CS839- Data Science Project- Stage 1

Team Members:

- Sri Harshal Parimi (sparimi@wisc.edu)
- Shebin Roy Yesudhas (royyesudhas@wisc.edu)
- Sankarshan Umesh Bhat (sbhat6@wisc.edu)

Recognized Entity - Person name
Dataset - IMDB movie review dataset
(www.cs.cornell.edu/people/pabo/movie-review-data/scale_whole_review.tar.gz)

Examples:

- In the English language version, **<person>**Dudley Moore**</person>** is the narrator….
- **<person>**Bruce Willis**</person>**, one of our most uneven actors……
- As **<person>**Andy's**</person>** electrician (**<person>**Willem Dafoe**</person>**) explains to **<person>**Basquiat**</person>**....
- As the show opens, 22 year old **<person>**Basquiat**</person>** is….
- The story is told in flashback as **<person>**Amy**</person>**, a part Indian, part English writer...

Number of documents and mentions:

	Set-I (Training set)	Set-J (Testing set)
Number of documents marked up	217	105
Number of mentions made	4855	2434

Features used:

The following features have been used for training the model:

- The number of words in the example
- Are all words starting with capital letters?
- Number of words starting with capital letters
- Is the example surrounded by parantheses?(Our dataset had lots of occurrences of the movie cast appearing within parantheses)
- POS tags of the n-gram example
- POS tag of the previous word in the document
- POS tag of the next word in the document
- Are all letters capitalized in the word? (The movie title was capitalized in most of the documents in the dataset)
- Is the next/previous word starting with capital letter?
- Does the example contain stray characters?

- Does the example start with a relation? (Ex- Father, Brother, Mother etc..)
- Features to check if the example contains an article, preposition, adverb, pronoun, adjective etc.
- Does the example contain a language or reference to a particular race like Hispanic, English, Jew etc…?
- Does the example contain stop-words?

Metrics achieved:

1) Cross-validation on the training data(set-I) for the initial model M:

ML model	**Precision**	**Recall**	**F1-score**
Decision Tree	0.83	0.76	0.80
SVM	0.78	0.75	0.76
Logistic Regression	0.82	0.69	0.73
Random forest	**0.84**	**0.76**	**0.80**

Out of all the classifiers, Random Forest Classifier performs best with the recall exceeding the required 60%. However the precision falls a bit below the required limit of 90%.

2) Debugging of model M to improve accuracy on training data(set-I)

1) Iteration 1:

The goal was to reduce false-positives so that precision can be improved. Though we used POS tags as features for our n-gram candidate examples, we noticed the inclusion of certain words which were not nouns in the prediction stage. This could be attributed to the high volume of such words in our text dataset and the presence of capitalized movie names containing verbs, adjective and adverbs. So we included 5 additional features like **"contains_articles", "contains_pronouns", "contains_prepositions", "contains_adjectives", "contains_adverbs"** that checks for the presence of the respective POS tags so as to provide more negative weightage to the false positive examples. The following were the results when we included these features to our best performing model (Random forest):

Random Forest Classifier

Max Precision: 0.8553921568627451
Max Recall: 0.7689075630252101
Max fscore: 0.8052805280528054

2) Iteration 2:

There was a slight improvement in the precision, however we noticed that we were still getting false positives and there were a majority of location words which being nouns(infact, proper nouns) were predicted as positive by the model. To eliminate these false positives arising due to location words like California, Silicon Valley etc, we compiled a stop-word list from the dataset ranking them based on their frequency and we included **a feature to check if a word being a noun actually belonged to the stop-word list**. Also we had a large occurrence of words related to race, language, nationality etc. which we included as a separate feature. The addition of these new features led to a marked increase in the precision of our model M (**Random forest classfier**) on the training dataset I. We fix this model X for our final step of prediction the test dataset.

Random Forest Classifier

Max Precision: 0.9104859335038363
Max Recall: 0.7747368421052632
Max fscore: 0.8212226066897347

3) Testing the final accuracy of the obtained model X on test dataset(set-J).

Random forest on Test Set

Precision on Test Set: 0.9059621909840039
Recall on Test Set: 0.7647299509001637
FScore on Test Set: 0.8293765254049257