/錯/了/./請/幫我/消/掉/./謝/謝	1.5099	4.3960
3	SnowNLP	3	請問這個跟剛才2組3980的有何不同呢？	請/問/這/個/跟/剛/才/2/組/3980/的/有/何/不/同/呢/？	1.1632	5.5592
4	SnowNLP	4	阿公說女神剪短髮好俏麗，像女大學生👩	阿/公/說/女/神/剪/短/髮/好/俏/麗/，/像/女/大/學/生/👩	1.4970	7.0562
...	...	...	...	...	...	...
9995	SnowNLP	9995	7S-77 珍珠天絲 醋酸纖維(乙酸)\n\n生產工藝\n醋酸纖維取材於可再生的木漿與...	7S/-77/珍/珠/天/絲/醋/酸/纖/維/(/乙/酸/)/生/產/工/藝/醋/酸/纖/維/取...	16.1581	9522.1345
9996	SnowNLP	9996	△停~止~關~鍵~字~△\n持續分享+按讚👍👍👍\n△停~止~關~鍵~字~△\n持續分享...	△/停/~/止/~/關/~/鍵/~/字/~/△/持/續/分/享/+按/讚/👍/👍/👍/△/停/...	2.8210	9524.9555
9997	SnowNLP	9997	多美丽的玫瑰花\n多可爱的玫瑰花\n我就这样深深的爱上她\n多美丽的玫瑰花\n多可爱的玫瑰花...	多/美丽/的/玫瑰/花/多/可爱/的/玫瑰/花/我/就/这样/深深/的/爱/上/她/多/美丽/...	21.3666	9546.3221
9998	SnowNLP	9998	△停~止~關~鍵~字~△\n持續分享+按讚👍👍👍\n△停~止~關~鍵~字~△\n持續分享...	△/停/~/止/~/關/~/鍵/~/字/~/△/持/續/分/享/+按/讚/👍/👍/👍/△/停/...	3.4831	9549.8052
9999	SnowNLP	9999	△停~止~關~鍵~字~△\n持續分享+按讚👍👍👍\n△停~止~關~鍵~字~△\n持續分享...	△/停/~/止/~/關/~/鍵/~/字/~/△/持/續/分/享/+按/讚/👍/👍/👍/△/停/...	4.0698	9553.8750

10000 rows x 6 columns

# Jieba 斷詞結果

	套件	id	原始留言	斷詞結果	執行時間 (ms)	累計執行時間 (ms)
0	Jieba	0	幫我改8118舖四雙人、8202舖四加大	幫/我/改/8118/舖/四/雙人/、/8202/舖/四/加大	0.1731	0.1731
1	Jieba	1	我兒子👦超愛60細棉夏被的，都早早去睡了	我兒子/👦/超愛/60/細棉/夏/被/的/，/都/早早/去/睡/了	0.2277	0.4008
2	Jieba	2	昨天下的單子打錯了。請幫我消掉。謝謝	昨天/下/的/單子/打錯/了/./ /請/幫/我/消掉/./ /謝謝	0.1469	0.5476
3	Jieba	3	請問這個跟剛才2組3980的有何不同呢？	請問/這個/跟/剛才/2/組/3980/的/有何/不同/呢/?	0.1249	0.6726
4	Jieba	4	阿公說女神剪短髮好俏麗，像女大學生👩	阿公/說/女神/剪短/髮/好/俏麗/，/像/女大學生/👩//👩	0.1242	0.7968
...	...	...	...	...	...	...
9995	Jieba	9995	7S-77 珍珠天絲 醋酸纖維(乙酸)\n生產工藝\n醋酸纖維取材於可再生的木漿與...	7S/ /-/77/ /珍珠/天絲/ / / /醋酸/纖維/(/乙酸/)\n\n生產/...	0.5858	659.3027
9996	Jieba	9996	△停~止~關~鍵~字~△\n持續分享+按讚👍👍👍\n△停~止~關~鍵~字~△\n持續分享...	△/停/~ /止/~ /關/~ /鍵/~ /字/~ /△\n\n持續/分享+/按/讚/👍/ /👍/ /👍/ ...	0.3080	659.6107
9997	Jieba	9997	多美丽的玫瑰花\n多可爱的玫瑰花\n我就这样深深的爱上她\n多美丽的玫瑰花\n多可爱的玫瑰花...	多/美丽/的/玫瑰花/\n\n多/可爱/的/玫瑰花/\n\n我/就/这样/深深/的/爱/上/她/...	0.5338	660.1446
9998	Jieba	9998	△停~止~關~鍵~字~△\n持續分享+按讚👍👍👍\n△停~止~關~鍵~字~△\n持續分享...	△/停/~ /止/~ /關/~ /鍵/~ /字/~ /△\n\n持續/分享+/按/讚/👍/ /👍/ /👍/ ...	0.3531	660.4977
9999	Jieba	9999	△停~止~關~鍵~字~△\n持續分享+按讚👍👍👍\n△停~止~關~鍵~字~△\n持續分享...	△/停/~ /止/~ /關/~ /鍵/~ /字/~ /△\n\n持續/分享+/按/讚/👍/ /👍/ /👍/ ...	0.4070	660.9046

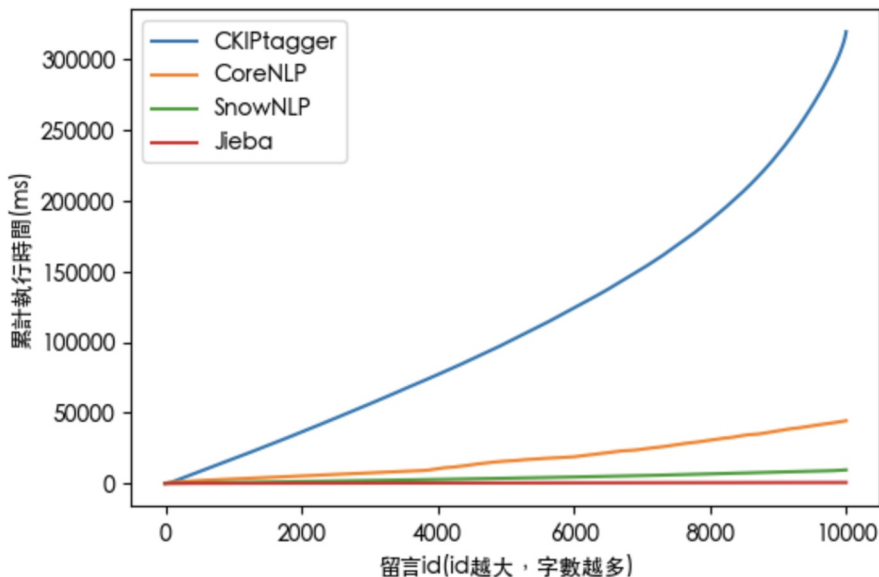
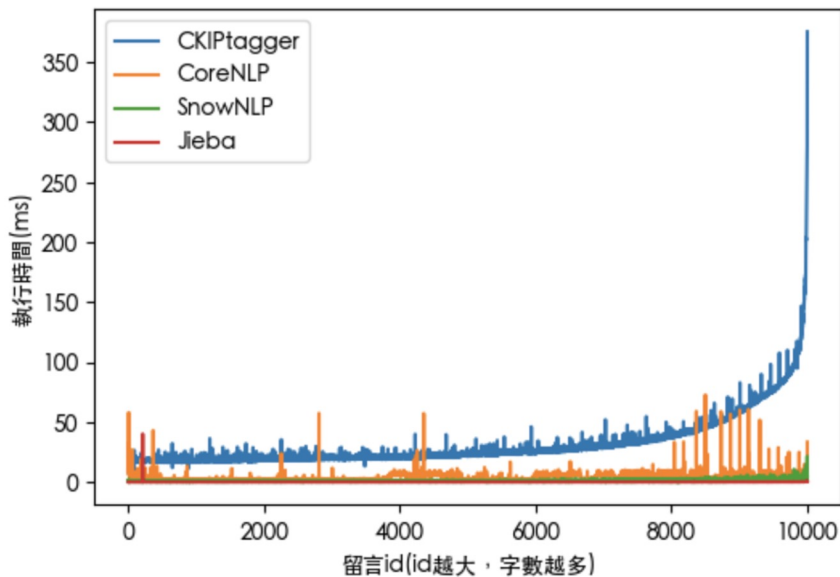
10000 rows x 6 columns



# 比較：斷詞執行時間

- 隨著留言字數的增加，執行時間有增加的趨勢，且 CKIPtagger > CoreNLP > SnowNLP > Jieba，其中又以 CKIPtagger 的增幅最大

- 整體執行時間 CKIPtagger > CoreNLP > SnowNLP > Jieba，且呼應左圖，CKIPtagger 在越長的留言所需執行時間增幅最大（觀察斜率）



# 比較：正確率

- 從10,000則留言中隨機抽樣10則進行人工標記，用以比較斷詞結果的正確性。
- 將斷詞結果儲存成區間表示，並找出交集，計算出 precision, recall, f1\_score

	斷詞結果	區間	重合部分
正確斷詞	'結婚/的/和/尚未/結婚/的'	[1, 2], [3, 3], [4, 4], [5, 6], [7, 8], [9, 9]	[1, 2], [3, 3], [9, 9]
套件斷詞結果	'結婚/的/和尚/未結婚/的'	[1, 2], [3, 3], [4, 5], [6, 8], [9, 9]	

Precision = 0.6

Recall = 0.5

F1\_score = 0.545

index	套件	id	原始留言	斷詞結果	正確斷詞	precision	recall	f1_score
0	279	CKIPtagger	279	7315一整組啦，不要多整頭套了😂😂😂😂	7315/一/整組/啦/，/不要/多/整頭套/了/😂/😂/😂/😂	0.6	0.75	0.666667
1	637	CKIPtagger	637	床上的帝王棉特大鋪包+2 改90x200	床/上/的/帝王/棉特/大/鋪包/+2 / 改/90/x/200	0.5	0.545455	0.521739
2	1074	CKIPtagger	1074	請問兩用被，有沒有沒有鋪棉的可以塞冬被嗎？	請問/兩/用/被/，/有沒有/沒有/鋪/棉/的/可以/塞冬/被/嗎/?	0.461538	0.6	0.521739
3	1150	CKIPtagger	1150	1049，1020，9149特大床罩各一組	1049，1020，9149/特大/床罩/各/一/組	0.833333	0.625	0.714286
4	1243	CKIPtagger	1243	8235單人薄三+1\8212加大薄三+1	8235/單人/薄三/+1\8212/加大/薄三/+1	0.857143	0.75	0.8

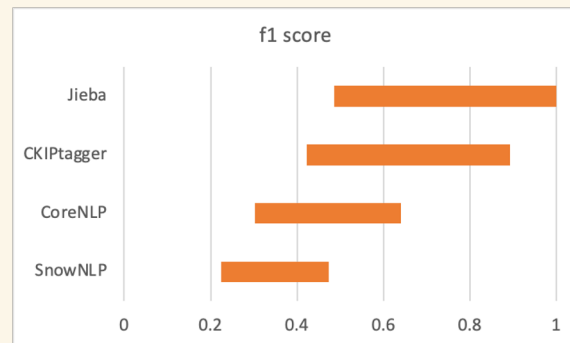
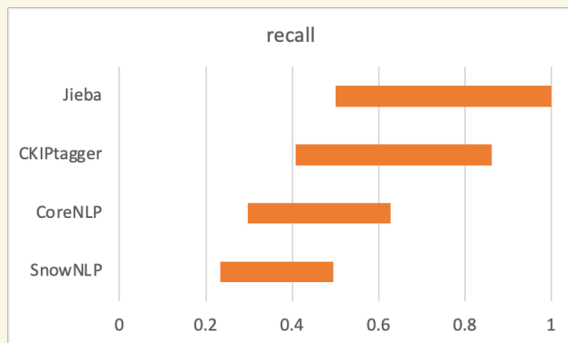
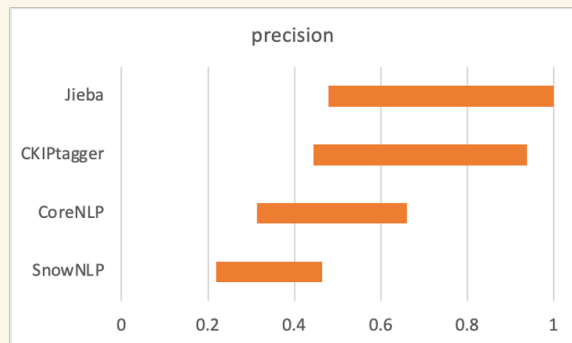
# 比較：正確率

- 不同套件下的 precision, recall, f1\_score 平均數、標準差與抽樣誤差
- CKIPtagger 與 Jieba 表現較好，Jieba > CKIPtagger > CoreNLP > SnowNLP

	precision			recall			f1_score		
	平均數	標準差	抽樣誤差	平均數	標準差	抽樣誤差	平均數	標準差	抽樣誤差
SnowNLP	0.3419	0.1881	0.1222	0.3644	0.2200	0.1302	0.3489	0.1985	0.1247
CoreNLP	0.4867	0.2733	0.1739	0.4623	0.2810	0.1652	0.4721	0.2756	0.1687
CKIPtagger	0.6907	0.2150	0.2468	0.6346	0.2135	0.2268	0.6580	0.2105	0.2351
Jieba	0.7445	0.1762	0.2660	0.7780	0.1579	0.2780	0.7569	0.1624	0.2704

# 比較：正確率

- 不同套件下的 precision, recall, f1\_score 平均數、標準差與抽樣誤差
- CKIPtagger 與 Jieba 表現較好，Jieba > CKIPtagger > CoreNLP > SnowNLP



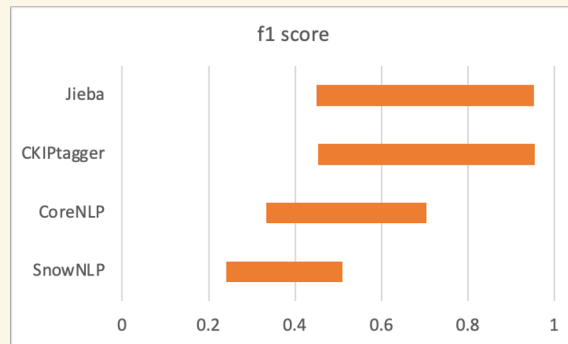
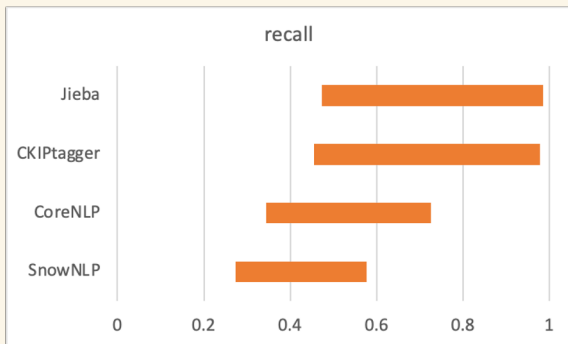
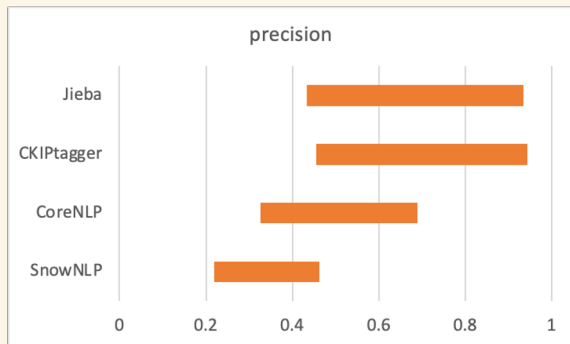
# 比較：正確率

- 不同套件下的 precision, recall, f1\_score 平均數、標準差與抽樣誤差
- 不考慮標點符號與表情符號斷詞結果
- CKIPtagger 與 Jieba 表現較好，CKIPtagger >= Jieba > CoreNLP > SnowNLP

	precision			recall			f1_score		
	平均數	標準差	抽樣誤差	平均數	標準差	抽樣誤差	平均數	標準差	抽樣誤差
SnowNLP	0.3408	0.1963	0.1218	0.4254	0.2415	0.1520	0.3754	0.2150	0.1341
CoreNLP	0.5087	0.2720	0.1818	0.5349	0.2834	0.1911	0.5193	0.2758	0.1855
CKIPtagger	0.6992	0.2234	0.2498	0.7168	0.2215	0.2608	0.7040	0.2172	0.2507
Jieba	0.6841	0.2237	0.2444	0.7299	0.2075	0.2561	0.7017	0.2127	0.2515

# 比較：正確率

- 不同套件下的 precision, recall, f1\_score 平均數、標準差與抽樣誤差
- 不考慮標點符號與表情符號斷詞結果
- CKIPtagger 與 Jieba 表現較好，CKIPtagger  $\geq$  Jieba  $>$  CoreNLP  $>$  SnowNLP



# 比較：常見詞與關鍵詞

- 將同一個套件的所有斷詞存成一個字串，並把這四個大字串存成list（視為四個文本），並統計所有出現過的字詞的**詞頻**與**TF-IDF**
- 4個套件裡前10名高分的常見詞

	package	rank_1	rank_2	rank_3	rank_4	rank_5	rank_6	rank_7	rank_8	rank_9	rank_10
0	CKIPtagger	停止	尺寸	關鍵字	60	天絲	可以	編號	舖包	薄包	請問
1	CoreNLP	停止	關鍵	尺寸	60	可以	編號	床罩	舖包	天絲	加單
2	SnowNLP	鍵字	四件	尺寸	60	兩用	方式	80	加大	可以	規格
3	Jieba	停止	關鍵	四件	天絲	尺寸	60	三件	可以	編號	薄包

可以發現常見詞都與**床鋪**相關產品或品質名詞有關

SnowNLP有出現較不符合預期的斷詞結果，如「鍵字」應指「關鍵字」

# 比較：常見詞與關鍵詞

- 將同一個套件的所有斷詞存成一個字串，並把這四個大字串存成list（視為四個文本），並統計所有出現過的字詞的**詞頻**與**TF-IDF**
- 4個套件裡前10名高分的關鍵詞

	package	rank_1	rank_2	rank_3	rank_4	rank_5	rank_6	rank_7	rank_8	rank_9	rank_10
0	CKIPtagger	停止	關鍵字	尺寸	天絲	編號	60	鋪包	請問	可以	關鍵
1	CoreNLP	關鍵	停止	尺寸	60	編號	可以	鋪包	天絲	加單	床罩
2	SnowNLP	鍵字	四件	尺寸	60	兩用	方式	80	加大	羅天	可以
3	Jieba	關鍵	停止	四件	天絲	尺寸	編號	60	三件	可以	加單

可以發現結果與常見詞類似，且關鍵詞都與**床鋪**相關產品或品質名詞有關



# 比較：專有名詞的斷詞

- 因為觀察到許多床鋪相關的專有名詞，因此想來比較這四個斷詞方式在專有名詞的差異
- 以下列資料為例

80支**紗臻絲棉**100%純棉n皇后棉，高端純棉經過二道絲光亮晶晶，將棉提昇到具有絲綢的滑順手感與光澤，且面料色澤明亮，久洗不變色；帝王棉是**緞紋織法**，**臻絲棉**是**絲綢織法**。可以把好的棉紗的柔順發揮到極致。古代說「綢緞」就是這種。古代皇后的衣就是這種材做的質。

專有名詞	紗臻絲棉	臻絲棉	緞紋織法	絲綢織法	
CKIP	紗臻絲棉	臻絲棉	緞紋/織法	絲綢/織法	切得最好且合理
CoreNLP	紗/臻/絲/棉	臻/絲棉	緞/紋/織法	絲/綢/織法	沒有特別被切出
SnowNLP	紗臻/絲/棉	臻/絲/棉	緞/紋/織法	絲/綢/織法	除了切出紗臻以外，其他都不太行
Jieba	支紗臻/絲棉	臻/絲/棉是	緞紋織法	絲綢織法	好像會跟前後的些詞切在一起