DL cup1 心得

Cup01
組員名單：110065508李丞恩 109062676劉廷哲 110062592姜宏昀 110062539 古之恒

如何選擇model?

我們有試過基本上lab出現過的model包括

- LogisticRegression
- RandomForeset
- Bayesian Classifier
  等等

但最後我們在查資料發現xgboost不錯且相當多人使用，所以最後基本上都是用xgboost

Feature selection如何選擇

一開始我們以為只是單純的文本分類任務，所以只使用tf-idf，但分數相當慘，調一大堆參數都沒有用，後來在網路上查資料、問同學才發現是要處理html裡面的tag，拿來當feature，
後來我們也有在網路上查到 **https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity#** 這個資料集，應該就是我們比賽的原始dataset，他也有公布他所使用的特徵說明，所以我們就照說明多產生了以下data

3. n_tokens_content: Number of words in the content
4. n_unique_tokens: Rate of unique words in the content
5. n_non_stop_words: Rate of non-stop words in the content
6. n_non_stop_unique_tokens: Rate of unique non-stop words in the content
7. num_hrefs: Number of links
8. num_self_hrefs: Number of links to other articles published by Mashable
9. num_imgs: Number of images
10. num_videos: Number of videos
11. rate_positive_words: Rate of positive words among non-neutral tokens
12. rate_negative_words: Rate of negative words among non-neutral tokens
13. avg_positive_polarity: Avg. polarity of positive words
14. min_positive_polarity: Min. polarity of positive words
15. max_positive_polarity: Max. polarity of positive words
16. avg_negative_polarity: Avg. polarity of negative words
17. min_negative_polarity: Min. polarity of negative words
18. max_negative_polarity: Max. polarity of negative words
19. title_subjectivity: Title subjectivity
20. title_sentiment_polarity: Title polarity
21. abs_title_subjectivity: Absolute subjectivity level
22. abs_title_sentiment_polarity: Absolute polarity level
23. num tag: 一些像是b, strong, em等強調的tag

還有一些weekday，author，channel等
文字特徵的部分有試過topic、內文的tfidf/hash vector

遇到的問題 or 發現

我們後來再做大概可以train model在cross_validation分數可以到0.6，但丟上去表現就沒有那麼好，可能是還漏了一些重要的feature吧!?
然後雖然是說是新聞分類，但其實不用用到文字的feature，其實分數就可以到0.58-0.59表現已經很不錯了，再加上topic 的tfidf可以到0.6，也是我們最好的成績了，所以可以知道這種任務feature selection超級重要!

另外遇到的問題是
我們針對training data中的作者根據他們發過文章的平均受歡迎分數產生新的feature，但是這樣坐在training時可以提升很高的表現，但是丟到kaggle卻退步相當多，這我們也不太知道原因

實驗結果

關於實作上的細節以及所使用的package，可參考我們的notebook。

Method	cross_val_score	Kaggle score
SVM-linear (example)	0.5	0.5
MNB	0.54	0.526
Adaboost classifier(by 廷哲)	0.531	0.51437
logistic regression+文本hash vector+作者+星期+圖片數量(by 丞恩)	0.527	X
logistic regression(C=0.001)+關鍵字hash vector (2048)+作者+星期+圖片數量(by 丞恩)	0.528	X
logistic regression(C=0.001)+關鍵字hash vector (1024)+作者+星期+圖片數量(by 丞恩)	0.557	X

channel, weekday, author 三個one-hot encoding, topic (hash vector) ,img count,media count 以及這些
`np.expand_dims(df_test['n_tokens_title'].values, axis=-1),`
`np.expand_dims(df_test['n_tokens_content'].values, axis=-1),`
`np.expand_dims(df_test['n_unique_tokens'].values, axis=-1),`
`np.expand_dims(df_test['n_non_stop_words'].values, axis=-1),`
`np.expand_dims(df_test['n_non_stop_unique_tokens'].values, axis=-1),`
`np.expand_dims(df_test['num_hrefs'].values, axis=-1),`
`np.expand_dims(df_test['num_self_hrefs'].values, axis=-1),`
`np.expand_dims(df_test['day_of_month'].values, axis=-1),`
`np.expand_dims(df_test['month'].values, axis=-1),`
`np.expand_dims(df_test['hour'].values, axis=-1),`
	0.591	0.57940

channel, weekday, author 三個one-hot encoding, topic (hash vector) ,img count,media count 以及這些
`np.expand_dims(df_test['n_tokens_title'].values, axis=-1),`
`np.expand_dims(df_test['n_tokens_content'].values, axis=-1),`
`np.expand_dims(df_test['n_unique_tokens'].values, axis=-1),`
`np.expand_dims(df_test['n_non_stop_words'].values, axis=-1),`
`np.expand_dims(df_test['n_non_stop_unique_tokens'].values, axis=-1),`
`np.expand_dims(df_test['num_hrefs'].values, axis=-1),`
`np.expand_dims(df_test['num_self_hrefs'].values, axis=-1),`
`np.expand_dims(df_test['day_of_month'].values, axis=-1),`
`np.expand_dims(df_test['month'].values, axis=-1),`
`np.expand_dims(df_test['hour'].values, axis=-1),`
+
`'global_sentiment_polarity' : global_sentiment_polarity,`
`'global_subjectivity' : global_subjectivity,`
`'title_subjectivity' : title_subjectivity_list,`
`'title_sentiment_polarity' : title_sentiment_polarity_list,`
`'abs_title_subjectivity' : abs_title_subjectivity,`
`'abs_title_sentiment_polarity' : abs_title_sentiment_polarity,`
	0.590	0.5840

(topic+title) sentiment detect + (pub_date, channel)用EDA分析找popular時段和channel + (author)用average_popular設定權重 + (Page Content)用feature hasing [1024/2048]-> 用corr()找出和(Popularity)的corr() top 100 features + [一些像是bold, strong, emphasize的html tags + ohe author]-> 直接用houyu的xgb, num_boost_round = 1000, max_depth = 3, subsample = .75 (by Ku)	0.594/0.598/[0.61]	0.55/X/[0.56]
上圖所有feature+關鍵字hash(1024個bucket)	0.597	0.5616