建製新聞分類器：基礎篇

重要知識點

- 學習建立 BOW 的方式，包括中文斷詞、用 POS 結果排除無意義的詞性，並且利用 BOW 和 Group Mean Vector 來做新聞分類

實作提示

請搭配 Jupyter Notebook 使用本教材

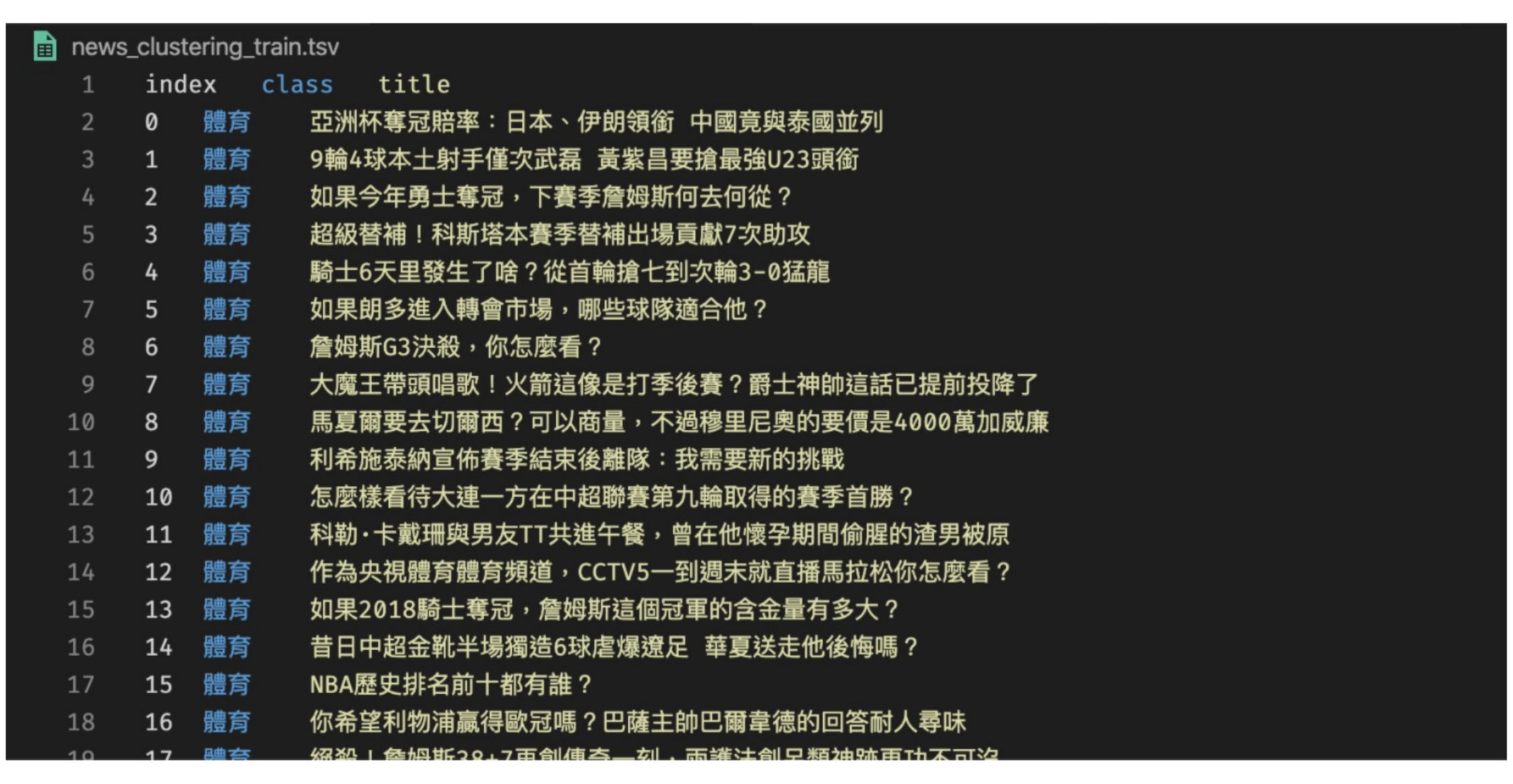
新聞分類器

本系列是實務專題，讓大家可以利用先前學的知識來建立一個新聞分類器

認識資料集

- news_clustering_train.tsv 中有 1800 篇新聞，六種類別的新聞各 300 篇
- news_clustering_test.tsv 中有 600 篇新聞，六種類別的新聞各 100 篇
- 六種類別：體育、財經、科技、旅遊、農業、遊戲

資料內容包含：新聞 index、新聞類別、新聞標題



news_clustering_train.tsv

```
1   index  class  title
2   0   體育   亞洲杯奪冠賠率：日本、伊朗領銜 中國竟與泰國並列
3   1   體育   9輪4球本土射手僅次武磊 黃紫昌要搶最強U23頭銜
4   2   體育   如果今年勇士奪冠，下賽季詹姆斯何去何從？
5   3   體育   超級替補！科斯塔本賽季替補出場貢獻7次助攻
6   4   體育   騎士6天里發生了啥？從首輪搶七到次輪3-0猛龍
7   5   體育   如果朗多進入轉會市場，哪些球隊適合他？
8   6   體育   詹姆斯G3決殺，你怎麼看？
9   7   體育   大魔王帶頭唱歌！火箭這像是打季後賽？爵士神帥這話已提前投降了
10  8   體育   馬夏爾要去切爾西？可以商量，不過穆里尼奧的要價是4000萬加威廉
11  9   體育   利希施泰納宣佈賽季結束後離隊：我需要新的挑戰
12  10  體育   怎麼樣看待大連一方在中超聯賽第九輪取得的賽季首勝？
13  11  體育   科勒·卡戴珊與男友TT共進午餐，曾在他懷孕期間偷腥的渣男被原
14  12  體育   作為央視體育體育頻道，CCTV5一到週末就直播馬拉松你怎麼看？
15  13  體育   如果2018騎士奪冠，詹姆斯這個冠軍的含金量有多大？
16  14  體育   昔日中超金靴半場獨造6球虐爆遼足 華夏送走他後悔嗎？
17  15  體育   NBA歷史排名前十都有誰？
18  16  體育   你希望利物浦贏得歐冠嗎？巴薩主帥巴爾韋德的回答耐人尋味
```

定義問題

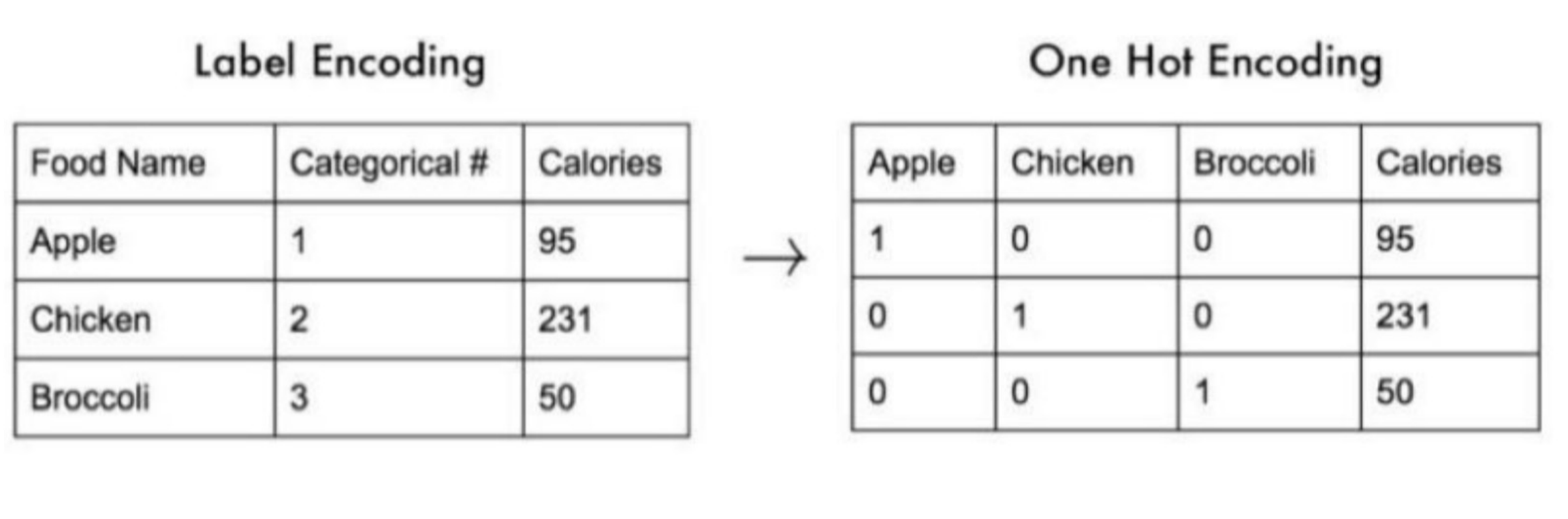
- 使用 news_clustering_train.tsv 中 1800 篇新聞標題及其類別訓練一個分類器
- 這個分類器要能夠把 news_clustering_test.tsv 中的 600 篇新聞給分類正確
- 分類器
  - Input： 新聞標題
  - Output：屬於六種類別中的哪一種？

斷詞＋POS

- 我們待會想要為每一句新聞標題給予一組 Bag of Words (BOW)，在這之前我們必須要進行斷詞
- 這裡嘗試使用 CKIP 的斷詞和 POS，POS 的目的是用來排除不具意義的詞性

Bag of Words (BOW)

- 在 ML 的世界我們需要將字詞轉換成電腦看得懂的數值
- 最簡單的 word embedding 方式就是將每一個單字以一個數值表示(Label encoding)，再轉換為 One hot encoding 的方式呈現 (因為不同文字之間，通常沒有強弱之分)



Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

→

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

資料來源：**初學Python手記#3-資料前處理(Label encoding、 One hot encoding)**

- 現在我們已經將文字轉換為 One hot encoding 的格式，那表達一句話或一篇文章最簡單的方式，就是將這句話或這篇文章中出現過的文字全部加起來，這也就是我們的 Bag-of-words

排除較無意義的詞性

- 如果直接使用斷詞的結果來造 BOW 會有以下問題：
  a. 標點符號會被考慮進去
  b. 無意義的語助詞會被考慮進去
  c. 無意義的代名詞會被考慮進去
  d. ... 等等

- 有些詞性對我們新聞分類而言是不具意義的，甚至會混淆判斷
- 所以我們可以藉由 POS 的結果來將這些詞性給篩選掉，不要讓他進來造 BOW

Cosine Similarity

很適合用來衡量兩個 BOW 是否相似

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

範例：

	我	愛	你	運動	在	思考
句一	1	1	1	0	0	0
句二	1	1	0	1	0	0
句三	1	0	0	0	1	1

使用 cosine similarity 去分析就會發現：句一和句二較為接近，而句一和句三較為遠

使用 Group Mean Vector 來進行新聞分類

- 我們已經有每篇新聞的 BOW 了，那接下來怎麼使用它們來進行新聞分類呢？
- 我們可以將每種類別底下新聞的 BOW 做平均，來代表特定種類的新聞向量，我稱此向量為 Group Mean Vector
- 在 Inference 時，我們將一篇新聞的 BOW 和六個類別的 Group Mean Vector 計算 Cosine Similarity，看這篇新聞跟哪個類別相似，就預測那個類別

知識點回顧

運用以下所學來完成此專題：CKIP 斷詞和 POS、BOW、Cosine Similarity

下一步：完成作業