一、資料描述與探索 基本資料

train.csv

- 14,869 筆資料(1.3 MB)
- Column=2(class & tweet)

class	tweet
1	[9-1-13] 2:50 pm "son of a bitch ate my mac n cheese" http://t.co/My5oJYZ8w9
1	RT @BryceSerna: Don't be a pussy grab the booty. Love the booty. Appreciate the booty.
2	RT @ClicquotSuave: bunch of rappers boutta flood the internets w/ trash remixes

test.csv

- 9,914 筆資料(918 KB)
- Column=2(id & tweet)

id	tweet
0	!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby4life: You ever fuck a bitch and she start to cry? You be confused as shit
1	!!!!!!!! RT @C_G_Anderson: @viva_based she look like a tranny
2	!!!!!!"@__BrighterDays: I can not just sit up and HATE on another bitch .. I got too much shit going on!"

一、資料描述與探索 基本資料

CLASS分類：

- 0=Hateful
- 1=Offensive
- 2=Clean

資料有稍微不太平衡的狀況

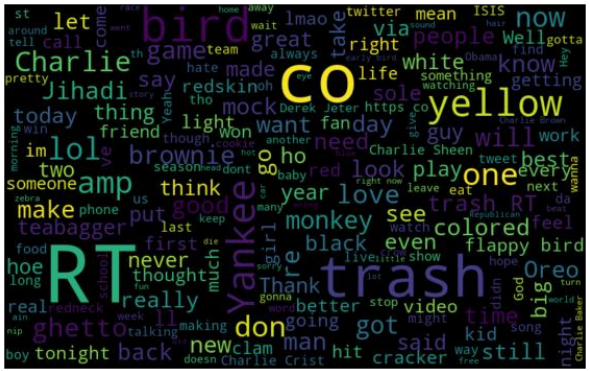
class	出現次數	比例
0	863	0.06
1	11491	0.77
2	2515	0.17
total	14869	1

一、資料描述與探索 文字雲

class=1
(Offensive)

class=0
(Hateful)

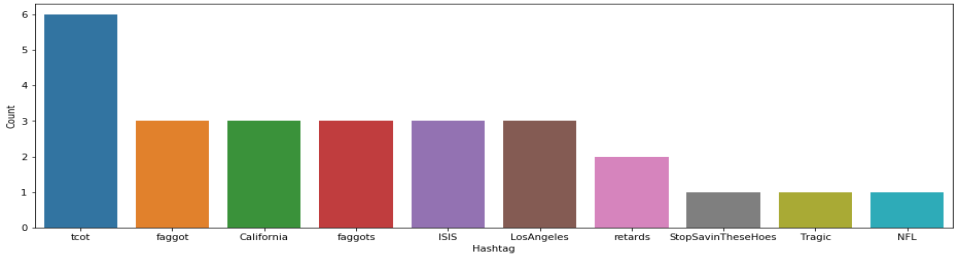
class=2
(Clean)

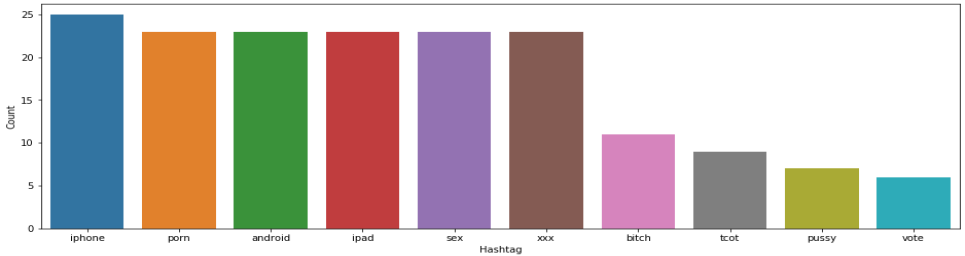


整體

一、資料描述與探索 hashtag

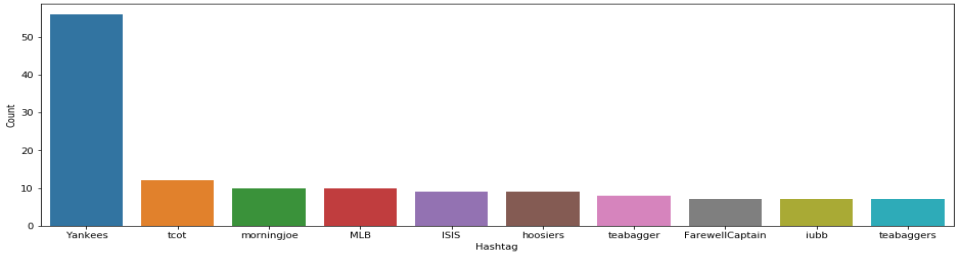
class=0
(Hateful)



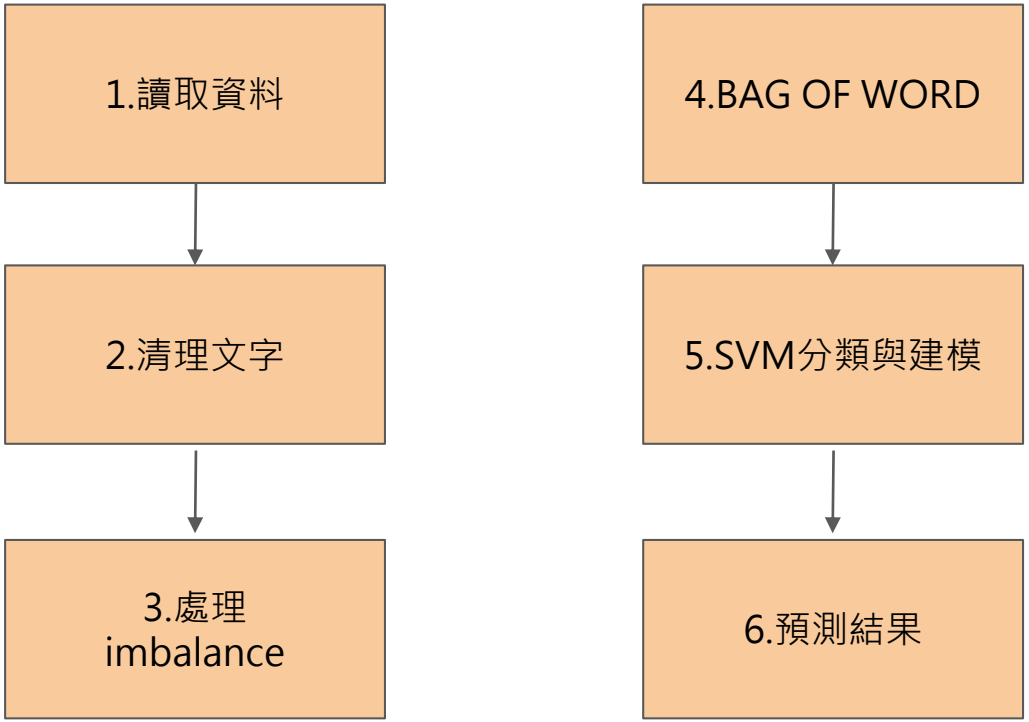


class=1
(Offensive)

class=2
(Clean)



二、文字探勘分析流程圖



二、文字探勘分析流程圖 文字清理

	class	tweet
0	0	!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby...
1	0	!!!!!!!!! RT @C_G_Anderson: @viva_based she lo...
2	0	!!!!!!"@__BrighterDays: I can not just sit up ...
3	0	!!!!“@selfiequeenbri: cause I'm tired of...
4	0	" @rhythmixx_ :hobbies include: fighting Maria...
5	0	" Keeks is a bitch she curves everyone " lol I...
6	0	" So hoes that smoke are losers ? " yea ... go...
7	0	" bitch get up off me "
8	0	" bitch nigga miss me with it "
9	0	" black bottle & a bad bitch "
10	0	" got ya bitch tip toeing on my hardwood floor...

清理前

	class	tweet
0	0	rt urkindofbrand dawg rt sbabylife you ever f...
1	0	rt cganderson vivabased she look like a tranny
2	0	brighterdays i can not just sit up and hate on...
3	0	selfiequeenbri cause im tired of you big bitch...
4	0	rhythmixx hobbies include fighting mariambitch
5	0	keeks is a bitch she curves everyone lol i w...
6	0	so hoes that smoke are losers yea go on ig
7	0	bitch get up off me
8	0	bitch nigga miss me with it
9	0	black bottle amp a bad bitch
10	0	got ya bitch tip toeing on my hardwood floors

清理後

二、文字探勘分析流程圖 處理Imbalance，使用Upsample方式

Class	出現次數	比例
0	863	0.06
1	11491	0.77
2	2515	0.17
total	14869	1

處理前

Class	出現次數	比例
0	2965	0.13
1	11491	0.50
2	8526	0.37
total	22982	1

處理後

把class != 1的分成一組、class== 1分成另一組

再將兩組資料 Upsampling 到和 class ==1 的一樣多的資料筆數

二、文字探勘分析流程圖 BAG OF WORD

```
from sklearn.feature_extraction.text import CountVectorizer
vectorizer=CountVectorizer(min_df=4)
vectorizer.fit(text_train) #建模

X_train=vectorizer.transform(text_train) #配適
X_test=vectorizer.transform(text_test) #配適
```

反覆調整與修改min_df，讓出現少於4次的詞被過濾掉。

當min_df=1時，分數為0.7454；當min_df=4，分數為0.7495，而min_df再大，分數就開始下降。

二、文字探勘分析流程圖 SVM分類與建模

```
clf = svm.SVC(C=0.1, kernel='linear', degree=5, gamma='auto')
clf.fit(X_train,y_train)
print(clf.score(X_test,y_test))  #0.8465019143752175
print(clf.score(X_train,y_train))  #0.8684149454629845
```

調整懲罰係數C。

C越大表示越不能容忍出現誤差（但易overfitting），C越小則容錯越大。

我們發現當C=0.1時，結果最好。

三、其他模型嘗試

模型	預測分數
LogisticRegression	0.63~0.64
DecisionTree	0.65~0.67
Multinomial Naïve Bayes	0.60~0.63
SGD Classifier	0.73~0.74
ensemble.ExtraTrees	0.5~0.55
SVM	0.7495

我們有使用其他模型來做預測，其中DecisionTree跟LogisticRegression有搭配AdaBoostClassifier來預測。

以上為各種模型的嘗試結果。

四、未來展望

資料前處理

- 把意思相近的文字（例如：bitch和bitches、nigga和nigger）視為同一類來處理，減少資料處理的複雜度。

模型

- 嘗試用其他深度學習模型，來得到更好的預測結果