Project Proposal

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, Bidisha Paul, will be working alone in the project. My Net id is bidisha3.

1. 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 sentiment analyses. My project will focus on comprehending the sentiment of customer review for different products in Amazon. Potential customers always want to know the opinion of existing users before they purchase a product. It would be interesting to know the positive and negative sentiment/opinion associated with a product to get a better understanding regarding the product. This helps amazon to understand which products are superior or inferior which further assists in making business decisions.

I will scrape the customer reviews from Amazon website. The extracted reviews and ratings of the product will be saved in a csv file. This will be followed by exploratory data analysis, correlation analysis and feature engineering. Then I will split the data into train- test and try to fit Naïve Bayes using Python’s Scikit learn package. I will try to check bias-variance trade-off and model performance on the test data. I will try to develop additional models with different architecture like Regression, Random Forest on the data and compare the performance to check the best performing models. The expected outcome is to have a robust model with good performance on both training and testing dataset so that it can be used for prediction. To evaluate my model, I will check different performance metrics like precision, recall and F1 score.

1. Which programming language do you plan to use?

I plan to use python for my project.

1. 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.

I will be executing the entire project myself. The following are steps to be taken with relative hours required:

* 1. Quick look at the data and exploratory data analysis – 3 hours
  2. Data Visualization and Train/Test Split –2 hours
  3. Correlation analysis and setting target variables – 1.5 hour
  4. Feature extraction and fitting a Naïve Bayes algorithm – 2 hour
  5. Testing the current model output – 1.5 hour
  6. Develop and compare the current model output with other approaches (Random Forest, Logistic Regression etcetera) – 4 hours
  7. Selection of the best performing model and hyper parameter optimization of the selected model to enhance its performance further - 3 hours
  8. Writing detailed steps taken and the project’s outcome and conclusion – 3 hours