SDML HW1 Task2 Report

B04902016 曾奕青
B04902103 蔡昀達
B04902105 戴培倫

October 15, 2018

Contents

1 Preprocess

Title and Abstract

1.1 Tokenize

nltk 的 RegexpTokenizer

1.2 Stopwords

nltk.corpus.stopwords('english')

1.3 Stemming

- stemming
  nltk 的 PorterStemmer
- lemmatize
  nltk 的 WordNetLemmatizer

由於 title 資料較不足，使用 doc2vec 的模型訓練之後，title 的 inference vector 部分無法從訓練資料中找出自己，說明訓練結果不佳，但在使用 stemming 的情況下，訓練結果大幅改善。

2 Feature Engineering

2.1 TF-IDF

- term frequency $\in [0.05, 0.95]$
- sublinear term frequency scaling

2.2 Node Features

- degree centrality
- betweenness centrality
- load centrality
- katz centrality
- pagerank

2.3 Similarity

- Jaccard Coefficient
  兩個 node Title 文字的 Jaccard Coefficient。
  兩個 node Abstract 文字的 Jaccard Coefficient。
- Cosine Similarity
  兩個 node Title Embedding 的 similarity。
  兩個 node Abstract Embedding 的 similarity。
- Correlation
  兩個 node Title Embedding 的 correlation。
  兩個 node Abstract Embedding 的 correlation。

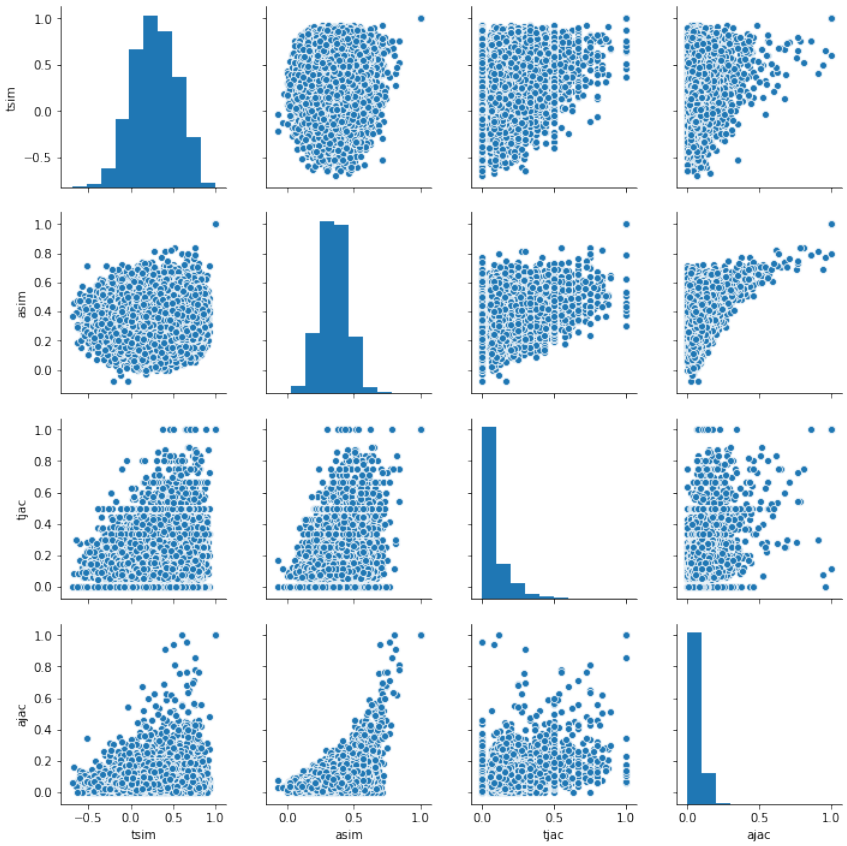


Figure 1: Test data: cosine similarity and jaccard's correlation

從圖中 (左上爲 (0,0)) 可以看出 cosine similarity 和 jaccard similarity 的關係。如 (0,2) 的 graph，可以看到當 title jaccard 趨近於 0 時，還是有非常多 title cosine similarity 很高的點，能看出 jaccard 可以補足許多 cosine similarity 沒辦法提供的訊息，可看出 jaccard 的重要性非常高。

3 Embedding

3.1 DeepWalk

- 使用 t2-train.txt 訓練 deepwalk embedding ，dimension=128。
- 沒有出現過的 node 則使用 0 向量

3.2 Doc2Vec

- Title 跟 Abstract 各做一個 dm=1 的 Model，dimension=100。
- Title 跟 Abstract 各做一個 dm=0 和 dm=1 的 Model。
  加上 dm=0 的 Model 沒有什麼差別因此之後沒有用。

3.3 LSTM AutoEncoder

跟 task3 的 model 都從上星期開始訓練，至今還未收斂，敬請期待。

4 Negative Sampling

使用拓樸結構上 path length 當作 negative samples 的依據，並檢查和 test data 文本相似度的 correlation。我們也嘗試過使用 similarity 優先調整 sampling weight，結果不理想。

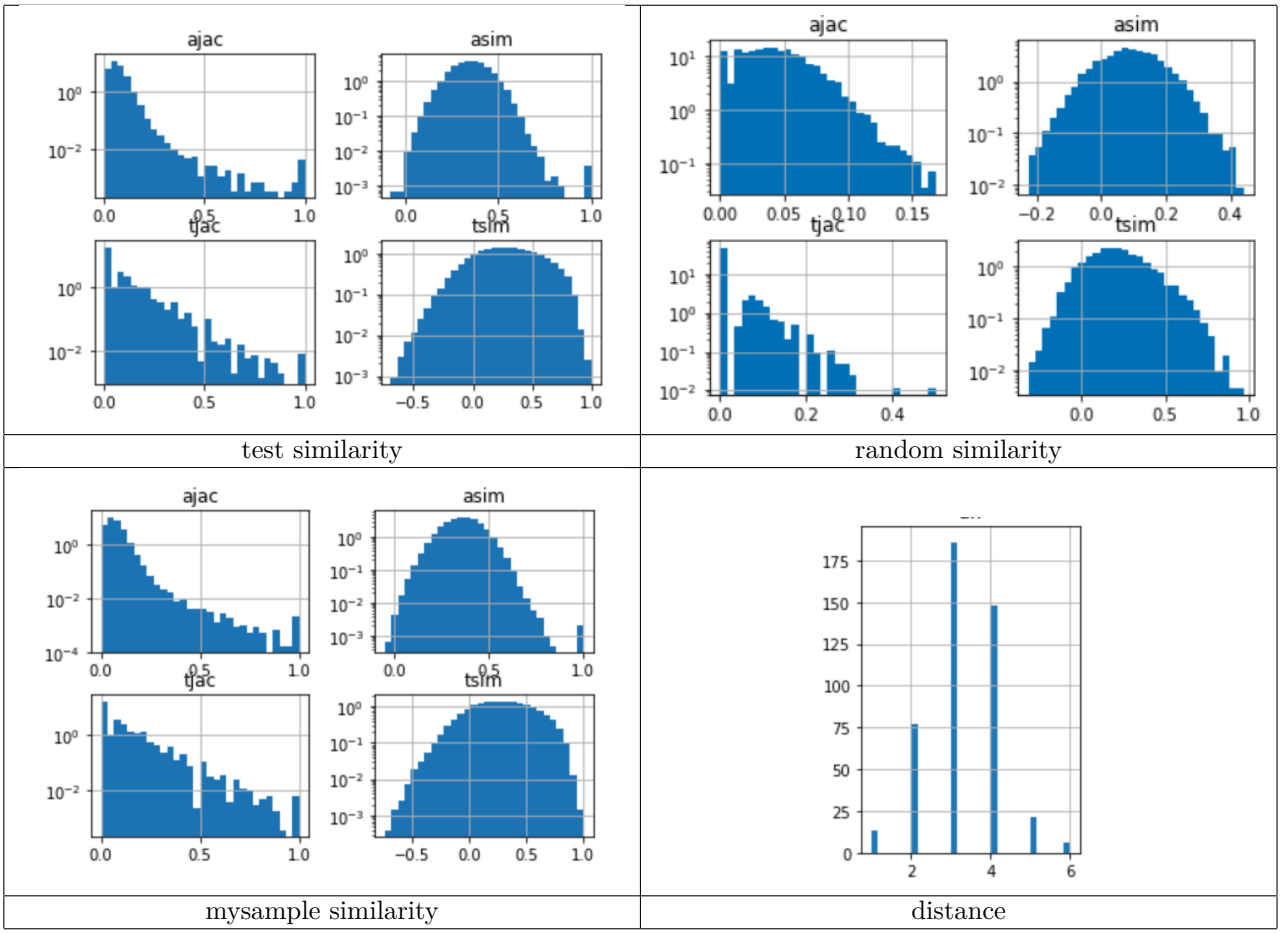
tsim: title cosine similarity
asim: abstract cosine similarity
tjac: title jaccard similarity
ajac: abstract jaccard similarity

	tsim	asim	tjac	ajac
0.15	0.010592	0.250086	0.000000	0.027273
0.25	0.102416	0.285553	0.000000	0.037267
0.5	0.285110	0.351998	0.000000	0.058252
0.75	0.473514	0.419230	0.095238	0.084615
0.95	0.707485	0.514501	0.250000	0.135593
mean	0.285212	0.352524	0.060260	0.065000
後5%	-0.076429	0.214870	0.000000	0.019193
前5%	0.644688	0.490927	0.222759	0.127097

Figure 2: test similarity

	tsim	asim	tjac	ajac
0.15	0.010082	0.261868	0.000000	0.028571
0.25	0.099207	0.297665	0.000000	0.038835
0.5	0.281507	0.363840	0.000000	0.061224
0.75	0.470833	0.429733	0.111111	0.088608
0.95	0.700070	0.523751	0.272727	0.142857
mean	0.282568	0.363613	0.068447	0.068066
後5%	-0.076427	0.226498	0.000000	0.019925
前5%	0.639808	0.500431	0.238579	0.133123

Figure 3: mysample similarity



test similarity | random similarity

mysample similarity | distance

從數值和圖表可以看出我們的 negative samples 在 cosine 和 jaccard similarity 上跟 test data 是相近的。Negative samples 主要是由 shortest path 2 - 4 步的 node 組成的。

5 Classifiers

5.1 Light gbm

- n_estimators = 2000
- objective = "binary"
- max_depth = 10
- early_stopping_round = 8
- learning_rate=0.1

訓練速度非常快。

5.2 Random Forest

- n_estimators = 3200
- min_samples_leaf = 20
- class_weight = {0: 1, 1: 1.15}

random forest 在訓練時 positive 和 negative 準確率常會有 gap，爲了消除這個 gap 我借助了原來用於 unbalance training 的參數，以消除 positve 和 negative 的 accuracy gap。

5.3 XGBoost

- n_estimators = 2000
- objective = "binary: logistic"
- max_depth = 10
- evaluation
  - rmse
  - mae
  - error

5.4 Strategy

在這次的作業中，困難之處在於 validation 和 evaluation。validation set 會因爲 negative sampling 好壞的影響，較難以判斷 overfit 或是 underfit，在調整 model 和評鑑結果上對我們造成了很大的障礙。Evaluation 因爲 error surface 非常陡峭，對於使用 early stopping 的調整也提升了難度。

我們除了調整 negative sampling 的作法外，使用了多種 evaluation metric 來作爲 overfit 和 underfit 的指標，因爲單一種評量方式太不穩定。當 rmse 開始上升時，mae 和 accuracy 可能都還在變好; 當 accuracy 下降，rmse 和 mae 可能都在還在下降，因此使用多種 evaluation

metrics 較能夠準確判斷，以有效調整模型。不過更精準的策略應該是設置足夠大的訓練次數，再經由人工檢視 error curve 來做判斷，缺點則是會花很多時間。

在這次的作業中，不論我們如何嘗試，傳上 kaggle 都得不到太好的結果，也許還有什麼我們沒發現的重要線索，藏在 testing data 的細節裡。

6 Results

6.1 Single Model

all models we tried got 50% acuuracy

6.2 Ensemble

Weighted average of 44 results.

- public: 0.50183
- private: 0.50267