目前位置

D23 : Naive Bayes 實作（鄭純興氏）

重要觀點

- 鄭純興氏的 scikit-learn - Naive Bayes API

讀入 airline-sentiment Dataset

- 資料 Tweet-news
- 目的：Dataset 內容是否對應關鍵詞情緒？（評價類型分類）或是內容是否對應到「負面」、「中性」、「正向」？

```
In [1]: !pwd
/home/jovyan/work
In [2]: ls -al ./airline-sentiment/
total 1
drwxrwxrwx 1 root root 4096 Jun 10 05:24 .
drwxrwxrwx 1 root root 4096 Jun 10 05:24 ..
-rwxrwxrwx 1 root root 1144 Jun 10 05:24 .ipynb_checkpoints
-rwxrwxrwx 1 root root 1144 Jun 10 05:24 Airline-Sentiment-2-w-neg.csv
-rwxrwxrwx 1 root root 1144 Jun 10 05:24 Airline-Sentiment-2-w-neu.csv
-rwxrwxrwx 1 root root 1144 Jun 10 05:24 Airline-Sentiment-2-w-pos.csv
-rwxrwxrwx 1 root root 1144 Jun 10 05:24 Airline-Sentiment-2-w-neg-shuffled.csv
-rwxrwxrwx 1 root root 1144 Jun 10 05:24 Airline-Sentiment-2-w-neu-shuffled.csv
-rwxrwxrwx 1 root root 1144 Jun 10 05:24 Airline-Sentiment-2-w-pos-shuffled.csv
-rwxrwxrwx 1 root root 1144 Jun 10 05:24 Airline-Sentiment-2-w-neg-shuffled.csv.bak
-rwxrwxrwx 1 root root 1144 Jun 10 05:24 Airline-Sentiment-2-w-neu-shuffled.csv.bak
-rwxrwxrwx 1 root root 1144 Jun 10 05:24 Airline-Sentiment-2-w-pos-shuffled.csv.bak
```

```
In [3]: import numpy as np
        import pandas as pd
```

```
In [4]: data = pd.read_csv("./airline_sentiment/")
```

```
In [5]: data.head()
```

- 如們查看預設的 column 內容 text、airline_sentiment、則可發現 為 資訊資料、airline_sentiment 為 label

```
In [7]: data[['text', 'airline_sentiment']]
```

```
Out[7]:
                                                text  airline_sentiment
0                @VirginAmerica What dialogue will...            neutral
1  @VirginAmerica plus you've added commands to ...           positive
2  @VirginAmerica I didn't today... must mean I ...            neutral
3  @VirginAmerica it's nice to have a direct line ...           negative
4                          @VirginAmerica yes, nearly every time I fly, I see ...           negative
5  @VirginAmerica really hard thing about flying ...           positive
6  @VirginAmerica you know my name, you change my ...           negative
7  @VirginAmerica we love to fly with you. We'll ...           positive
14640 rows x 2 columns
```

```
In [8]: data[['text', 'airline_sentiment']]
```

取出 X、Y

- 如們將之取得到每筆 text 、轉換至取出 Dataframe 資料

```
In [9]: label_to_index = {'negative': 0,
                          'neutral': 1,
                          'positive': 2}
```

```
In [10]: # 以字典的方式呈現
         # 字典的 keys 為 airline_sentiment 的 Series 物件的 'negative', 'neutral', 'positive' 的 Series 內容
         # 字典的 values 則為 0, 1, 2 的 Series 物件
         # 這樣 .map() 會把 airline_sentiment 的 Series 內容
         # 一一對應到 label_to_index 字典的 keys 的 values 上
         # 產生一個 Series 物件 labels 來儲存 airline_sentiment 的 Series 內容
         labels_to_index = {'negative': 0,
                            'neutral': 1,
                            'positive': 2}
```

```
In [11]: # 以 airline_sentiment 的 Series 物件 data['airline_sentiment'] 的 Series 內容
         # 對 labels_to_index 字典的 keys 的 values 上
         # 產生一個 Series 物件 labels 來儲存 airline_sentiment 的 Series 內容
         labels = data['airline_sentiment'].map(labels_to_index)
         # 印出 labels 的 Series 內容
         labels
```

```
In [12]: # 以 data 的 Series 物件 data['text'] 的 Series 內容
         # 產生一個 Series 物件 texts 來儲存 text 的 Series 內容
         texts = data['text']
         # 印出 texts 的 Series 內容
         texts
```

文字處理

- 如前取出後考之資料，經斷詞後即可將處理了。

```
In [13]: from sklearn.model_selection import train_test_split
         from sklearn.feature_extraction.text import CountVectorizer
         from sklearn.naive_bayes import MultinomialNB
         from sklearn.metrics import classification_report
         from sklearn.metrics import confusion_matrix
         from sklearn.metrics import accuracy_score
         import numpy as np
         import pandas as pd
         import matplotlib.pyplot as plt
         %matplotlib inline
         # 設定 matplotlib 繪圖時要內嵌顯示圖形
         # 這是 Jupyter Notebook 的 magic command
         # 參考 http://stackoverflow.com/questions/43027980/how-to-display-plots-in-a-jupyter-notebook-in-a-docker-container
```