**Text Analytics of Business Insider Articles**

Charles Marshall

Prof. Klabjan

IEMS 308

8 March 2020

**Overall Preprocessing:**

The first steps I took in this assignment were to perform some overall preprocessing on the Business Insider corpus. The goal of this preprocessing was to remove fragments of the corpus which would not be useful as features in an NER classifier and not be candidates for extraction for any of the entities. With these fragments removed, it would allow further processing to be quicker and more efficient because there would be less to be considered. With these goals in mind, I decided it was best to remove all Unicode characters and the special character \*. I decided to keep other special characters such as periods, commas, dashes, colons, and percent signs in the corpus because I saw them as crucial for parts of some of our examples candidates. For instance, a company names such as Wall-Mart might have a dash in it and a percentage such as “0.5%” would have both a period and percent sign in it. Next, the corpus was tokenized by sentences. This preprocessing does not pose an inherent performance advantage but is advantageous because individual sentences are easier to examine than individual articles. Lastly, I decided to remove all stop words from the corpus because these are words which are very common throughout the Business Insider corpus yet add very little value in terms of quality features or candidates for extraction. With this in mind, it is best to remove them in order to most efficiently process the data.

**CEO Classifier:**

**Regex:**

The following regex expression was used to extract CEO candidates from the data set:

**r'[A-Z]\w+ [A-Z]\w+'**

The idea behind this expression is that the vast majority of CEOs in the corpus, as shown by the ceo.csv file, should follow the Firstname Lastname for names. In terms of regex, this means that both words should start with capital letters, shown by [A-Z] at the start of each word, and should be followed by letters, shown by \w, the metacharacter for a word character. The plus (+) after the metacharacter means that the word characters repeat. There is a space between the two expressions because all first and last names have a space between them. This regex resulted in 450,405 candidate CEOs.

**Additional Preprocessing:**

No additional preprocessing was performed for the CEO classifier with the thought in mind that the regex expression and chosen features would mainly contribute to extracting CEO names from this point forward.

**Feature Selection:**

**Company Classifier:**

**Regex:**

**Additional Preprocessing:**

No additional preprocessing was performed for the company classifier with the thought in mind that the regex expression and chosen features would mainly contribute to extracting company names from this point forward.

**Feature Selection:**

**Model Selection:**

**Model Performance:**

**Percentages Classifier:**

**Regex:**

**Additional Preprocessing:**

No additional preprocessing was performed for the percentages classifier with the thought in mind that the regex expression and chosen features would mainly contribute to extracting percentages from this point forward.

**Feature Selection:**

**Model Selection:**

**Model Performance:**