Group Name: MBgroup

Name: Marija Babić

Email: marija0408@gmail.com

Country: Croatia

Company: Bonsai technologies

Specialization: NLP

Problem description

The dataset contains 18,828 messages and each message belongs to one of 20 newsgroups. The task is to use the current dataset to create the model which will do the document classification, that is, for each new message, the model will assing the message to one of 20 newsgroups.

EDA

After creating the dataset it was time to build a model. First thing that was done was creating two datasets, train and test dataset. Train dataset was used for training the model. First step was creating the message/words matrix that was used in multinomial Naive Bayes classifier which is suitable for classification with discrete features. After the model was trained, it was tested on the test dataset.