Kalyssa A. Owusu

NLP Lab 4 Report

Justification of chosen libraries

1. Scikit-learn - This library was chosen due to it's applicability to machine learning and data mining. In addition, it features data preparation tools as well as the Naive Bayes and Logistic Regression algorithms, and is compatible with Pandas data frames, which I used to prepare my data.
2. Pandas - A software library that aids in "data manipulation and analysis" by placing data in the form of tables, which can be interrogated. This was convenient because the labelled data could be sorted into columns of sentiment and label. It was also more convenient than constantly reading from a text file.
3. Stop_words - A library containing common stop words in a variety of languages. In this case, the English stop words were used. This is helpful in normalization as it helps remove noisy words.

Evaluation Metrics & Results

I chose the metric accuracy, as well as precision, recall and f-measure to evaluate the models after they were split, trained and tested using my original data.

	Logistic Regression		**Naive Bayes**	
	Normalised	Unnormalised	Normalised	Unnormalised
Accuracy	0.82	0.782	0.82	0.80
Precision	0.82	0.789	0.84	0.83
Recall	0.84	0.79	0.81	0.77
F-measure	0.83	0.79	0.82	0.80

Table 1. The results of the various tested models

These measures were chosen as they give an overview of how much of a correlation exists between the predicted results and the true data.

Kalyssa A. Owusu

NLP Lab 4 Report

Discussion

Expectations & Evaluation

In undertaking this lab, it was expected that the following results would yield the following characteristics:

- A significant distinction between the normalized and unnormalized datasets, with the normalized dataset yielding better metrics overall.
- A notable difference in the metrics between logistic regression and naive bayes

Of these expectations, the following observations were made:

- The normalized data outperformed the unnormalized data sets, due to the fact that normalized data is more consistent and convenient. The normalization was done by making all text small caps, removing punctuation, and removing stop words.
- There was not a significant difference between the results of the logistic regression and naive bayes models, although the normalized logistic regression model outperformed the other models when the F1 metric was the major indicator, followed by normalized naive bayes, which were then followed by unnormalized logistic regression and unnormalized naive bayes respectively.
- In terms of accuracy, the normalized logistic regression and naive bayes model performed equally well.
- The naive bayes model generally performed better than logistic regression when it came to the unnormalized data set, which is most likely because of naive bayes' suitability when it comes to text data (bag of words model assumption) and sentiment analysis.

Potential Improvements

- Incorporating stemming & lemmatization in text normalization to gauge the effect of further normalization on the models metrics.