Part B

Data preparation

In part B the Naive Bayesian Classifier was implemented. The blogs were split that were provided in test- and training datasets. The test dataset consists of blogs that were assigned as test blogs (e.g M-test3.txt). We put the test blogs in a folder called “test” and the training blogs in a folder called “train”. These newly-created folders are converted to datasets that can be used by our NBC implementation. Both datasets consist of normalized tokens following the manual. Thus a sentence “Hello, everyone. Do you like the new layout?” will be : [hello, everyone, do, you, like, the, new, layout]. To reduce the size of the dictionary, a filter is used that only tokens are used that have a length between 4 and 10. A sample of 40% of this new dictionary is used to calculate the conditional probabilities for the NBC.

Training

The NBC algorithm is implemented as stated in Probabilistic Inference and Bayesian Classification, blz. 19. Our algorithm is slightly differently implemented than described in the document. The difference is that we put the second loop outside the first loop in the context of performance issues.

Testing

After completing the training phase, the NBC is tested on the created test dataset. The implementation of the application of the NBC is done following the algorithm as stated in Probabilistic Inference and Bayesian Classification, blz. 20. The output of the test phase is the accuracy of the NBC applied on the test dataset. The result of the accuracy of the performed test varied between 70 - 76%.

Usage

To see the whole NBC in action, main.py needs to be run. The only required dependency is NumPy.

Authors

Tjeerd Jan Heeringa, s1497324

Joshua van Kleef, s1385801