CS 410 Group ZZZRQ

Project Progress Report

1. Which tasks have been completed?

   Data Collection and Cleaning:

   1. Filter the review data from 2018-08-01 to 2018-08-31. This period presents the general date review data.
   2. Filter the review data from 2021-08-01 to 2021-08-31. This period presents the covid-19 date review data.
   3. Filter the review data during the Thanksgiving period from 2018 to 2021.
   4. Construct the relation between and business name and business I'd
   5. Filter the review text json, Pick the first 3 frequency reviews json which contains review_id, business_id and user_id,text,star and name them based on the business name

   Model Training:

   1. Using random selection, select 200 reviews across the entire set of review data.
   2. Isolate the review words and star review entries to treat as data points for model training.
   3. Trained a custom text classification model for predicting the rating of a review based on the input text.
   4. Extracted a dataset with 100 of each 1, 3, and 5 star reviews for future training due to the inaccuracy of the current model.
   5. Studied the APIs of Azure Sentiment Analysis model for analysis review in batch.

2. Which tasks are pending?
   1. Data Analysis according to the result from the model based on the date or restaurants and so on.
   2. Training of a more accurate model using the new dataset
   3. Writing a script to work with the Azure Sentiment Analysis model API and extracting the results for analysis.

4. Analyze and evaluate the text classification model results and come up with a conclusion.

3. Are you facing any challenges?
    1. Because the original dataset is too large, which will cause a memory error, we cannot use these data directly. So we split the dataset into several files of 20000 review entries each and applied the query to each file.
    2. The Azure Sentiment Analysis model appears not to be very scalable due to requiring manual importing and tagging. We are looking into automating the import process.
    3. The first training set, even though employing random selection, resulted in a data set that was skewed towards the extreme scores: either 1 or 5 stars. We decided to have our extraction script to still employ random selection but keep on selecting until 100 of each 1, 3 and 5 stars reviews are retrieved.