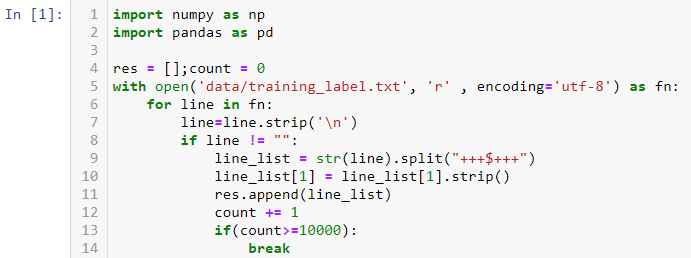
資料探勘研究與實務 Data Mining

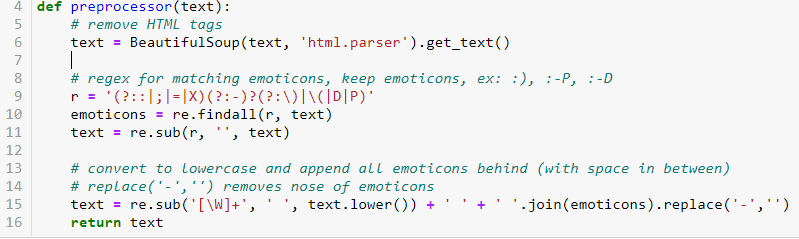
0853426陳紹雲

1. 資料前處理-讀入資料

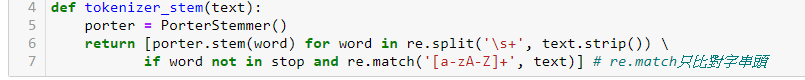
利用迴圈以行依序讀入串列中，分別對Train, Test Data做處理，最後轉成DataFrame 

1. 資料前處理- Data Cleaning

主要對資料做三種保留方式，

* 1. 用bs4套件去除 html tags。
  2. 用re抓emoji，保留在text
  3. 最後再把所有單詞轉成小寫把emoji 附加在最後面

1. 資料前處理- Word Stemming

製作一個可以拆解句子成token的函式，且跟StopWord比對

1. CountVectorizer

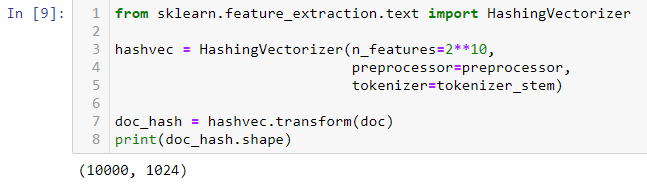
利用字頻建立Token of Bags.

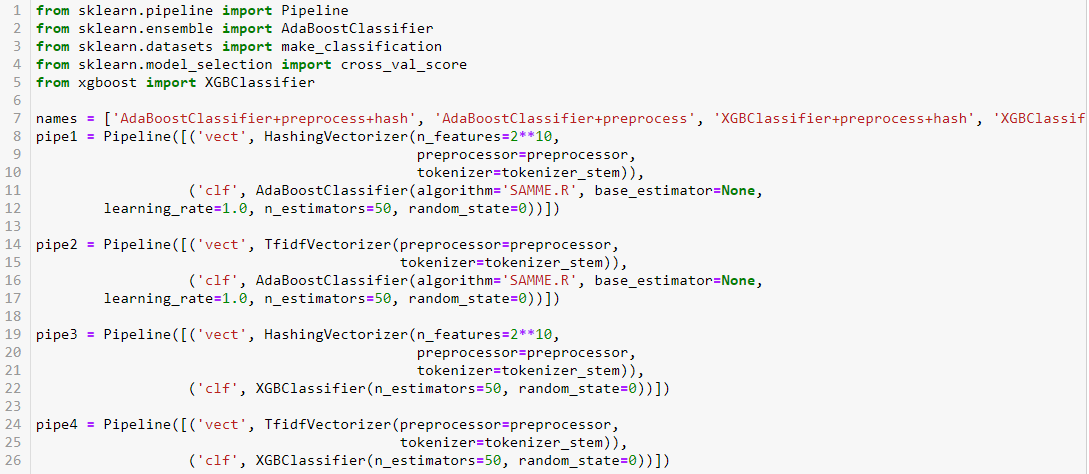


1. TfidfVectorizer

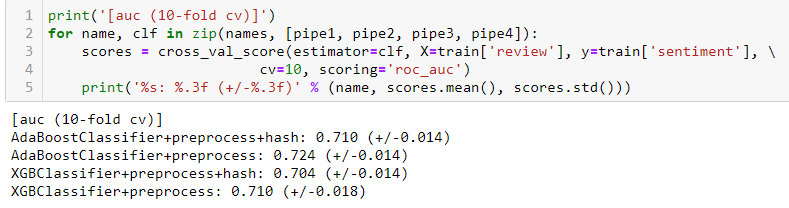
建立idf scores vector ，其中考慮term-frequency (TF) as BoW 也考慮到the document-frequency (DF)

1. Feature Hashing

將每個詞彙散列到具有固定數量的bucket的table中來減少維度詞彙空間。(降低內存所需，但是犧牲掉IDF加權) 

七、利用Pipeline建模

1. 交叉驗證(10-fold)



這邊發現Adaboost加上有做過預處理的效果最好。

1. 結果(抓兩個模型+預處理的做比較)

