自然語言處理 HW1

組員 :

110590450 歐佳昀(40%) => 作業架構、流程、分析、改善

110590452 莊于潔(60%) => 作業模型挑選、參數調整

環境 : 使用 Jupyter Notebook	語言 : python

Classification results

- Precision = 0.77
- Recall = 0.77
- F-measure = 0.77
- Accuracy = 0.77

經使用不同模型後，幾個方法最好數值一致

<table>
<tr>
<td>Logistic Regression with TfidfVectorizer
- Precision = 0.77
- Recall = 0.77
- F-measure = 0.77
- Accuracy = 0.77
LogisticRegression
precision recall f1-score support

0 0.78 0.76 0.77 30969
1 0.77 0.79 0.78 31029

accuracy 0.77 61998
macro avg 0.77 0.77 0.77 61998
weighted avg 0.77 0.77 0.77 61998

Precision: 0.77
Recall: 0.77
F-measure: 0.77
Accuracy: 0.77</td>
<td>BernoulliNB with TfidfVectorizer
- Precision = 0.77
- Recall = 0.77
- F-measure = 0.77
- Accuracy = 0.77
BernoulliNB
precision recall f1-score support

0 0.77 0.75 0.76 30969
1 0.76 0.78 0.77 31029

accuracy 0.77 61998
macro avg 0.77 0.77 0.77 61998
weighted avg 0.77 0.77 0.77 61998

Precision: 0.77
Recall: 0.77
F-measure: 0.77
Accuracy: 0.77</td>
</tr>
<tr>
<td>Logistic Regression with CountVectorizer
- Precision = 0.77
- Recall = 0.77
- F-measure = 0.77
- Accuracy = 0.77
LogisticRegression
precision recall f1-score support

0 0.79 0.75 0.77 30969
1 0.76 0.80 0.78 31029

accuracy 0.77 61998
macro avg 0.77 0.77 0.77 61998
weighted avg 0.77 0.77 0.77 61998

Precision: 0.77
Recall: 0.77
F-measure: 0.77
Accuracy: 0.77</td>
<td>BernoulliNB Regression with CountVectorizer
- Precision = 0.77
- Recall = 0.77
- F-measure = 0.77
- Accuracy = 0.77
BNBmodel
precision recall f1-score support

0 0.77 0.75 0.76 30969
1 0.76 0.78 0.77 31029

accuracy 0.77 61998
macro avg 0.77 0.77 0.77 61998
weighted avg 0.77 0.77 0.77 61998

Precision: 0.77
Recall: 0.77
F-measure: 0.77
Accuracy: 0.77</td>
</tr>
</table>

資料預處理

Way	Expected Use	Final Use
Convert text to lowercase	V	V
Cleaning URLs	V	V
Removing punctuation and odd symbols	V	V
Replacing consecutive repeating characters	V	V
Cleaning numbers	V	V
Cleaning single characters	V	V
Lemmatizing words	V	V
Cleaning non-English words	V	V
Cleaning extra spaces	V	V
Word Cloud Visualization	V	V
Set Unknown Word	V	X, too long time

特徵提取

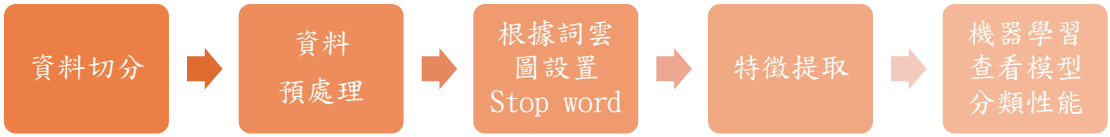
Way	Expected Use	Final Use
TfidfVectorizer	V	V
CountVectorizer	V	V

模型

Way	Expected Use	Final Use
Logistic Regression	V	V
Gaussian Naive Bayes	V	V
Bernoulli Naive Bayes	V	V
Multinomial Naive Bayes	V	V
SVM	V	X,too long time
k-Nearest Neighbors	V	X,too long time

課外知識使用 : word cloud、TfidfVectorizer

流程

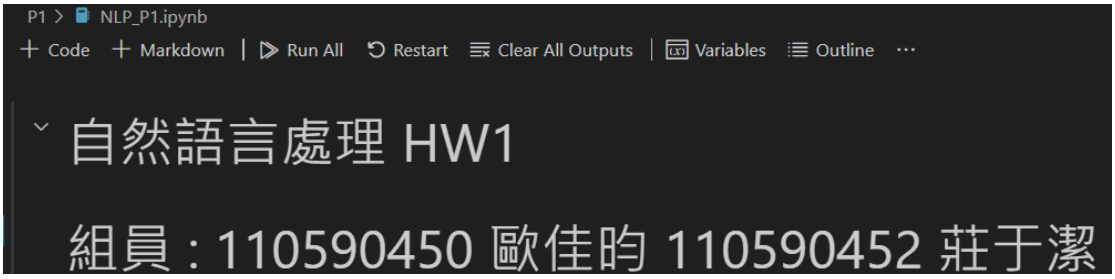


完整運作 ：

先將 input 的檔案放入 data 資料夾

```
.
└── HW1/
    ├── data/
    │   ├── test_62k.txt
    │   └── train_150k.txt
```

下載好各個需要 import 的包後，直接執行 Run All 即可



最終應出現以下檔案架構

```
.
└── HW1/
    ├── data/
    │   ├── test_62k.txt
    │   └── train_150k.txt
    ├── other_data      /
    │   ├── dataset.csv
    │   ├── feature_names_CountVectorizer.csv
    │   └── feature_names.csv
    ├── NLP_HW1.ipynb
    ├── readme.md
    └── HW1.pdf
```

**如.ipynb 無法操作，也附上.py 的版本可供使用