CS 839 - Data Science
Project Stage 1

Team Members: Abhinav Garg (garg26@wisc.edu)
Bidyut Hota (hota@wisc.edu)
Chirayu Garg (cgarg2@wisc.edu)

Entity Type: We have taken BBC sports articles from UC Davis dataset (http://mlg.ucd.ie/datasets/bbc.html) and **marked name of persons in the articles**. 300 documents were selected out of more than 700 available. Entity are marked using a pair of square opening and closing brackets. Example:
[[Roger Federer]]
[[Serena Williams]]

Entity Count and Distribution:

Total 300 documents marked containing **3936** entities. The documents can be looked up at https://github.com/abhinav4192/Data-Science/tree/master/stage1/documentPool

- **Set-I**: 200 documents containing **2719** marked entities available at https://github.com/abhinav4192/Data-Science/tree/master/stage1/train

- **Set-J:** 100 documents containing **1217** marked entities available at https://github.com/abhinav4192/Data-Science/tree/master/stage1/test

Classifier Selection:

- **M-Classifier:** Random Forest classifier was selected after first round of cross validation on Set-I with following metrics:

Metric	**Score**
Precision	0.64
Recall	0.47
F1	0.54

- **X-Classifier:** Final selected classifier after debug was Random Forest with enhanced number of estimators and tree depth with following metrics:

Metric	**Score**

Precision	0.92
Recall	0.65
F1	0.76

Prediction on Set-J: Selected Random Forest classifier gave following results on set-J:

Metric	**Score**
Precision	0.91
Recall	0.62
F1	0.74

More Information:

- We have not done any rule-based post processing.
- After selecting Random Forest classifier as M-Classifier after first round of validation, we debugged the false positive/negative examples as follows :
    - We create a few new features to provide more information to the classifier
    - We found that country names and football teams were being falsely classified as positive as they have same semantics as names. As the country names and english football teams list are easily available and is a limited set, we added them to pruning list. This helped in improving the precision
    - We also tried different combination of hyper-parameters for all the algorithms to improve the precision/recall