Data Mining and Warehousing [CS-2376]
Prof. Anirban Mondal
Spring 2021, Ashoka University

Decision Support and Visualisation

Rishabh Goswami, Mohammed Zahir Ali, Shreyash Naman

Scraping:

Reviews for 60 hotels were scraped from Agoda using Python Selenium webdriver. These 60 hotels consisted of 20 hotels each from Singapore, Bangkok and Kuala Lumpur. For each hotel, over 100 reviews (in chronological order) were scraped, making for a total of 7120 reviews. Following data points were scraped for each review:

- Hotel City
- Hotel Name
- Hotel Address
- Overall Hotel Rating
- Review Rating
- Review Date
- Review Title
- Review Comment

Clustering:

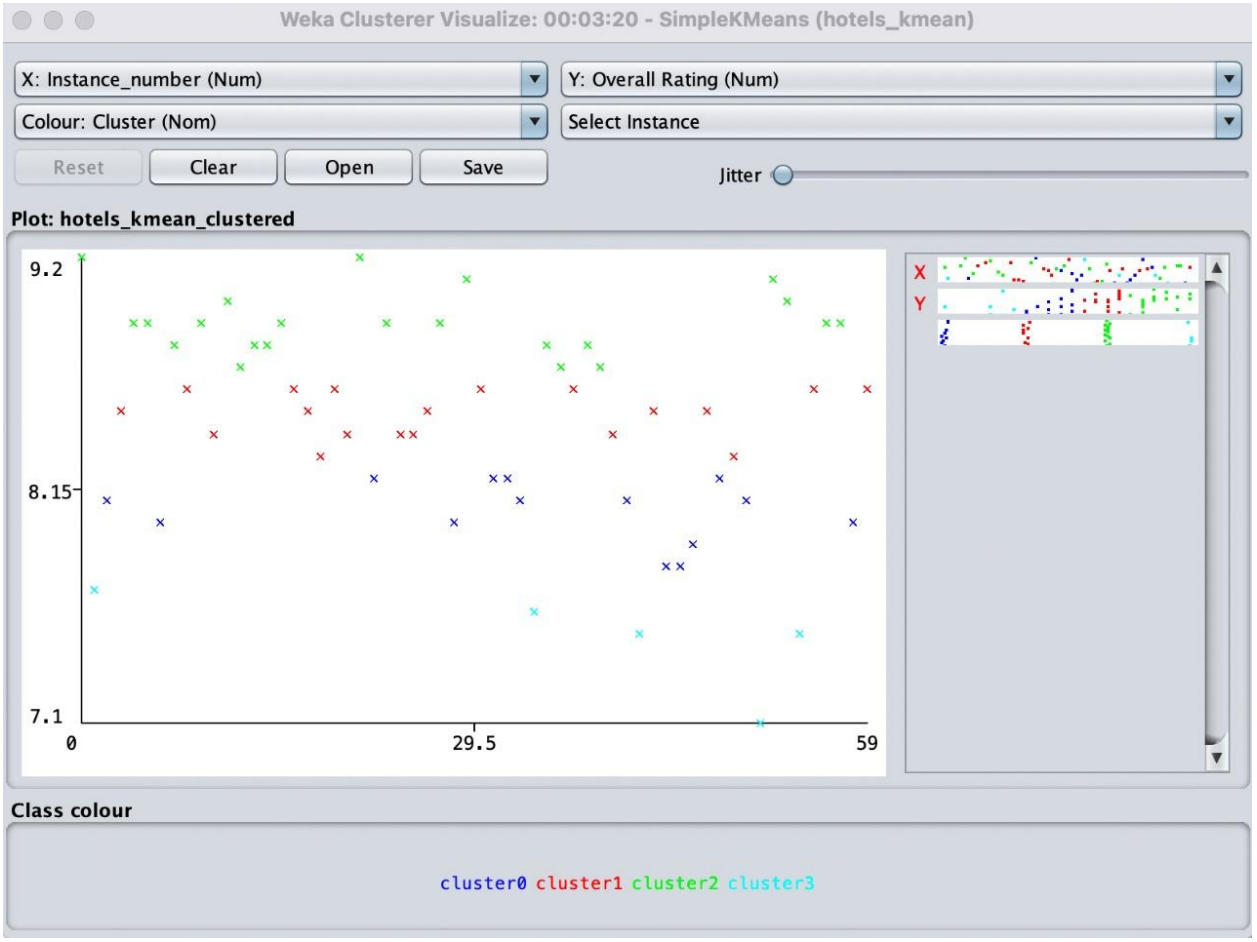
Following descriptive statistics were calculated for the overall rating of all 60 hotels:

- Minimum: 7.1
- Maximum: 9.2
- Mean: 8.466
- Standard deviation: 0.462

We used WEKA to perform clustering on the overall rating of 60 hotels. K-means was the method of choice and the value of K was set at 4.
Values of cluster centres:

- Cluster 1 (22 hotels, Green) : 8.9045
- Cluster 2 (19 hotels, Red) : 8.4895
- Cluster 3 (14 hotels, Blue) : 8.05
- Cluster 4 (5 hotels, Cyan) : 7.48

We found that most hotels were clustered towards the upper end of the scores, with the green cluster containing the most number of hotels, followed by the Red and Blue clusters. These 3 clusters were relatively closer to the mean rating of 8.466, while Cluster 4 (Cyan) saw a steep drop off at 7.48. Cluster 4 was an outlier, consisting of hotels with exceptionally low scores. These included 3 hotels from Kuala Lumpur (Arte Plus, Silka Cheras, Avani Sepang Goldcoast Resort), 1 from Bangkok (Royal River Hotel) and 1 from Singapore (Village Hotel Bugis).



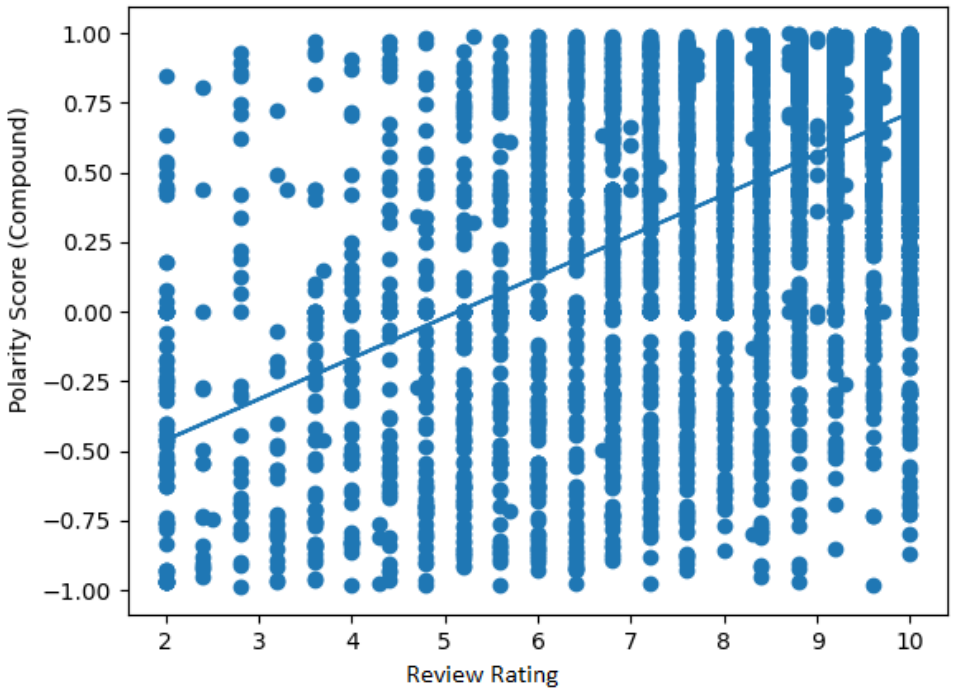
Sentiment Analysis:

We used Python's NLTK package to perform sentiment analysis on all 7120 reviews. This resulted in 4 polarity scores for each review: Negative, Neutral, Positive, and Compound. The Compound score was the score of interest which was used for further analysis.

The average compound polarity score for each hotel was calculated by computing the mean compound polarity of all 100+ reviews for that hotel, and were added to the dataset along with the review rating and overall rating for comparison.

Correlation:

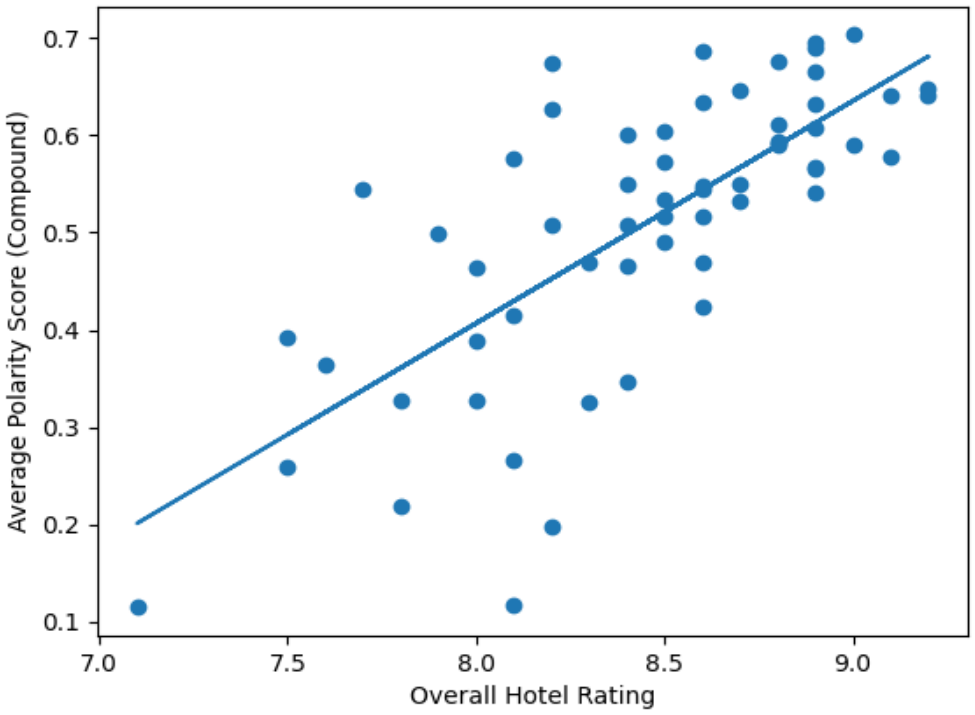
A correlation plot was generated between the review rating and compound polarity score for each review. This plot has 7120 points.



Given the plot density, it was hard to draw conclusions, so another correlation plot was generated between overall rating for a hotel and the average compound polarity score from all reviews for that hotel. This plot has 60 points, one for each hotel.

A regression analysis was performed which **revealed a positive correlation** and the following statistics:

- Covariance: 0.049
- Pearson's correlation: 0.73
- Spearman's correlation: 0.733
- Slope: 0.23



Following observations were made:

- There was a disconnect between polarity score and review rating for some reviews, owing to the nature of NLTK module. Sentiment analysis in NLTK works by tokenizing words, and several reviews had inconsistencies and spelling errors as they were not written by native speakers. Additionally, some reviews were not in English.
- Reviews were scraped chronologically, so for each hotel, only the latest 100+ reviews were scraped. Some hotels had 10000+ reviews. So the reviews scraped did not form a large enough sample size to accurately represent all reviews, and they were not randomly sampled either owing to their chronological order.
- Upon inspecting the reviews for a few hotels, we found that there was a steady drop in the quality of service (which was reflected in the reviews) post the COVID pandemic. The pandemic being a novel event, does not significantly affect the overall rating of the hotel (which is calculated from thousands of reviews). But the recent negative reviews skew the overall polarity score and average review rating for the latest 100+ reviews.

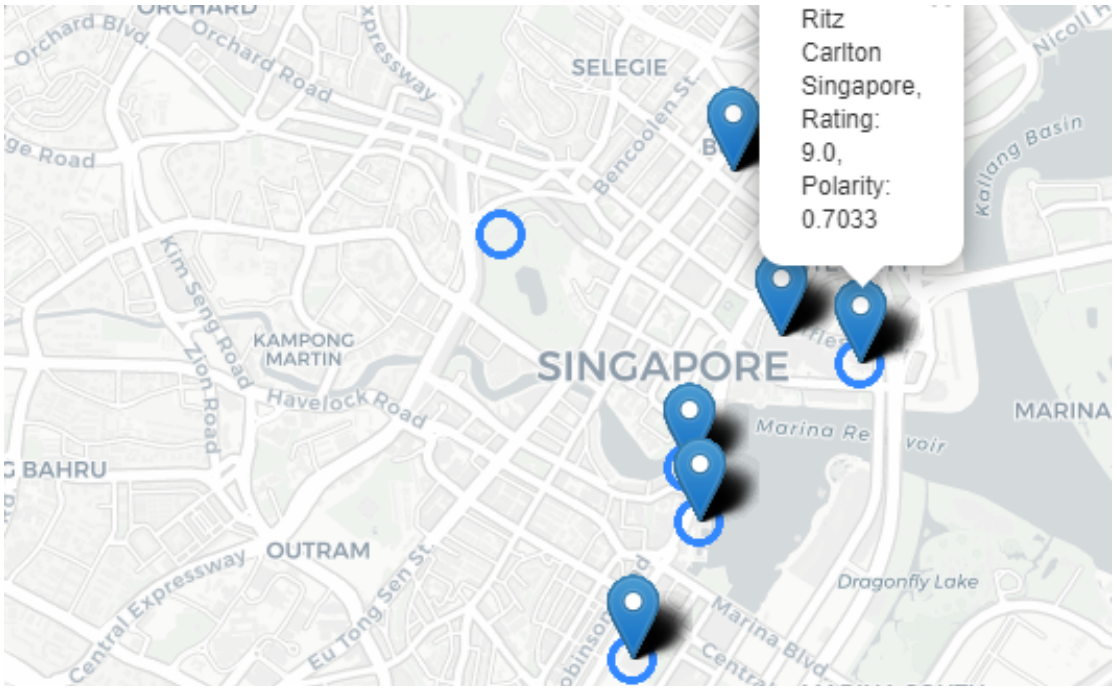
Geospatial Mapping

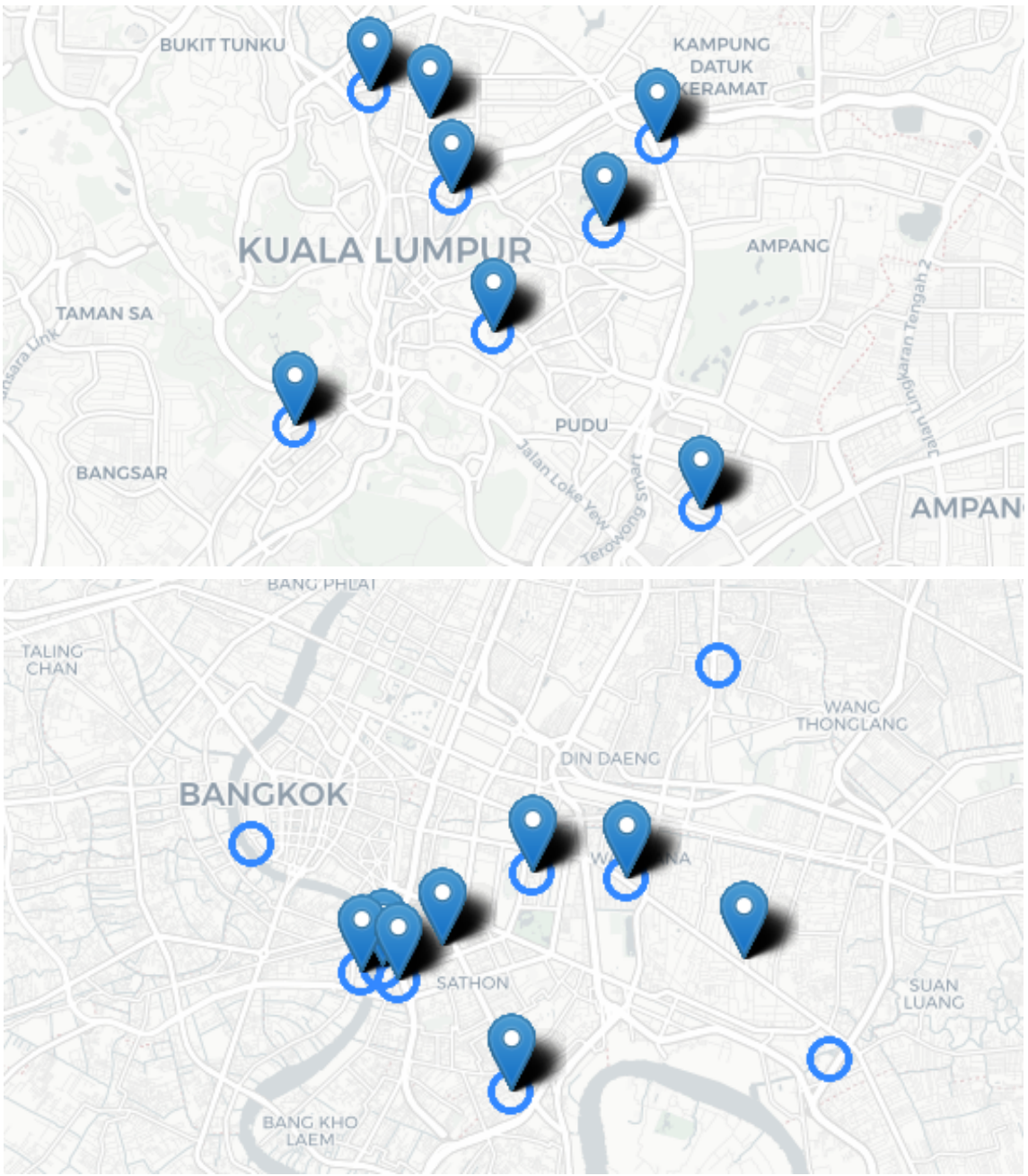
Top 10 hotels in each of the 3 cities (based on average polarity score and overall rating) were mapped. Nominatim module was used to geolocate these hotels (convert physical addresses into geographic coordinates) which were then plotted with Folium.

Circles indicate the top hotels based on highest polarity score.

Pins indicate the top hotels based on highest overall rating.

There were a lot of overlaps, as some hotels were in the top 10 both on the basis of polarity score, and overall rating. This indicated that largely, recent 100+ reviews still accurately represented the hotel. But there were a few hotels that were unique to each leaderboard. For example, *Hotel Fort Canning Singapore* (Rating: 8.6, Polarity: 0.6862) was in the top 10 on the basis of polarity score but not Overall rating, *InterContinental Singapore* (Rating: 8.8, Polarity: 0.5940) was in the top 10 on the basis of overall rating but not polarity score, while the *Ritz Carlton* was in the top 10 for both.





Files:

Accompanying folder **'Data'** contains two dataset files.

Polarity_allrev.csv - All mined reviews for 60 hotels (7120 reviews) with the polarity scores for each review calculated, and concatenated into the same dataset.

Polarity_final.csv - Average polarity score calculated for each hotel.

Accompanying folder **'Code'** contains the code files for Scraping, Sentiment Analysis, Regression Analysis and Correlation, and GeoMapping.