Project Stage 1
CS 839, Spring 2019

Team members:
Akshaya Kalyanaraman
Elena Milkai

Entity type:
Person Names like: “Norman Mineta”, “Michael Boyd” and “Jamelia Davis”

Dataset: http://mlg.ucd.ie/datasets/bbc.html

Total number of mentions:
2445

Number of files in Set I and mentions:
200 files
1752 mentions

Number of files in Set J and mentions:
100 files
693 mentions

Cross Validation Results:

Classifier	Precision	Recall	F1
Decision Tree	0.7027	0.6562	0.6786
Random Forest	0.7223	0.6777	0.6996
SVM	0.6911	0.5801	0.6375
Linear Regression	0.4700	0.3822	0.42
Logistic Regression	0.5426	0.4245	0.4763

Rule-Based post-processing:
We didn’t do any Rule-Based post-processing

Final Results:
Random Forest gives us the best results:

Precision	Recall	F1
0.7223	0.6777	0.6996

Possible reasons that we have not reached the required precision and recall:

We have restricted the names that we extract very much. We only return full names with name and surname. Most of the mentions in the data-files mention people by their surname without using their “small name”. However, with the features that we are using we cannot separate clearly the entities with full names and those without full names. More specific, the features that we use:

- if it is 1-gram, 2-gram etc. or not
- if it is followed by ‘s or ‘
- if it is a noun
- if it is a verb
- if the previous token is a prefix (ex: Mr, Mrs,... etc)
- if it contains prefixes
- if it contains words related to days, dates etc.
- if it contain words related to countries, cities.. etc
- if is the first word after full-stop
- if it contains stop-words

All these features cannot clearly separate the one word names from full names and they are not even enough to extract the full names. Also, the space of negatives examples was very big in contrast to the positive examples, although we keep only the words (1-grams and 2 grams) with their first letter to be capital (which we also select not to be stop-words or prefixes).