IEMS308 HW3 Text Analytics

Taige Hong

Feb 2019

**Executive Summary**

BusinessInsider is a portal for business news, and we are using the scraped articles for year 2013 and 2014 to perform text analytics to extract the CEO names, company names and percentages from them.

**Process**

Preprocessing:

We are using three different preprocessing method for the three classes. For the CEO names, we see that there are many formats of the names, and we are transforming them into ‘Firstname Lastname’ and store them in a tuple. For the company names, we are simply removing the wrapping white spaces and periods from the names. If we are to include the periods, we will be including a lot of period as the end of sentence as candidates, as the letter following the period is very likely to be capital letters. For the percentages, we find out that the training samples include many of pure numbers which could affect the effect of the model, so we decide to use regex expression to filter the percentages.

Training model for CEO name and company name:

For the CEO names, we find out that there are a lot of words that are very unlikely to appear in people’s names (locations, numbers, etc). So, we decide to expand the stop words to filter out more candidates. For the company names, we see that many names actually include the stop words, so we decide not to filter out them. Instead, we are filtering out the single English words as they are a huge portion of the candidates but are very unlikely to be company names.

The regex expression I used for finding out CEO names is: r'([A-Z]\w+) ([A-Z]\w+)' This is simply two consecutive words with capital letter in the beginning. The regex expression I used for finding out company names is: r" (?:[A-Z]+[A-Za-z0-9']\* ?)+" This is a certain number of consecutive words with capital letter in the beginning or mixed lowercase and uppercase.

After finding out the candidates, we use the criteria of whether or not they are in the training set to determine whether or not they should be marked positive. It turns out positive labels are way less than negative ones, as we expected. So, we decide to keep them balanced as we are using an equal number of training labels to train the data. This prevents the model being skewed and inflating the accuracy rate.

For the CEO names, I used the following features:

* If “ceo” is in the sentence
* If “chief”, “executive”, “officer” is in the sentence respectively
* If there are numbers in the sentence
* The length of the sentence (number of words)
* Length of the first name
* Length of the last name
* Proportion of capital letter
* If the previous word has capital letter in the beginning
* If the next word has capital letter in the beginning

The features I used to find company name include:

* If company is in sentence
* If keywords are in the list of words. (keywords are 'inc', 'group', 'ltd' etc. as a good indicator of company names)
* If there are numbers in the sentence
* Number of words in the sentence
* Number of words of the candidate
* Number of letters of the candidate
* Proportion of capital letters
* If previous word has capital letter in the beginning
* If next word has capital letter in the beginning

I used a logistic regression to fit the CEO names and got an accuracy rate of 80.86%. Considering we are using a manually-balanced training set, this result is rather satisfactory.

I used support vector classifier to fit company names and got an accuracy rate of 86.78%.

Considering we are using a manually-balanced training set, this result is rather satisfactory.

We found out 495 different CEO names and 4529 different company names.

Training model for the percentage:

As mentioned in the preprocessing, the quality of the given training label set is questioning, so we are using regex expression to filter out the percentages. The expression I used is: r'(([0-9]+)|(one)|(two)|(three)|(four)|(five)|(six)|(seven)|(eight)|(nine)|(ten))[.[0-9]\*]?(%| percentage point| percent)+s?'. It might be missing out some of the values but we found out it is a small proportion. We found out 627 different percentages.