CS 839 Spring 2018, Project Stage 1

Team members:
Arpith Neelavara (neelavara@wisc.edu)
Bhargav Tangirala (btangirala@wisc.edu)
Aribhit Mishra (amishra28@wisc.edu)

Entity Type:
We have chosen Car Make as our entity type. Some of the examples are BMW, Toyota, Mercedes and KIA.

Total Markup Done:
The total number of markups are 1235.

Documents in Set I:
Total number of documents in Set I (Developement Set) are 220 and the number of mentions are 829.

Documents in Set J:
Total number of documents in Set J (Test Set) are 110 and the number of mentions are 406.

Cross Validation scores of different classifiers on Set I:
Random Forest : 87.953
SVM : 84.97
Logistic Regression : 87.975
Linear Regression : 87.66
Decision Tree : **87.976**

Classifier (M) performance on Set I:
We chose to use Decision Tree Classifier and the scores are below:
Precision: 97.32 %
Recall: 73.65 %
F1: 83.85 %

Classifier (X) performance on Set J:
We decided to use Decision Tree Classifier and the scores are below:
Precision: 92.76%
Recall: 67.93%
F1: 78.43%

We did not do any rules based post-processing as we already reached required scores.

Note :

We are considering all possible positive examples obtained from marked up words. For negative examples, we are pruning a word if it has number in it or if it does not start with a capital letter. We used a small dictionary of 20 Car Makes (positive_list in the code) as a feature, on Professor Anhai's suggestion, in addition to the other features that are based on grammar and syntax.