Gbètondji Jean-Sébastien Dovonon

Lab 4 report

The project includes two classifiers: one Naïve Bayes classifier and one Logistic Regression classifier that can both be trained with normalized data or unnormalized data.

Both classifier were built using the python library sklearn because of its completeness (includes processing tools, multiple classifier and metrics), the fact that it is open source and flexible.

The Naïve Bayes classifier was built using the sklearn Naïve Bayes MultinomialNB(). We are using the multinomial classifier because of its suitability for sparse data such as word counts or word frequency. Since it can take word frequencies, it is appropriate for this text classification task.

The Logistic Regression was built using the linear multinomial Logistic Regression classifier which uses the cross-entropy loss function.

The inputs taken by both classifiers are bag of words. During training, we generate a tf-idf matrix representation of the data that is used for training.

By using multinomial classifiers in both cases, we made sure that both classifiers can take the same input data (tf-idf features)

Another library used is pandas for data processing.

The classifiers were evaluated using their accuracy with 25% of the data as test set.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Classifier | NB-U | NB-N | LR-U | LR-N |
| Accuracy | 88 | 91 | 88 | 90 |

We can notice that normalization of the data increases the accuracy. It is therefore possible that more sophisticated normalization methods would have increased the accuracy. Also, the classifier classes imported from the libraries come with some features that help increase the accuracy. For instance, the logistic regression has L2 regularization by default.

Also, Logistic Regression heavily depends on the features used. For instance, in this case the features used where tf-idf features. Proper feature engineering would most likely increase the accuracy of the logistic regression because of its discriminative nature. It is however less sure for the Naïve Bayes model.