Richa Meherwal
Project Proposal
CS 410 Fall 2020

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.**

   Individual Project
   Name- Richa Meherwal
   Netid- meherwa2
   Captain-Richa Meherwal

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

   Using Topic Mining and Sentiment Analysis to compare customer level satisfaction in Airbnb vs Hotels

   Problem Statement

   Ever wondered traveling to a city and choosing the best place to stay for a full blown and complete travelling experience? With the rise of Airbnb, travellers are usually drawn to this new type of accommodation. But does it satisfy what they are looking for. Similarly some travellers always chose to stay in expensive Hotels despite the price, why? What is it that attracts tourists to hotels over Airbnb.

   I chose this topic because I was interested in learning what level of customer satisfaction present in Airbnb users vs Hotels. I particularly wanted to compare the customer level satisfaction features involved with these two types of accommodation using topic extraction and sentiment analysis.

Task

To compare the customer satisfaction between Airbnb and Hotels using LDA and Sentiment Analysis

Datasets

https://www.kaggle.com/mrinaal007/reviews
https://zenodo.org/record/1219899#.X5Uic0Izba4

Tools

jupyter notebook
Nltk toolkit
genism

Approach

1. Complete the pre-processing of the datasets. This includes tokenisation, removing stop words, normalisation.
2. Use LDA to first extract the common topics that the customers review about in both the datasets.
3. Do a sentiment analysis on each of the sentences containing the extracted topic and assign a sentiment to it. Gather the associated sentiment and the topic over each dataset.
4. Now under each accommodation type we should be able to visualise the topics that it is positively credited for by the reviewers and also negatively.

My expected outcome is to show topic and sentiment level comparisons for each type of accommodation (Airbnb or Hotel). The idea is to see which topics are associated positively or negatively with each of the accommodation types and to gain insight into where each of these services perform better than the other.

Evaluation

To check the sentiment analyser, I would compare the sentiment associated with the sentence/review to the rating given.
To check topic extraction worked well, I will use wor2vec on all the

common topics and see if they separate well.

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.**

   Tasks and hours

   1. PreProcessing data - 5-6 hrs
   2. Research methods for Sentiment Analysis and Topic Extraction and deploy it - 9-12 hrs
   3. Post Processing and visualising the final data - 6-7hrs
   4. Cleaning code ,demo, documentation - 4-5 hrs

   N=1

   Total estimated work hours- 24- 30 hrs