

自然語言處理 HW1

組員：

110590450 歐佳昀(60%) => 作業架構、流程、分析、改善

110590452 莊于潔(40%) => 作業模型挑選、參數調整

環境：使用 Jupyter Notebook	語言：python
------------------------	-----------

Classification results (Best performance => *CRF)

- Accuracy = 0.936
- Precision = 0.927
- Recall = 0.936
- F-measure = 0.928

```
=====

*CRF :
Accuracy: 0.936
Weighted Average Precision: 0.927
Weighted Average Recall: 0.936
Weighted Average F1 Score: 0.928

=====
```

經使用不同模型後，在此作業分類表現 CRF(補充方法)> NLTK > spaCy

(以下依作業執行順序)

spaCy train - test - Precision = 0.619 - Recall = 0.570 - F-measure = 0.573 - Accuracy = 0.511	spaCy train - (test+dev) - Precision = 0.641 - Recall = 0.587 - F-measure = 0.611 - Accuracy = 0.521
--	--

```
=====

spaCy :
Accuracy:0.511
Weighted Average Precision: 0.619
Weighted Average Recall: 0.570
Weighted Average F1 Score: 0.573

=====
```

```
=====

spaCy :
Accuracy:0.521
Weighted Average Precision: 0.641
Weighted Average Recall: 0.587
Weighted Average F1 Score: 0.611

=====
```

spaCy k -fold (5-fold) <ul style="list-style-type: none"> - Precision = 0.3357 - Recall = 0.3619 - F-measure = 0.3465 - Accuracy = 0.2267 <pre> ===== spaCy : Accuracy: 0.22675619834710745 Weighted Average Precision: 0.3357 Weighted Average Recall: 0.3619 Weighted Average F1 Score: 0.3465 ===== </pre>	spaCy (train for command - dev for command - test): <ul style="list-style-type: none"> - Precision = 0.879 - Recall = 0.825 - F-measure = 0.850 - Accuracy = 0.781 <pre> ===== spaCy : Accuracy: 0.781 Weighted Average Precision: 0.879 Weighted Average Recall: 0.825 Weighted Average F1 Score: 0.850 ===== </pre>
---	--

NLTK (train - test): <ul style="list-style-type: none"> - Precision = 0.882 - Recall = 0.891 - F-measure = 0.878 - Accuracy = 0.891 <pre> ===== NLTK : Accuracy:0.891 Weighted Average Precision: 0.882 Weighted Average Recall: 0.891 Weighted Average F1 Score: 0.878 ===== </pre>	*CRF : <ul style="list-style-type: none"> - Precision = 0.936 - Recall = 0.927 - F-measure = 0.936 - Accuracy = 0.928 <pre> ===== *CRF : Accuracy: 0.936 Weighted Average Precision: 0.927 Weighted Average Recall: 0.936 Weighted Average F1 Score: 0.928 ===== </pre>
--	---

資料處理：

Test: Section F Development: second half of Section H Training:
everything else (資料切分)

- 論文作者之 github:
https://github.com/GateNLP/broad_twitter_corpus/blob/master/README.md

文字內容處理：

- 篩去：篩去不合 iob 格式之資料列
- 格式轉換：iob 格式轉換，其符合 spaCy、NLTK、CRF 之輸入格式
- *CRF：大小寫轉換