Project Proposal: Names: Elizabeth Lin (el18@illinois.edu), Nirali Rai (narai2@illinois.edu)

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?

Our topic is to do sentiment analysis of movie scripts because we want to investigate language tokenization within the entertainment industry. Movie scripts are pieces of writing that are unique in medium and may portray different types of sentiments in various parts of a movie. We found that it would be interesting to uncover sentiment and character analysis for our project.

Our approach for this project is to mine data from various movie scripts and clean the text data which includes tokenization, POS tagging, and leverage language models to interpret movies' sentiment analysis. We are primarily utilizing the data set on Kaggle (https://www.kaggle.com) and will use Jupyter notebooks to load the data. We will use numpy libraries to parse the data and interpret movie sentiments. We plan to use bag of words methods as a baseline for our data analysis and explore advanced models as necessary.

Our expected outcome of this project is to determine movie dialogue sentiment using language parsing methodologies we have learned in Text Information class. This will help us leverage the end-to-end cycle of data pipeline and explore hands-on data. We will evaluate our work by comparing expected outcome of movie sentiment to our model results. We will compute accuracy of our model using precision and recall.

2. 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. el18

Elizabeth Lin (Captain)	el18@illinois.edu
Nirali Rai	narai2@illinois.edu

3. Which programming language do you plan to use?

Python

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.

1) Initial Research: 4 hours
    a) Kaggle
    b) Reading descriptions of datasets and format
2) Dataset Collection: 4 hours
    a) Collect sample datasets to test with
    b) Evaluate and compare datasets
3) Data Parsing: 6 hours
    a) Parse datasets dependent on program needs
4) Model Design: 10 hours

    a) Create structure/algorithm
5) Model Implementation: 10 hours
    a) Code up selected model(s)
    b) Verify model functionality against given dataset
6) Testing & Evaluation: 3 hours
    a) Data Validation Tests
    b) Computing Accuracy with Precision and Recall
7) Documentation: 3 hours
    a) Document how model runs and design reasoning

Total: 2 * (N=2) = 40 hours estimated work