自然語言處理 HW1

組員 :

110590450 歐佳昀(70%) => 作業架構、流程、分析、改善

110590452 莊于潔(30%) => 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

|  |  |
| --- | --- |
| Logistic Regression with TfidfVectorizer  - Precision = 0.77  - Recall = 0.77  - F-measure = 0.77  - Accuracy = 0.77 | BernoulliNB with TfidfVectorizer  - Precision = 0.77  - Recall = 0.77  - F-measure = 0.77  - Accuracy = 0.77 |
| Logistic Regression with CountVectorizer  - Precision = 0.77  - Recall = 0.77  - F-measure = 0.77  - Accuracy = 0.77 | BernoulliNB Regression with CountVectorizer  - Precision = 0.77  - Recall = 0.77  - F-measure = 0.77  - Accuracy = 0.77 |

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 |
| k-Nearest Neighbors | V | X,too long time |

課外知識使用 : word cloud、TfidfVectorizer

流程圖

完整運作 :

一張含有 文字, 字型, 螢幕擷取畫面, 數字 的圖片

自動產生的描述先將input的檔案放入data資料夾

Data 下載 link: test\_62k.txt、train\_150k.txt

[SentimentAnalysisBert/data at main · cblancac/SentimentAnalysisBert · GitHub](https://github.com/cblancac/SentimentAnalysisBert/tree/main/data)

一張含有 文字, 螢幕擷取畫面, 字型, 行 的圖片

自動產生的描述

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述下載好各個需要import 的包後，直接執行Run All 即可

一張含有 文字, 字型, 螢幕擷取畫面, 數字 的圖片

自動產生的描述最終應出現以下檔案架構

\*\*如.ipynb無法操作，也附上.py 的版本可供使用