ML1010 – Independent Project 3
Michael Vasiliou

Problem:
When I first started with Machine Learning I realized that an important and time consuming aspect of the process was to manage the data files. Ensuring that source data remains intact, cleaned, lemmatized, stemmed, encoded, balanced, train/test split, etc. There were a lot of different steps and a lot of different data files created that had to be managed. An important aspect of comparing models is to ensure that the data is identical in every experiment.
During the Machine Learning lifecycle a number of artifacts and data points are created that are referenced continually: trained model, train/test data sets, graphs/charts, and results from various tests. During this process repetitive tasks are performed as well providing rise to common utilities.

To solve this problem I started a framework to help with an end to end Machine Learning lifecycle where you are able to create/store/save and retrieve all aspects of your Data Experiment without having to re-run everything. It keeps track of what you have done, all your artifacts, and can be reloaded to start where you left off without having to re-run all your work.

The framework is ever evolving. I'm currently re-writing the Data storage component, integrating Tensorflow models, and working on saving generated graphs.