Phase 4 Part A: Summary

Part A.1

In Phase 2, we stemmed all words in the review_texts. Stemming chops off prefixes and suffixes, so it may produce words that are not in the English dictionary, which is problematic for the word vectorization steps we conduct (and discuss) in Phase 4 Part B. Lemmatization returns the base root of the word, so it does not have this issue. We re-staged our dataset so that all review_texts use lemmatization instead of stemming, loaded this restaged data into the data mart, and then transferred the data from our data mart into a csv file (we excluded keys) before importing it into a dataframe. We generated data visualizations, including histograms for columns with numeric/continuous values (our 'rating' column) and bar graphs for columns with categorical values (such as our 'branch_name,' 'month,' 'quarter,' and 'year' columns). We discovered that our dataset is skewed towards positive ratings (particularly, 5 star ratings) based on the histogram for the 'rating' column. We examined the distribution of our data using the bar graphs; there was a very small number of reviews written before 2010. We made a scatterplot of review rating vs. number of reviews by branch and observed that Universal Studios Florida has the most 5 star ratings. We also made a boxplot for the ratings of each Disneyland and Universal Studios branch. Except for Disneyland Paris, 1 and 2 star ratings were regarded as outliers for all branches. Note that the visualizations we created in Part A.1 overlapped with the visualizations we constructed for Phase 3 Part B; for more comprehensive visualizations, please see the work we conducted for Phase 3 Part B.

Part A.2

We pre-processed our data for the three classifier models we would be building in Part B of this phase: the Decision Tree, Gradient Boosting, and Random Forest classifiers. From all the visualizations and summarizations we did in Part A.1, two trends in the data stood out to us: an extremely small number of reviews were written before 2010 and the reviews are skewed towards positive ratings (particularly 5 star ratings). The first trend is directly related to our first pre-processing step (the second trend will come into play during Phase 4 Part B). While we had already dealt with missing values during Phase 2, our Universal Studios Singapore attendance data for the years 2006-2009 included null values (since the branch did not open until 2010). We decided to delete all reviews that were written before 2010 (and their corresponding attendance data) because none of the classifiers accept null values when training. As previously indicated, our data summarization in Part A.1 shows that we have extremely little review data prior to 2010. Just 491 of the 90846 reviews in our data mart were written before 2010; eliminating them would have little influence on our classification tasks and training. We then performed one-hot encoding on each of our 'branch_name' columns (which are categorical attributes that are strings) so that they can be accepted by the classifiers in Part B. We standardized each 'attendance (millions)' column to ensure that these features have a similar impact on the analysis and that data mining methods are not biased toward features with larger scales. The range of attendance values is relatively narrow (for the entire dataset, these values range between 4-15 million), so we chose standardizing over normalizing. Feature selection is done when there is an abundant number of features, which is not the case here. We finished this phase by exporting our pre-processed data so that our models could use it.