Theme 5: Free Topics

Requirements

In your proposal, please answer the following questions:

1. What are the names and NetIDs of all your team members? Who is the captain? The captain will have more administrative duties than team members.

I am working on this project alone. My name is Edward Zhao and my netID is etz3.

2. What is your free topic? Please give a detailed description. What is the task? Why is it important or interesting? What is your planned approach? What tools, systems or datasets are involved? What is the expected outcome? How are you going to evaluate your work?

My free topic is I will perform sentiment analysis on movie reviews to determine if they are positive or negative. It is interesting because for movie reviews, the star rating (out of 5 stars) review method is somewhat subjective. Also, sometimes people may just give an incorrect star rating whether intentionally or accidentally. Determining how good a movie is using sentiment analysis on the actual reviews themselves is a different and interesting take. My planned approach is to find a dataset to train on, probably from a movie review site such as Rotten Tomatoes or IMDb. Then I will clean the data, utilizing Python libraries such as numpy and pandas. Then I will do the sentiment analysis, maybe using something like TF-IDF and other methods. The expected outcome is that I can have a program that will be able to tell if a movie review is positive or negative. I will evaluate the model with standard classification evaluation metrics such as precision, recall, f-score, and accuracy.

3. Which programming language do you plan to use?

I plan on using Python. I think the Python libraries will be very helpful for this project.

4. Please justify that the workload of your topic is at least 20*N hours, N being the total number of students in your team. You may list the main tasks to be completed, and the estimated time cost for each task.

The initial data cleaning will probably take a good amount of time. I would estimate about 5 hours depending on how messy the dataset is and the cleaning I need to do. I would say building the model would take about 10 hours. It will take some time for me to figure out the best methods to use. The evaluation and fine-tuning part would probably take about 5 hours. And finally, I would need to figure out some sort of front end to visualize and present my program/model. I would say this is the hardest to estimate in terms of time, but I would say 5 hours at least. I think my workload will easily exceed 20 hours.