Assignment 3

Zijie Huang

March 2, 2020

This assignment is to extract company names, numbers involving percentages and name of CEOs from scraped articles for year 2013 and 2014 of BusinessInsider, a portal for business news.

This assignment consists of the following three steps:

1. Candidates Extracting

The preprocessing of the articles includes sentence segmentation, tokenization, removing stop words and normalization. The NER tool spaCy is then used to extract People names, Organization names and percent numbers as candidates(dataset) for the model.

Since spaCy has inner module for segmentation, tokenization and removing stop words, the manually preprocessing codes are removed from the submitted files.

The extract candidates are saved as file *CEO_candidates.csv, Companies_candidates.csv, Percent_candidates.csv.*

2. Feature Selection

Samples of CEOs have following features:

(1) Capital Letters: The number of capital letters.
(2) Space: The number of spaces.
(3) Length: The length of this sample.
(4) Frequency: Occurrences in the article
(5) Chief in sentence: If string *chief executive* or *Chief Executive* appears in ALL the sentences where the sample locates, this feature is labeled as 1, else as 0.
(6) Chief in article: If string *chief executive* or *Chief Executive* appears in the article where the sample locates, this feature is labeled as 1, else as 0.
(7) CEO in sentence: If string *CEO* or *ceo* appears in ALL the sentences where the sample locates, this feature is labeled as 1, else as 0.
(8) CEO in article: If string *CEO* or *ceo* appears in the article where the sample locates, this feature is labeled as 1, else as 0.

Candidates with the same name that appear in different places in the same article are recorded as one sample. Candidates with the same name that appear in different articles are recorded as different samples.

Samples of companies have following features:

(1) Capital Letters: The number of capital letters.
(2) Space: The number of spaces.
(3) Length: The length of this sample.
(4) Frequency: Occurrences in the article

Samples of percentages have following features:
(1) Space: The number of spaces.
(2) Length: The length of this sample
(3) Has .: If character '.' appears in sample, this feature is labeled as 1, else as 0.
(4) Has %: If character '%' appears in sample, this feature is labeled as 1, else as 0.
(5) Has percent: If string *percent* appears in sample, this feature is labeled as 1, else as 0.

3. Classification

Logistic Regression is selected as the classification model in this assignment. The labeled dataset is saved as file: *Companies_candidates_labeled.csv, CEO_candidates_labeled.csv, Percent_candidates_labeled.csv.*

Considering the fact that the amount of dataset is sufficient, 50% of the dataset is used as training data, while the remaining 50% is used as test data. The results of the classification model are as follow:

```
X = Percent_cdd.iloc[:, :-1]
Y = Percent_cdd.iloc[:, -1]
train_X, test_X, train_Y, test_Y = train_test_split(X, Y, test_size = 0.5, random_state = 42)



LR = LogisticRegression(C = 1.0, penalty = 'l2')
LR.fit(train_X, train_Y)
test_prediction= LR.predict(test_X)
print(classification_report(test_Y, test_prediction))
```

```
              precision    recall  f1-score   support

           0       0.90      0.92      0.91     10671
           1       0.95      0.93      0.94     15896

    accuracy                           0.93     26567
   macro avg       0.93      0.93      0.93     26567
weighted avg       0.93      0.93      0.93     26567
```

Figure 1 Result of percentage extraction

	precision	recall	f1-score	support
0	0.94	1.00	0.97	25553
1	0.28	0.03	0.05	1586
accuracy			0.94	27139
macro avg	0.61	0.51	0.51	27139
weighted avg	0.90	0.94	0.91	27139

Figure 2 Result of CEO extraction

	precision	recall	f1-score	support
0	0.87	1.00	0.93	30249
1	0.42	0.02	0.03	4560
accuracy			0.87	34809
macro avg	0.65	0.51	0.48	34809
weighted avg	0.81	0.87	0.81	34809

Figure 2 Result of company extraction

Regardless of precision, record, or accuracy, the performance of the classification model is satisfactory.