**NLP Assignment 3 Report**

1. Develop sentiment analysis based on maximum entropy classifier on movie reviews data

The Max Entropy classifier is a probabilistic classifier which belongs to the class of exponential models. The Max Entropy does not assume that the features are conditionally independent of each other. The MaxEnt is based on the [Principle of Maximum Entropy](http://en.wikipedia.org/wiki/Principle_of_maximum_entropy) and from all the models that fit our training data, selects the one which has the largest entropy. The Max Entropy classifier can be used to solve a large variety of text classification problems such as language detection, topic classification, sentiment analysis and more.

**Experiment Setup:**

Training Data Size: 75% of the whole corpus

Test Data size: 25% of the whole corpus

Proportion: Equal proportion of Positive and Negative reviews

**Experiment Result:**

Using original corpus:

train on 1500 instances, test on 500 instances

==> Training (5 iterations)

Iteration    Log Likelihood    Accuracy

---------------------------------------

1          -0.69315        0.500

2          -0.69252        0.953

3          -0.69190        0.953

4          -0.69128        0.954

Final    -0.69066        0.954

Total accuracy: 0.724

1. **Lemmatization**

**Experiment Setup**:

Training Data Size: 75% of the whole corpus

Test Data size: 25% of the whole corpus

Proportion: Equal proportion of Positive and Negative words

Replacing all the words with their lemmatized words

**Experiment Result:**

Using Lemmatization:

train on 1800 instances, test on 200 instances

==> Training (5 iterations)

Iteration    Log Likelihood    Accuracy

---------------------------------------

1          -0.69315        0.500

2          -0.69258        0.934

3          -0.69200        0.934

4          -0.69143        0.934

Final          -0.69086        0.933

Total accuracy: 0.705

1. **Removing Punctuations**

**Experiment Setup**:

Training Data Size: 75% of the whole corpus

Test Data size: 25% of the whole corpus

Proportion: Equal proportion of Positive and Negative reviews

Removing all the punctuations from the corpus

**Experiment Result:**

Removing punctuations marks:

Train on 1500 instances, Test on 500 instances

==> Training (5 iterations)

Iteration    Log Likelihood    Accuracy

---------------------------------------

1          -0.69315        0.500

2          -0.69253        0.951

3          -0.69190        0.952

4          -0.69128        0.952

Final          -0.69067        0.953

Total accuracy: 0.718

1. **Removing Stop Words**

**Experiment Setup**:

Training Data Size: 75% of the whole corpus

Test Data size: 25% of the whole corpus

Proportion: Equal proportion of Positive and Negative reviews

Removing all the stop words from the corpus by using nltk.corpus.stopwords.

**Experiment Result:**

Removing stop words

train on 1500 instances, test on 500 instances

==> Training (5 iterations)

Iteration    Log Likelihood    Accuracy

---------------------------------------

1          -0.69315        0.500

2          -0.69253        0.952

3          -0.69191        0.952

4          -0.69129        0.952

Final          -0.69067        0.953

Total accuracy: 0.724

1. **Experiment with Unbalanced Collection**

**Experiment Setup**:

Training Data Size: 50% of the whole corpus

Test Data size: 50% of the whole corpus

Proportion:

For Training:

Positive words: 75%

Negative words: 25 %

For Testing:

Positive words: 25%

Negative words: 75 %

**Experiment Result:**

Using original corpus (Unbalanced Collection):

train on 1000 instances, test on 1000 instances

==> Training (5 iterations)

Iteration    Log Likelihood    Accuracy

---------------------------------------

1          -0.69315        0.250

2          -0.68960        0.750

3          -0.68610        0.750

4          -0.68265        0.750

Final          -0.67925        0.750

Total accuracy: 0.25

**Conclusion:**

1. By taking a large training data set percentage, Max Entropy classifier give superior accuracy
2. We get lower accuracy by lemmatizing the corpus, ideally we should have higher accuracy but since we are training data on small training set with less variations we get lower accuracy.
3. By Removing the punctuations and stop words we get almost similar accuracy.
4. In Unbalanced collection we get low accuracy as in the training set we are taking more of positive words than negative words and in the test we are doing the opposite .i.e. more of negative words so the corpus is getting trained on less negative words and thus giving lower accuracy.