CS 839 Project Stage 1

Team Members:
Phil Martinkus and Abhijeet Kamble

Entity We Tried to Extract:

- We have attempted to create a model that can extract People Names from political articles found on CNN.
- We only consider the actual name as an entity. For example, we do not include “Doctor,” “Mr.” or any other prefixes as part of the name. Additionally, we do not consider grammatical endings as part of a name. For example, if an article contains the phrase “President Obama’s administration” we would only extract “Obama.” The apostrophe ‘s’ (‘s) is not considered part of the name.
- We marked up the names in each document using a <n> starting tag and a </n> ending tag.

Total Number of Entities Marked Up:
In all 300 documents we found 1932 names.

Training Set (Set I):
The training set contains 200 documents and contains 1284 actual names.

Test Set (Set J):
The test set contains 100 documents and contains 648 actual names.

First Classifier Selected:
After our first end to end pipeline, we chose to debug the random forest model. After ten fold cross validation, it had the following statistics:

- Precision = 0.602190
- Recall = 0.400411
- F1 = 0.480996

While the SVM model had a higher precision, it had a much lower recall so we decided that the Random Forest would be the best model to debug.

Final Classifier Before Rule-Based Post-Processing:
After our updated end to end pipeline, we again chose the random forest model. After splitting the training data into another training set I and a test set J, we recieved the following scores for our model:

- Precision: 0.7638888888888888
- Recall: 0.6062992125984252
- F1: 0.6760316066725197

Final Classifier After Rule-Based Post-Processing:
After applying our post-processing rules to the random forest predictions, we ended with the following scores:

- Precision = 0.9362549800796812
- Recall = 0.6317204301075269
- F1 = 0.754414125200642

Finally, we were able to apply our model and post-processing rules to the test set and we got these final scores on the test set:

- Precision: 0.9078014184397163
- Recall: 0.6047244094488189
- F1: 0.7258979206049149

We just wanted to note that our post-processing rules are completely overfitting the training data. For our main rule, we pretty much just created a list of all words that were found in the false positives that would almost never be found in an actual name. During the post-processing step, if a word was marked as an actual name, but contains any word from the list, then we switch the prediction from true to false. Luckily for us, many of the articles covered similar topics so these post-processing rules were also effective for the test set. However, our model may not be as effective as time goes on and different topics are covered in the news.