**Dataset**

The dataset is titled “Trip Advisor Hotel Reviews” based on hotels from online reviews of hotels around the world and collected in TripAdvisor.com database and queried in 2018 as part of research for a paper. The dataset contains two columns “Review” and “Rating” and has 20,491 unique rows of text and ratings scaled 1-5.

**Analysis**

We gathered the count of all ratings in the dataset. The largest Rating count by far is a Rating of “5” at over 9,000 rows. We have approximately 1,200 ratings of “1”, 1,800 ratings of “2”, and approximately 6,000 ratings of “4”. Generally, the population seems to share positive experiences.

Chart

Description automatically generated

After lemmatizing the reviews, the text count ranges from 10 to 2,744 text in a review.

Graphical user interface, text, application

Description automatically generated

As a whole we decided to capture the most frequent words through out all of the ratings. These words such as room, crime, police, pool, breakfast, and drink all seem to fit our project features. These features included cost, crime history, amenities, flexibility, parking, indoor facility, room space, air-conditioning, and neighborhood.

Text, application, chat or text message

Description automatically generated

Next, we gathered the most common Bigrams and split them by rating class from 1 to 5.

Graphical user interface, application, table, Excel

Description automatically generated

Graphical user interface, table, Excel

Description automatically generated

Graphical user interface, application, table, Excel

Description automatically generated

Graphical user interface, application, table, Excel

Description automatically generated

Table

Description automatically generated

These bigrams and unigrams will be the root of our topic tagging as we decide which bigrams will be most valuable to tagging our reviews.

Even after text cleaning there are some tokens that managed to escape. We will fine tune our text cleaning, and perform an additional cleaning step involving token removal less than two characters, and decide which NTLK method is best for removing the least valuable tokens.