Rosanne Harrison

IEMS 308 – Homework 3

**FEATURES USED**

\*Below, “Uppercase” refers to a word that starts with an uppercase letter, and “lowercase” is a word that starts with a lowercase letter

**CEOs**

The following are the four regular expressions that I used to identify possible names of CEOs

* “Uppercase Uppercase CEO”
* “CEO Uppercase Uppercase”
* “lowercase Uppercase CEO” – only Uppercase is part of name
* “CEO Uppercase lowercase” – only Uppercase is part of name

After pulling out these potential names, I transformed these variables into features that counted how many times a potential name showed up in the files. For each of the four regular expressions, I created feature variables that counted whether the potential name (with ‘CEO’ and any lowercase words removed) appeared at least 1, 2, 5, 10, 15, 20 and 25 times in the documents.

**Companies**

The following are the four regular expressions that I used to identify possible names of Companies. For finding potential companies, I identified a list of “indicator” words that were usually attached to company names.

Below, the list of indicator words includes all of the following: Corporation, Corp, Inc, Co, Ltd, LLC, Group, Financial, Management, Capital, Bank, Limited, Association, Entertainment, Software, Advisors, Holdings, Labs, Lab, Company.

* “Uppercase Uppercase … (unlimited Uppercase) … Indicator\_Word”
* “Uppercase Indicator\_Word
* “UppeRcase Indicator\_Word” – uppercase word has a random capitalized letter in the middle of it
* “CAPITALIZED Indicator\_Word” – first word is in all CAPS

For each of the four regular expressions, I created feature variables that counted whether the potential company name appeared at least 1, 2, 5, 10, 15, 20, 25, 30, 50 or 100 times in the documents.

I also created features that counted the number of words in the potential company name: whether the number of words was less than 4, 6, 8, 10 or greater than 10.

**Percentages**

The following are the ten regular expressions that I used to identify possible Percentages. [0-9]+ refers to a number with digits, [a-z]+ refers to a word that is (hopefully) a number.

* “[0-9]+%”
* “[0-9]+ %” – space between number and percent sign
* “[a-z]+ percent”
* “[a-z]+-[a-z]+ percent” – two+ words have dash in the middle (ie. ninety-nine)
* “[0.9]+.[0-9]+ percent” – number with a decimal point (and word percent)
* “[0.9]+.[0-9]+%” – number with a decimal point (and percent symbol)
* “-[0-9]+%” – negative number %
* “-[0.9]+.[0-9]+%” – negative number with a decimal %
* “[0-9]+ percent” – digit with the word percent
* “[0.9]+.[0-9]+ percent” digit with decimal and word percent

For each of the four regular expressions, I created feature variables that counted whether the potential percentage (with % or ‘percent’ attached) appeared at least 1, 5, 10, 25, 50, 100, 500 or 1000 times in the documents.

**PROCESS TAKEN**

1. Load all text files into a Corpus
2. Pre-processing:
   1. For CEOs and Companies: Removed all stop words and all punctuation
   2. For Percentages: Removed all stop words ONLY (punctuation was needed for identifying percentages using regular expressions)
3. Create dictionaries of already identified CEOs/Companies/Percentages in data frames (and remove duplicates from the data frames)
4. Create regular expressions to identify CEOs/Companies/Percentages and pull out all potential named entities. Move all potential named entities into data frames (one for CEOs, one for Companies and one for Percentages).
5. Create a column that counts the occurrences of each potential CEO/Company/Percentage. Use this column to create binary features related to number of occurrences.
6. Merge “label” column onto the data frame using the training data – this column will be what we attempt to predict in the logistic regression
7. Run a logistic regression to predict whether a potential CEO/Company/Percentage was found in the corresponding training file.
8. Identify the best cutoff value for the prediction and record the results

**SELECTED CLASSIFICATION MODEL**

The classification model I decided to use was a logistic regression, with the prediction being whether or not a potential named entity was in the training file. I ran a logistic regression using all the predictor variables in the model; only some of the variables were significant. I then used the regression results to identify the which CEO/Companies/Percentages were real.

**MODEL PERFORMANCE**

**CEOs**

For CEOs, I used a cutoff of 0.5 – if the probability of being a CEO was at least 0.5, then I marked it as a CEO. The CEO regression results are seen in the table below:

CEO.predict **No Yes**

**No** 729 146

**Yes** 140 1168

Overall, 1,168 CEOs were correctly identified. The overall accuracy of the prediction was **86.9%**.

**Companies**

For Companies, I used a cutoff of 0.25– if the probability of being a Company was at least 0.25, then I marked it as a Company. Lowering the cutoff from 0.5 to 0.25 had a minimal impact on the overall accuracy, but greatly increased the number of positive ‘Yes’s. The Company regression results are seen in the table below:

Company.predict **No Yes**

**No** 9419 894

**Yes** 337 1699

We see that 1,699 Companies were correctly identified. The overall accuracy of the prediction was **90.0%.**

**Percentages**

For Percentages, I used a cutoff of 0.3– if the probability of being a percentage was at least 0.3, then I marked it as a percentage. Again, lowering the cutoff from 0.5 to 0.3 had a minimal impact on the overall accuracy, but greatly increased the number of positive ‘Yes’s. The percentage regression results are seen in the table below:

Percent.predict **No Yes**

**No** 1554 363

**Yes** 875 3474

We see that 3,474 percentages were correctly identified. The overall accuracy of the prediction was **80.2%.**