1. 統計111 H24076029 劉米婷
2. 競賽敘述與目標：

推特仇恨言論偵測，從推特中的留言中判斷這段話是否帶有歧視或冒犯等意味，若帶有hateful的意味即被分類為’”0”，若帶有offensive的意味即被分類為”1”，若沒有帶有任何仇恨的意味即被分類為”2”，我們須從train.csv中以分類好的言論進行整理分析，並預測tesr.csv中的言論是屬於哪一分類。

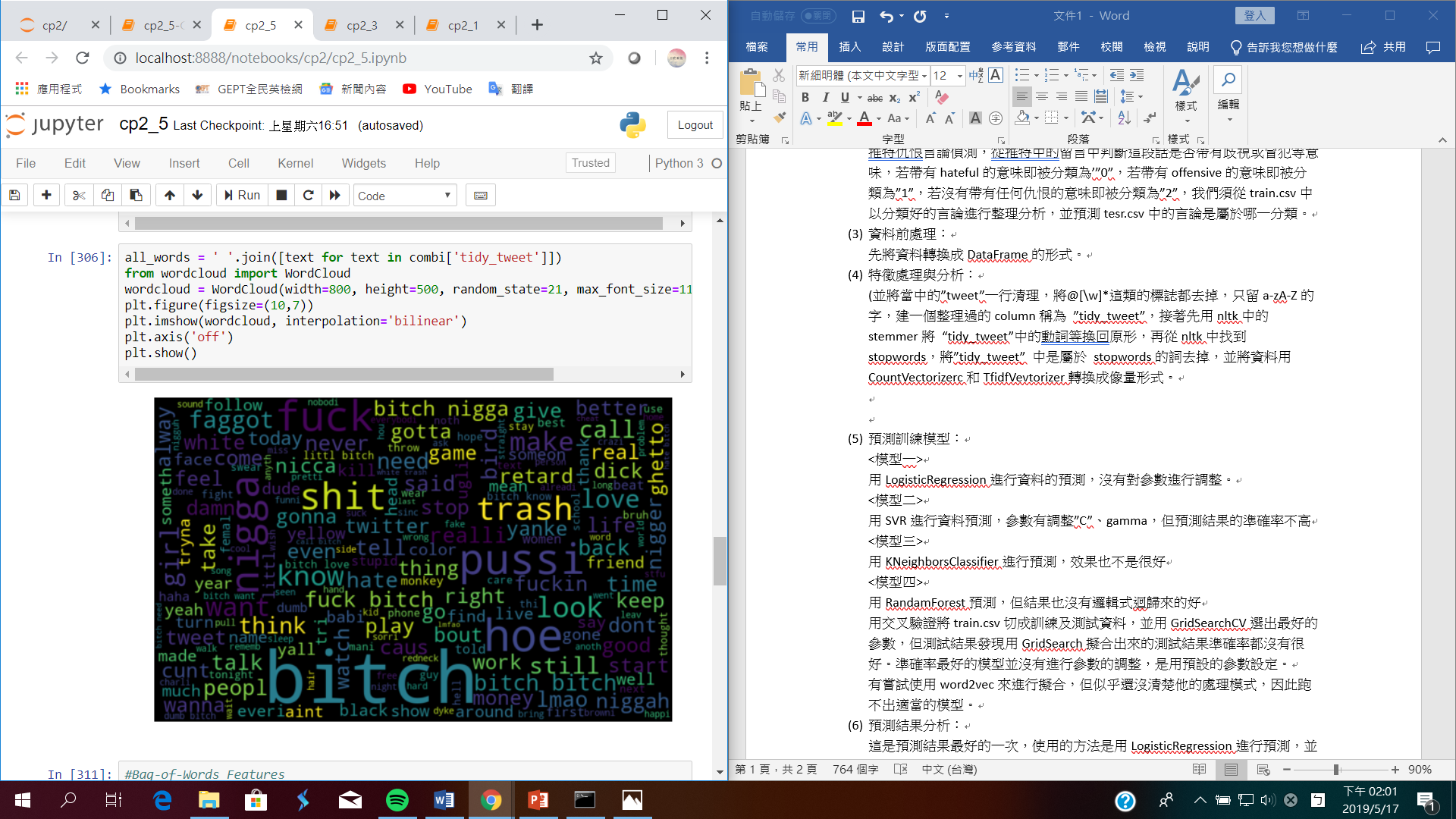
1. 資料前處理：

先將資料轉換成DataFrame的形式。

1. 特徵處理與分析：

(並將當中的”tweet”一行清理，將@[\w]\*這類的標誌都去掉，只留a-zA-Z的字，建一個整理過的column稱為 ”tidy\_tweet”，接著先用nltk中的stemmer將 “tidy\_tweet”中的動詞等換回原形，再從nltk中找到stopwords，將”tidy\_tweet” 中是屬於 stopwords的詞去掉，並將資料用CountVectorizerc和TfidfVevtorizer轉換成像量形式。

這是我用WordCloud跑出來的文字所占比重



1. 預測訓練模型：

<模型一>

用LogisticRegression進行資料的預測，沒有對參數進行調整。

<模型二>

用SVR進行資料預測，參數有調整”C”、gamma，但預測結果的準確率不高

<模型三>

用KNeighborsClassifier進行預測，效果也不是很好

<模型四>

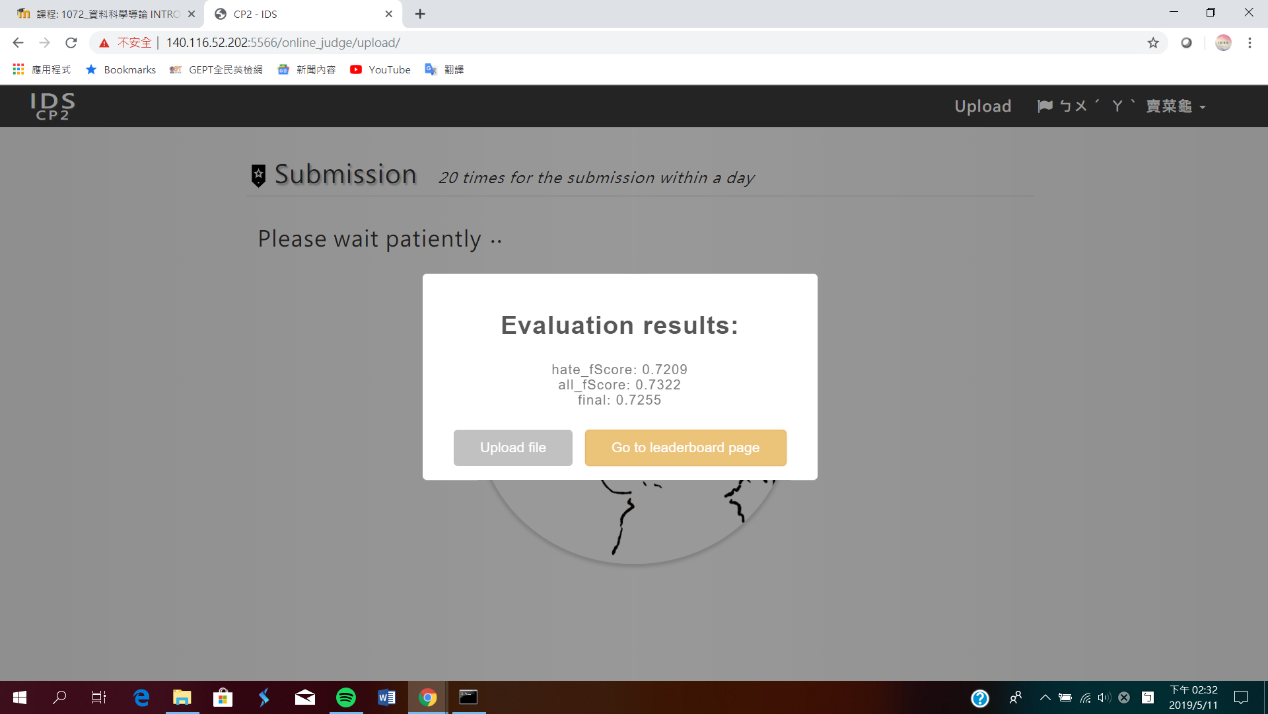
用RandamForest預測，但結果也沒有邏輯式迴歸來的好

用交叉驗證將train.csv切成訓練及測試資料，並用GridSearchCV選出最好的參數，從train.csv中切出來的資料進行預測都能達到0.8以上的準確率，但使用test.csv測試結果發現用GridSearch擬合出來的測試結果準確率都沒有很好，大概都停留在0.66左右。準確率最好的模型並沒有進行參數的調整，是用預設的參數設定。

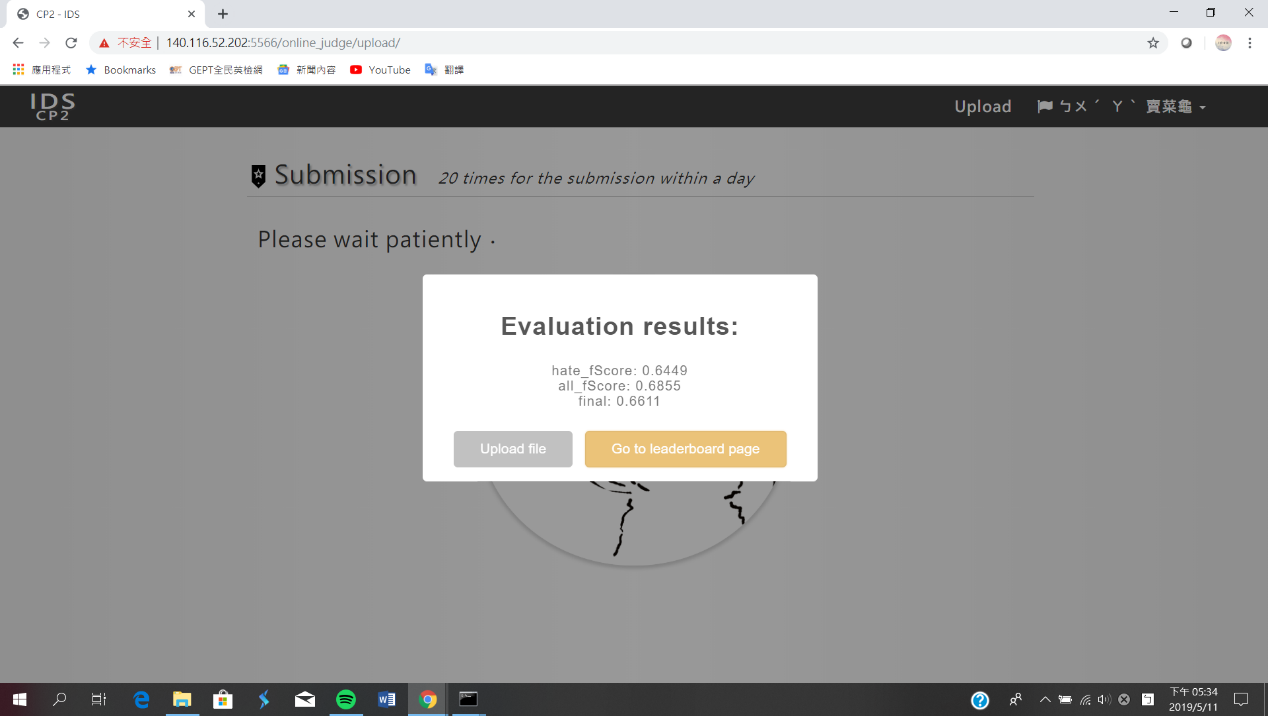
有嘗試使用word2vec來進行擬合，但似乎還沒清楚他的處理模式，因此跑不出適當的模型。

1. 預測結果分析：

這是預測結果最好的一次，使用的方法是用LogisticRegression進行預測，並且沒有對參數做過多的調整。



這是使用GridSearchCV進行參數的選擇後的準確率，不論使用哪種模型，都維持在0.6-0.7之間，沒有辦法突破0.7



我覺得會有這樣的現象有可能是因為資料的特徵處理與分析沒有做好，有可能有些詞同時在被分類為”0”跟”1”的資料中都存在，可是又不屬於stopwords中的詞，所以導致判斷錯誤，因為後面模型的訓練已經用交叉檢驗及GridSearchCV等方式避免果擬合的現象了，我下次如果再處理類似的資料應該要將不同模型跑出來的資料進行比對，再跑出最終的結果，但前提是使用的模型種類要夠多。

最後上船的模型是使用LogisticRegression使用原預設參數的模型。

1. 感想與心得：

統計111 H24076029 劉米婷

我覺得這次做起來比cp1稍微順手一點，而切上次聽完前五名分享他們的預測方式後真的收穫很多，有些方法感覺蠻簡單的但我們都沒有想過，上次的困難是想用很多模型，但不太了解他的模式所以不會配，這次就有比較熟悉了，這次遇到的困難是蠻多想嘗試的但礙於時間關係沒辦法做完，下次我覺得可以再改進的是除了使用模型預測，也要再分析一下預測出來的結果有什麼問題再針對模型的問題進行改良，而不是一直套用各種方法。這次嘗試用交叉檢驗法跟GridSearchCV但好像沒有得到比較好的結果有點出乎我的意料，希望下次能先做好模型訓練前的資料分析及處理。