Google Play Application Comments Sentiment Analysis

問題定義

這次期末專題想試試看簡單的 NLP 任務，決定做 sentiment analysis，會從 google play 的網站上挑選幾個應用程式，並抓取應用程式的評論幫作資料集，評論有一顆星到五顆星，我的想法是把是把 1~2 星的評論當作 negative，3 星為 neutral，4~5 星為 positive。訓練一個模型讓模型使模型能預測某評論是正向、負面，或是中立。

資料蒐集

我在網路上有找到一個 python library 叫 google-play-scraper，他有提供一些方便的 API 可以將 google play 上面的應用程式的資料爬下來以 json 檔的形式呈現，這邊附上這個 library 的 github 連結:
https://github.com/JoMingyu/google-play-scraper

爬下來的資料大概會長這樣，要使用的部分是 content 還有 score

```
{
    "userName": "EasyJet 123",
    "userImage": "https://lh3.googleusercontent.com/a-/AOh14GhE3-Fsq5KDs_kmCRGcifbNUQTOtK5DpZkJ2
    "content": "Easily my favorite game. Relaxing, with easy controls, no purchase necessary to
    "score": 5,
    "thumbsUpCount": 79,
    "reviewCreatedVersion": "1.14",
    "at": datetime.datetime(2020, 2, 12, 8, 42, 41),
    "replyContent": None,
    "repliedAt": None,
    "reviewId": "gp:AOqpTOHyQo9QEPtxefmvjNuqR9VmFyBaj2FNXLvHsuH19de9bC0dT_voHWSKNGFcc10jv077wOdz
},
```

爬下資料後將有用的資料轉成 pandas dataframe 即完成自製的 dataset。
這次期末專題我只會抓取英文的留言，考量到自己的能力，對 NLP 梅很懂，先不要碰 multilingual 的 task 比較好，預計抓取相同類型的應用程式的留言，預計抓取 5~10 個應用程式的資料，讓訓練資料不要太少。

預期成果與效益

希望模型的準確率能達到 85 左右，並試圖用教授教的資料分析方法做結果評估 (precision, recall, f1score 之類的方法)。想利用這次個 人期末專題學一些基礎的 NLP，預計會使用 BERT 相關模型來完成，同時學習一些 NLP 的資料前處理技巧，進而對 NLP 有進一步的認識。