Nanis Kanana

Lab 4 Report

To implement the naïve Bayes and logistic regression classifiers, I used pandas (“Python Data Analysis Library”) for data processing. Pandas is an open source, free to use (under a BSD license) library providing high-performance, easy-to-use data structures and data analysis tools for the [Python](https://www.python.org/) programming language according to pandas documentation. I also used [Scikit-Learn](http://carrefax.com/new-blog/2017/6/28/naive-bayes-document-classifier-with-scikit-learn); a free software machine learning software that includes tools for many of the standard [machine-learning tasks](http://strata.oreilly.com/2013/09/gaining-access-to-the-best-machine-learning-methods.html)(such as clustering, classification, regression, etc.) which is what I needed. Using [Scikit-Learn](http://carrefax.com/new-blog/2017/6/28/naive-bayes-document-classifier-with-scikit-learn) also allows the user to create charts for data analysis, evaluation and visualization by creating matplotlib charts although I did not implement that feature. Moreover, [Scikit-Learn](http://carrefax.com/new-blog/2017/6/28/naive-bayes-document-classifier-with-scikit-learn) is an open source library thus it’s convenient and readily available. For Naïve Bayes classification I used [Scikit-Learn](http://carrefax.com/new-blog/2017/6/28/naive-bayes-document-classifier-with-scikit-learn) Multinomial Naive Bayes because it is one of the standard classic algorithms which is used in text categorization (classification). Each event in text classification represents the occurrence of a word in a document. For Logistic Regression classification I used linear multinomial Logistic Regression because according to the sklinear documentation, it can handle both dense and sparse input. Multinomial Logistic Regression can also address different types of classification problems. Where the trained model is used to predict the target class from more than 2 target classes.

The inputs to the Naïve Bayes MultinomialNB and multinomial logistic regression are the features we have in the dataset i.e. the reviews which in both classifiers we fit the estimator and transform the vectors to tf-idf.

I evaluated the classifiers by calculating accuracy for each of them using the tdifVectorizer which uses the bag of words assumption (i.e. position of words does not matter) . I used 30% of the data as a testing set and the results are as shown in below:

![A screenshot of a cell phone

Description automatically generated]()The accuracy results show that normalization of the data increases accuracy in both classifiers