Project Stage 1
CS 839, Spring 2019

Team members:
Akshaya Kalyanaraman
Elena Milkai

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

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

Total number of mentions:
3927

Number of files in Set I and mentions:
200 files
2676 mentions

Number of files in Set J and mentions:
100 files
1251 mentions

Cross Validation Results (M classifier trained with I and tested with J):

Classifier	Precision	Recall	F1
Decision Tree	1	0.7984	0.8879
Random Forest	1	0.7984	0.8879
SVM	1	0.7984	0.8879
Linear Regression	1	0.2318	0.3764
Logistic Regression	1	0.7984	0.8879

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

Final Results:
All classifiers except for linear regression have the same results:

Precision	Recall	F1
1	0.7984	0.8879

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

Also, the space of negatives examples was very big in contrast to the positive examples, so we reduced it by keeping 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).