Assignment 1

Homework assignments will be done individually: each student must hand in their own answers. Use of partial or entire solutions obtained from others or online is strictly prohibited. Electronic submission on Canvas is mandatory.

1. **Document Classification** (100 points) In this homework, you need to classify text paragraphs into three categories: *Fyodor Dostoyevsky*, *Arthur Conan Doyle*, and *Jane Austen* by building your own classifiers. The data provided is from Project Gutenberg. Please follow a few steps as below:

   - (5pts) **Preprocess data**: remove punctuations, irrelevant symbols, and common words etc.
   - (5pts) **Construct examples**: Divide each document into multiple paragraphs. Each paragraph will be one example.
   - (5pts) **Data split**: Sample these paragraphs into training and testing data.
   - (5pts) **Feature extraction**: Build a vocabulary to represent each paragraph using only training data. Consider TF-IDF features for each input example.
   - (60pts) **Build** two classifiers (described below).
   - (5pts) **Plot** training loss and validation loss at each epoch.
   - (5pts) Using **cross-validation** on the training data, report the classification test/validation error (or accuracy) for each category.
   - (10pts) **Compare** both classifiers and provide an analysis for the results.

   (a) Implement a Logistic Regression model using both gradient descent and stochastic gradient descent.

   (b) Implement a Multilayer Perceptron (MLP) model using backpropagation and mini-batch gradient descent.