Project Stage 1
CS839 Data Science

Team :

1. Shruti Nambiar
2. Ushmal Ramesh
3. Anuja Golechha

Entity

Our choice of entity is person names.
We have included prefixes like Mr, Dr, Sir etc and suffixes like Jr, Sr etc to be part of the name but not position names like President, Prime Minister etc.
We have considered full names only as the entities and not partial names. So "Charles Kennedy" is considered an entity but at the same location, Charles is not considered an entity, neither is Kennedy.

Dataset

We have chosen an open BBC dataset, with news articles about political leaders.

Total number of mentions -

- The number of documents in training set : 200
- The number of mentions in training set : 2008
- The number of documents in test set : 109
- The number of mentions in test set : 1146

Steps followed

1. *Collection of ground truth*
   Since our task is a supervised learning task, for our first step we markup the text file to obtain labelled data that our classifier will use as ground truth. Every person name we come across is enclosed within <n> and </n> tags. This is inclusive of prefixes like Mr, Ms, St, Sir and suffixes like Jr, Sr etc.

2. *Splitting the raw data*
   We write a program that randomly splits the marked up files into test and training sets in the ratio 1:2.

3. *Tokenizing*
   We take in the data files as input and tokenize the string sequences into unigrams, bigrams and trigrams. We use a sliding window to get bigrams and trigrams to ensure every sequential combination.

4. *Labelling*
   Every n-gram if fully contained between the tags is named positive and the rest are named negative. Partial names are also named negative.

5. *Preprocessing*
   In the preprocessing step, we filter out the n-grams that have stopwords or other parts of speech that are unlikely to occur in names.

6. *Classifying*
   Our choice of classifier is a Decision Tree Classifier. This was selected after running cross validation on multiple ML algorithms.

7. *Post-processing rules*
   We have added 6 famous politician names - Tony Blair, Alan Milburn, Charles Kennedy, David Blunkett, George Bush, Gordon Brown, Michael Howard - to a whitelist as a post processing rule.

Results

1. Results for Decision Tree Classifier on **training** data before post processing
   Precision : 0.89
   Recall : 0.57
   F1 : 0.69

2. Results for Decision Tree Classifier on **training** data after post processing
   Precision : 0.90
   Recall : 0.67
   F1 : 0.76

3. Results for Decision Tree Classifier on **test** data before post processing
   Precision : 0.93
   Recall : 0.54
   F1 : 0.68

4. Results for Decision Tree Classifier on **test** data before post processing
   Precision : 0.94
   Recall : 0.64
   F1 : 0.76