Data Mining and Warehousing: Assignment on Visualization

Report submitted by
Adya Sharma, Dhruv Sabharwal and Shweta Prasad

For this assignment, we scraped data for 20 hotels each from Bangkok, Singapore and Kuala Lumpur, from booking.com. Using Selenium we mined 100 reviews per hotel, performed sentiment analysis on them using NLTK to obtain polarity scores, compared them against the overall ratings of the hotel and obtained the results shown below. We used WEKA to perform clustering and obtain visualizations of the clusters of hotels in a city with similar ratings. We used Mapbox to project and visualize the top-ten rated hotels in each of the three cities.

Data Scraping using Selenium

We used the Selenium WebDriver for Python to perform the scraping. A script was written, which takes in as input a list of city names (three in our case) and outputs a csv file containing 100 reviews per hotel (20) per city. In total we get 100 x 20 x 3 = 6000 reviews. The script takes about 1.5 hours to completely execute. It manipulates the DOM elements to send inputs to the website, navigate through the pages, and extract the final outputs. The code can be found in the file Web_Scraping.ipynb.

Sentiment Analysis using NLTK

We have used NLTK's Vader model to carry out sentiment analysis. Vader's SentimentIntensityAnalyzer() takes in a string and returns a dictionary of scores in each of four categories: negative, neutral, positive and compound (computed by normalizing the previous 3 scores). We wrote a script that takes in as input the csv file containing the reviews and returns as output another csv file with the polarity scores (range [-1,1]), normalized polarity scores (range [0,10] to match booking.com) and the mean of the polarity scores for each hotel. The code can be found in the file sentiment_analysis.ipynb. Before we start discussing the results, it's important to understand certain problems with the vader model and scraped data that lead to deviations between the polarity score assigned by our algorithm and the actual rating provided by the user.

The Vader model that we have used for sentiment analysis has two shortcomings:

1. Since the algorithm is pre-trained it does not fit perfectly to our data. If we had a training set on a distribution similar to the reviews set, we could have gotten better results.
2. The model simply outputs 0 (here 0 means neutral as it is exactly between -1 and 1) for out of vocabulary words. Since the data is not perfectly clean, there are several words that have been misspelt and often customers have used abbreviations (example gr8 for great) to express themselves. This leads to some problems with the scoring.

There are two more inherent drawbacks in our scraped data itself that lead to wrong scores being assigned to the reviews.

1. On booking.com some of the reviewers have put the positives and negatives separately. For example, a review may be: *Positives: Food, view, swimming pool. Negatives: Staff.* When we scrape this data we get one single string: *"Food, view, swimming pool. Staff."*. This review now does not make much sense, even though it did make sense on booking.com.

2. On several occasions the review does not align with the actual rating provided by the user. For example: review *"good service"* is given a rating of 10, while the review *"fantastic service"* is given the rating of 8. These small subjective changes from review to review lead to deviations between the polarity score assigned by our algorithm and the actual rating provided by the user.

After this point all the results are presented city-wise. For this purpose, the report is split into 3 sections: Bangkok, Kuala Lumpur and Singapore.

<u>City 1: Bangkok</u>

The first task was to obtain the overall rating of 20 hotels in Bangkok and cluster hotels based on the ratings. For this, we made use of WEKA, a free ML and visualization software provided by Waikato University. Using simple KMeans Clustering on the data with k = 2, we were able to obtain the following results:

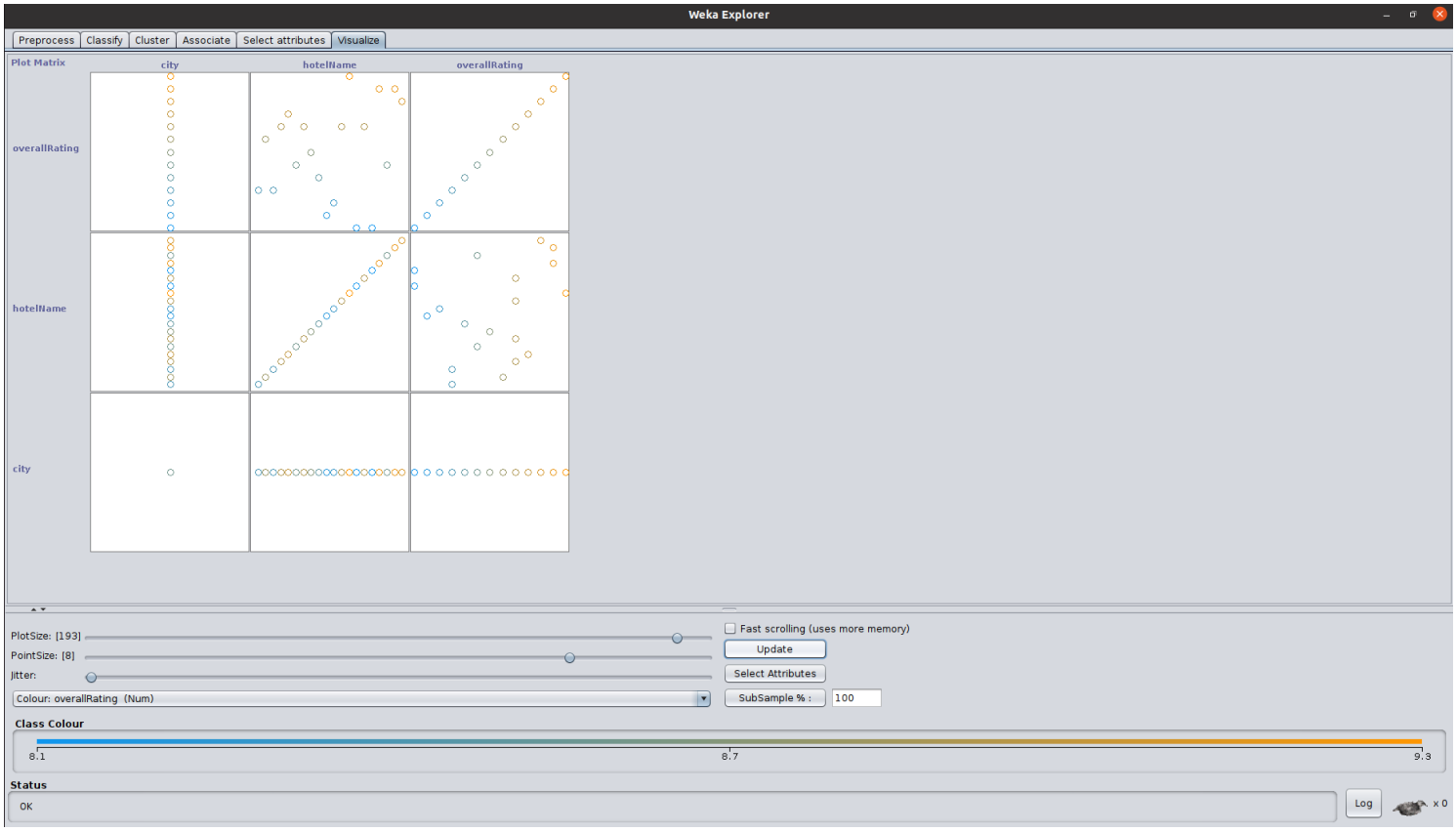
Clusterer output

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Relation:     hotels_with_polarity_score_final - Bangkok-20-overall
Instances:    20
Attributes:   3
              city
              hotelName
              overallRating
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 18.696263227513224

Initial starting points (random):

Cluster 0: Bangkok,'Ibis Bangkok Riverside',8.1
Cluster 1: Bangkok,'Evergreen Place Siam by UHG',8.4

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                                  Cluster#
Attribute                                           Full Data            0                        1
                                                       (20.0)          (6.0)                   (14.0)
=========================================================================================================
city                                                  Bangkok          Bangkok                  Bangkok
hotelName                          Evergreen Place Siam by UHG  Millennium Hilton Bangkok Evergreen Place Siam by UHG
overallRating                                           8.705           8.2667                   8.8929




Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       6 ( 30%)
1      14 ( 70%)
```

Log x 0

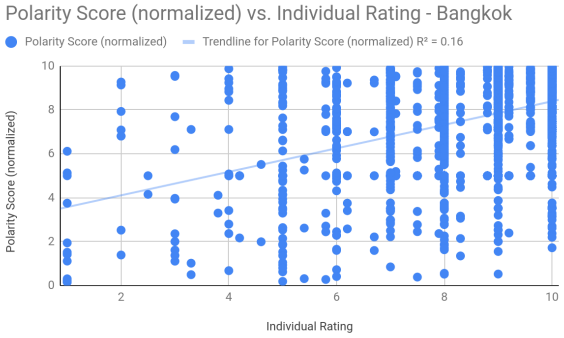
Mean overall rating: 8.7
Min. overall rating: 8.1
Max. overall rating: 9.3
Std. Deviation: 0.373

Cluster 0 had 30% of the hotels, centered at Millenium Bangkok with a rating of 8.2667, and Cluster 1 had 70% of the hotels, centered at Evergreen Place Siam by UHG, with a rating of 8.8929. The coloring scheme in the visualization provided by the software follows a gradient and that does not reflect the two separate clusters that have been identified as a



result of the clustering. However, it gives an idea about which hotels are in what rating range and who is closer to the other, relatively, in terms of ratings. The plot of interest here is the overallRating vs hotelName one, which will cluster hotels based on overall ratings.

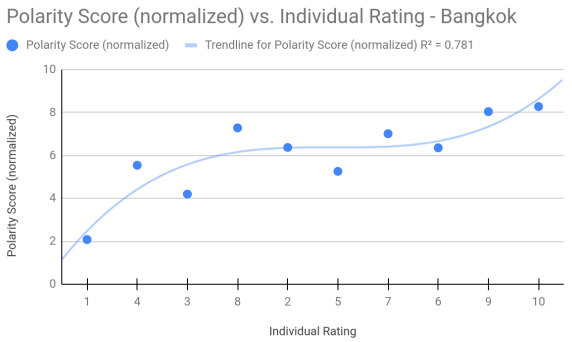
On running NLTK on the scraped reviews for all the 20 hotels, we were able to obtain polarity scores for the most recent reviews given by customers of the hotels. The polarity scores are in the range of -1 to 1 and need to be normalized so that we can compare them against the overall rating of the hotel in question, which was obtained by two means: directly mining them, and taking the mean of all the individual ratings mined for that particular hotel.



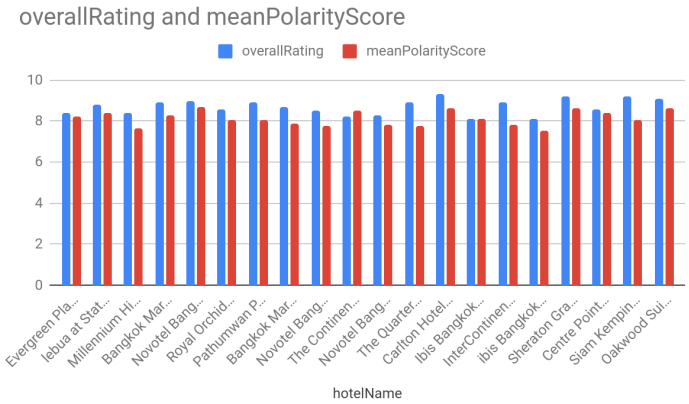
The following graphs explore if there is any correlation between the polarity scores mined from the customer's review and the individual rating provided by the customer.

On the left, we plotted all the 2000 points pertaining to Bangkok. We see that there is not much correlation between the mined polarity of the reviews provided by the customer and the rating given by them,

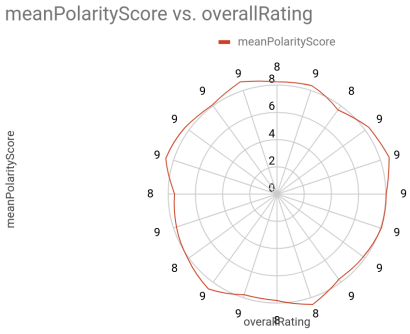
although a linear trend can be observed. Inorder to see if such a trend does in fact exist, we aggregated all points from one hotel and performed the same test to obtain the trend below. The best fit without overfitting was a polynomial of degree 3 and a pearson's correlation coefficient of 0.781 was obtained. The corresponding graph is shown on the left.

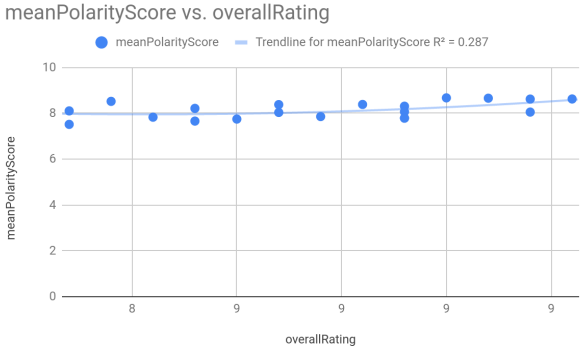


While correlations are a useful measure, visually, the more informative plot in this case are the following two plots, which show how different the overall rating given vs. the mean polarity score is for each of the twenty hotels. Below, we also plot the mean polarity score vs the over -all rating of the hotels and fit a regression polynomial to it.

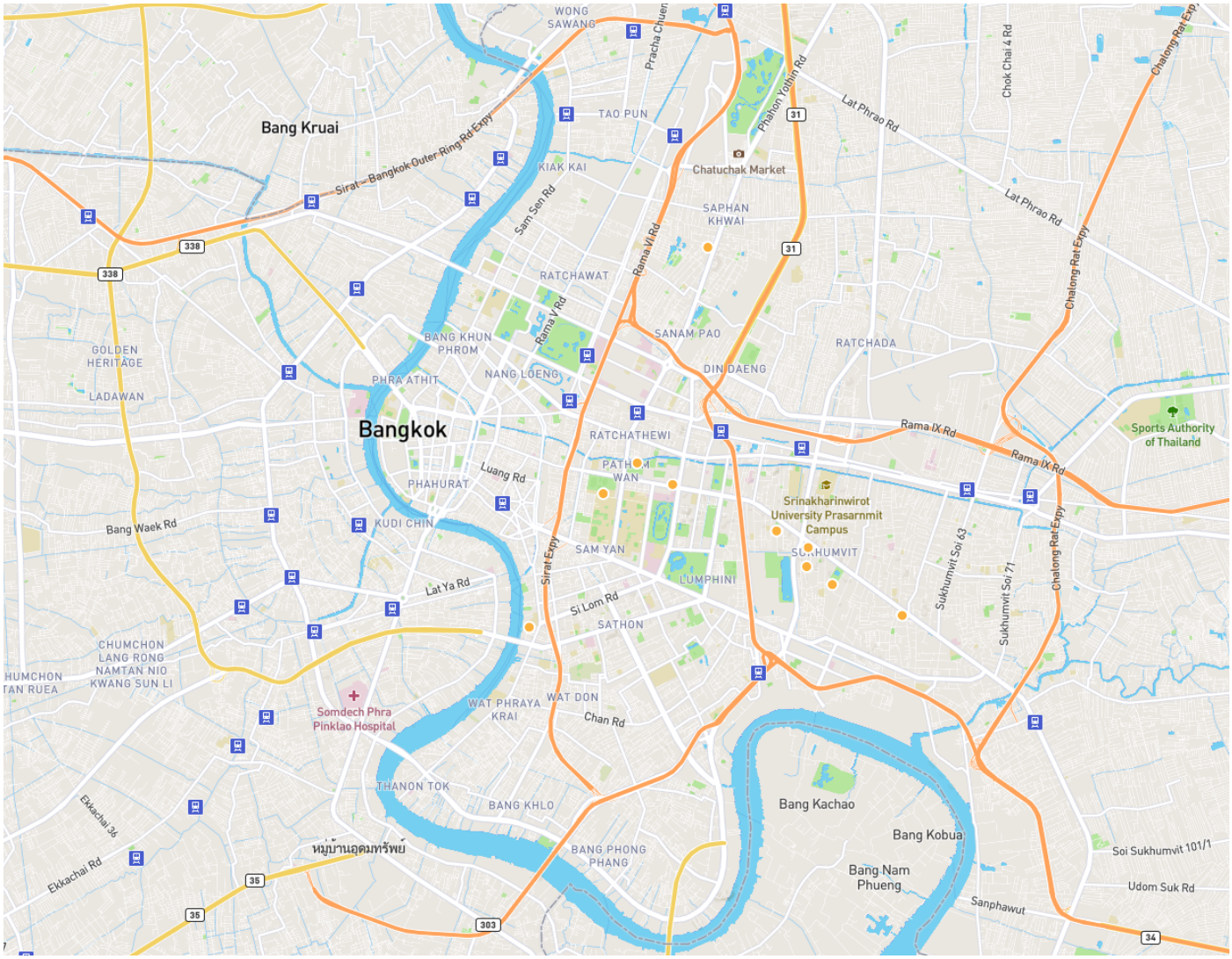


From the plot on the left, it is clear that there is not much of a difference between the aggregates of the polarity score we found versus the mean overall rating we obtained directly from the site. Were there to be overlarge differences in the aforementioned fields, it would be clear that the hotels have meddled with their ratings and they are not to be trusted.



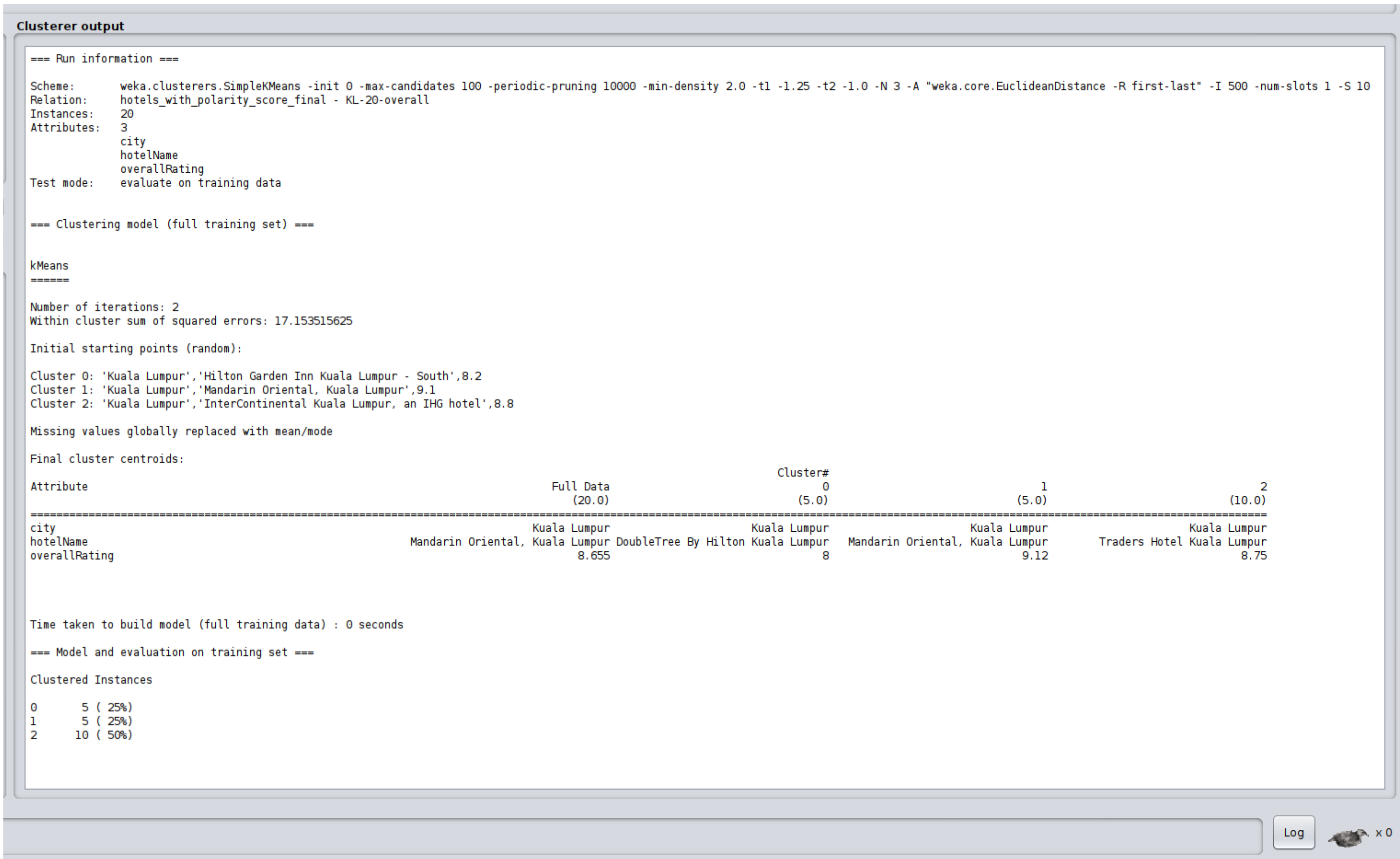


Below, we see the top 10 rated hotels in Bangkok. This visualization was obtained using MapBox's free services. Our data was fed in with the latitude and longitude information for the top ten hotels, and the following plot was obtained. The orange points demarcate the locations of the top ten hotels.


Bang Kruai
Sirat Bangkok Outer Ring Rd Expy
WONG SAWANG
TAO PUN
KIAK KAI
Sam Sen Rd
Pracha Chuen
Phahon Yothin Rd
Chatuchak Market
Lat Phrao Rd
Chok Chai 4 Rd
Chalong Rat Expy
SAPHAN KHWAI
Rama VI Rd
RATCHAWAT
BANG KHUN PHROM
GOLDEN HERITAGE
LADAWAN
PHRA ATHIT
NANG LOENG
SANAM PAO
DIN DAENG
RATCHADA
Bangkok
RATCHATHEWI
PATHUM WAN
Rama IX Rd
Sports Authority of Thailand
Luang Rd
PHAHURAT
KUDI CHIN
Bang Waek Rd
Srinakharinwirot University Prasarnmit Campus
SUKHUMVIT
SAM YAN
LUMPHINI
Sirat Expy
Lat Ya Rd
Si Lom Rd
Sukhumvit Soi 63
Sukhumvit Soi 71
SATHON
CHUMCHON LANG RONG NAMTAN NIO KWANG SUN LI
Somdech Phra Pinklao Hospital
WAT PHRAYA KRAI
WAT DON
Chan Rd
THANON TOK
BANG KHLO
Bang Kachao
Bang Kobua
Bang Nam Phueng
BANG PHONG PHANG
Ekkachai 36
Ekkachai Rd
Soi Sukhumvit 101/1
Udom Suk Rd
Sanphawut
338
31
35
303
34

City 2: Kuala Lumpur

The same procedures as in the case of Bangkok were followed for Kuala Lumpur.



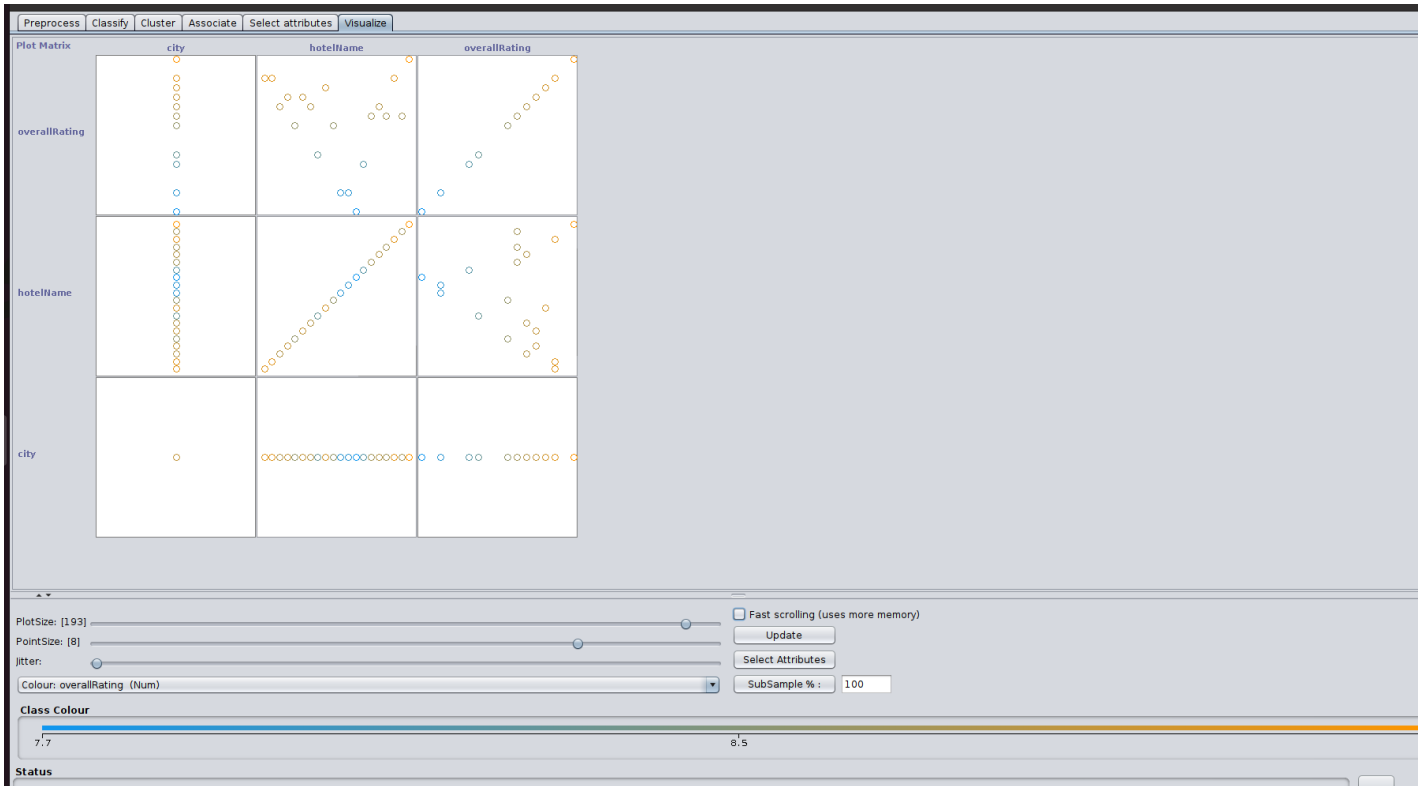
Clusterer output

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Relation:     hotels_with_polarity_score_final - KL-20-overall
Instances:    20
Attributes:   3
              city
              hotelName
              overallRating
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 17.153515625

Initial starting points (random):

Cluster 0: 'Kuala Lumpur','Hilton Garden Inn Kuala Lumpur - South',8.2
Cluster 1: 'Kuala Lumpur','Mandarin Oriental, Kuala Lumpur',9.1
Cluster 2: 'Kuala Lumpur','InterContinental Kuala Lumpur, an IHG hotel',8.8

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                                   Cluster#
Attribute                                   Full Data                     0                              1                          2
                                               (20.0)                 (5.0)                          (5.0)                     (10.0)
=====================================================================================================================================
city                                     Kuala Lumpur          Kuala Lumpur                   Kuala Lumpur               Kuala Lumpur
hotelName             Mandarin Oriental, Kuala Lumpur DoubleTree By Hilton Kuala Lumpur   Mandarin Oriental, Kuala Lumpur       Traders Hotel Kuala Lumpur
overallRating                                   8.655                     8                           9.12                       8.75



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       5 ( 25%)
1       5 ( 25%)
2      10 ( 50%)
```

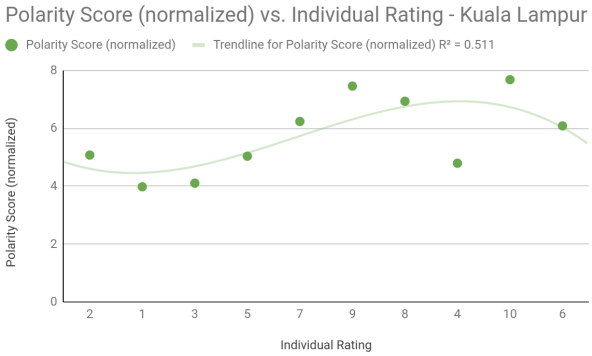
Log x 0

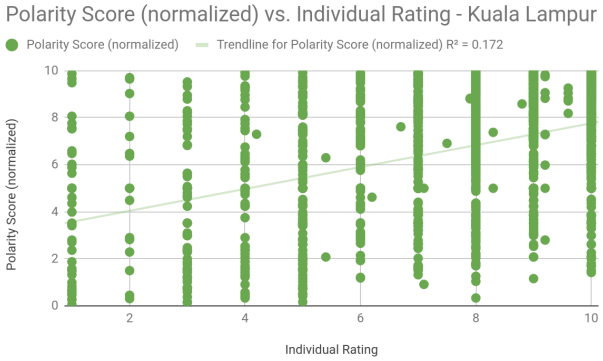
The following (and above) are the results we obtained from running KMeans Clustering on WEKA.
Mean overall rating: 8.65
Min. overall rating: 7.7
Max. overall rating: 9.3
Std. Deviation: 0.442

This time, we used k=3 since preliminary analysis indicated that the ratings could be broadly split into three groups.
Cluster 0 has 5 hotels with DoubleTree by Hilton at the center with a rating of 8; Cluster 1 has 5 hotels with Mandarin Oriental at the center with a rating of 9.12 and cluster 2 has 10 hotels with Traders Hotel at the cluster center, with a rating of 8.75. This indicates that a larger number of hotels are in the 8.75 rating range. Here are the plots we obtained from WEKA:



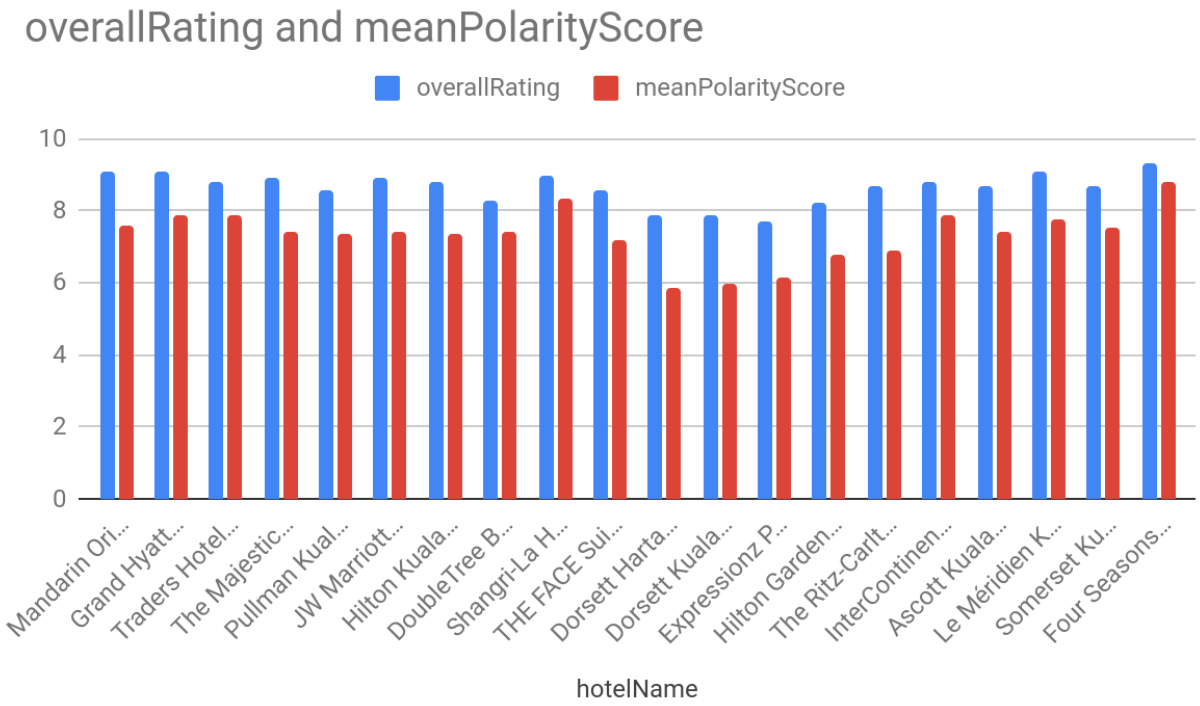
The graphs below plot the individual reviews' polarity score obtained using sentiment analysis vs the actual individual rating given by the customers. We observe that there is no good correlation observable between them, but when plotted as aggregates (the plot on the left), a pattern emerges: the best fit polynomial was a degree 3 polynomial with a pearson's



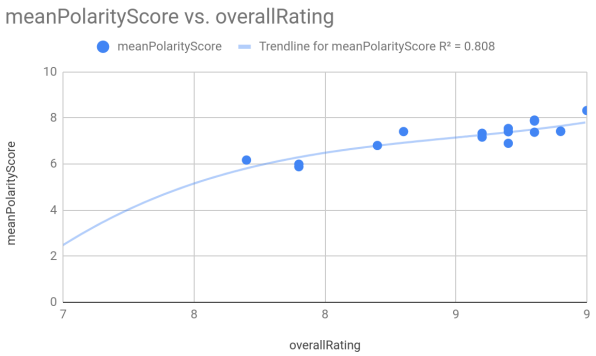


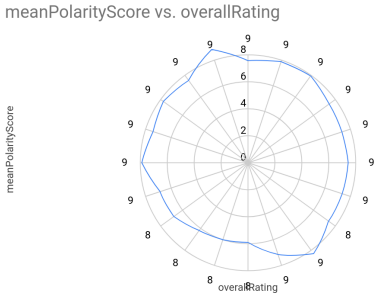
correlation coefficient of 0.511. This correlation is worse than the one we saw for Bangkok.

We can thus expect that there will be a large difference in the overall ratings vs mean polarity score for each hotel. That is exactly what we see below. This should indicate that there is either something faulty with NLTK's analysis, or the hotels themselves have meddled with the ratings, therefore giving rise to the gross misrepresentation of satisfaction as construable from reviews.

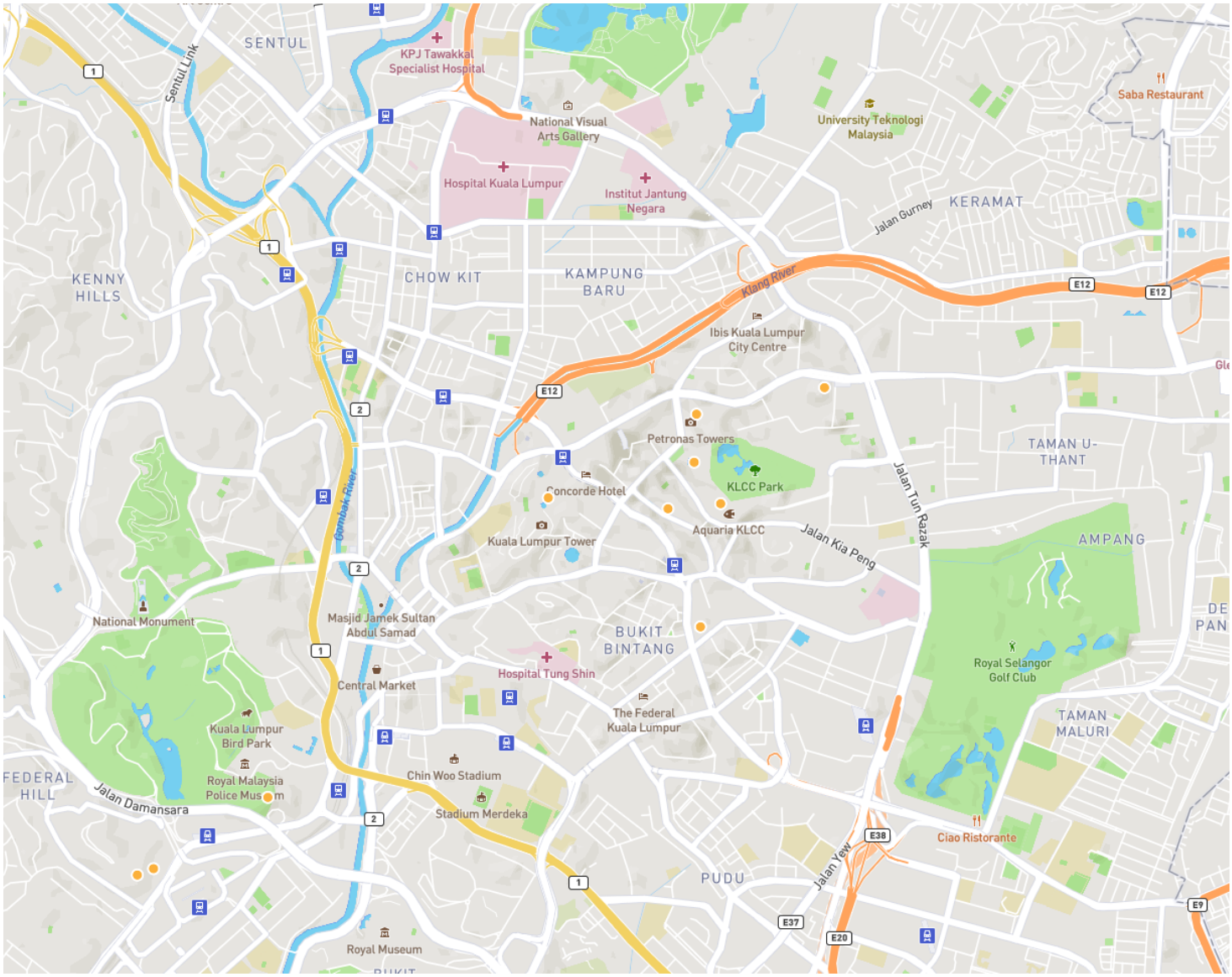


Below, we see the correlation plots for mean polarity score vs overall rating of the 20 hotels.





Below, we plot the top 10 hotels in Kuala Lumpur using MapBox (orange points).


SENTUL
Sentul Link
KPJ Tawakkal Specialist Hospital
National Visual Arts Gallery
Hospital Kuala Lumpur
Institut Jantung Negara
University Teknologi Malaysia
Saba Restaurant
Jalan Gurney
KERAMAT
KENNY HILLS
CHOW KIT
KAMPUNG BARU
Klang River
Ibis Kuala Lumpur City Centre
E12
Petronas Towers
KLCC Park
TAMAN U-THANT
Concorde Hotel
Gombak River
Kuala Lumpur Tower
Aquaria KLCC
Jalan Kia Peng
Jalan Tun Razak
AMPANG
National Monument
Masjid Jamek Sultan Abdul Samad
BUKIT BINTANG
Hospital Tung Shin
Central Market
Royal Selangor Golf Club
The Federal Kuala Lumpur
TAMAN MALURI
Kuala Lumpur Bird Park
FEDERAL HILL
Royal Malaysia Police Museum
Chin Woo Stadium
Jalan Damansara
Stadium Merdeka
Ciao Ristorante
PUDU
Jalan Yew
E38
E37
E20
E9
Royal Museum

City 3: Singapore

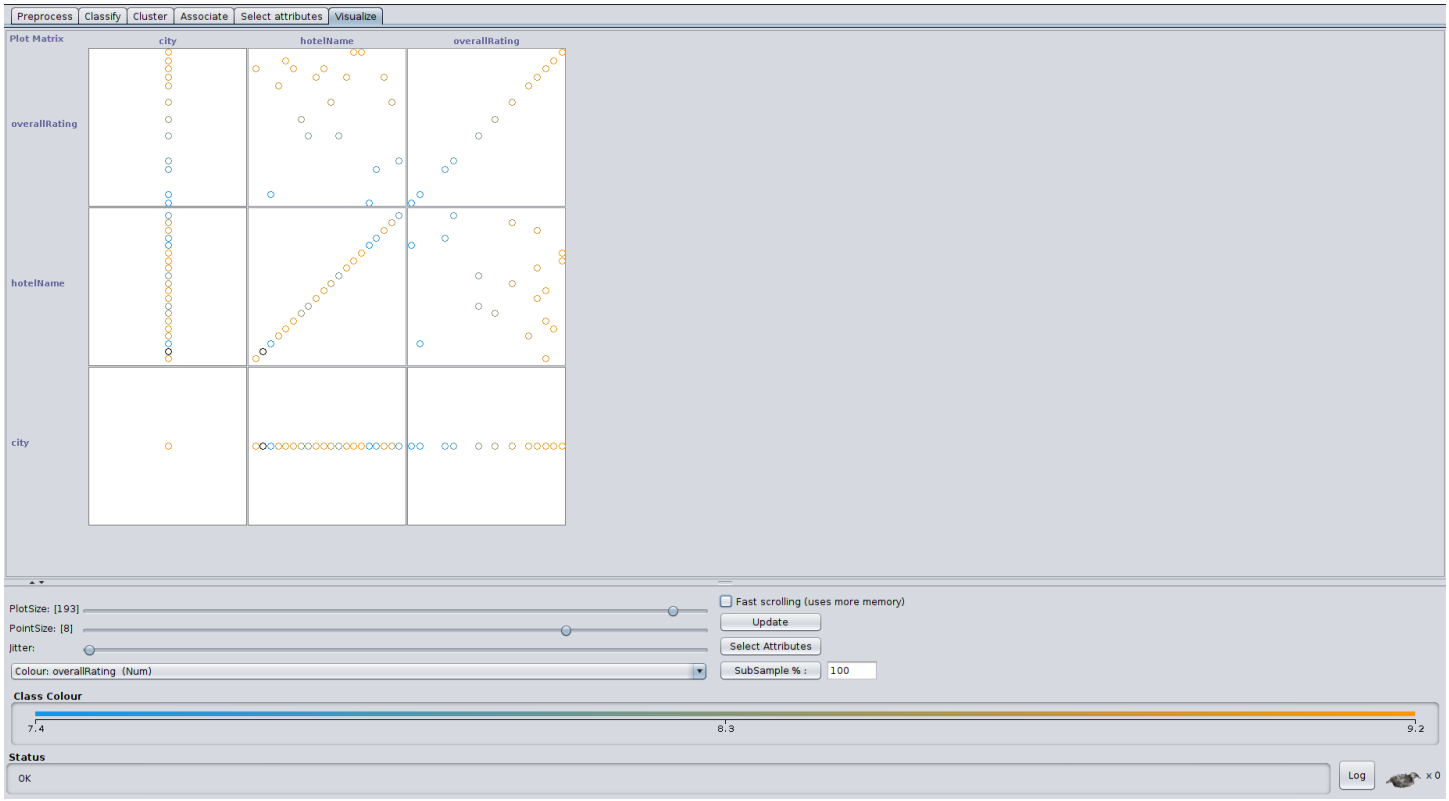
Following the exact same procedure as above, we get the following results from clustering hotels in Singapore with similar ratings using WEKA. Preliminary analysis showed that 2 groups of hotels existed, so clustering was done with k = 2.

Clusterer output

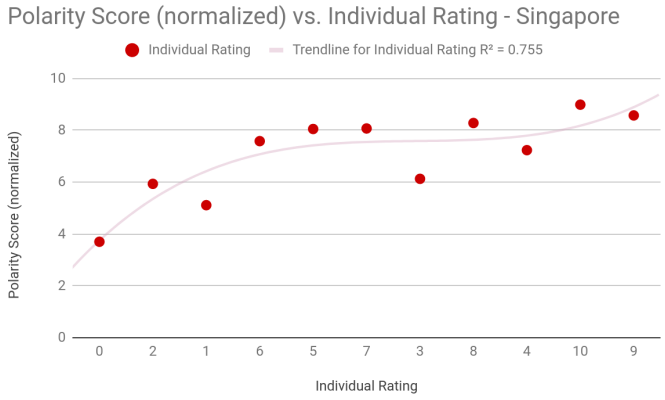
```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Relation:     hotels_with_polarity_score_final - Singapore-20-overall
Instances:    20
Attributes:   3
              city
              hotelName
              overallRating
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 6
Within cluster sum of squared errors: 18.74903959052928

Initial starting points (random):

Cluster 0: Singapore,'The Ritz-Carlton, Millenia Singapore (SG Clean)',9.2
Cluster 1: Singapore,'Marina Bay Sands (SG Clean)',9

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                                                         Cluster#
Attribute                                                         Full Data                     0                     1
                                                                     (20.0)                (12.0)                 (8.0)
========================================================================================================================
city                                                              Singapore             Singapore             Singapore
hotelName                                          Marina Bay Sands (SG Clean)   Shangri-La Marina Bay Sands (SG Clean)
overallRating                                                        8.5579                8.8965                  8.05




Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      12 ( 60%)
1       8 ( 40%)
```

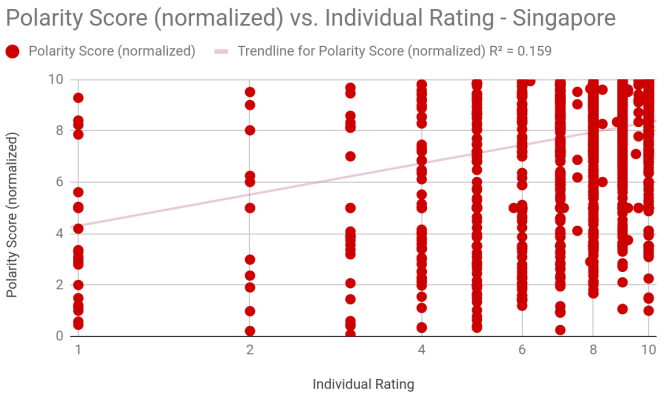
Mean overall rating: 8.558
Min. overall rating: 7.4
Max. overall rating: 9.2
Std. Deviation: 0.571

Here are the visualizations WEKA provided after clustering:

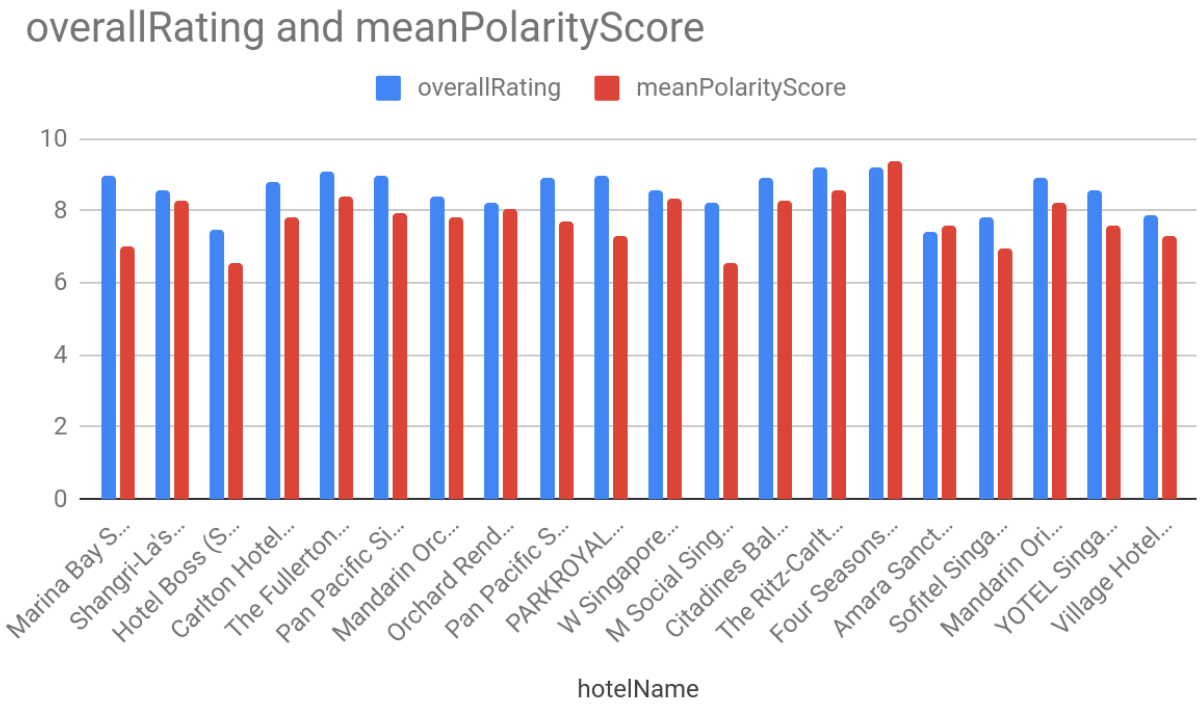


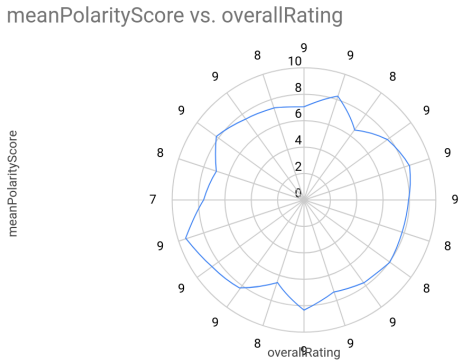
The graphs below plot the individual reviews' polarity score obtained using sentiment analysis vs the actual individual rating given by the customers. We observe that there is no good correlation observable between them, but when plotted as aggregates (the plot on the left), a pattern emerges: the best fit polynomial was a degree 3 polynomial with a pearson's coefficient of 0.755. However, we do see that most reviews are on the upper right corner, implying that the polarity scores do reflect the individual rating the higher they get.

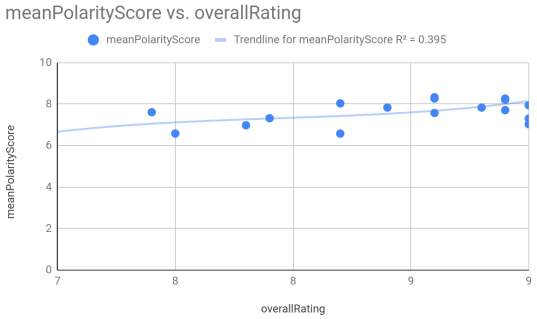




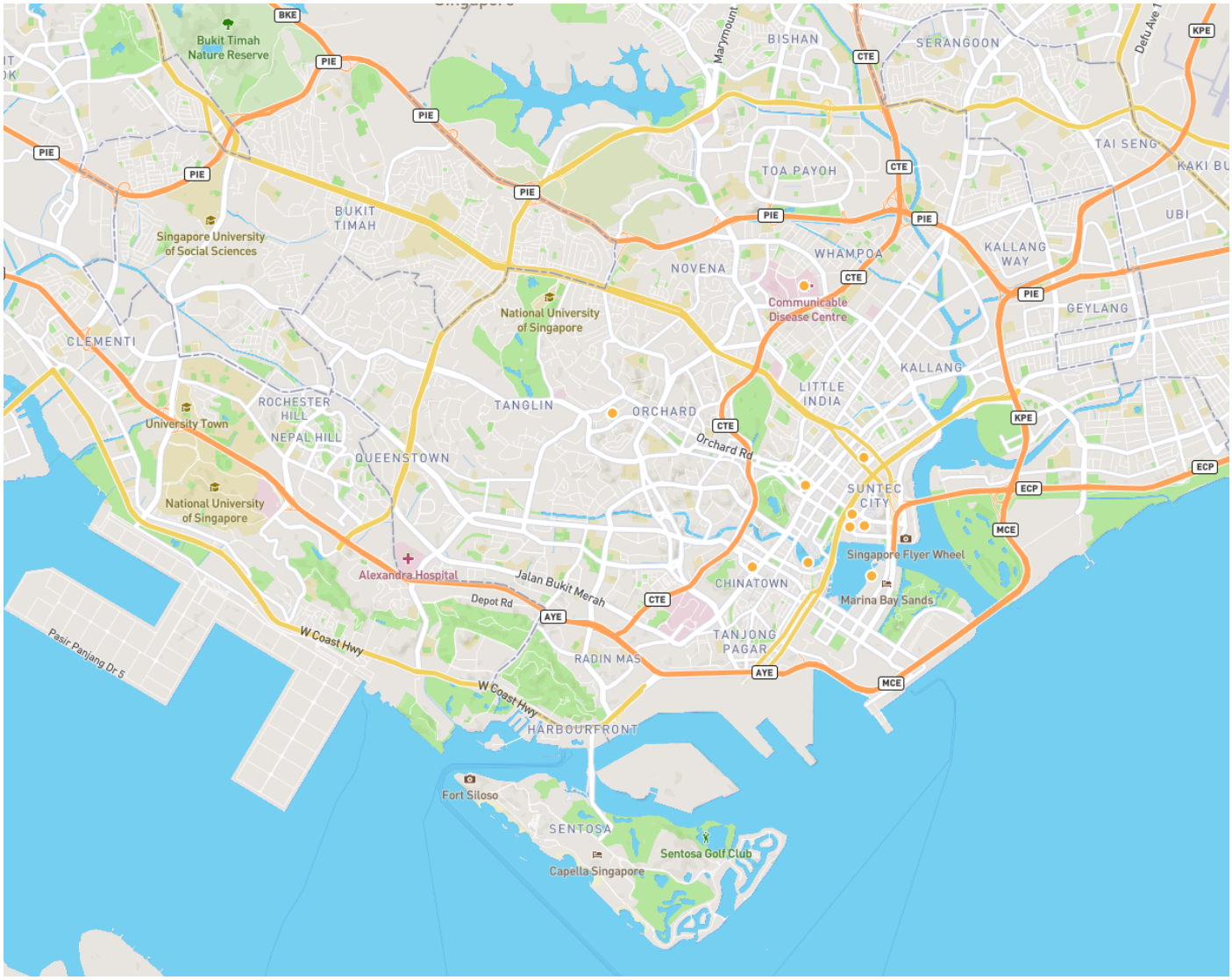
The following graphs talk about the mean polarity scores versus the overall rating of each of the 20 hotels in Singapore. Thi







Top 10 hotels in Singapore plotted using MapBox. The orange points indicate the locations of the 10 hotels.


Bukit Timah Nature Reserve
BKE
PIE
BISHAN
Marymount
SERANGOON
CTE
KPE
Defu Ave 1
TAI SENG
KAKI BU
TOA PAYOH
UBI
BUKIT TIMAH
Singapore University of Social Sciences
NOVENA
WHAMPOA
KALLANG WAY
Communicable Disease Centre
GEYLANG
National University of Singapore
CLEMENTI
KALLANG
LITTLE INDIA
ROCHESTER HILL
TANGLIN
ORCHARD
University Town
NEPAL HILL
QUEENSTOWN
Orchard Rd
SUNTEC CITY
ECP
National University of Singapore
MCE
Singapore Flyer Wheel
Alexandra Hospital
Jalan Bukit Merah
CHINATOWN
Depot Rd
Marina Bay Sands
AYE
W Coast Hwy
TANJONG PAGAR
Pasir Panjang Dr 5
RADIN MAS
HARBOURFRONT
Fort Siloso
SENTOSA
Sentosa Golf Club
Capella Singapore
