Project Proposal CS-410 / Fall 2021
TripAdvisor
JunyangWang

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.**
   Team Member: Junyang Wang (Jw111) as individual

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: Topic Mining and Sentiment Analysis for Airline reviews.

   Why is it important: Customer satisfaction is always top of mind for airlines. Unhappy or disengaged customers naturally mean fewer passengers and less revenue. It's important that customers have an excellent experience every time they travel.

   Task: Using Topic Mining and Sentiment Analysis to determent customer satisfaction level for each airline company.

   Datasets:
   https://www.kaggle.com/efehandanisman/skytrax-airline-reviews
   https://github.com/quankiquanki/skytrax-reviews-dataset/blob/master/data/airline.csv

   Tools, Approach:

   - Complete the datasets pre-processing with Nltk library. Includes text tokenization, lowercasing, stop words removal and text stemming.
   - Topic Modeling with Gensim. Find the common words in the customer airline reviews.
   - Complete topic-based sentiment analysis.
   - Finally, we should able to monitor the airline brand reputation based on the varies topics that is positively or negatively credited by the reviewers.

   Outcome: My expected outcome is to show topic and sentiment level comparisons for each airline companies.

   Evaluate: I will compare the sentiment associated with the review to the given rating.

3. **Which programming language do you plan to use?**
   I will use python programming language for this project.

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.**

   Dataset analyzing and preprocessing. – 4 hours
   Extract topics from preprocessed review dataset. –6 hours
   Design of sentiment analysis code. –6 hours
   Post processing and code cleaning. –5 hours