Melissa Melnick
Homework 3
IEMS 308
03/07/21

Introduction:

The overall task for this project was to extract all CEO names, all Company names, and all numbers involving percentages from a series of Business Insider articles from the years 2013 and 2014. To accomplish this, a number of steps had to be taken.

First, the text files were read into Jupyter Notebook. The articles for 2013 and 2014 were held in separate files, so each group was read in separately, then combined. Next, using the NLTK package, the large file of text was tokenized into its individual sentences. From there, the extraction process could begin.

Percentages:

The percentages portion of the task seemed to be the most simple to solve. All numbers involving percentages were extracted through four different regular expressions detailed as follow:

1. Percentage made up of digits, a decimal point, and a percent sign (e.g. 9.5%)
    a. Regex used: **[0-9]*[\.]?[0-9]+[\%]**

2. Percentage made up of digits and the word percent (e.g. 9.5 percent)
    a. Regex used: **[0-9]*[\.]?[0-9]+ percent**

3. Percentage made up of digits and the words percentage points (e.g. 9.5 percentage points)
    a. Regex used: **[0-9]*[\.]?[0-9]+ percentage point[s]?**

4. Percentages written in English
    a. Regex used: **(?:one|two|twe|thr|thi|for|fou|fif|fiv|six|sev|eig|nin|ten|zer|hun|hal|thou|quart)[a-z]+ percent**

The most difficult type of percentages to extract were by far the ones written in plain English. Unfortunately the regex written for those was not perfect, and garnered many false positives. However, with time and some simple adjustments this can easily be adjusted to more accurately extract all of the percentages written in English.

CEO and Company Names

Unlike the percentages, extracting all of the CEO names from the Business Insider articles required a more complex method with several more steps.

The first step taken was to read in the provided training data sets for both CEOs and Companies. Both of these data sets were cleaned by stripping white space. (In retrospect, at this stage, duplicates should have been removed as well, but time was a limiting factor which prevented it from happening).

Another regular expression was used at this point to pull segments out of each sentence that could serve as a potential candidate for a CEO name

Regex used: **[A-Z][a-z]* [A-Z][a-z]***

The next step was to decide what features of a given sentence might signal that the name of a CEO may be contained inside of it. After careful consideration, the following features were chosen:

- Number of characters in sentence
- Number or words in sentence
- Number of words in Potential CEO
- Number of characters of potential CEO
- Number of Capital Letters in potential CEO
- Number of Capital Letters in sentence
- Number of punctuation characters in potential CEO
- If the sentence contains a word pertaining to CEO
- If the sentence contains a word pertaining to a company

Two lists of words were created: one that indicated an entity IS a CEO name and one that indicated an entity is NOT a CEO name. These features were indicated as binary variables within the feature vector. Furthermore, an attempt was made to filter out any entities that fell into the category of NOT indicating a CEO name.

From these features we were able to create a vector for each potential CEO name. It is from these vectors that we were eventually able to predict whether or not they were actually CEOs.

Before creating a model to predict, we had to create both positive and negative examples of CEO names. For this we used the provided training data of CEOs and Companies. We created the same feature vectors as we did for the CEO Candidates, and then added an additional feature equal to 1 if it is a CEO (positive match) or 0 if it is a Company (Negative Match)

Finally we were able to incorporate our model: a Decision Tree Classifier. The model was trained using the positive and negative matches created in the previous step, and then used to predict the final attribute of the test data set. Those with a predicted value of 1 for CEO was extracted and saved to an external CVS.

The exact same steps were taken to extract the company names as the CEO names with the exception of the Regex used and the assignment of CEOs to negative matches and Companies to positive matches.

Regex used: **([A-Z][a-z]+(?=\s[A-Z])(?:\s[A-Z][a-z]+)+)**

Findings:

This model is far from accurate. Although time restrictions prevented exact accuracy and precision of the models from being calculated, from a brief overview of the extracted entities, it is easy to observe that for both companies and CEOs there are a significant number of false positives. Areas of improvement certainly exist in both the Regexes used, the methods used to filter the data, and the development of the model. With time, all of these shortcomings can certainly be fixed.