IEMS 308
Renee Probetts
Text Analytics -Third Homework

Process Taken:

1. Read in excel files and txt files
2. Create corpus of txt files
3. For each item (CEO/Company/Percentage), write the regular expressions for the features
4. Create data table of the training data and features
5. Train a logistic model based on this data
6. Extract potential candidates for CEO/Company/Percentage from the corpus using regular expressions
7. Create data table of these potential values and the features
8. Predict whether they are in fact that item or not using the logistic regression model

Features Used (** All Regular Expressions can be found in the source code file)

CEO

1. All Characters
2. Two Words
3. Between 8 and 25 characters long
4. Both Words Uppercase
5. CEO Nearby
6. Chief Nearby

COMPANY

1. All Capital Letters
2. Ends in "Inc."
3. Ends in "Co"
4. Ends in "Ltd"
5. Ends in "Group"
6. Uppercase Word
7. Capital Letter in the Middle of Word
8. Three Uppercase Words in a Row

PERCENTAGE

1. Digit.Digit
2. Digit.Digit %
3. Number - written out
4. Number percent - both written out
5. Digit percent

Selected Classification Model: logistic regression