CS 839 – Data Science Project – Stage 1

Team Members:

- Madan Raj Hari – mhari2@wisc.edu;
- Raghavan Vellore Muneeswaran – velloremunee@wisc.edu
- Shadana Subramanian – ssubramani23@wisc.edu

Data Set:

BBC News Articles - http://mlg.ucd.ie/datasets/bbc.html
The dataset includes documents from BBC sport website corresponding to sports news articles in five tropical areas from 2004 – 2005. The documents labelled under cricket, football, rugby and tennis is considered for this project.

Platform:

The processing was done in python where scikit-learn was used for machine learning model and pandas for data processing.

Labels and Formats:

The file are in '.txt' format. The occurrences of names are tagged by using the <name> and </name> tags. Names of <FirstName> or <FirstName LastName> formats are tagged. A few examples from the tagged documents are:

- <name>Stephen Fleming</name> chose to put Australia ..
- <name>Boje</name> finishes off with 1-34.
- <name>Ferrero</name> insists he is feeling positive after chicken pox

The detailed rules of the entity types are:

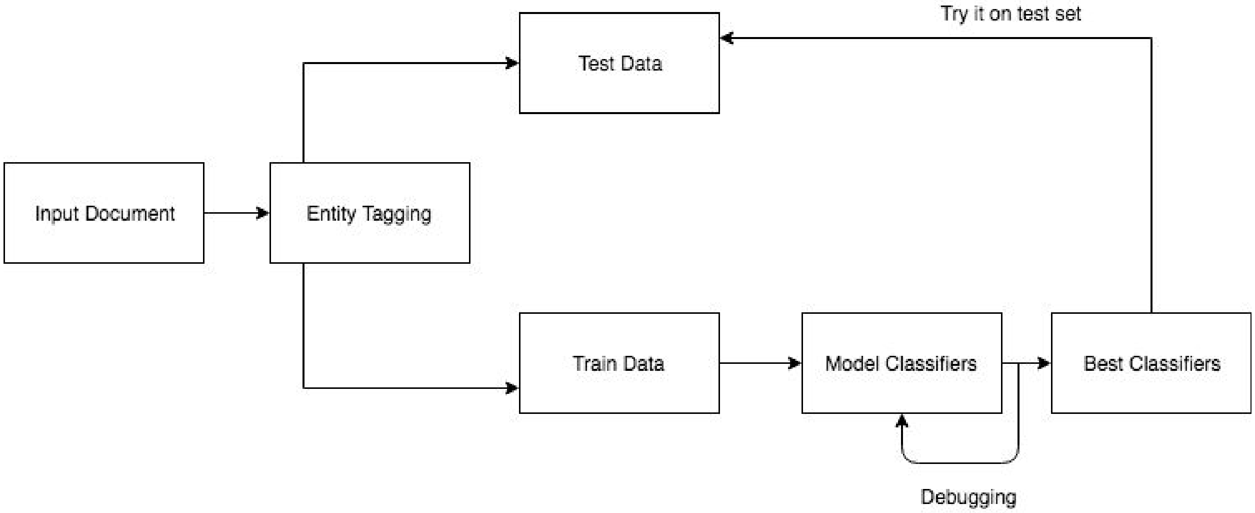
- Prefix titles like Mr., Mrs., Sir are not included
- Names with suffix – 's are included

Documents:

Data	**Number of documents**	**Number of tags**
Test Data	100	1719
Train Data	201	3215
Total	301	4934

Process Flow:

The documents are subjected to the following phases.



Features Used:

The following are the features used by the model to identify the names:

- Is the token's first letter capitalized?
- Is the token preceded by any of the keyword?
- Is the token a noun?
- Is the token succeeded by any of the keyword?

- Is the token succeeded by -ed verbs? (example: <name>Yousuf Youhana</name> batt**ed**,<name>John Howard</name> toss**ed** a coin to start the match
- Is the token stopword?
- Tf for previous word - the term frequency of a word – number of occurrences of the word in a document / total number of words in the document, the inverse document frequency – number of occurrences of a word across all the documents / number of documents
- Tf for successor word
- Parts of speech tagging for previous word
- Parts of speech tagging for next word

Classifier - Metrics:

The Precision, Recall and F1 Scores of various classifiers on the training data set is shown below:

Classifier	**Precision**	**Recall**	**F1 Score**
Decision Tree Classifier	0.82868461154	0.8343207256	0.8242609508
Linear SVM	0.82535437539	0.863350298184	0.836776303997
Linear Regression	0.70775199923	0.8456479098655	0.748415113551
Logistic Regression	0.95504690199	0.8034367269545	0.855933075669
Random Forest	0.91732315791	0.8320554177432	0.864901194736

Based on the data above we consider the – Logistic Regression classifier to be the best.

Final Metrics:
On Test Data:

Classifier	**Precision**	**Recall**	**F1 Score**
Logistic Regression	0.94338384624	0.7572193008937	0.819932130194