#1:
Netids: mbosc2, anishm2, kevinez2
Names: Matthew Bosch, Anish Meka, Kevin Zhang
Captain: Anish Meka

#2:

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

Theme - Free topic

a, b, e. Sentiment analysis using NLTK library. We will also choose (some, not all) from the following libraries: Numpy, Pytorch, or scikit-learn.

Here are our ideas for data sources:Kaggle: Your Machine Learning and Data Science Community

Please note: We plan to proceed with the first idea if we get it approved in the proposal feedback. Otherwise, we'd like to opt for the second or third idea in that respective order.

First idea: Analyze amazon product reviews for a specific category from their website. Example categories include technology, whole foods/grocery, or toiletries in that respective order (we'll narrow this down based on the amount of data we're retrieving). Our foremost preference is technology, and we'll narrow this down based on the amount of data we're retrieving. For example, if our data structures cannot fit in memory, we can obtain reviews for just computers, for example. Here is an example Kaggle source: https://www.kaggle.com/datasets/kritanjalijain/amazon-reviews?select=train.csv. We will evaluate each source for the features it provides and choose the best accordingly.

Second idea: Perform sentiment analysis on political speeches and compare and contrast differences (for example, republican vs democratic speeches). An example source is https://millercenter.org/the-presidency/presidential-speeches. We will narrow this down by time period, starting with presidents from the 2000s, for example.

Third idea: Make an abstract analyzer where the user will provide the products/services and reviews they want to analyze with our program. This will require a great deal of data preparation/filtering and will thus require an interactive/adaptive UI.

c, f. This project is important because it's a summary of overall sentiment that can save users minutes or even hours of time from browsing through thousands of reviews. If users already

have an opinion, they may want to compare it with the online community, so this summary will serve a multi fold purpose.

d. After text preprocessing/wrangling and feature extraction (we will perform these data preparation stages as needed), we will conduct a train-test split. We will then use scikit-learn to train a Naive Bayes classifier. If this doesn't prove effective based on our evaluation, we will use scikit-learn to convert text data into TF-IDF vectors and then apply Support Vector Machines as a machine learning algorithm for sentiment classification.

g. In order to evaluate our work, we will test our sentiment analysis classifier on a set of testing data. The testing data will include randomly selected reviews that we will evaluate our sentiment analyzer on by comparing our results to the actual sentiment score that the review originally had. From this, we will be able to obtain an accuracy rate on the testing data which will serve as the primary metric to which we evaluate our work.

#3:
Python (several libraries as listed above and below) / Javascript (for web app if time permits)

#4:
Our team is 3 people so we will need approximately 60 hrs of work. Our main deliverable is the sentiment analyzer using NLTK and machine learning libraries. If we have time, we can turn this into a web application with both a frontend and backend where each portion would require ~20 hours of work. If we have time for a full stack application, each group member will contribute to all three sections and will need to collaborate to successfully complete the project. If we're able to complete the full stack stretch goal, we may use Javascript for frontend and Flask or Django for backend. Please note that a lot of time will need to be spent learning new technologies to implement the project and none of our group members are familiar with NLTK and machine learning. None of our team members are familiar with Javascript, Flask, or Django.