**EXECUTIVE SUMMARY**

The purpose of this study is to analyse the English written online reviews of **59 hotels in Jamshedpur .**The goal of this analysis grouping the customer review document.

**Case study**

We conducted a case study by analysing the English written reviews of different hotels in Jamshedpur. These hotels were selected through TripAdvisor which includes **59** domestic hotels. The analysis is focused on the hotel reviews written in English. The data set includes **300** hotel reviews written in English. The customer review along with rating is collected; rating is set from 1-5 where 1 is the worst and 5 is the best.

We obtained customer reviews on www.TripAdvisor.in for hotels in Jamshedpur, from December 10, 2012, through April 7, 2018. The year wise review summary is given below.

|  |  |
| --- | --- |
| **Year** | **#of Reviews** |
| 2018 | 100 |
| 2017 | 140 |
| 2016 | 49 |
| 2015 | 4 |
| 2014 | 5 |
| 2013 | 1 |
| 2012 | 1 |
| **Total** | **300** |

We use text mining, cluster analysis & topic modelling methods to identify patterns and insights from these English written reviews. Below is a description of the steps we followed to collect and analyse the data set

**Collect Raw Text**

The most critical step in analysing unstructured text documents is transforming free form text into a structured form

1. We extract all the reviews related to the **59 hotels** in Jamshedpur from tripadvisor.in
2. We parse unstructured text and imposes a structure for further analysis and then saved the reviews into an excel spreadsheet.

**Pre-processing and text representation.**

The following steps are performed to transform document into “bag of words representation of text.

1. Read customer review data into data frame
2. Remove newline character from the review text
3. We conducted natural language pre-processing to clean the text comments such as special characters like (@,#,| etc),punctuation, numbers white spaces and number
4. Remove words composed of fewer than three characters, custom stop word (than, which, but, this, though, and etc) and English stop words from the Corpus
5. Expand contractions in an English-language source. For example, if the text contains contractions such as we’ll, the application ill split tis two words we and will.
6. Preserve inter word dash like Wi- Fi, state-of-the -art etc so that it can be treated as single token
7. Reduce number of dimension by using lemmatization techniques (finds correct dictionary base form of word). With these techniques, words such as is, are, am, and were can be mapped to the same term “be”
8. Transform all text into lowercase
9. Removing words that occur either too frequently (50%) or rarely (1%)
10. Picks and selects the most commonly occurring words in the corpus i.e. the words having the highest frequencies and plots them, the more the frequency of a particular word the greater is the size of the word in the word-cloud.

The word cloud for the Corpus is given below.



**Grouping the Documents**

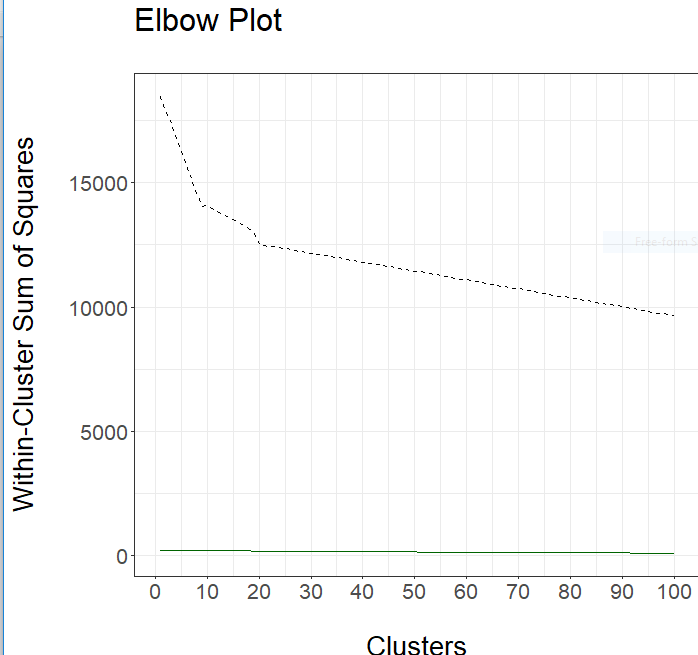
There are **300** (Document) hotel reviews written in English. Document grouping is achieved with K mean clustering.

The goal of K-Mean clustering algorithm is to partition a given dataset into distinct, exclusive clusters so that the data points in each group are quite similar to each other.

**Analysis of K-Mean Clustering**

* One of the key decisions to be made when performing K-Means clustering is to decide on the numbers of clusters to use. We randomly chose the number of clusters to be 2 for illustration purposes only.
* To identify the optimal number of cluster we used the Elbow method and it involves observing a set of possible numbers of clusters relative to how they minimise the within-cluster sum of squares
* We have identified 10 clusters from the elbow plot.

The Elbow plot is given below.



From the above elbow plot we can see that, WSS is greatly reduced when k increases from 10 to 11. Another substantial reduction in WSS occurs at k = 10. However, the improvement in WSS is fairly linear for k > 10 Therefore, the k-means analysis will be conducted for k = 10. The process of identifying the appropriate value of k is referred to as finding the "elbow" of the WSS curve.

The size of each cluster is given below.

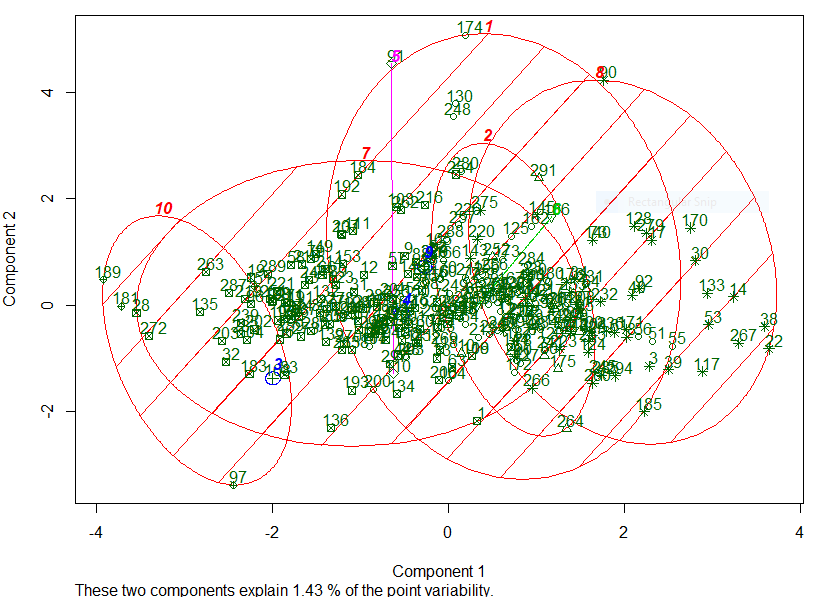
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cluster | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| # of documents | 91 | 14 | 1 | 1 | 2 | 2 | 129 | 55 | 1 | 4 |

Since no of documents in cluster 3,4,5,6,9,10 is very low, we ignore these cluster for further analysis.

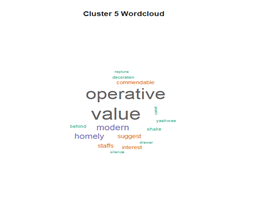
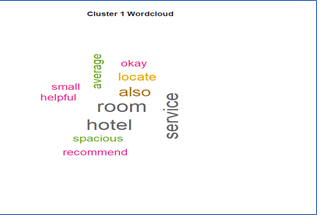
We propose to group the features into 10 groups by applying a statistical clustering technique, e.g., k-means

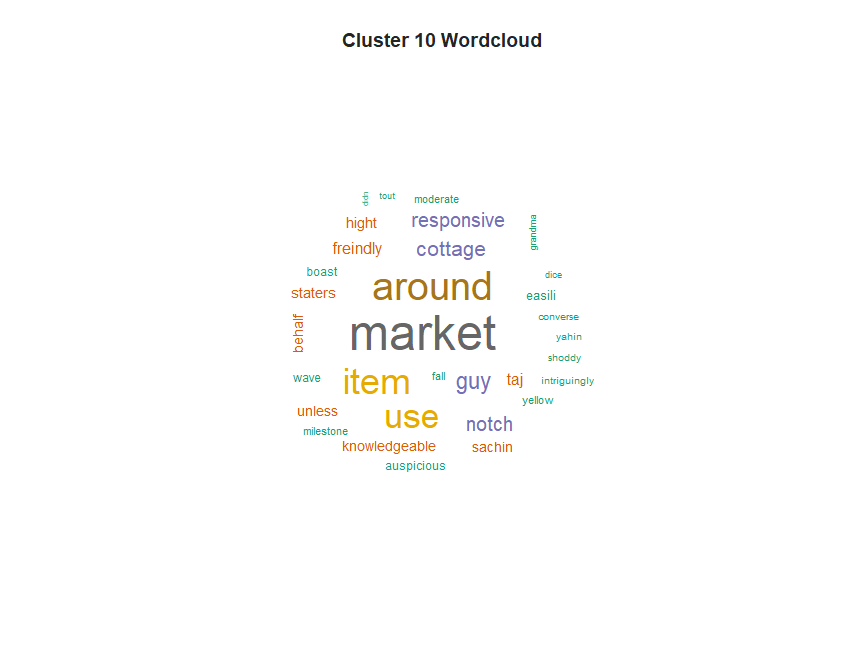
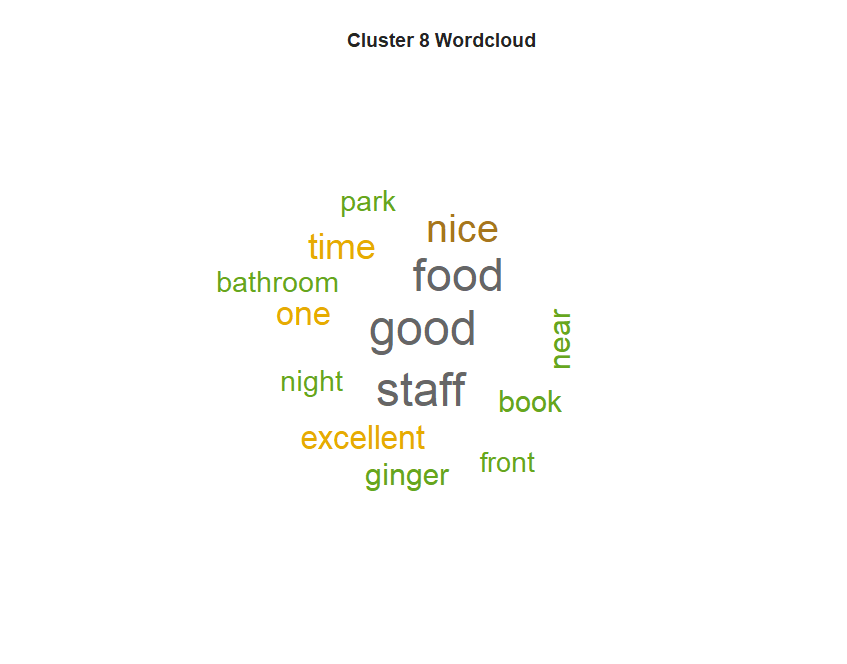
* Construct the review matrix for the review
* **USE Latent semantic analysis** (**LSA**) is a technique in natural language processing, of analysing relationships between the set of documents and the terms
* Compute the frequency of occurrence for each feature
* Apply k-means clustering technique with k = 10 for the data set of frequencies of features computed
* Created wordcloud for each of the cluster

Below is the cluster plot, showing which terms belong to which cluster. All clusters are overlapped.



The wordcloud for the cluster 1,5,7 , 8 &10 is given below





Cluster1: This cluster tells bout service features like “service”, ‘helpful”, “recommend”, “ok” etc

Cluster5: This cluster tells about value for money

Cluster7: This cluster tells about overall experience of the customer

Cluster8: This cluster tells about food and staff experience.

Cluster10: This cluster tells about surrounding of the hotel

The distribution of the rating of the review across the cluster is given below

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Cluster | #of documents | #of 5 ratings | #of 4 ratings | #of 3 ratings | #of 2 ratings | #of 1 ratings |
| 1 | 30 | 29 | 1 | 0 | 0 | 0 |
| 2 | 48 | 19 | 18 | 10 | 1 | 0 |
| 3 | 74 | 16 | 16 | 15 | 8 | 19 |
| 4 | 5 | 4 | 1 | 0 | 0 | 0 |
| 5 | 2 | 2 | 0 | 0 | 0 | 0 |
| 6 | 67 | 19 | 24 | 16 | 4 | 4 |
| 7 | 20 | 8 | 11 | 1 | 0 | 0 |
| 8 | 13 | 3 | 7 | 2 | 1 | 0 |
| 9 | 14 | 9 | 4 | 0 | 0 | 1 |
| 10 | 27 | 2 | 5 | 15 | 2 | 3 |

From the above table, we can find that

* The cluster 1 contains 29 “5-star review “out of 30 reviews. All the good hotels seem to be in the cluster 1.
* The cluster 3 contains 19 “1-star review comments” which is higher than all other cluster. It seems all the poor hotel service provider are in the cluster 3.
* The cluster 10 contains 15 “3-star review comments” out of 27 reviews. All the hotels in the cluster seems to be average.

**Topic Modelling using LDA**

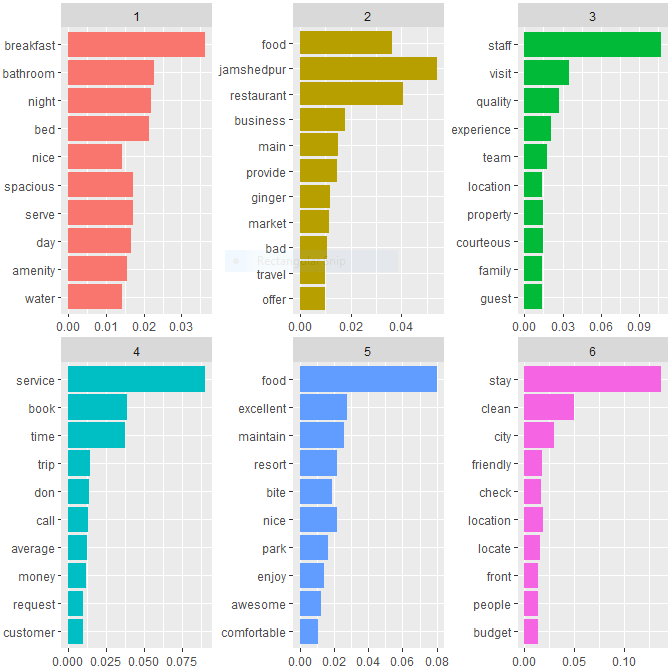
We use topic modeling which automatically organize, search, understand, and summarize large text (customer review) by discovering hidden topics or themes found within a set of documents.

It examines words within a corpus., determine the themes over the text, and discover how the themes are associated. It defined as a distribution over a fixed vocabulary of words. LDA assumes that there is a fixed vocabulary of words, and the number of the latent topics is predefined and remains constant

We use latent Dirichlet allocation (LDA) along with Gibbs sampling for the computation

Here is the most likely words in each topic

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Term | Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 | Topic 6 |
| 1 | breakfast | food | staff | service | food | stay |
| 2 | bathroom | Jamshedpur | visit | book | excellent | clean |
| 3 | night | restaurant | quality | time | maintain | city |
| 4 | bed | business | experience | trip | resort | friendly |
| 5 | nice | main | team | don | bite | check |
| 6 | spacious | provide | location | call | nice | location |
| 7 | serve | ginger | property | average | park | locate |
| 8 | day | market | courteous | money | enjoy | front |
| 9 | amenity | bad | family | request | awesome | people |
| 10 | water | travel | guest | customer | comfortable | budget |



**Topic 1**: Amenities. The likely terms associated with the amenities provided by the hotel include breakfast, bed, water supply & bath room conditions.

**Topic 2:** Location. Terms associated with the location or vicinity of the hotel include market, business centre, and restaurant.

**Topic 3:** People. Terms associated with a guest’s overall experience bout hotel staff and their behaviour.

**Topic 4**: Value. Terms associated with guests’ perceived value for money that include price, overall service

**Topic 5**: Experience. The terms associated with a guest’s overall experience include generally positive descriptive terms, such as, excellent, awesome, comfortable and terms that pertain to particular characteristics, such as staff, food etc

**Topic 6**: Transactions. Transaction terms connected to guest’s stay include check in time, friendly behaviour of front desk staff

**Conclusion and Implications**

The customer feedback is vital for the hotel industry’s efforts toward continuous improvement, a comprehensive characterization of the customer experience remains difficult to achieve. The ability to analyse text contained in online reviews presents an opportunity for greater clarity, as we demonstrate in this study. We show that text mining algorithms that extract the main topics of discussion can provide valuable support to management. We extract the main topics of discussion can provide valuable support to management

**Limitations and future research**.

There are several scopes for future analysis, some of them based on our study’s limitations

1. By measuring each review’s positive and negative sentiment (using sentiment analysis), we can find that the overall sentiment of a review
2. Quantify the review sentiment
3. We have collected data from TripAdvisor. The reviews on that site could be bias. To correct for that possibility, future studies should analyse reviews from other online review websites and social media sites
4. The sample size is very small & we consider only single city for our study. Future studies should consider another city as well.

**ANNEXTURE**

1. R Code(Text\_Mining\_Assignment-II\_Rajib\_Mandal\_TA17002.R)
2. Data File(hotelreview.csv)
3. Program Output (Text\_Mining\_Assignment-II\_Console\_Rajib\_Mandal\_TA17002.txt)