Text Analytics

Executive Summary

Articles written in 2013 to 2014 provided by Business Insider are scraped. Articles can be used for further uses such as market sentiment analysis so that investors' mood can be gauged. Other methods to utilize the data include developing a Q&A system. However, prior to further analysis, name entity recognition (NER) should be used so that information can be extracted from unstructured text.

Text mining is an analytical strategy that can be used to train NER models. Training dataset and test dataset were split 60-40 before any prior knowledge. Regular expressions were then used to extract potential candidates for CEO, Company names and Percentage. Given the labeled dataset for samples, a set of features were created and positive and negative samples were gathered with 1-1 ratio. Logistic Regression and Random Forest models are examined. Random Forest, with a high accuracy, now are used to predict training dataset. For percentage extraction, regex gives a really accurate extraction and since the sample dataset got irrelevant numbers, no further steps were taken.

Data Exploration for Labeled dataset

After reading the dataset for labeled ceo and company, the method of drop duplicates are used to see characteristics of words contain ceo and company. Ceo names usually are two words with capitalized letters. Stop words don't appear in the names so that excluding stopwords in the potential candidate should be done. Stop words should also be expanded to exclude non-related objects. For example, cities with two words are excluded as well as well-known companies with two-words. In addition, both words should be noun.

For Company names, "the" is vastly used in company names so excluding all stopwords shouldn't be feasible. I constructed a customized stopwords list for better prediction purpose. Out of 2592 company names, only 8 of them doesn't start with capitalized letter. On the other hand, if all lowercase samples are included, the model accuracy can drop drastically. Therefore, lowercase samples are ignored in the regex expression.

Regex Expressions

For Percentage:

The end of the word is either %, percentage,percent,percentile,percentage points
For the numeric values beforehand, it's either in numeric or in words
For words version, it can be one to nine, or two-digits
For regex expression, it's easy to aggregate them with (), however, python re.findall extract groups as extra columns so that all these regex expressions are examined individually.

#numeric digits
r1 =r"['-]?\d+\.?\d+\s?%"
r2= r" ['-]?\d+\.?\d+\spercent"
r3= r" ['-]?\d+\.?\d+\spercentile"
r4= r" ['-]?\d+\.?\d+\spercentage"
r5= r" ['-]?\d+\.?\d+\spercentage points"

#words-one digit (one to nine)
n1= r"(?:f(?:ive|our)|s(?:even|ix)|t(?:hree|wo)|(?:ni|o)ne|eight|zero)\spercent"
n2=r"(?:f(?:ive|our)|s(?:even|ix)|t(?:hree|wo)|(?:ni|o)ne|eight|zero)\spercentage"
n3=r"(?:f(?:ive|our)|s(?:even|ix)|t(?:hree|wo)|(?:ni|o)ne|eight|zero)\spercentile"
n4=r"(?:f(?:ive|our)|s(?:even|ix)|t(?:hree|wo)|(?:ni|o)ne|eight|zero)\spercentage points"

#word version in two-digits
n5=r"(?:fif|six|eigh|nine|(?:tw|sev)en|(?:thi|fo)r)ty['-]?(?:f(?:ive|our)|s(?:even|ix)|t(?:hree|wo)|(?:ni|o)ne|eight)\spercent"
n6=r"(?:fif|six|eigh|nine|(?:tw|sev)en|(?:thi|fo)r)ty['-]?(?:f(?:ive|our)|s(?:even|ix)|t(?:hree|wo)|(?:ni|o)ne|eight)\spercentage"
n7=r"(?:fif|six|eigh|nine|(?:tw|sev)en|(?:thi|fo)r)ty['-]?(?:f(?:ive|our)|s(?:even|ix)|t(?:hree|wo)|(?:ni|o)ne|eight)\spercentile"
n8=r"(?:fif|six|eigh|nine|(?:tw|sev)en|(?:thi|fo)r)ty['-]?(?:f(?:ive|our)|s(?:even|ix)|t(?:hree|wo)|(?:ni|o)ne|eight)\spercentage points"

For CEO:
r" [A-Z][a-z]+\s[A-Z][a-z]+"

For Company:
r"[A-Z][\w-]+(?:(?:\s+[A-Z][\w-]*)+)"

Features
For CEO's:
Condition:

- Not in the extended stop words list
- Both words are NNP (noun)

Criteria :

- If the original sentence contains 'CEO'
- The position of the first word's index
- Length of the first word

- If the original sentence contains ‘chief’
- If the original sentence contains ‘executive’
- If the original sentence contains ‘officer’

For Company:

Condition:

- Not in the customized stop words list
- Not in the CEO sample list (because CEO names are easy to be categorized as potential candidate)

Criteria:

- If there are specific key words in the sentences: {'Inc', 'Group', 'Ltd', 'Co', 'Corp', 'Corporation','Management','Company'}
- If the original sentence contains lowercase ‘company’
- The index of the first word
- Number of words in the regex captured string (i.e. number of words for ‘Vianet Group’ is 2)
- Number of letters in the captured string (i.e. number of letters for ‘Vianet Group’ is 11)

Training Dataset Size Readjustment

One thing to notice is that the ratio of positive sample versus negative sample. By using groupby to sum numbers of positive sample and negative sample, we can see that the numbers of negative samples are way larger than the positive samples in both CEO and Company case. Therefore, we adjust to a 1-1 ratio so that the model will not be biased.

Models

After obtaining the adjusted feature matrix for training dataset, we then standardize the matrix to run logistic regression and random forest. Since Random Forests model’s precision is way higher than logistic one (as illustrated in the following figure), only test result for random forest model is displayed in the following.

```
Logistic Regression for ceos: accuracy score is 0.6445856855236667
Logistic Regression for ceos: precision is 0.3951831138372722, recall is 0.655936517442731, and f1 score is
0.4932169997185477.
Random Forest for ceos: accuracy score is 0.6890766254786626
Random Forest  for ceos: precision is 0.6714114197927592, recall is 0.719334632762972, and f1 score is 0.6634415606806023.
```

Figure 1: Logistic versus Random Forest

Random Forest Model testing sample accuracy

Summary	Accuracy	Precision	Recall	f-1
CEO	0.689	0.671	0.719	0.663
Company	0.972	0.652	0.556	0.579

The baseline of the model accuracy should be 0.5 so that both Random Forest models past the baseline and the model for Company has a really high accuracy.
We then use these fitted model to predict the entire dataset. Outputs are captured in the following three csv files: ‘output_ceo.csv’, ‘output_company.csv’, ‘output_perctage.csv’

Next Step
In terms of improving the predictability of the model, we can expand the stop words dictionary for CEO so that unrelated two-word objects can be excluded from model training. Another thing to implement for next step is to try different supervised learning model. Random Forest Model is easy to interpret, while others such as Naive Bayes can be further explored.