Project Stage 1
CS 839, Spring 2019

Team members:
Akshaya Kalyanaraman
Elena Milkai

Entity type:
Person Names like: *"Norman Mineta"* --full name , *"Boyd"*--surname and *"Jamelia"* --name.
More specifically, we consider as person names: full names, surnames or even only the name without the surname.

Dataset: http://mlg.ucd.ie/datasets/bbc.html
It includes articles from 5 different categories: News, Sports, Entertainment, Politics, Business and Tech.

Total Number of Files we extracted:
300 files

Total number of mentions:
3891 in 300 files

Number of files in Set I and mentions:
200 files
2652 mentions

Number of files in Set J and mentions:
100 files
1239 mentions

Cross Validation Results:

```
-----------------SVM results----------------
True Positives: 757
True Negatives: 2443
False Positives: 232
False Negatives: 284
Total number of data: 3716
Accuracy: 0.861141011841
Precision: 0.765419615774
Recall: 0.727185398655
F1: 0.745812807882
-----------------RF results----------------
True Positives: 570
True Negatives: 2568
False Positives: 107
False Negatives: 471
Total number of data: 3716
Accuracy: 0.844456404736
Precision: 0.841949778434
Recall: 0.547550432277
F1: 0.663562281723
-----------------DT results----------------
True Positives: 820
True Negatives: 2478
False Positives: 197
False Negatives: 221
Total number of data: 3716
Accuracy: 0.887513455328
Precision: 0.806293018682
Recall: 0.787704130644
F1: 0.796890184645
-----------------LogR results----------------
True Positives: 743
True Negatives: 2465
False Positives: 210
False Negatives: 298
Total number of data: 3716
Accuracy: 0.86329386437
Precision: 0.779643231899
Recall: 0.713736791547
F1: 0.745235707121
-----------------LinearR results----------------
True Positives: 1031
True Negatives: 596
False Positives: 2079
False Negatives: 10
Total number of data: 3716
Accuracy: 0.437836383208
Precision: 0.331511254019
Recall: 0.990393852065
F1: 0.496747771621


Best classifier M --> Decision Tree
F1: 0.796890184645
Precision: 0.806293018682
Recall: 0.787704130644
```

More organised:

Classifier	Precision	Recall	F1
Decision Tree(DT)	**0.8062**	**0.7877**	**0.7968**
Random Forest(RF)	0.8419	0.5475	0.6635
SVM	0.7654	0.7271	0.7458
Linear Regression	0.3315	0.9903	0.4967
Logistic Regression	0.7796	0.7137	0.7452

The best classifier (M) after the cross validation is the Decision Tree:

Classifier M	Precision	Recall	F1
Decision Tree	**0.8062**	**0.7877**	**0.7968**

Debugging Results:

Before debugging:

```
-----------------DT results----------------
True Positives: 452
True Negatives: 1746
False Positives: 99
False Negatives: 342
Total number of data: 2639
Accuracy: 0.832891246684
Precision: 0.820326678766
Recall: 0.569269521411
F1: 0.672118959108
```

After debugging:

```
-----------------DT results----------------
True Positives: 601
True Negatives: 1732
False Positives: 113
False Negatives: 193
Total number of data: 2639
Accuracy: 0.884046987495
Precision: 0.841736694678
Recall: 0.756926952141
F1: 0.797082228117
```

More organised:

Debugging	Precision	Recall	F1
Before	0.8203	0.5692	0.6721
After	0.8417	0.7569	0.7970

Final classifier X (after debugging):

Precision	Recall	F1
0.8417	0.7569	0.7970

Note: After each debugging stage the Decision Tree classifier remained the best classifier in the Cross Validation process.

Rule-Based post-processing:

In order to create our post-processing rules we used the False Positives examples. We created a list with 56 words which can be found in directory “/code” in file *“rules.csv”*. These words are common words that were falsely classified as names (ex: *Bridge, Indoor, Professional,*
White, Five, Six etc.)

Final Results:

Decision Tree results after rule-based post-processing:

```
-----------------DT results---------------
True Positives: 927
True Negatives: 2853
False Positives: 157
False Negatives: 312
Total number of data: 4249
Accuracy: 0.889621087315
Precision: 0.855166051661
Recall: 0.74818401937
F1: 0.798105897546
```

More organised:

Precision	Recall	F1
0.7981	0.7481	0.8551

Why we didn't reach the required Precision?
Almost all our False Positives are person names that were mistakenly **NOT** tagged from us, during the labeling process, as person names. However, the classifier predicts them correctly as person names but we cannot accept them as True Positives since they are not tagged with the <person>...</> tag during the labeling process. This means that we need to be very careful during the labeling process.

Features we used:

Our candidates are 1-grams, 2-grams and 3-grams with capitals in the beginning which are also not stop-words.
The features we extracted are:

1. Is candidate a 1-gram? - (0 or 1)
2. Is candidate a 2-gram? - (0 or 1)
3. Is candidate a 3-gram? - (0 or 1)
4. Is candidate a Noun? - (0 or 1)
5. Is candidate a Verb? - (0 or 1)
6. Is candidate an Adjective? - (0 or 1)
7. Is candidate the first word of a sentence? - (0 or 1)
8. Is candidate followed by "'s"? - (0 or 1)
9. Is candidate followed by a verb? - (0 or 1)
10. Is there a prefix (ex: Mr, Mrs etc.) before the candidate? - (0 or 1)
11. Does candidate include countries or cities? - (0 or 1)
12. Does candidate include dates? - (0 or 1)
13. Does candidate include words associated with sports (ex: Dynamo Kiev, Manchester United etc..)? - (0 or 1)
14. Number of capitals in candidate - Numeric
15. Number of vowels in candidate - Numeric
16. Number of consonants in candidate - Numeric
17. Has the candidate symbols? (ex: -, $, ! etc..) - (0 or 1)
18. Has the candidate numbers?- (0 or 1)