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.

Data cleansing

Creating a dataset with three columns: id, message and class.
Before inserting the data into the dataframe, from and subject parts of each message (the only fixed parts) are removed. Also new lines and tabs are removed from the data.