CS 839 Project Stage 1

Team Members:
Shuoxuan Dong sdong34@wisc.edu
Ziyun Zeng ziyun.zeng@wisc.edu
Chushi Shi cshi46@wisc.edu

Data Source:
Sports articles for objectivity analysis Data Set from UCI machine Learning repository:
http://archive.ics.uci.edu/ml/datasets/Sports+articles+for+objectivity+analysis

Examples:

- In 2012, <Name>Wilson</Name> attempted more than 25 passes in only five games.
- <Name>Marshall</Name> took advantage of his openings with a nice blend of dunks and mid-range jumpers.
- <Name>Zab Judah</Name> has a great chance to win.
- This is boxing and anything can happen," <Name>Matthysse</Name> said.

Mark up:
946 names were tagged in the 300 text file dataset.
Set I: 200 files with 693 tags for training purpose
Set J: 100 files with 253 tags for testing purpose

Feature Generation:

The features are:
[1] 0/1 boolean - whether contain a name suffix (e.g. Jr, II, III)
[2] integer - number of words in a sentence substring
[3] 0/1 boolean - whether capitalized
[4] integer - index
[5] integer - length of the sentence
[6] 0/1 boolean - whether the word before this substring is 'the'

Structured data with feature vectors
Examples:
(substring, containNameSuffix, numWord, isCapitalized, index, sentenceLength, prevWord, isName)
'Big', 1, 0, 1, 1, 10, 0, 0
'East', 1, 0, 1, 2, 10, 0, 0
'the same', 2, 0, 9, 20, 0, 0

'same stage', 2, 0, 10, 20, 1, 0
'Naomi Lang', 2, 0, 1, 0, 6, 0, 1
'Peter Tchernyshev', 2, 0, 1, 3, 6, 0, 1
'Tanith Belbin', 2, 0, 1, 1, 11, 0, 1

Cross Validation of Four Classifiers:

- Decision Tree
- SVM
- Logistic Regression
- Random Forest
- Linear Regression

model\performance	Recall	Precision	F1 score
Decision Tree	90.187590 %	88.152327 %	89.158345 %
SVM	92.063492 %	87.277702 %	89.606742 %
Logistic Regression	99.567100 %	88.348271 %	93.622795 %
Random Forest	99.567100 %	88.348271 %	93.622795 %
Linear Regression	88.144187 %	88.196408 %	88.170290 %

Best Classifier on training data:
Logistic Regression has the best cross-validation precision equal to 89.124294 %
Logistic Regression train recall = 99.567100 %
Logistic Regression precision = 88.348271 %
Logistic Regression F1-score = 93.622795 %

Apply the trained Logistic Regression model to test set, we can get
test recall = 98.814229 %
test precision = 95.785441 %
test F1-score = 97.276265 %