**Data Mining Final Project**

題目：預測旅館評論之正負面情緒

組員:張巧欣、楊雅雯、吳宛儒、羅廷君

## 壹、動機

若是能有效分析出評論的正負面情緒，之後有新的評論出現，便可以分析出其評論是否好或壞，能讓我們快速了解其旅館評價，不需要先親自花大量時間看完全部評論。

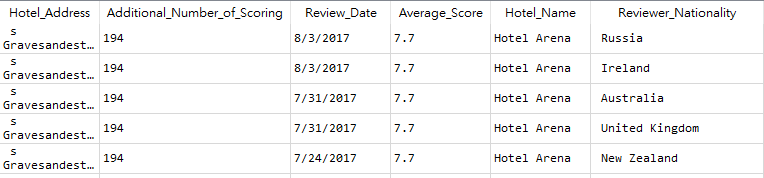
## 貳、資料集敘述

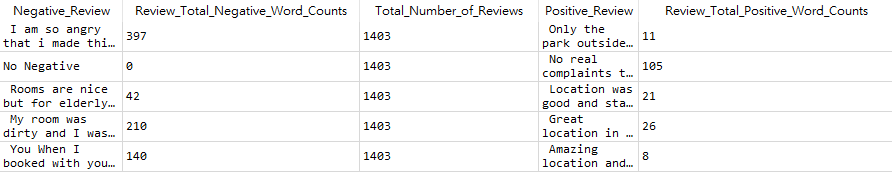
資料名稱：**515K Hotel reviews data in Europe**

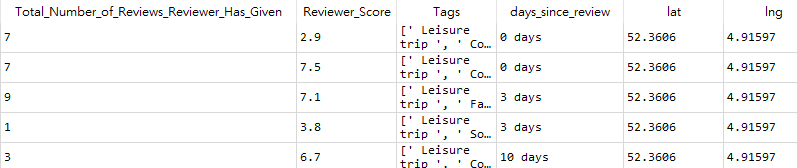
此資料集共有525,000位客戶在歐洲對1496間旅館的評論與評分

共有17個欄位，以下是對資料欄位的說明：

* Hotel\_Address: 旅館住址
* Review\_Date: 評論posted的時間
* Average\_Score: 旅館過去一年平均評價
* Hotel\_Name: 旅館名稱
* Reviewer\_Nationality: 評論者國籍
* Negative\_Review: 評論者對旅館的負面評價。如果評論者沒有給出負面評論，則是"沒有負面評價"
* Review\_Total\_Negative\_Word\_Counts: 負面評論檔案的總字數
* Positive\_Review: 評論者對旅館的正面評價。如果評論者沒有給出正面評論，則是"沒有正面評價"
* Review\_Total\_Positive\_Word\_Counts: 正面評論檔案的總字數
* Reviewer\_Score: 評論者依照自己的經驗給出分數
* Total\_Number\_of\_Reviews\_Reviewer\_Has\_Given: 評論者之前評論的數目
* Total\_Number\_of\_Reviews: 評論這間旅館有效的評論數目
* Tags: 評論者給的標籤(標誌)
* days\_since\_review: 評論日期到擷取日期
* Additional\_Number\_of\_Scoring: 只有給分數而未做評論的客人
* lat: 旅館位置的緯度
* lng: 旅館位置的經度







## 參、分析工具

LogisticRegression、Adaboost、DNN、RMDL

## 肆、實作與評估方法

1. preparing dataset

* 標label
  + negative review : no negative的review -->1
    - 佔全部24.78%
  + positive review : no positive 的review -->0
    - 佔全部6.96%
* 30% testing and 70% trainning dataset

2. Clean and Preprocess

* + 確認所有的reviews都是英文
  + 是否保留過短或過長的reviews : score 和 reviews 長度的相關性

3. Text Processing

* + Removing Stop Words
  + (Normalization)
  + Stemming/Lemmatization

4.N-grams

* + unigrams-bigrams

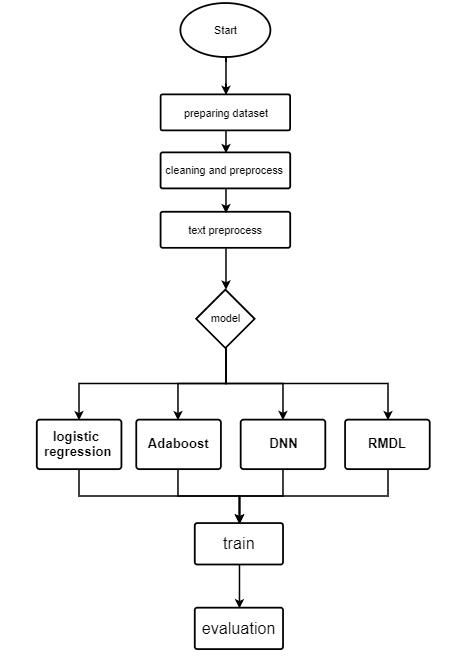
5.Vectorization(特徵向量化)

* (Countvectorizer)
* term frequency-inverse document frequency (TF-IDF):
  + TfidfVectorizer
  + where words that appear in many documents have a value closer to zero and words that appear in less documents have values closer to 1.

6.LogisticRegression、Adaboost、DNN、RMDL

7.Predict

## 伍、流程(分析流程圖、結果截圖等)

****

## 

## 陸、分析結果與結論(或其他補充內容)

|  |
| --- |
| **降維比較(用logistic regression)** |
| **1.SVD** |
|  |
| **2.LDA** |
|  |

|  |
| --- |
| **model比較** |
| **1.logistic regression + tfidf + ngram** |
| **調整c值，當c值等於2時accuracy=0.9591045140797753最高** |
| **2.Adaboost** |
| 1.Average precision-recall score: 0.92  2.accuracy: 0.9153380978559648 |
| **3.DNN** |
| * 在train了很多次後發現，非常容易overfitting(trainning set的accuracy可以到99%)所以用了以下方法去避免overfitting * dropout * l2 regulariation(用過l1但表現差於l2) * 最佳的accuracy : 95.67% |
| **4.RMDL** |
| * 論文(2018/5/31): * <https://arxiv.org/pdf/1805.01890.pdf> * 程式碼: * <https://github.com/kk7nc/RMDL> * <https://github.com/kk7nc/RMDL/tree/master/Examples> * RMDL簡單從架構上來看就是把DNN、RNN、CNN接在一起 * 可接受的input :text, video, images, and symbolic * 訓練時間長 * 目前accuracy最高為95.89% * 由於ram容易爆掉，所以如果ram足夠的話，可以突破96%順利奪冠 |

全部總結：

* Test accuracy
  + logistic regression : 95.9%
  + SVD : 85.51%
  + LDA : 91.01%
  + DNN : 95.67%
  + RMDL : 95.89%
* 有時候很簡單的方法其實搞不好效果會較佳