**The 20 Newsgroups data set**

The [20 Newsgroups](https://archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups) data set is a collection of approximately 20,000 USENET newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. The dataset, originally available from the [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/index.php), has become a popular data set for experiments in text applications of machine learning techniques, such as text classification and text clustering. There are a number of versions of this dataset.

**Organization**

The data is organized into 20 different newsgroups, each corresponding to a different topic. Some of the newsgroups are very closely related to each other (e.g. **comp.sys.ibm.pc.hardware / comp.sys.mac.hardware**), while others are highly unrelated (e.g **misc.forsale / soc.religion.christian**). Here is a list of the 20 newsgroups, partitioned (more or less) according to subject matter:

|  |  |  |
| --- | --- | --- |
| comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x | rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey | sci.crypt sci.electronics sci.med sci.space |
| misc.forsale | talk.politics.misc talk.politics.guns talk.politics.mideast | talk.religion.misc alt.atheism soc.religion.christian |

**Data**

You will use [Jason Rennie’s version of the 20 newsgroups data](wone.com/~jason/20Newsgroups/), in particular the one called 20news-bydate collection: 20 Newsgroups sorted by date; duplicates and some headers removed (18846 documents). This "bydate" collection is sorted by date into training (60%) and test (40%) sets, does not include cross-posts (duplicates) and does not include newsgroup-identifying headers (Xref, Newsgroups, Path, Followup-To, Date).

For convenience I have extracted the data from its original .tar.gz format, and place it in a zip file. Please note that these are not just single train and test files. There is a folder for each type (train and test), and within them 20 subfolders corresponding to each newsgroup category. In each newsgroup categorysubfolder you have document files containing each of the newsgroup text msgs (a total of 18846 docs)**Assignment**

1. Build a text classification model in scikit-learn using a Bag of Words (use sklearn CountVectorizer) and a Logistic Regression algorithm. As this is a multiclass classification problem Report accuracy and the confusion matrix. Note: you will have to set the solver parameter to ‘multinomial’.

For details check:

<https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html>

1. Build a text classification model in Keras using a Bag of Words (use Keras’ Tokenizer) and a feed forward neural network architecture of your choice. Once again, as this is a multiclass classification problem, report accuracy and the confusion matrix.

*Note 1: The code you write should be able to load the data from the directory structure and organize it in a pandas dataframe / numpy array (whatever you prefer). The exercise continues from there onwards similar to what we have covered in class.*

*Note 2: This is one of most highly used datasets in the text mining field, which means that there are all types of solutions readily available on the web, including how to read the data and load it to a dataframe. I cannot stop you from using one of those, but it would be better for you to try to do it on your own (you will definitely learn more).*