# CS 839: Data Science

Project Stage 1 – Report

## Team Members


- Arun Jose (*jose4@wisc.edu*)
- Kundan Kumar (*kkumar36@wisc.edu*)
- KN Sushma (*kudlurnirvan@wisc.edu*)


## Entity Type

For the Entity type, we have chosen to extract **locations** from the set of documents.
We have included continents, countries, states and cities in our model. We have NOT included location-based terms which are used out of context (for example: the American government) as well as other locations such as lakes/oceans/parks/museums/Universities.

## Summary of the steps followed

- In this project stage we will perform information extraction (IE) from natural text documents, using a supervised learning approach. Here are the steps that we followed:
- Collected 300 text documents from which we will extract mentions of LOCATION entity type. These documents contain well-formed sentences (such as those in news articles).
- We went through documents and marked up the mentions of location entity type. After the mark up of these locations, we had around 1100 mentions (distributed among 300 documents).
- We split these 300 documents into a set I of 200 documents and a set J of the remaining 100 documents. The set I will be used for development (dev set), and the set J will be used for reporting the accuracy of our extractor (the test set). The goal here is to develop an extractor that achieves a precision of at least 90% and as high recall as possible, but at least 60% in recall.
- Next we performed Cross Validation (CV) on the set I to select the best classifier. We considered the following classifiers: decision tree, random forest, support vector machine, linear regression, and logistic regression. We used the scikit-learn package for this CV purpose.

- After this step, Linear Regression model performed better when compared to other classifiers. We got an accuracy of 88% using Linear Regression. This is the classifier M.
- Then we debugged M using the same set I. For this, we further split up set I into two sets P and Q, trained M on P, applied M to label examples in Q, then identified and debugged the false positive/negative examples.
- Once this debugging was done, we repeated CV to see if another classifier provided a better accuracy now.
- Here we observed that the Random Forest classifier performed better which provided 92% precision and 77% recall. As we were able to achieve required precision of 90% and recall of 60%, we did not perform any Rule-based post-processing steps. This is the model X(and Y).
- Then we applied Y to the set-aside test set J and found that model M reported 91% precision and 80% recall which represents the final accuracy measure of our model.

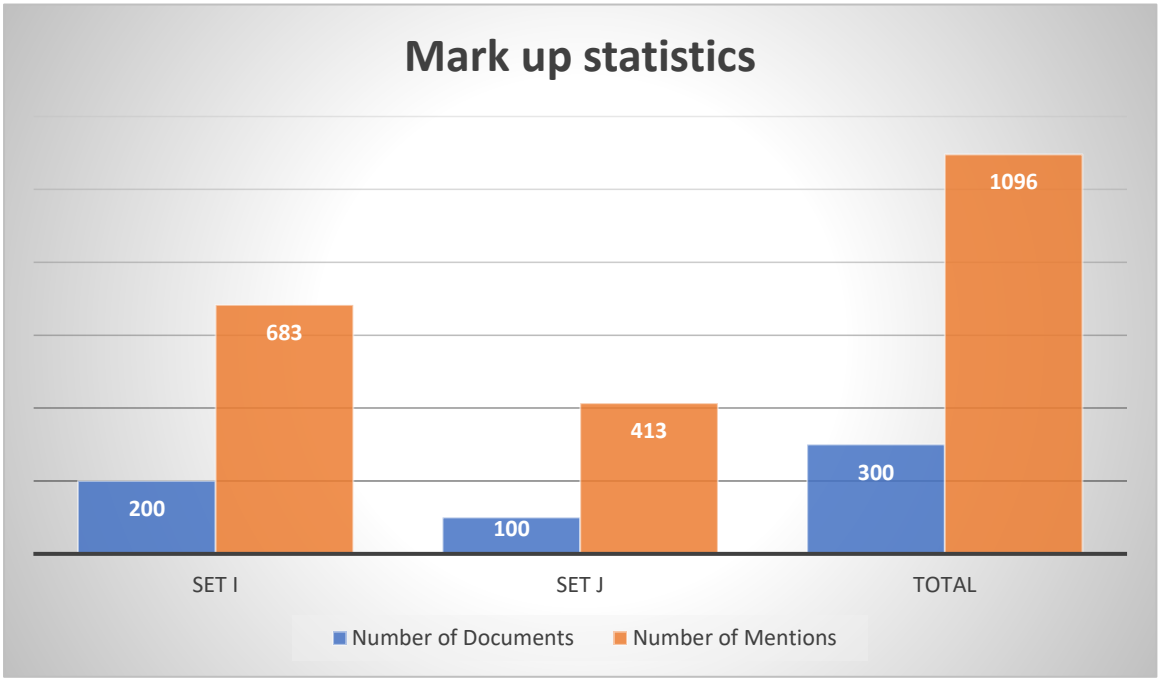
These steps have been elaborated in some detail below.

## Mentions Marked Up

We collected around 340 documents from the following online resources:

Kaggle (https://www.kaggle.com/datasets) and http://www.textfiles.com/news/ which were primarily news articles and hence consisted of well-formed sentences with locations that we could extract.

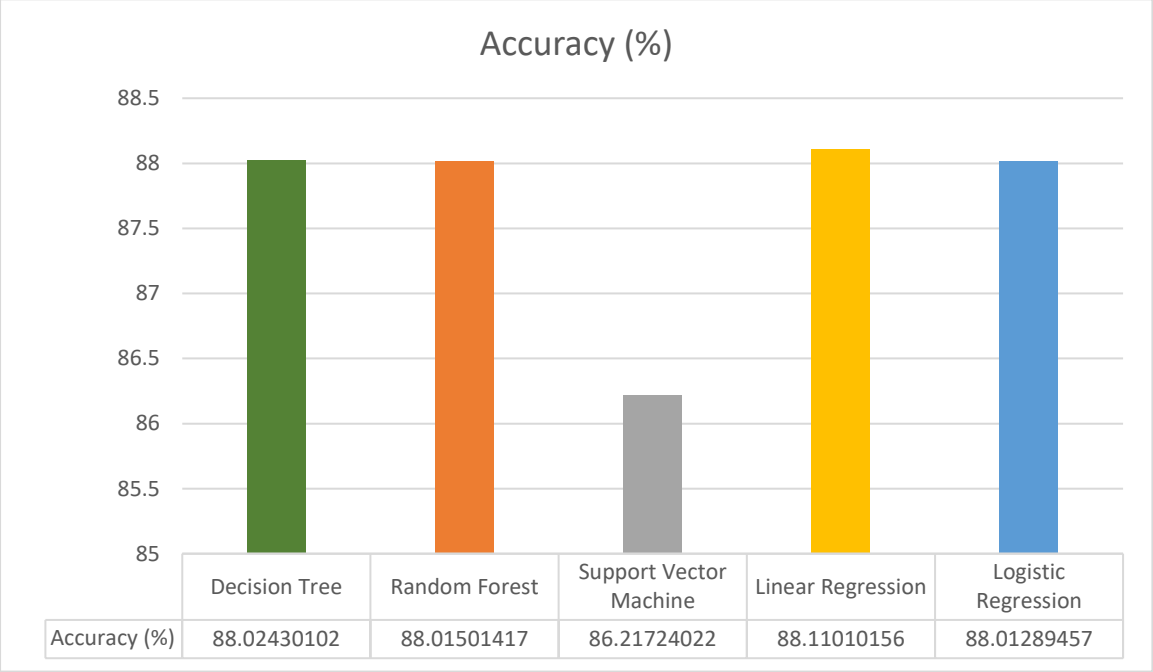
After labelling this dataset, we discarded the scarcely labelled and skewed documents to ensure a good uniform distribution of labelled words in the 300 documents.



## Classifier after cross validation (M)

After extracting the features from our corpus of text, a 10-fold cross validation was performed using the following five models to compute accuracy.

1. Decision Tree accuracy:  88.02430101841882
2. Random Forest accuracy:  88.0150141725025
3. Support Vector Machine accuracy:  86.21724022322228
4. **Linear Regression accuracy:  88.11010156330749**
5. Logistic Regression accuracy:  88.01289457243401




The highest accuracy model was observed to be Linear Regression.

**M: LINEAR REGRESSION**

This model M was run on Set I to obtain the values for Precision, Recall and F1 score as follows:

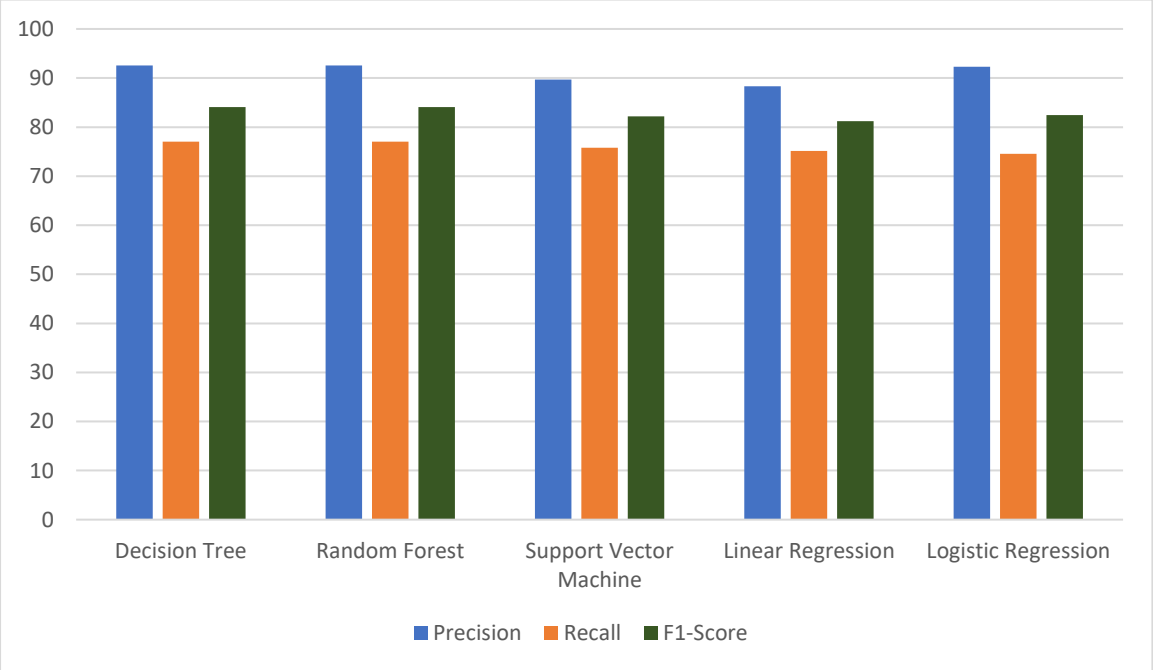
| | |
|---|---|
| **Precision** | 88.32116788321168 |
| **Recall** | 75.15527950310559 |
| **F1 score** | 81.20805369127517 |

## Classifier before the post-processing step (X)

The obtained classifier M(Linear Regression) is now debugged on the set I by splitting up I into two parts P, with 150 documents and Q, with 50 documents.

After debugging the Set I to improve the Precision and Accuracy of the model, Cross Validation was redone to check if any other classifier performs better.

During this check, we observed that the **Random Forest** model outperforms the other models.




Random Forest and Decision Tree had similar scores with respect to Precision and Recall:

| | |
|---|---|
| **Precision** | 92.53731343283582 |
| **Recall** | 77.01863354037267 |
| **F1 score** | 84.06779661016948 |

However the overall accuracy was observed to be higher for Random Forest:

1. Decision Tree accuracy: 87.46103019550804
**2. Random Forest accuracy: 87.78800077423377**
3. Support Vector Machine accuracy: 85.69075912702504
4. Linear Regression accuracy: 87.35815966227243
5. Logistic Regression accuracy: 87.56967395323034

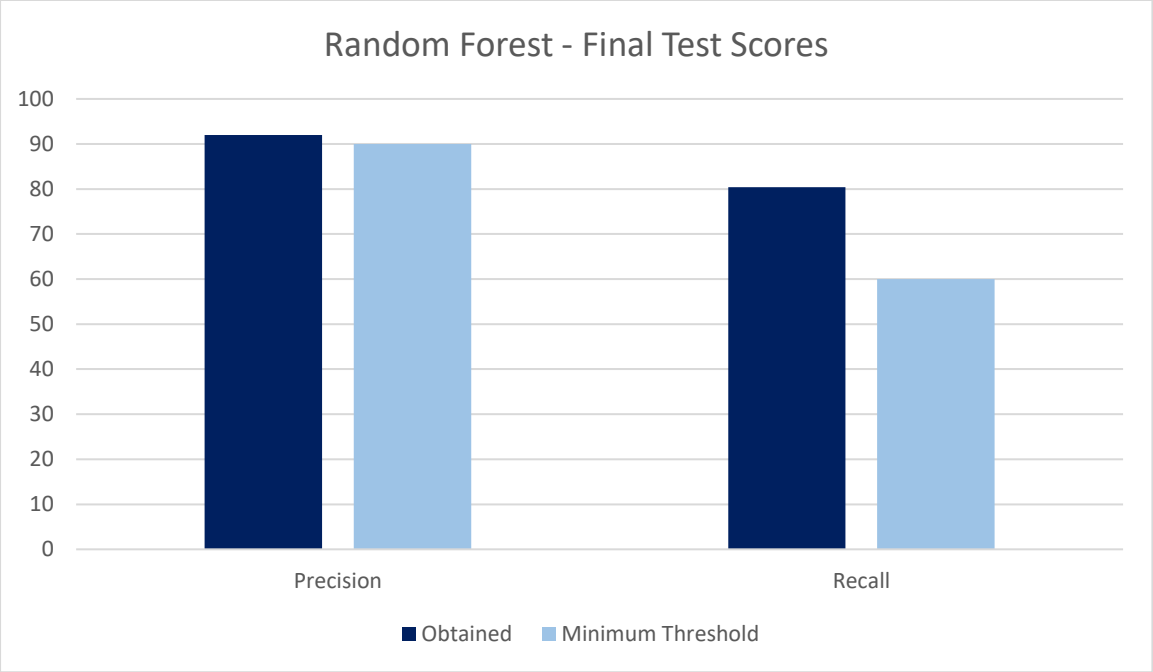
As a result, we chose Random Forest as our classifier X.

**X: RANDOM FOREST**

## Rule-Based Post Processing Steps

We ran X (the Random Forest) on the final untouched test Set J to check if the model met the required thresholds for Precision and Recall. The following results were obtained:

| | |
|---|---|
| **Precision** | 91.96675900277008 |
| **Recall** | 80.38740920096852 |
| **F1 score** | 85.78811369509043 |
| **Accuracy** | 87.78800077423377 |




As can be observed above, since this already meets and exceeds the given threshold values of 90% for Precision and 60% for Recall, **NO rule-based post processing** steps have been added.

Therefore, the final model Y is the same as X.

**Y: RANDOM FOREST**

## Conclusion

We have successfully performed Information Extraction to extract location information from a given set of natural text documents. Our model delivers an acceptable 91.97% precision and 80.39% recall.