Python Pioneers: Project Proposal

1. **What are the names and NetIDs of all your team members?**
   Ethan Shen: ethans11
   Matt Macrides: mbm12
   Ajay Ramsunder: ajayr4
   Huyen Lai: hlai9

2. **Who is the captain?**
   Ethan Shen

3. **What is your free topic? Please give a detailed description.**
   Create a sentiment analysis model. that scrapes Amazon reviews and assigns a positive, neutral or negative sentiment. 4 or 5 star reviews will be assigned a positive sentiment, 3 star reviews will be assigned a neutral sentiment and 1 or 2 star reviews will be assigned a negative sentiment. We will also create a Flask UI. The user will be able to type in a URL or URLs associated with an Amazon product or products. The UI will then return a random sample of reviews (assuming these products have thousands of reviews), the number of stars associated with the review, and the predicted sentiment from our model.

4. **What is the task?**
   The task that needs to be implemented is taking in different Amazon Reviews and then providing whether a review is positive, neutral, or negative.
   - Data Mining Task
     - Take in different Amazon reviews using a web scraping tool and make sure they have ratings as that will be how we can evaluate whether the right tag has been put on the review.
   - Data Analysis Task
     - We will be using a Jupyter file to analyze the reviews and fully classify them and integrate them with a Flask file
   - Evaluation
     - We will use many different evaluation metrics along with accuracy scores and F1, precision, and recall metrics as well.

5. **Why is it important or interesting?**
   Analyzing sentiment in Amazon reviews has real-world applications for understanding customer opinions and preferences on their products. Having a sentiment analysis can provide more nuanced insights, which could help potential customers decide how they view the product. It will hopefully give the customers a voice.

6. **What is your planned approach?**
   a. Begin with web scraping Amazon reviews and their respective stars (1-5)

b. Utilize the NLTK toolkit for sentiment analysis, split data into training and testing and evaluate
c. Create a Flask UI that will return reviews, the number of stars, and the predicted sentiment

7. **What tools, systems or datasets are involved?**
   The tool we are using is Python. We will use the natural language package, NLTK, for sentiment analysis and evaluation and a web-scrape package, Scrapy, to extract the reviews from Amazon. This scraped dataset will be used to train the sentiment analysis model.

8. **What is the expected outcome?**
   The two main pieces of this project are the sentiment analysis model and the Flask UI. The expected outcome is to have both of these parts to be one cohesive unit.

9. **How are you going to evaluate your work?**
   We will split the data into a training and test set. We will train the sentiment analysis model on the training set and then evaluate the model on the test set. To evaluate the model, we will use metrics such as accuracy, AUC, precision, recall and F1 score.

10. **Which programming language do you plan to use?**
    We are going to use Python because the team is most familiar with this language and since it's very versatile for web scraping and natural language processing.

11. **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.**

Outline Plan/Choose toolkits	3 hours
Web scraping/ETL/Pre-processing	23 hours
Sentiment Analysis/ Generate insights/Evaluation	23 hours
Python UI - Flask	23 hours
Finalize Project for Submission	8 hours