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 understanding

Each message is in one of 20 folders that represent newsgroups. First thing I have to do is to create a csv that contains two columns: message, newsgroup. After that, data preparation (data transformations and cleansing) can take place.

What type of data you have got for analysis

The data are 18,828 text files which contain From, Subject and message parts. There are no problems in the data because the dataset is already prepared (duplicate values are removed and only parts that are needed for the modeling are left).