Assignment 1: Text Classification

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. **Part A: Binary Text Classification** (80 points) In this homework, you need to classify text messages of SMS into two categories (binary classification): *ham* (not-spam), and *spam*, by building your own Logistic Regression classifier. The data can be found in "a1-data/SMSSpamCollection", which is provided is from UCI SMS Spam Collection Data Set. Please follow the steps below:

   - (20pts) **Data Preprocessing**:
     - (5 pts) Load data into sentences and labels, split into training, validation, and test set. Report the data distribution in a table.
     - (5 pts) Remove punctuation, urls, and numbers. Change text to lower case.
     - (5 pts) Tokenize input text into tokens, including word stemming and removing stopwords.
     - (5 pts) Feature extraction: build your TF-IDF feature extractor for the provided dataset.
   - (50pts) **Build a logistic regression classifier**.
     - (10 pts) Given the objective function of a logistic regression (LR) model with L_2 regularization:

$$J = \sum_{i=1}^{n} L(\mathbf{x}_i, \mathbf{y}_i \mid \mathbf{w}, \mathbf{b}) = \underbrace{-\frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}\log\hat{\mathbf{y}} + (1-\mathbf{y})\log(1-\hat{\mathbf{y}}))}_{\text{binary cross entropy loss}} + \underbrace{\lambda\sum_{j=1}^{d}\mathbf{w}_j^2}_{\text{regularization}} \tag{1}$$

       where
       - $\hat{\mathbf{y}} = \sigma(\mathbf{x}\mathbf{w} + \mathbf{b})$
       - n: Number of training samples
       - d: Dimension of $\mathbf{w}$
     - (10 pts) Derive the gradient of the objective function of LR with respect to $\mathbf{w}$. Please write down detailed steps.
     - (10 pts) Implement this Logistic Regression model. This step includes writing code for initialization, objective function, gradient, and gradient descent.
     - (5 pts) Implement evaluation metrics, including accuracy, precision, recall, and F1 score.
     - (10 pts) Stochastic Gradient Descent (SGD): fill the code for the function of SGD.
     - (10 pts) Mini-batch Gradient Descent: fill the code for the function of mini-batch GD.
     - (5 pts) Evaluation your model on the test set.
   - (5 pts) **Cross-validation**: Use cross-validation to choose the best λ using the validation set.
   - (5 pts) **Conclusion**: Analyze the results.

2. **Part B: Multi-class Text Classification** (20 points) In this homework, you are given sentences from three authors *Arthur Conan Doyle*, *Fyodor Dostoyevsky*, and *Jane Austen*. The sentences can be found in "a1-data/books.txt". You need to classify each sentence into their corresponding author by building

your own Logistic Regression classifier. Please follow the steps similar to Part A. Most of the code can be reused from Part A.

Note that you do not need to submit a detailed derivation for the categorical cross-entropy loss function. Please calculate the gradient just in the code.

Please follow the below instructions when you submit the assignment.

1. You are NOT allowed to use packages for implementing the classifier required in this assignment.
2. Your submission should consist of a zip file named Assignment1_LastName_FirstName.zip which contains:
   - Two jupyter notebook files (.ipynb) for part A and B: They should contain the code and the output after execution. You should run all the code provided in the jupyter notebook. The notebook should show your outputs. You should also include detailed comments. The derivation of gradients can be included in the notebook using Markdown.