Stage 1 Analysis and Findings

Team members: Niko Escanilla, Derek Hancock, Srujith Poondla
Team name: Data Vaders

I. Topic

For this project stage, we chose to perform information extraction (IE) from natural text documents from ESPN's archived NBA articles written in the first two weeks of February, 2018. During this time of the year, the NBA reaches its halfway point of the season. This means that during the time we extracted the text documents, the trade deadline (February 8, 2018) had occurred. In turn, we were curious to see what player names (and names in general) showed up. Therefore, for this stage, the entity type we chose were *person names*.

The following are some examples of this entity type:
"Lebron James"
"Adrian Wojnarowski"
"Elfrid"
"Larry Nance Jr."

II. Summary of Data

We marked up a total of 7,489 mentions of names in 300 documents. We split them into two sets. Set I had 200 documents with 4,883 positive examples out of 22,415 generated examples. Set J had 100 documents with 2,606 positive examples out of 12,020 generated examples.

III. Features and Results for Set I

We developed about 20 features to help our classifier distinguish between person names and other common capitalized words in ESPN articles. For example, some features we developed were:

- If the string was preceded or followed by a capitalized word
- The number of words in the string
- Whether the string contained a common first name (using census data)
- Whether the string was preceded or followed by a common first person name
- Whether a string contained an NBA team name or city
- Whether the string contained a common pronoun, such as days of the week or months
- Whether the string contained part of an NBA player's name
- Whether the string contained a full NBA player's name
- etc.

The varied features we developed as well as the features using dictionaries of common person names, NBA jargon, etc. helped our classifier achieve good initial performance.

After performing 10-fold cross validation (CV) on Set I with a decision tree, random forest, support vector machine, linear regression, and logistic regression, our best classifier was the random forest. The following are the cross-validated scores for the random forest:
Precision: 0.9172
Recall: 0.9121
F1 Score: 0.9145

IV. Results for Set J

Choosing the random forest as our classifier, we then trained it on Set I and tested it on Set J. The following are our results:
Precision: 0.9258
Recall: 0.9239
F1 Score: 0.9248

Because we achieved good results, we did not perform any rule-based postprocessing.

V. Additional Experiments

At the request of Professor Doan, we ran some experiments with our random forest classifier while removing some features. For this first experiment, we removed any features that relied on a dictionary of NBA players. On set J, our classifier was able to obtain the following results:
Precision: 0.8931
Recall: 0.88880
F1 Score: 0.8905

By increasing the threshold for positive prediction, we were able to keep the classifier's precision above 90% and get the following results:
Precision: 0.9038
Recall: .8292
F1 Score: .8594

We then ran the same experiment but also removed the feature relying upon a dictionary of the 30 basketball team names and US cities. Thus this experiment used no domain knowledge of the NBA in its features. It achieved the following results:
Precision: 0.8293
Recall: 0.8130
F1 Score: 0.8207