Meep Meep Go NLP - Sentiment Analysis

Problem Statement – Airbnb is the leading and rapidly growing alternative to the traditional hotel networks. Travelers always hop on the websites to find location, neighborhood, and top stay places for their vacation. To make their much-deserved vacation a good experience stay is a very integral part and hence the travelers spend a large amount of time going over the reviews for the listings.

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.**

First Name	Last Last	NetID	Position
Joel	Rosales	joelar2	Captain
Abhinav	Abhay	aabhay3	
Syed	Peerzada	syedp3	
Prasanna	Muralimanohar	pkm7	
Piyush	Gambhir	piyushg2	

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?**

   Our team decided on a Natural Language Processing (NLP) - Sentiment Analysis project on an Airbnb dataset. Realizing we did not have much experience with Sentiment Analysis, we decided this would be one of the best opportunities to learn. Through this project we wish to provide a summarization of those listings and reviews to the travelers so that they get more concise, intelligent, and knowledgeable data to make their decision about their stay. Happy Vacation!

   Our primary goal using the Airbnb data is to provide location and price based insights to users to aid their decision making while booking an Airbnb reservation. In addition, we are also interested in listing useful metrics that would further help users unfamiliar with the city to decide on neighborhood location and pricing.

   Our first step would be to implement Data Cleaning/Pre-processing such as tokenization, punctuation removal, stopword removal, lower casing, lemmatization and implementing a TF-IDF matrix. The second step would be Classification and Sentiment Analysis on our chosen dataset. Concurrently, we would also have a working backend infrastructure at this point to store our data and begin the algorithm implementation step where we would set our parameters. Those parameters would include # of reviews, price, subjectivity, dates, positive terms, and negative terms. Lastly, we will implement a frontend that will communicate our expected outcome by reporting a correlation between the location of an

Airbnb listing and likelihood of getting a positive review. In addition, we also expect to find a similar correlation between price of listing and the likelihood of getting a positive review. With our experience level, we plan to work on testing and improving every step to include incorporating additional functionality pending time constraints.

3. **Which programming language do you plan to use?**

   Python and the following libraries: Numpy, Pandas, Scikit-learn, Pickle, nltk, seaborn, matplotlib, plotly, and spaCy. Javascript or PyScript for the frontend.

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.**

Task	Estimate Hours
Research and Outline	5
Tool Familiarity	5
Step 1 Data Cleaning / Pre-processing	15
Backend	5
Step 2 Classification / Sentiment Analysis	25
Implement Algorithms	10
Step 3 Frontend	20
Testing / Improvements	10
Administrative / Documentation	10

Total:	105