Nutifafa Cudjoe Amedior

Natural Language Processing

Lab 4 – Report

09/11/2018

1. An explanation for your choice of libraries

PANDAS – This library provides fast data structures designed to make working with “labelled” data easy and intuitive. I chose this library because it enables me to express labelled data in a tabular form such as in an Excel spreadsheet.

NLTK – The Natural Language Toolkit (NLTK) is a library for statistical natural language processing. I chose this library because is enables text tokenization and normalization of the features of labelled data.

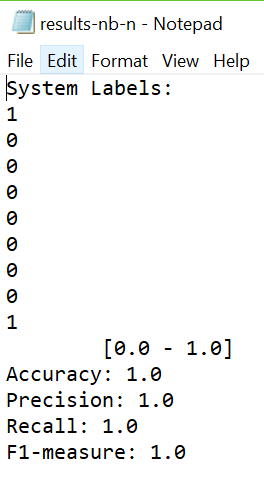
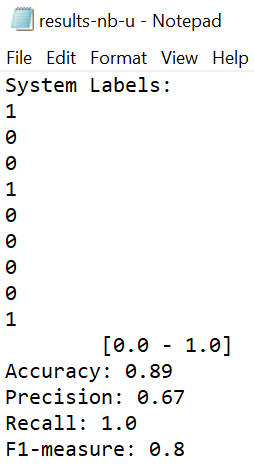
SYS – This library provides access to variables used by the interpreter. I chose this library to enable me to access variables passed in the command line that determine which functions in my program to run.

NUMPY – This library is a package for scientific computing with python. I chose this library because it allows me to represent features of labelled data as arrays paired with their labels.

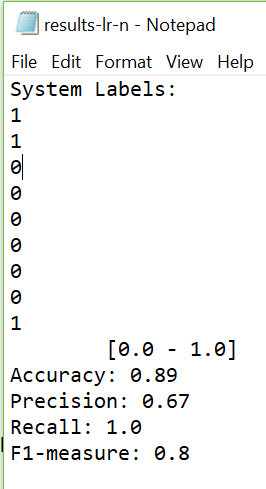
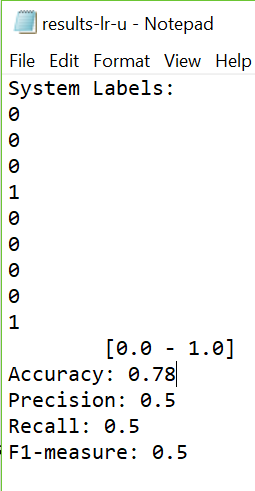
SKLEARN – I chose this library because it provides me with tools for the classifiers, extracting features for the classifiers, training and testing the classifiers.

1. How you evaluated your classifiers and your results. You should have results for each version of each classifier.

I evaluated my classifiers by giving them the same data to test on. I complemented this with using a contingency table to view accuracy, recall, precision and F1-measure.

Normalized Naïve Bayes Unnormalized Naïve Bayes

Normalized Logistic Regression Unnormalized Logistic Regression

1. A discussion of your results. Why do you think you got the results that you did? Compare the results.

The normalized Naïve Bayes had 100% accuracy, 1.0 Precision, Recall and F1-Measure.

The unnormalized Naïve Bayes had 89% accuracy 0.67 Precision, 1.0 Recall and 0.8 F1-Measure.

The normalized Logistic Regression had 89% accuracy, 0.67 Precision, 1.0 Recall and 0.8 F1-Measure.

The unnormalized Logistic Regression had 78% accuracy, 0.5 Precision, Recall and F1-Measure.

The normalized Naïve Bayes had the best performance overall. The normalized Logistic Regression and the Unnormalized Naïve Bayes had the same performance. The Unnormalized Logistic Regression had the worst performance among all the classifiers.

The reason why the normalized classifiers had better performance compared to their unnormalized versions is because the normalization reduces the empirical error rate on classifying document features from the test data sets (Raschka, 2014)..

References

Raschka, S. (2014, July 14). *About Feature Scaling and Normalization – and the effect of standardization for machine learning algorithms*. Retrieved from sebastianraschka: https://sebastianraschka.com/Articles/2014\_about\_feature\_scaling.html