IEMS 308

Jay Jung

11/24/2016

Text Analytics

**EXECUTIVE SUMMARY**

Business Insider (BI) is “an American business, celebrity and technology news website launched in February 2009 and based in New York City”. In this project, we were provided with daily BI articles for the years of 2013 and 2014. This is the first step of designing a simple Question-Answer system. For this deliverable, the goal is to design an algorithm that can extract all the names of the CEOs that were mentioned in the articles along with companies and percentage numbers.

The analysis was performed through following procedures: sentence segmentation, word tokenization by stop words, POS Tagging, Stemming, Data extraction and training data set creation.

CEO’s and company lists given were used to create training data sets. Based on the training sets created, the company name’s and CEO name’s were extracted. The training model was evaluated based on the performance.

**Problem Statement**

The goal of this project is to extract names of CEOs, company names and percentages from Business Insider articles. We trained a naive Bayes classifier on articles from 2013 and testing on 2014 articles. The trained model is going to be used to predict Q&A analysis, the measurement to predict how company performed in the next assignment when either the name of the company or the CEO is given.

**DATA EXPLORATION**

First, as mentioned above, a dataset of Business Insider articles for the years 2013 and 2014 was given. The articles were given in text files. Each file was published in a single day. Secondly, a list of CEO names, company names, and percentages was given to train with. The “CEO” file originally contained 2657 words. Some of the world contained “,”; some of them only had last names in the file.

**Assumption**

* ‘NLTK’ package is performing well enough so that we can ignore the error. ‘nltk’ package was used to process the data. Sentence Segmentation, word tokenization, POS tagging and stemming was performed by the package. It definitely has some errors. However, some of the data they failed to process was assumed to be ignorable.
* The duplicate data on Company and CEO lists were not abbreviated. It might be possible that more than two CEOs and companies have same name.

**Method**

Sentence Segmentation

Sentence segmentation was performed by ‘nltk’ packages. All the articles was uploaded as the corpus. All rows in corpus was segmented into sentences.

Word Tokenization

The word tokenization was performed by ‘nltk packages. The packages performed this by tokenizing the world by stop words.

POS Tagging

The first step was to generate part of speech (POS) tags for each word in the corpora. NNP(proper noun) is our interest. The words tagged as NNP are likely to be either company name or CEO name.

Stemming

The stemming was performed by ‘nltk’ package. Stemming will shorten the data processing time and increase the accuracy. Stemming was performed after POS tagging. This is because we would like to keep the correct POS tag that was performed on original text. This was done by regular expression, regex = r'[^"()\']\*\w\*[^.,\!:"()\']' .

Training Set Creation

The given CEO and company lists were used to create dictionary. The attributes for the dictionary and training sets were created using some standards of those given lists.

Tag(=NNP),FirstUpperLetter,FirstLowerLetter,SecondUpperLetter,SecondLowLetter were the attributes of the training set and dictionary. Training set was assessed by ‘sklearn package’. Y(classification if it is CEO or Company or not) values were predicted based on X(Attributes). The words that yield Y value of 1 was analyzed by key attributes in dictionary of CEO and company lists. CEO and company lists were extracted.

**Analysis**

Due to the time constraint, Part 3 can not be completed. However, company names and CEO names were extracted appropriately. The lists of words are provided below. The R squared of the training set was 0.973857605575, which was pretty high. It means that the model has the mean squared value of 2.6%. Since the lists were too large, the lists were attached and uploaded as the separate file.

0.973857605575

precision recall f1-score support

0 0.98 0.99 0.99 1775163

1 0.00 0.00 0.00 36186

avg / total 0.96 0.97 0.97 1811349

**Conclusion**

Based on dictionary created by CEO and company lists, training model was created and evaluated. The model itself fits the training data well. It has high support, f1-score, precision and recall.

Some of the extracted CEO and file names were not company name nor CEO names. For example, country names such as ‘China’ and ‘Scotland’ were extracted as the company names. Some none NNP that started at the beginning of the sentence was classified as the company name: “Or”. Some acronyms such as ‘Inc’ and ‘Co’ were treated as the company name.

However, the error seems to be minute part of the analysis, based on the accuracy measurement of the training data.

**Next Step**

Part 3 should be finished. Otherwise, it would be impossible to perform the Q&A analysis. Binary classifier and other attributes should be created in order to predict the performance of the company when either the name of CEO and company are given.