HotelGenie - Hotel Recommendation System

Anh The Nguyen
Arizona State University
United States
atnguy37@asu.edu

Deep Patel
Arizona State University
United States
dcpatel2@asu.edu

Hardik Shah
Arizona State University
United States
hrshah5@asu.edu

Jaykumar Vaghasiya
Arizona State University
United States
jvaghasi@asu.edu

Maunil Vyas
Arizona State University
United States
mrvyas@asu.edu

ABSTRACT

With the need for Visualization based recommendation systems for hotels, we are providing you, Hotel Genie, an intuitive and interactive visualization based, Hotel Recommendation System. We are providing a location and class of the hotel based filtering to provide a user with multiple hotels to choose from. Secondly, we are providing two time-series visualization of the hotels. First is the Heat map visualization which shows how all the 7 characteristics have behaved and changed over the period of time. Second is the line chart used to show how the overall rating of the hotel has changed over a period of time. Next, we present user top hotels in multiple criteria to compare the selected hotel above. With the selection of any hotel, a radar chart is used to show the comparison of all the 7 characteristics. Lastly, we provide the user with the sentiment analysis of the reviews based on word cloud and top reviews. We give top 5 reviews of the hotel out of more than 500+ reviews which also contains a similar count. This way we provide user just 5 reviews to get an overview of more than 500 reviews. Thus our hotel recommendation system gives an extensive experience to the user in their hotel search journey. Link: https://hotelgenie.ml/, Video Link: https://youtu.be/mlsq3dT5ys0

1 INTRODUCTION

According to the World Economic Forum, the world is producing 2.5 quintillion bytes of data every day, and 90% of all data has been created in the last two years. Having this much of data it becomes significantly hard to manage and make sense out of it. Ideally, it is now impossible for a single person to go through with the entire data and try to make sense out of it. Data proliferation can be managed as part of the data science process, which includes data visualization. Data visualization has become an indispensable part of the business world and ever-increasing part of managing our daily lives. The emergence of society interested in, and driven by the data has created a premium for quality and engaging data visualizations.

With the booming demand for qualitative and interactive visualization for diverse business problems. Here, we have tried to build a visualization system for Hotel Recommendation named Hotel Genie. Mainly our aim is to solve the four questions mentioned in figure 1 through our visualization.

For this work, we have used the dataset [1] form the TripAdvisor. The dataset consists of 878561 reviews from 4333 hotels crawled from. Figure 2 and 3 depicts the Json data format.

Figure 1: Queries our visulization addresses


```
{
  "hotel_class":4.0,
  "region_id":60763,
  "url":"http://www.tripadvisor.com/Hotel_Review-g60763-d113317-
  Reviews-Casablanca_Hotel_Times_Square-New_York_City_New_York.html",
  "phone":"",
  "details":null,
  "address":{
    "region":"NY",
    "street-address":"147 West 43rd Street",
    "postal-code":"10036",
    "locality":"New York City"
  },
  "type":"hotel",
  "id":113317,
  "name":"Casablanca Hotel Times Square"
}
```


Figure 2: Data Instance for the hotel information

In the following section will talk about the Motivation behind this work. Section 3 talks about the implementation details followed by section 4 that provides the reasoning behind the decisions we took in the implementation part. Afterwards, section 5 proposes the evaluation plan for validating our system and section 6 leads to the discussion and future work.

2 MOTIVATION

Hotel Recommendation System is the need of an hour. Today there are hundreds of hotels providing world-class facility and extensive experience to the hotel stayers. With hundreds of hotels to choose

```
{
  "ratings": {
    "service":5.0,
    "cleanliness":5.0,
    "overall":5.0,
    "value":5.0,
    "location":5.0,
    "sleep_quality":5.0,
    "rooms":5.0
  },
  "title":"\u201cTruly is \"Jewel of the Upper Wets Side\"\u201d",
  "text":"Stayed in a king suite for 11 nights and yes it cots us a
  bit but we were happy with the standard of room, the location and
  the friendliness of the staff. Our room was on the 20th floor
  overlooking Broadway and the madhouse of the Fairway Market. Room
  was quite with no noise evident from the hallway or adjoining
  rooms. It was great to be able to open windows when we craved fresh
  rather than heated air. The beds, including the fold out sofa bed,
  were comfortable and the rooms were cleaned well. Wi-fi access
  worked like a dream with only one connectivity issue on our first
  night and this was promptly responded to with a call from the
  service provider to ensure that all was well. The location close to
  the 72nd Street subway station is great and the complimentary
  umbrellas on the drizzly days were greatly appreciated. It is
  fabulous to have the kitchen with cooking facilities and the access
  to a whole range of fresh foods directly across the road at
  Fairway.\nThis is the second time that members of the party have
  stayed at the Beacon and it will certainly be our hotel of choice
  for future visits.",
  "author": {
    "username":"Papa_Panda",
    "num_cities":22,
    "num_helpful_votes":12,
    "num_reviews":29,
    "num_type_reviews":24,
    "id":"8C0B42FF3C0FA366A21CFD785302A032",
    "location":"Gold Coast"
  },
  "date_stayed":"December 2012",
  "offering_id":93338,
  "num_helpful_votes":0,
  "date":"December 17, 2012",
  "id":147643103,
  "via_mobile":false
}
```


Figure 3: Data Instance for the hotel review

from, a user finds himself in confusion state. Also, each hotel is known for its services and each user has their own preference towards it. This gives us an interesting problem statement to solve. There is a need for a hotel recommendation system which will give the user an upper hand in visualizing the best hotel and finalizing his/her choice.

Although there are many hotel recommendation systems available in the market but we have designed our project in such a way that it makes the user to deep dive into it. Many hotel recommendation systems provide map-based visualization and provide you the cost of the hotel, but we provide you the insights that will help the user to get a clear idea about how has the hotel has behaved and changed over the period of time. We are presenting a heat map and time series visualization to visualize the hotel graph over the period of the time.

Also in today‘s world, sentiments have become the primary area of research and many recommendation systems works on it. But there are very few Hotels Recommendation system that will provide sentiment based analysis. We have provided word cloud of the hotel reviews to give an idea about the number of words used in the reviews and whether they are positive or negative. Moreover, we are providing top 5 user comments out of more than 500+ comments of the hotel and these 5 comments have the count of similar comments to it that will save a lot of time of the user. User can read just 5 comments to get an idea of all 500 comments.

Thus to solve the problem of the need for interactive and intuitive visualization, we are providing you our project Hotel Recommendation system named, Hotel Genie.

3 VISUALIZATION DESIGN (IMPLEMENTATION)

3.1 Data Cleaning

The mentioned dataset [1] has mainly two parts named - offering set and review set as mentioned earlier. The major requirement in the cleaning and processing for the offering set was related to the missing entries for the hotels such as its star ratings, invalid or missing address details, missing or negative hotel category values etc. Largely as a part of the data cleaning process, we tried to avoid these entries. However, most issues related to the missing or corrupted hotel address is solved by Google Map API.

Secondly, for the review set, the main challenge was to map the reviews given by the users with the hotels mentioned in the offering dataset and filter out the remaining ones. Here, in terms of the cleaning aspect, the dataset contains missing reviews and missing entries in the different rating categories. Moreover, some of the reviews are in the french language. So, we have also filter those things out.

Following the above-mentioned strategies, we were able to get a proper dataset. Afterwards, as per the requirements of our different visualization models and with a view of making the website efficient, we try to organize our dataset mainly in the form of CSV files having processed data in it.

Considering the implementation part, we developed our entire system using the following libraries basic js functionality, D3 version 4, jQuery 3.3.1, Bootstrap 4.3.1, Font awesome, Chernoff Face js, Cloud to js, NLTK, and Allen NLP DecAtt Model.

3.2 Location and Class based hotel filtering

For location-based visualization, we have used the *Google Maps API* for Javascript. we made two filters for the map. First one is to provide a location filter based on the city. We have twenty-five cities data available. The second filter is based on hotel class, that provides the functionality to show only hotels of a specific class from 1 star to 5 star based on the selected city. Additionally, a button for showing hotels of all classes on the map is added. We used bootstrap buttons and column layout to format it. Star icons from font awesome icon were used to create a class filters.

On Map, Markers were created for each hotel for the selected city and class. We selected a customized marker icon for our representation. Furthermore, the marker‘s color was decided from the user rating of the hotel, light color marker represents less user rating and dark color marker represents higher user rating. So, this color gradient helps the user to differentiate hotel rating with a quick glance. On hovering of this marker, the info window is showed which includes more details about hovered marker. For example, hotel name, hotel address, average user rating, no of reviews and hotel class. On click of this marker the time series analysis of hotel is generated.

To fast access the data and loading marker quickly, we created a JSON file with the city as a key. Every city has data in the array, where each hotel data for that city is stored as an element of the array. We also stored hotel Id, latitude, longitude, class, average user rating and HTML of info window for each hotel. Because we

are directly storing info windows HTML it is faster than creating HTML for info window on run time. To give a personalized touch to the map visualization we added a custom theme to the map.

To locate the hotels on the map, their latitude and longitude were needed, however, they were not provided from the original data-set, therefore we used *OpenCage Geocoding API* to fetch those coordinates for each hotel in the data. This process took considerable time as the coordinates for each of the 4000 hotels were to be fetched from online service. Figure 4 depicts the map based visulization.

3.3 Time series Visualization

To depicts the change in the ratings for different categories or how the hotel's reputation has changed over the time span. We tried to use two visualizations, *Heat map* and *Line chart*. For the Heat Map, we are trying to deliver information about how different aspects of hotel changes over time as per its visitors ratings. For instance, sleeping experience, Room Quality Cleanliness of the Place etc. We have taken seven such characteristics of the hotel. On the other hand, the line chart deliveries information about how the overall ratings of the hotel changes over the time span. Figure 5 depicts the mentioned visualization.

Firstly for the Heat map, we are using a color gradient scheme of light purple to dark purple, where the color is adjusted according to the ratings from 1 to 5. The grey values represent that there is no data available for that particular year. Hovering on the Heat map will help the user to see the average ratings of the particular hotel feature for a specific year. Contrarily the line chart has been created with similar ideas to showcase how the hotel perform overall over the time span. Hovering on the line chart will present the overall average rating for that hotel in a given year.

3.4 Comparison based Visualization

For the selected city, we provided 4 categories like, most clean, best in service satisfaction, best in room quality, and best surrounding. For each of this category, we provide the top 4 hotels in this category that fulfill this criterion for the selected city. Best hotel for the specific category is determined by the highest average rating for that specific criteria. In this way, we provide 16 hotels that user can select and compare with their chosen hotel from the map. To compare two hotels different rating we used radar chart.

On each axis of the radar chart, seven different rating criteria like room quality, best service, cleanliness, sleep quality, surroundings, value for money and overall ratings are mapped. Rating starts from 0 and goes till 5, for all seven criteria, which makes a circle. On this circle, two hotels are compared with their ratings. A *Radar chart* is transparent so different rating of two hotels can be easily compared visually. On the Radar chart, hotel selected from the map has purple color while hotel selected from 4 categories has a sky-blue color. Figure 6 depicts the comparison based visualization

3.5 Text Mining

We believe that text mining is the core part of our Project and we intensively investigated all the aspects of it to make the user experience fruitful. Here, we have used Chernoff Face [5] to showcase the overall review sentiments, top 5 comments that summaries mostly the reviews provided by the visitors and a word cloud.

3.5.1 *Chernoff Face, Overall Review Sentiments :*

Figure 7 depicts the mentioned visualization. We have used NLTK [7] for getting the sentiments of the comments. The library provides the polarity scores from the range -1 to 1 for the sentiments. We kept a threshold that if a sentence has a polarity score more than 0, we consider it as a positive comment and on the other hand, if it has a score below 0, will admit it as a negative comment. With these classes, we tried to stack all the comments of a specific hotel and find its avg sentiment. Afterwards, we modify the Chernoff face according to average sentiment score. Where a dark Green color with a happy face represents that the reviews are more than 90% positive, light green color with a slight happy face represents the reviews are in the range from 50% to 90% positive range, and for the remaining interval of average sentiments we are providing a red, sad face.

3.5.2 *Finding top Comments :*

Here, we tried to seek only those comments that have the most similar other comments, and out of this set, we find the top five. For the implementation part, we have used two methods to make our comment extraction algorithm robust. Initially, we used term frequency and inverse term frequency [8] to find the similarity score among comments for a specific hotel by creating a gigantic correlation matrix. Thereafter, we seek similarity scores above 0.5 in the matrix. For example, if cell (i,j) has a score of 0.56, it means that comment i and j has a 56% similarity. Once the similarity matrix is computed, we sort comments based on how many number of similar comments they have.

Following the above-mentioned idea, we get a set of comments that are most similar to others. This was the first part of the comment extraction algorithm. Later, we used Natural Langauge Inference model named Decomposable Attention (DecAtt) [6] for getting entailment scores between comments. The proposed DecAtt Neural Network model is trained for classifying three classes Entail, Neutral and Contradict between two comments. So, we used the computed similar comments from the first part and verified the similarity by forward passing those comments in this trained model to get entailment scores. Note that we have given more priority to those comments that perform better in both similarity tasks.

After extracting these comments, we still believe that we have not perfectly solved the user demand. Therefore, with these top 5 comments, our visualization is also showing the time frame of when a comment is posted, the user profile who posted the comment, the user's helpful votes and the comment's helpful votes. Figure 8 shows the visualization for this part.

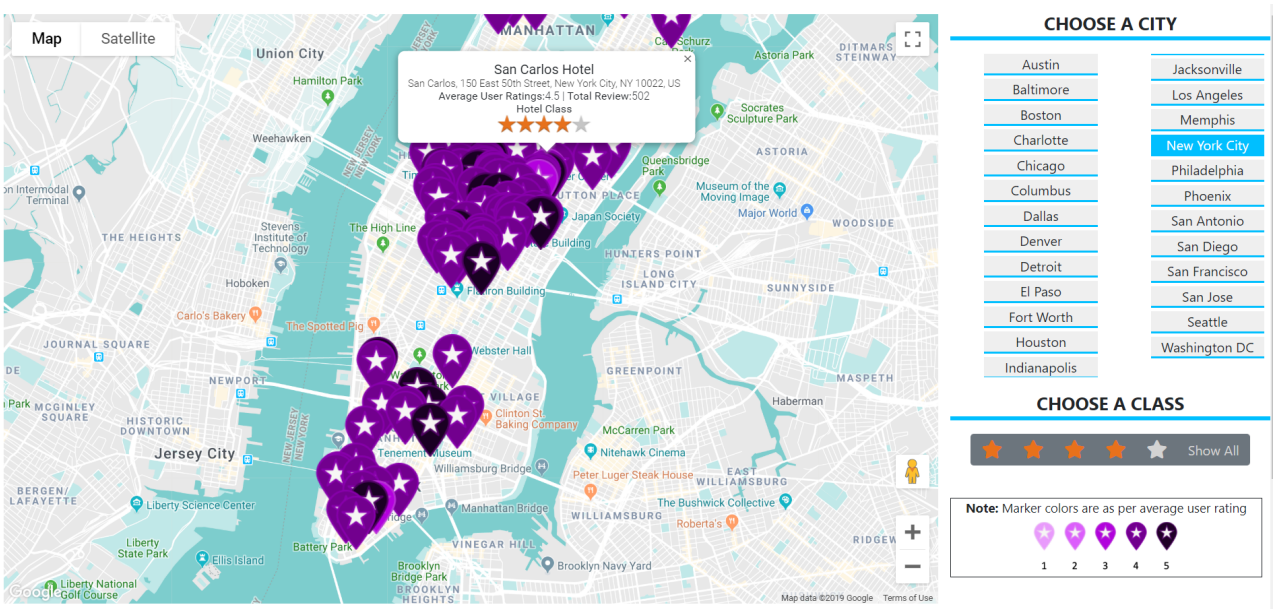

Figure 4: Map Visualization: Location and Class based hotel filtering

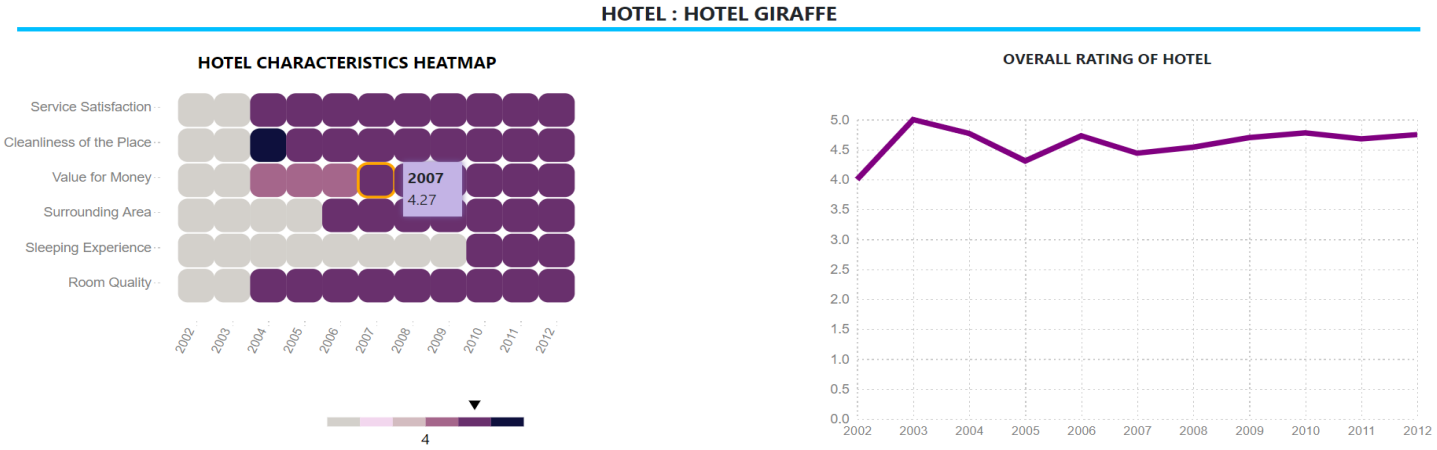

Figure 5: Time Series Visulization, Heat Map & Line Chart

3.5.3 *Word Cloud:*

Apart from extracting top comments and overall sentiment, we are also providing word cloud computed from the comments given to a specific hotel. For computing the positive and negative words, we used NLTK sentiment intensity analyzer, which gives polarity score to each word in comments. We kept a threshold zero here means If the word has a polarity score below zero, we classify it as a negative word and on the other hand, a word having a positive score will be classified as a positive word. Based on the occurrence of the word, we dynamically provide the word size. The green color is used for positive words and the red color is used for negative words. Figure 9 depicts the visualization.

4 METHODOLOGY

Our motto behind designing this visualization is to improve the user experience for hotel searching. Mainly, we are trying to solve the four major questions that a user has while searching for a hotel as mentioned earlier in the introduction section. We believe that our visualization has the capability of solving those questions. We extensively tried to provide information to the user from different dimensions and in our opinion, everything comes to so nicely that it helps the user to have trust in what we are showing.

Here, we try to reason how our data visualization can be used to solve the mentioned questions one by one.

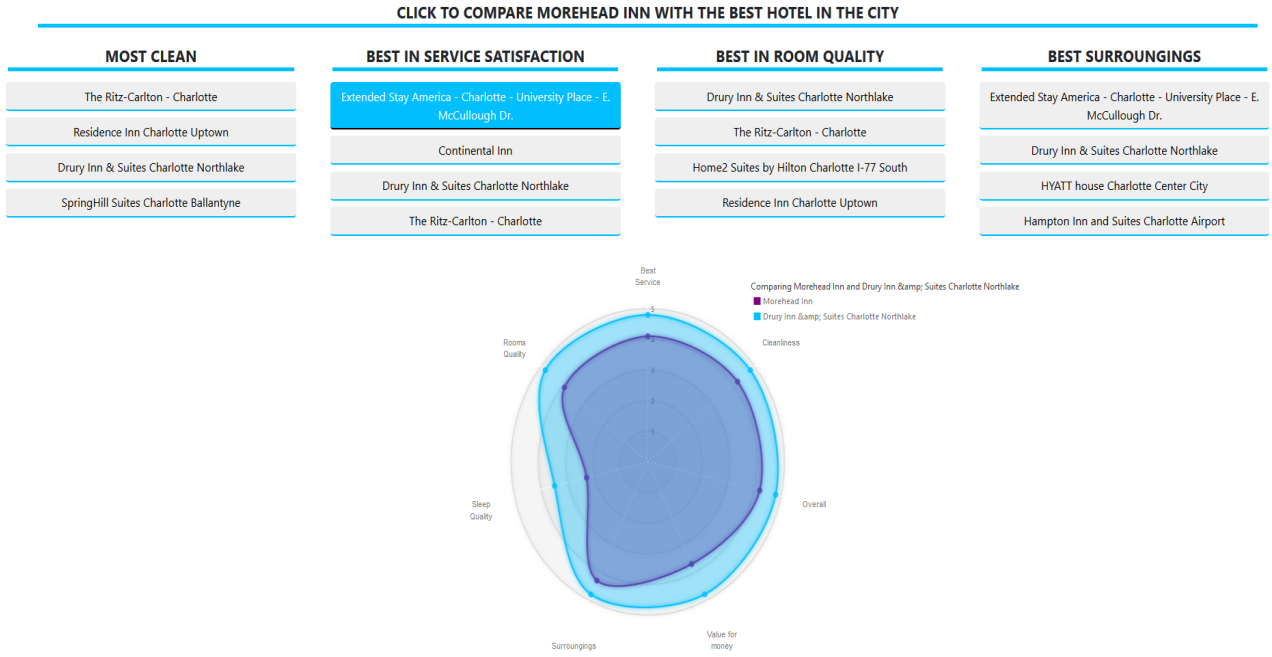

Figure 6: Comparison based Visualization, A Radar Chart

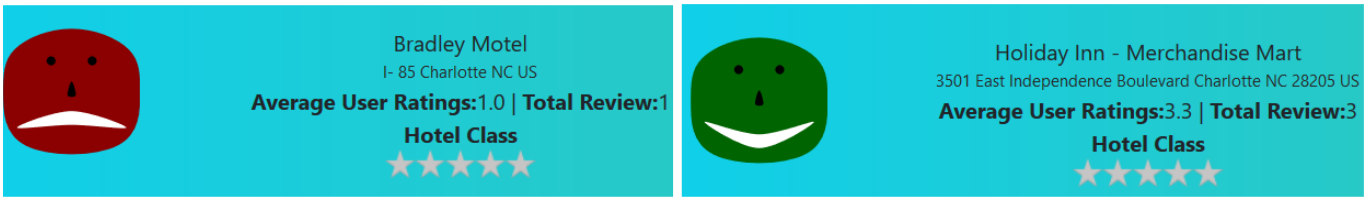

Figure 7: Text Mining, Chernoff Face for Comment Sentiments, Red shows Negative, Green shows Positive

4.1 Location and Class based hotel filtering

In our opinion, location and class-based filtering will help the user to reduce his/her search space. With these filtered outcomes we are also providing avg user rating information through color markers to make the user hotel selection process smoother.

Color Justification: For showing avg user ratings, we are using a color gradient of light purple to dark purple. Where light and dark purple color shows lower and higher user avg ratings respectively. The reason for selecting this colour gradient scheme is that it is consistent with our overall visualization theme. Secondly, it is eye-catching, it clearly separates the background and foreground color, also, the lighter and darker gradient can be clearly differentiable.

Size Justification: Visualization and Text size are decided based on the requirement of the overall system.

4.2 Time Series Visualization

To answer the question of how the hotel is performing concerning different attributes such as sleeping quality, room service etc over the time span. We used the Heat Map to showcase this information. The Heat Map helps the user to see different hotel attributes simultaneously. On the other hand, the line chart presents the overall rating trend, to facilitate certain users who are only interested in seeing the overall rating trend. With this in a view, we believe it will be easier for the user to see different components of the given hotel in one glance.

Color Justification: The line chart is in the purple color theme. For Heat Map, again the same light and dark purple gradient scheme were used. The color scheme is decided in a way that it helps to have consistency in our visualization. For missing values, we assigned a grey color.

4.3 Comparison based Visualization

Most of the times when we have two things and want to decide one out of them require comparison on a different aspect of the items. With this in a view, we offered a visualization that helps the user to perform such comparisons between hotels. Not only we are facilitating the users to compare two hotels, but we are providing an extensive list of the top hotel for different services to compare.

Figure 8: Text Mining, Finding top comments and its similar comments number

Let's take a scenario that, a user requires a hotel having good sleeping and room quality. But, he/she is not sure that the selected hotel the best among all the hotels for such requirements in the mentioned location. In such cases our comparison based visualization becomes ideal. Where we are helping the users by providing top hotels in each specific categories. So he/she can compare the selected hotel with the top hotel for a specific service to get a reference idea.

Reason for picking Radar Chart for Comparison We used the Radar chart to facilitate the comparison. As it can easily facilitate different attribute comparisons. Moreover, for two hotels it is easier to visualize with the Radar Chart. Hence, we prefer a Radar Chart over other comparisons based visualizations.

Color Scheme: For Radar chart, we are using two colors the hotel selected by the user has the same purple and the referenced hotel has a sky blue color. Again the scheme is used to maintain consistency with our visualization system.

4.4 Text Mining

Seeing all ratings for different attributes of the hotel does not completely convey the user that the mentioned hotel is good or bad. We believe to convey the users we need to have their trust in our system. Therefore, we worked towards text mining to infer the information about how people feel about a particular hotel by processing their comments.

4.4.1 Chernoff Face, Overall Review Sentiments

. The initial step with this motive is to get the sentiments of the comments given by the users. We have analysed many techniques to showcase the sentiments and later feel that Chernoff Faces are the best to convey this information.

Visualization Justification: The Chernoff Faces show sentiment via facial expressions. Most humans are very good at capturing the sentiment emotions via faces.

Color Justification: To bolster the sentiment score, we have filled colors in the Chernoff Face. For example, highly negative sentiment will have a red face, highly positive sentiment have a green face, and if the sentiment is neutral we are providing light green color face.

Figure 9: Text Mining, Word Cloud; Green shows positive words Red shows negative words

4.4.2 *Finding top Comments :*

To bolster the user trust with an aim of making it easier to see the overall idea of people's opinion for a particular hotel, we extracted the top five comments that mostly summaries all the given comments. We believe that this is the best visualization feature we have in our system.

To understand people's view on one specific hotel requires to read almost all the comments. However, many times people write similar things. This leads to information redundancy. Moreover, it is time-consuming and not a user-friendly idea to make the user read entire set of comments. Hence, we seek an approach that solves both these issues. In our opinion, the provided top comments and the score of how many other comments are similar to it, help saving the time of the user. Additionally, it is capable to give overall view of people's opinion about a particular hotel.

4.4.3 *Word Cloud :*

Often, it is ideal to showcase the sentiments by extracting positive and negative words from the comments, and counting the occurrence of these words. With the same intention, we tried to give a word cloud to support our other text-based visualization with a motto of showing the true sentiment of people for a particular hotel.

Color Justification: We have used Green color to showcase the positive words and red color to showcase the negative words. The color scheme is used based on the intrinsic property.

5 EVALUATION PLAN

We believe that the following are the key components of our visualization that can be used to assess our visualization.

- **Simplicity:** We tried to build our visualization considering its users are mostly laymen. We limit ourselves to not make complex visualization and to reduce information overflow. We believe that this is the core strength of our visualization.
- **Color Schemes:** The color scheme that we have used throughout in our system has a consistency in it. The colors are picked in a way that they contrast each other and provides a complete insight to the user.
- **Extensive Text Mining:** The proposed visualization system not only capable of providing diverse recommendation based on different aspects of the ratings and hotel popularity but also provides accurate insights from the comments given by the users. In our opinion, the extensive text mining approaches, not only enhance the user experience but plays a crucial role to gain the user trust on recommendations.
- **Supports in all the system:** Our visualization - Hotel Genie is hosted on git hub so it is not a system specific visualization, it can be accessed through any device having internet connection.
- **Responsiveness of the Visualization:** Although, our system is not perfectly responsive on diverse devices but it can be viewed in any devices irrespective of the screen size and our visualization doesn't get hamper.

6 DISCUSSION AND FUTURE WORK

There are many recommendations based visualization available online these days, and more or less each of them uses the similar ideas that we proposed in our system except the text-based visualizations. We believe that our unique text mining visualizations make us stand out from the conventional set of visualizations. In

our opinion, the proposed visualization is capable to address the four questions mentioned in the introduction part by using various dynamics of the data and providing a concise and accurate recommendation.

The future work can be extended to incorporate more data and to make the system available for other states and cities. There is also a scope of incorporating a learning agent in the recommendation system. That can learn the user demands by his/her hotel stay history and provides a recommendation based on that.

7 ACKNOWLEDGMENTS

We would like to thank Professor Dr Hsiao Sharon for her constant guidance and support throughout the course work. We would also like to extend our sincere gratitude to the Teaching Assistant, Yancy for his prompt response to our queries and providing constructive critiques in all our assignment.

REFERENCES

[1] Hotel-Review Datasets - Trip Advisor http://www.cs.cmu.edu/ jiweil/html/hotel-review.html
[2] Matthieu Viry. 2017. Radar Chart d3 v4. (June 2017). Retrieved April 28, 2019 from http://bl.ocks.org/mthh/7e17b680b35b83b49f1c22a3613bd89f
[3] d3noob. 2018. Simple line graph with v4. (December 2018). Retrieved April 28, 2019 from https://bl.ocks.org/d3noob/402dd382a51a4f6eea487f9a35566de0
[4] Tom May. 2019. Day / Hour Heatmap. (April 2019). Retrieved April 28, 2019 from http://bl.ocks.org/tjdecke/5558084
[5] Lars Kotthoff. 2015. Chernoff faces for D3. (October 2015). Retrieved April 28, 2019 from http://bl.ocks.org/larskotthoff/2011590
[6] Ankur P. Parikh, Oscar TÃďckstrÃűm, Dipanjan Das, and Jakob Uszkoreit. 2016. A Decomposable Attention Model for Natural Language Inference. (September 2016). https://arxiv.org/abs/1606.01933
[7] NLTK Python Library for Sentiment Extraction http://www.nltk.org/
[8] Term frequency, inverse document frequency Tutorial http://www.tfidf.com/