**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, 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:**

**Company Classifier:**

**Percentages Classifier:**