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.

   Name: Yasamin Tabatabaee
   Netid: syt3
   Team name: AmazonSentiment

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?

   Topic: Sentiment analysis on Amazon reviews for different product categories.

   Description: In this project, I plan to do sentiment analysis on a large collection of customer reviews for products at Amazon.com and compare reviews across different product categories. One goal of this project is to create a simple software that takes a product review as input and outputs its sentiment (i.e. good, bad, average) and rating on the scale of 1-5. Another goal is to get an insight on the impact of various factors, such as the review length, type of words that are commonly used in the reviews, type of product, etc on the sentiment of a review for different product categories. The final goal is to compare different machine learning classifiers (e.g. random forest, Naive Bayes, SVM) for sentiment analysis on this data, and use the most promising one in the final software.

   Importance of topic: Sentiment analysis of textual data provides insight into the opinions, emotions and attitudes of people towards objects and topics. In the case of product reviews, it can be useful for marketing researchers and producers to understand the needs of users and the factors that contribute to their satisfaction or dissatisfaction of a product. Additionally, when new users want to buy a product, they are influenced by the reviews and ratings of previous users, so the reviews that a product gets can have a determining effect on how it performs in the market and can also be important to investors. Therefore, developing methods for sentiment analysis of customer reviews can be useful from many aspects.

   Approach and tasks: I will use a large collection of Amazon reviews with ratings, from over the period of more than 20 years, to train classifiers for sentiment analysis. The main tasks for this project are as follows:

   - **Data cleaning and text preprocessing:** This stage includes cleaning the data of duplicate or inconsistent entries and entries with missing values. The rest is typical text processing of the reviews, such as removing punctuations, stopwords, lemmatization and stemming, tokenization, and finally getting a representation of the text data using TF-IDF and length normalization. I might also explore more advanced vector representations, such as GloVe.
   - **Exploratory data analysis:** This stage includes data analysis on the pre-processed reviews with the goal of 1) finding the common words or phrases in each type of review, 2) understanding the motivation behind the reviews, e.g. are people more likely to write a review when they find a serious problem with a product, when they are impressed, or when they find it average? The outcome of this part is plots and visualizations that show these trends.
   - **Machine learning classifiers:** In this stage, I will use different machine learning classifiers, such as random forest, Naive Bayes, and SVM to do sentiment analysis on the reviews and compare their accuracy on test data. If time allowed, I might also explore deep learning approaches such as LSTM-based methods.

- **Final sentiment analysis software:** I will finally implement a simple command-line software that takes a review from the user and uses the best classifier trained in the previous part to return its sentiment. This software will accompany a report that presents the findings of the two previous parts.

Tools, software, datasets:
The dataset I am planning to use is the Amazon Review Dataset from UCSD available at https://nijianmo.github.io/amazon/index.html that was published in 2018, and contains more than 233 million Amazon reviews between the years 1996 to 2018 from 29 different product categories, as well as additional information and metadata for each product. Since this dataset is very large and it will be difficult to use all of it to train classifiers in a timely manner, I will sample a smaller collection from each category (~50K reviews).
For software and tools, I am planning to use python, which has several useful libraries for analyzing text data. In particular, I am planning to use the nltk and pandas libraries for text analysis and the scikit library to implement the classifiers for sentiment analysis and finally seaborn and matplotlib for visualizations.

Expected outcome:
The expected outcome of this project is a software that takes a product review as input and predicts the sentiment of it, as well as a detailed report of data analysis on Amazon reviews that provides insight into factors such as the words that are more commonly used in different types of reviews (bad, good, neutral), which product categories are more likely to receive positive/negative reviews from the users, etc, as well as a comparison between the accuracy of machine learning classifiers used for sentiment analysis on the test data.

Evaluation:
For evaluation, I measure the accuracy of sentiment analysis classifiers using different metrics. The Amazon reviews dataset has ratings on the level of 1-5 for each product that shows the level of satisfaction of the users (1 : very unsatisfied, 5 : completely satisfied). These ratings will be used as ground-truth labels to train the classifiers and also for evaluating accuracy on the test dataset.

3. Which programming language do you plan to use?

   Python for the main code and potentially R for data analysis

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.

   I will work alone on this project, and below is a tentative breakdown of the timing:
   - Familiarity with the datasets and libraries (~2 hrs)
   - Data cleaning and text preprocessing (~4)
   - Exploratory data analysis (~5 hrs)
   - Training various classifiers for sentiment analysis (~8)
   - Evaluation and tuning the models built in the previous stage (~6)
   - Finalizing the software, writing the report and making the tutorial (~7)

Therefore, my estimation of the total time needed for this project (other than preparing the proposal/report/tutorial) is at least 25 hours.