CSE 4392 Special Topic: Natural Language Processing

Homework 4 - Spring 2025

Due Date: Feb 19, 2025, 11:59 p.m. Central Time

Problem 1 - 60%

Finish the todos for implementing logistic regression from scratch in **Logistic.ipynb**. The accuracy over the toy dataset should be 99% if your implementation is correct. After finishing upload your filled **Logistic.ipynb** and an equivalent **Logistic.pdf** which can be produced using a tool such as this.

Problem 2 - 40%

Your task is to build a document classification model to classify product descriptions from an E-commerce website into four categories: Electronics, Household, Books, and Clothing & Accessories.

You will use the provided dataset in CSV format, which contains two columns: the first column represents the class name, and the second column represents the corresponding product description. You can find the dataset here: Dataset Link

You will use the logistic model you implement from scratch (make a copy of the file called **Ecommerce.ipynb** and do away with the toy example). Your notebook should have five main sections aside from the logistic implementation: dataset loading and splitting, feature extraction, feature visualization (optional), model training, model evaluation (using Macro F1 score and accuracy).

The most important section will be feature extraction where you will implement a procedure to map any text document from the dataset into a numerical vector and then apply it on all train and test examples.

After finishing, upload your filled **Ecommerce.ipynb** and an equivalent **Ecommerce.pdf** which can be produced using a tool such as this.

Note

If your accuracy/F1 is not reasonable (or is exceptional!) then this may affect evaluation. Submitting on or before Feb. 18th grants you a 7% bonus.