***Project 3 Report:***

***Sentiment Analysis on movie reviews.***

**Developed By:** Snehal V Sutar.

**Net Id:** svs130130.

Subject: Natural Language Processing.

Prof: Dan Moldovan.

***Problem Statement:***

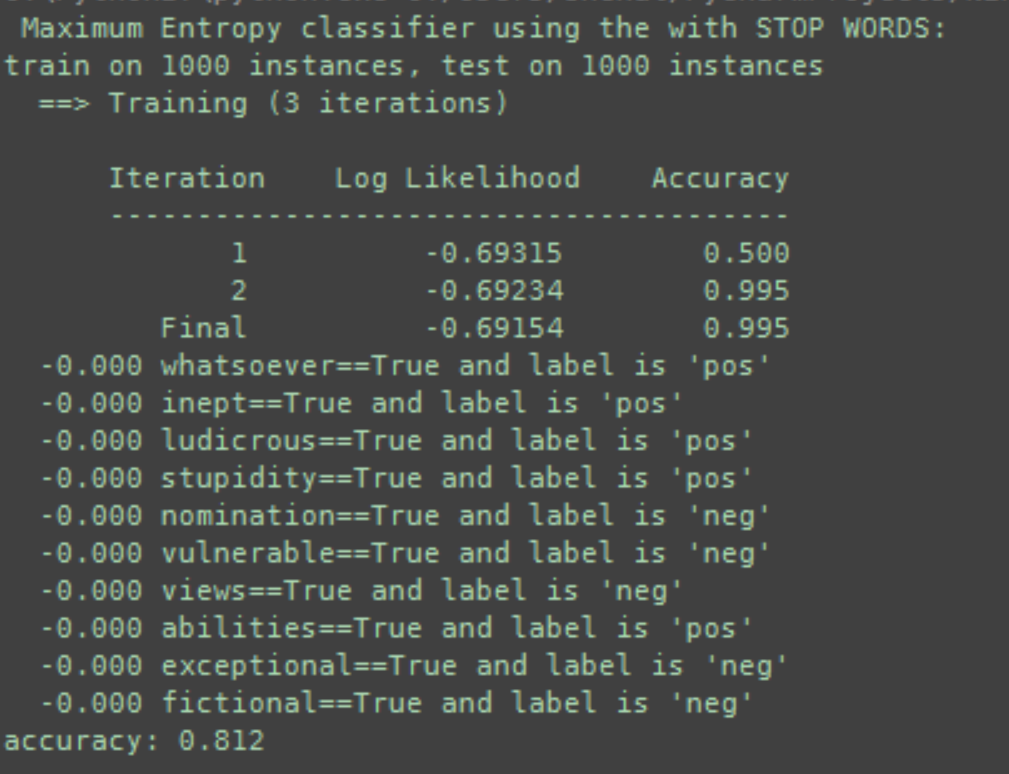
1. Develop sentiment analysis based on maximum entropy classifier on movie reviews data
2. Identify discriminating features, experiment with
3. lemmatization/no lemmatization (WordNet).
4. filtering out punctuation,
5. filtering out stopwords;
6. Experiment with unbalanced collection: change proportions of positive/negative examples in training data.
7. Submit: (1) source code for your experiments, (2) report, containing description for experiment setup (i.e. preprocessing/filtering, training data size, proportion), experiment results, and conclusion.

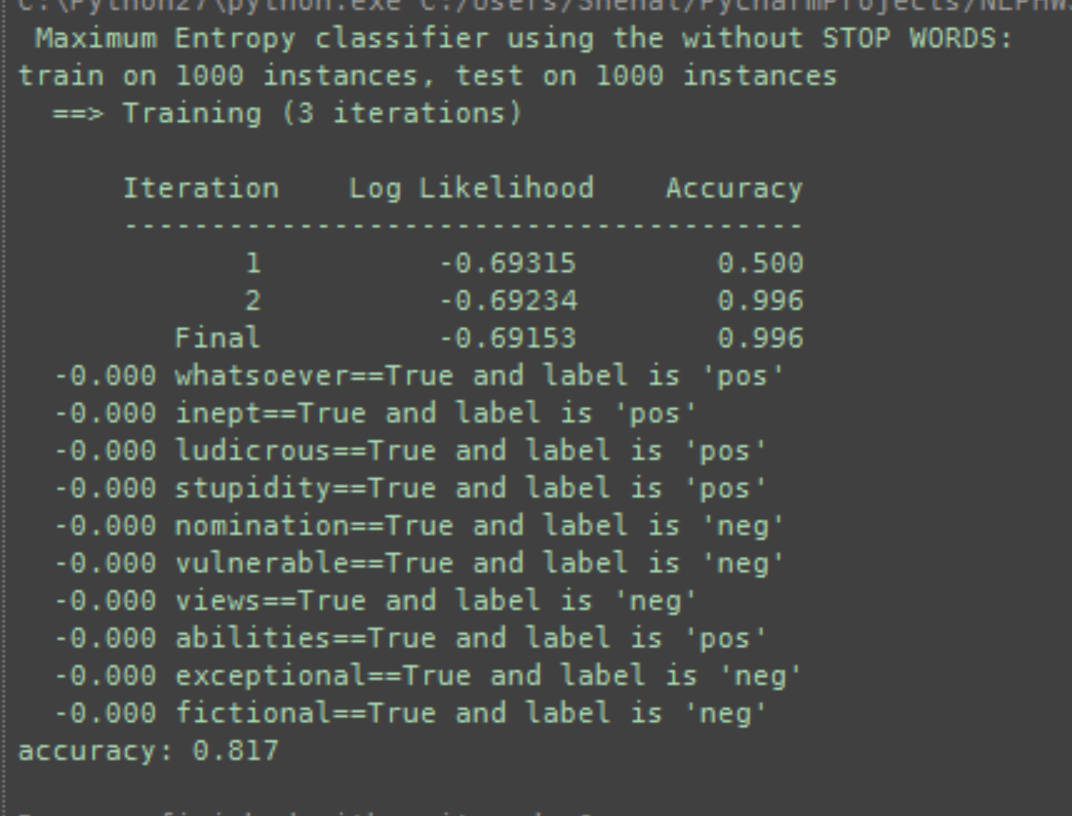
***Various test cases and their respective results:***

**Filtering out stop words:**

Movie test data is used which has 1000 positive and 1000 negative reviews. Out of that 50% (1/2) is used as training data and rest 50% (1/2) is used as testing data. This means out of total 1000 positive reviews 500 reviews will be used for training data and remaining 500 reviews will be used for testing data. Similar distribution will be applied for negative review data.

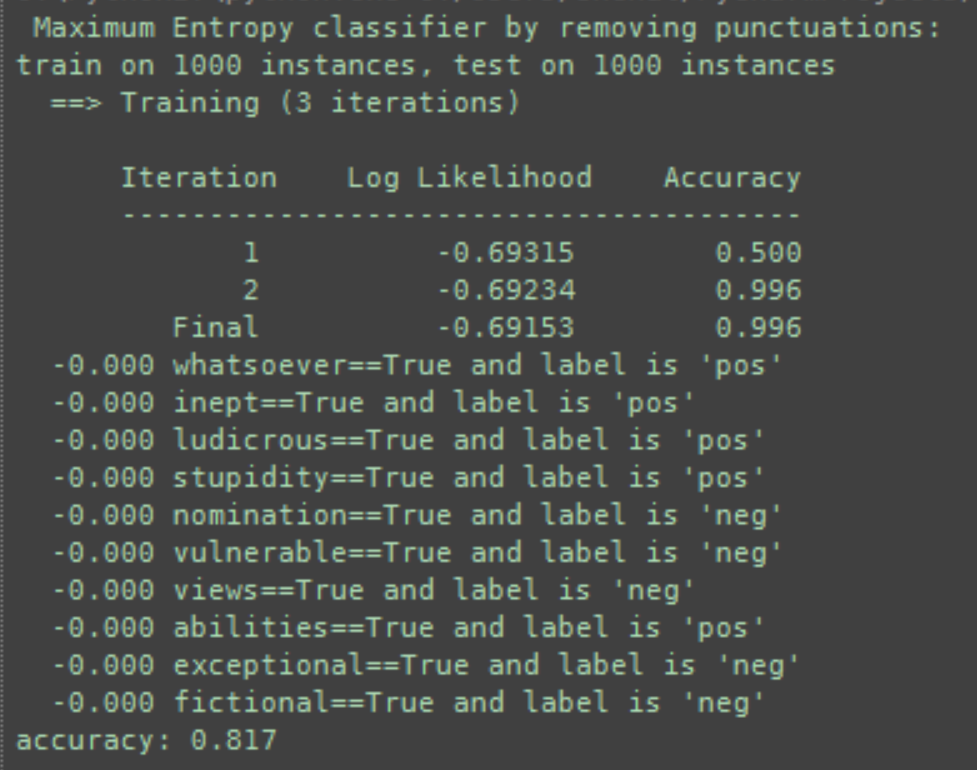
**Accuracy with stop words: 0.812**



**Accuracy without stop words** **accuracy: 0.817.**

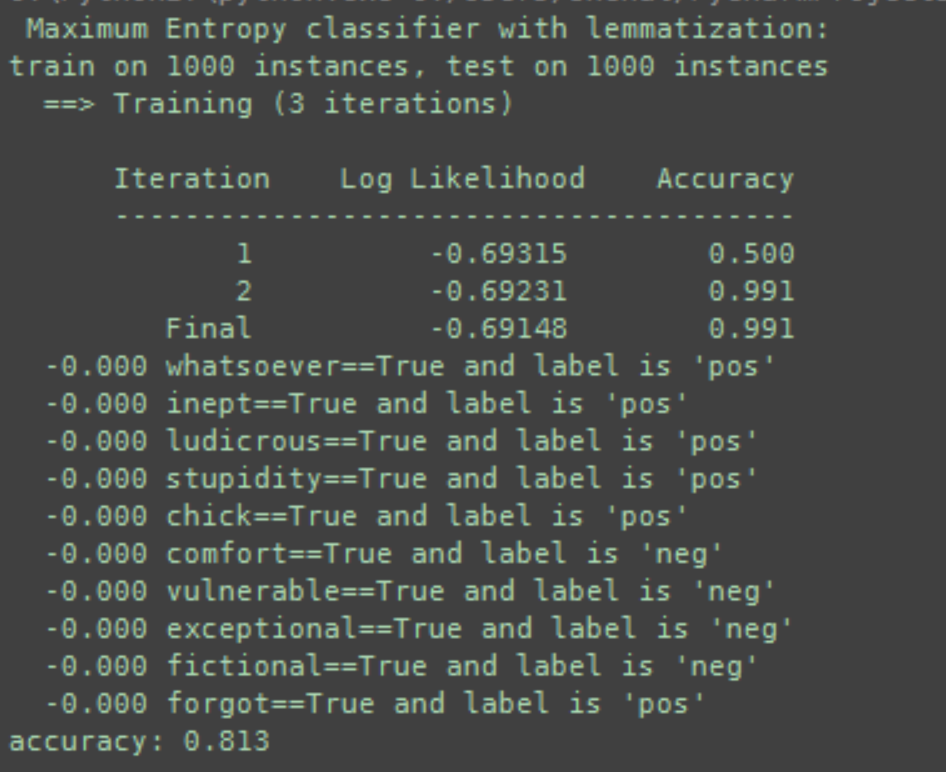
**Filtering out punctuation words:**

**Accuracy without stop words** **accuracy: 0.817.**



**With Lemmatization:**

**Accuracy with lemmatization: 0.813**



**Tabular format data - Considering all words:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Raw Accuracy** | **After removing stop words** | **After removing punctuations** | **With lemmatization** |
| **25% training data** | 0.812 | 0.813 | 0.812 | 0.802 |
| **50% training data** | 0.817 | 0.812 | 0.817 | 0.813 |
| **75% training data** | 0.722 | 0.724 | 0.718 | 0.714 |
| **80% training data** | 0.7275 | 0.7225 | 0.7225 | 0.73 |
| **90% training data** | 0.725 | 0.725 | 0.725 | 0.705 |

**Tabular format data - when only top 1000 words are considered for Accuracy calculation**:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Raw Accuracy** | **After removing stop words** | **After removing punctuations** | **With lemmatization** |
| **25% training data** | 0.5847 | 0.584 | 0.583 | 0.802 |
| **50% training data** | 0.605 | 0.608 | 0.604 | 0.813 |
| **75% training data** | 0.614 | 0.612 | 0.616 | 0.714 |
| **80% training data** | 0.6275 | 0.6225 | 0.6275 | 0.73 |
| **90% training data** | 0.64 | 0.64 | 0.64 | 0.705 |

**Tabular format data - when only top 500 words are considered for Accuracy calculation**:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Raw Accuracy** | **After removing stop words** | **After removing punctuations** | **With lemmatization** |
| **25% training data** | 0.542 | 0.556 | 0.5453 | 0.802 |
| **50% training data** | 0.556 | 0.566 | 0.556 | 0.813 |
| **75% training data** | 0.554 | 0.548 | 0.556 | 0.714 |
| **80% training data** | 0.57 | 0.56 | 0.57 | 0.73 |
| **90% training data** | 0.58 | 0.585 | 0.585 | 0.705 |

**Conclusion:**

* When top 500/1000 words are used for accuracy calculation lemmatized data is most accurate in all the divisions of data.
* When all words are used for training, the reviews are best guessed when only 25% of data is used as training data and remaining is used as testing data.
* The accuracy rate goes on decreasing – while we keep on increasing the size of training data compared to testing data, when all the words are considered.