MSA 8150 Final Project:
Topic Analysis

Quick Overview

Each team can have a maximum number of 3 members, and would need to pick 2 projects among the 4 possible projects. Notice that if you decide to team up with fewer people than 3, you are still responsible for the same number of projects.

This project deals with text data. You would be developing a model which can predict the topic category of a comment scraped from an online platform.

Problem Setup

The data of this project is scraped from an online platform, where people leave comments related to 40 categories. For data confidentiality, the category names are replaced with numerical labels. Notice that the data are raw, and the comments may contain unprofessional language.

You are provided a dataset with 900,000 textual comments. For each comment a Topic category which is an integer label between 1 to 40 is provided. The goal would be for you to train a machine learning model, which can predict the topic category from the comment.

As a research problem, you are encouraged to do a survey of the available techniques and set up your own techniques to address this problem. Simply downloading someone's code from Github and running it on the data is not an acceptable project completion. You would need to explore different directions, novelties and technologies, along with comparisons to ultimately produce the most reliable results for the competition.

Please make sure to communicate with the instructor and Piazza about any potential questions related to the data.