

## **CS 839 - Data Science Project - Stage 1**

### **Team Members**

- Aravind Soundararajan (soundararaj2@wisc.edu)
- Krishnan Rajagopalan (krajagopalan@wisc.edu)
- Palaniappan Nagarajan (pnagarajan3@wisc.edu)

### **Recognized Entity** - Location

**Data Set** - BBC News Articles (<http://mlg.ucd.ie/datasets/bbc.html>)

### **Examples**

- ...worked with <LOCATION>Europe</LOCATION>.....
- ...But <LOCATION>India</LOCATION> now faces...
- ...trading in <LOCATION>New York</LOCATION> on Friday...
- ...in <LOCATION>South Korea</LOCATION> use...
- ...<LOCATION>US</LOCATION> currency...

**Total Number of mentions made** - 1136

Set	Set I (Training Set)	Set J (Test Set)
Number of documents marked up	219	100
Number of mentions made	1326	836

### **Features used**

The following features have been used for modelling the learner.

- Is the token titleized?
- Is the token preceded by keywords “at”, “in” etc?
- Is the token succeeded by keywords “based”, “Kingdom”, “States”?
- Is the part of speech of the token ‘Noun’?
- Is the token suffixed by -berg, -shire,-cester, etc?
- Is the token preceded by words indicating direction (North, West, East, South)
- Position of the token in the sentence
- Part of Speech of preceding two tokens
- Part of Speech of succeeding two tokens
- Is the token a stopword?

## **Metrics**

After cross validation on Set I,

	<b>Precision</b>	<b>Recall</b>	<b>F1 - Score</b>
<b>Random Forest</b>	<b>0.87</b>	<b>0.63</b>	<b>0.73</b>
Decision Trees	0.71	0.69	0.69
Logistic Regression	0.74	0.45	0.56
Support Vector Machines	0.88	0.24	0.37

Before rule-based postprocessing on Set J,

	<b>Precision</b>	<b>Recall</b>	<b>F1 - Score</b>
<b>Random Forest</b>	<b>0.91</b>	<b>0.65</b>	<b>0.76</b>
Decision Trees	0.80	0.70	0.75
Logistic Regression	0.82	0.47	0.60
Support Vector Machines	0.92	0.28	0.43

**Postprocessing rules** - No post processing rules were used.

After rule-based postprocessing on Set J (same as the previous tabulation)

	<b>Precision</b>	<b>Recall</b>	<b>F1 - Score</b>
<b>Random Forest</b>	<b>0.91</b>	<b>0.65</b>	<b>0.76</b>
Decision Trees	0.80	0.70	0.75
Logistic Regression	0.82	0.47	0.60
Support Vector Machines	0.92	0.28	0.43