Melissa Cho
Iris Kan
CS 410
December 9, 2021

Documentation

1. Code Overview:

Our code consists of two parts: the web scraper and the sentiment analysis model.

The function of the web scraper is to scrape the reviews of hotels/apartments in Chicago and populate a csv file in preparation for the sentiment analysis. The csv file, which the user can specify, has 3 columns: Summary, Text, and Score. The summary contains the header of the review, the text contains the main body, and the score is how many stars the user gave.

Once the user uses the web scraper to populate the csv, the csv is then used to generate the sentiment analysis model via logistic regression. The model classifies which reviews are positive or negative. It disregards the neutral reviews of 3 stars. From this, a graph is generated to help the user visualize the sentiment distribution of the reviews. In the console, the classification report of the sentiment analysis model is then printed out, which displays the accuracy (precision, recall, F1 scores) of the model.

2. Implementation:

We used BeautifulSoup to web scrape the reviews and generated a dataframe to organize the reviews for sentiment analysis. When web scraping, we used a for loop to loop through all 55 pages of the reviews to gather a sufficient number of reviews to populate the csv. We then used Scikit-learn library to turn our dataframe of reviews into a bag-of-word model, to use our bag-of-word model to generate and fit a model using logistic regression, and to test the accuracy of our sentiment analysis model.

3. Instructions to Install/Run:

To run, you would need to have python installed as well as the libraries included in the beginning of both files.
Here are the following libraries you would need to install if you do not have already:

- requests
- urllib (HTTP Client)
- BeautifulSoup
- pandas
- numpy
- matplotlib

- plotly
- scikit-learn

4. Team Contributions:
Melissa worked on the web scraper and Iris worked on the sentiment analysis model.
Melissa web scraped the apartments/hotel website and generated a dataframe for the reviews, including the ratings given, the summary, and the text block.
Iris then used this dataframe to classify each review according to their rating, assign a sentiment rating, and use the data frame as a training and testing dataset for a sentiment analysis model. Both team members spent 20 hours each on this project.