Rumour Finder

The goal of this project is to classify rumours using Machine Learning. Various methods of classification and preprocessing will be compared.

Scripts

All scripts can be found in the scripts folder. Here is a small summary of each file.

data_reader.py: Reads the data from the CSV file, removes any duplicates and returns the tweet with its classification of whether it is a rumour (R) or not (NR).

machine_learner.py: Class that can be used to add classifiers and test them, by calculating the accuracy and the F1 score from the test-set. It gathers the sets from the preprocessor.

my_nn.py: Self-written Neural Network library that uses the Sklearn framework to be able to use the same functionality as the other Sklearn librariers.

preprocessor.py: Gathers the data from the *data_reader.py* and converts them to ready-to-use numpy arrays as features. Has multiple ways of preprocessing, such as using tfidf or countvectorizer. The preprocessor can split up the data randomly in a training-set (60%), cross-validation set (20%) and test-set (20%) or in a training-set (75%) and test-set (25%).

classification_tests.py: Contains two tests, the first test compares various settings for preprocessing and the second test compares different classification algorithms.