HOTEL REVIEWS: SENTIMENT ANALYSIS

ABSTRACT

This study analyses hotel reviews using sentiment analysis and topic modelling to identify key factors influencing customer satisfaction. The BING method classified reviews into positive, negative, and neutral, revealing trends in customer experience. LDA topic modelling highlighted key themes: room quality, staff service, and location for positive reviews, while cleanliness, staff behaviour, and pricing issues dominated negative feedback. Insights from this analysis help hotels enhance service quality and customer satisfaction using data-driven decision-making.

Shreshth Sharma

Contents

Note: *Check For explanation under Appendices Section for yellow highlighted data.*

---

1. Introduction:

This report presents an analysis of hotel customer reviews using topic modelling techniques to identify the key factors influencing customer satisfaction and dissatisfaction. By analysing a sample of 2000 reviews from a dataset of 10,000 observations.

2. Data Preparation:

These below are the first and last 5 observations of sampled data.

	Score	Review
1	5	Hotel muito bacana, bem localizado e com todos os serviço...
2	5	Very near the DLR station (Prince Regent I think). But as a pr...
3	5	I often stay here when in London on business. The location i...
4	4	My wife and I just returned from a 6 night stay at the Hilton ...
5	4	Прекрасно по месту расположения, номера замечательн...

	Score	Review
1996	3	Stayed here 3 nights and as the reason I stay here and prob...
1997	3	Aunque esta en el sur del Támesis, tiene una ubicación perfe...
1998	3	L'hotel è sicuramente in una buona posizione, vicino alla sta...
1999	5	Fabulous hotel, close to everything, metro at 5 minutes walk...
2000	4	Hemos disfrutado de varios días de vacaciones en este hote...

Filtering Data:

S. No.	Language	Count
1	Breton	1
2	Danish	5
3	Dutch	11
4	English	1516
5	Esperanto	1
6	French	71
7	Frisian	1
8	German	47
9	greek-iso8859-7	2
10	hebrew-iso8859_8	1
11	Italian	122
12	Malay	1
13	middle Frisian	3
14	Norwegian	3
15	polish	2
16	Portuguese	36
17	Romansch	1
18	russian-koi8_r	6
19	russian-windows1251	2
20	Sanskrit	12
21	Scots	71
22	Spanish	72
23	Swedish	11
24	Welsh	1

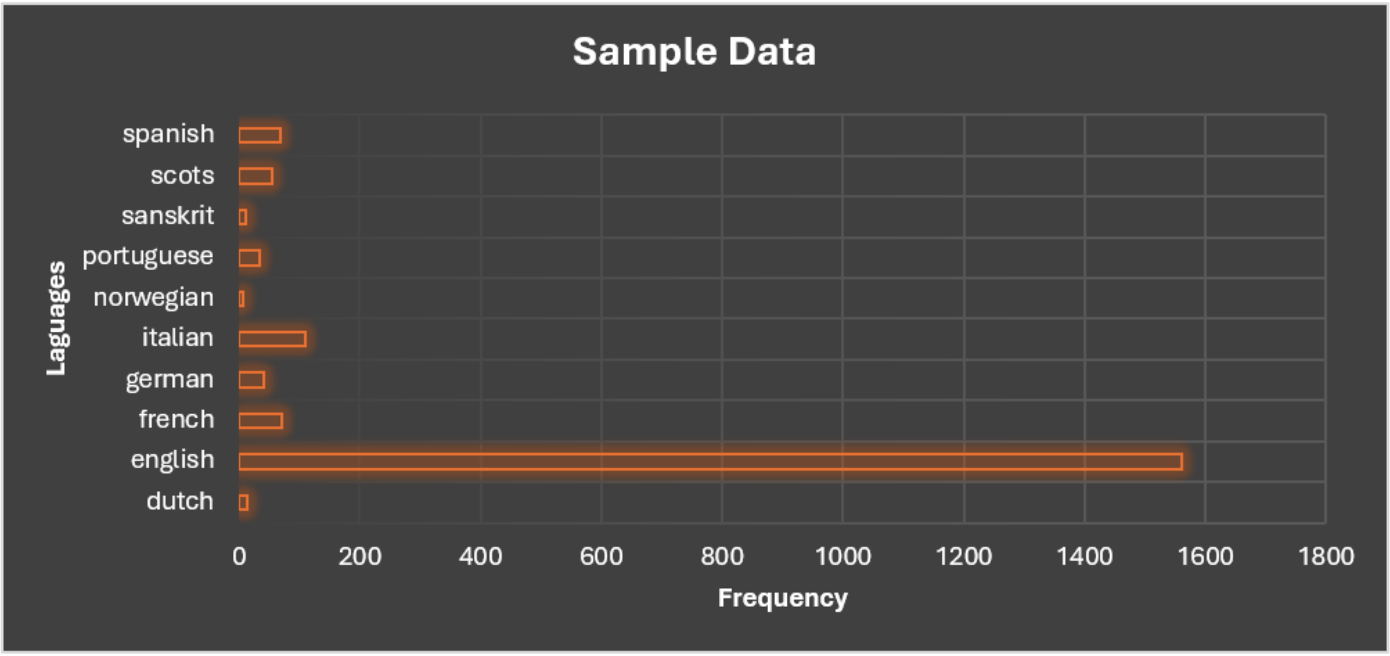
Data Preparation

Sample:

Using set.seed(788) random number is generated according to Student ID and which helps in selecting a randomized sample of 2000 observations.

Filtering Data for Sentiment Analysis:

The dataset initially contained reviews in multiple languages, which were filtered to include only English reviews. Stop words were removed to enhance the accuracy of the analysis.



Mean value of Score given by customers in the sample.

```
> sample %>% summarize(stars_mean = mean(Score))
  stars_mean
1       4.007
```

Score	Review	Language
3	All in all, this Hotel is not as bad as I have it in my mind. It w...	english
3	the room is small but very clean, the breakfast contain: Chee...	english
5	We arrived a bit early for our check-in. Instead of having to ...	english
3	Have stayed a many, many Hiltons. Am a Silver member of ...	english
2	We stayed at this hotel for a few nights in October. The loca...	english
3	This hotel is situated right next to Aldgate East tube station ...	english
5	This hotel is still owned by the Goring family, and it has all t...	english

Language	count
english	1516

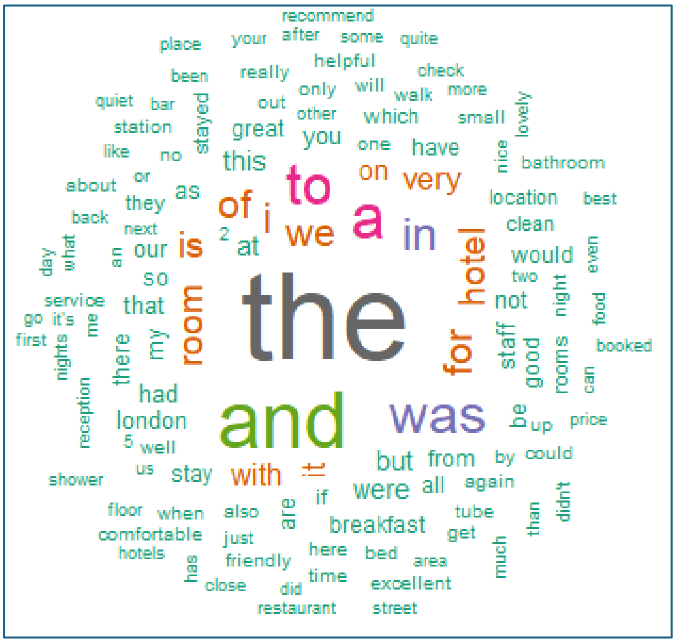
```
> print(language_counts2)
# A tibble: 1 × 2
  Language count
  <chr>    <int>
1 english   1516
```

```
> # Mean of Review Score of the Test Set
> Fsample %>% summarize(stars_mean = mean(Score))
  stars_mean
1   4.074538
```

Filtered Sample looks like this!

Now, after filtering the sample we just have 1516 observations left and these observations are just in English language. 1516 is a pretty good sample size to observe and do the analysis.

Mean of the filtered sample is almost same as the mean of the main sample.



Tokenization:

Now using unnest_tokens() function let's tokenize the words in review column of the filtered sample.

As we can observe in the worldcloud there are many words that are unnecessary for our analysis.

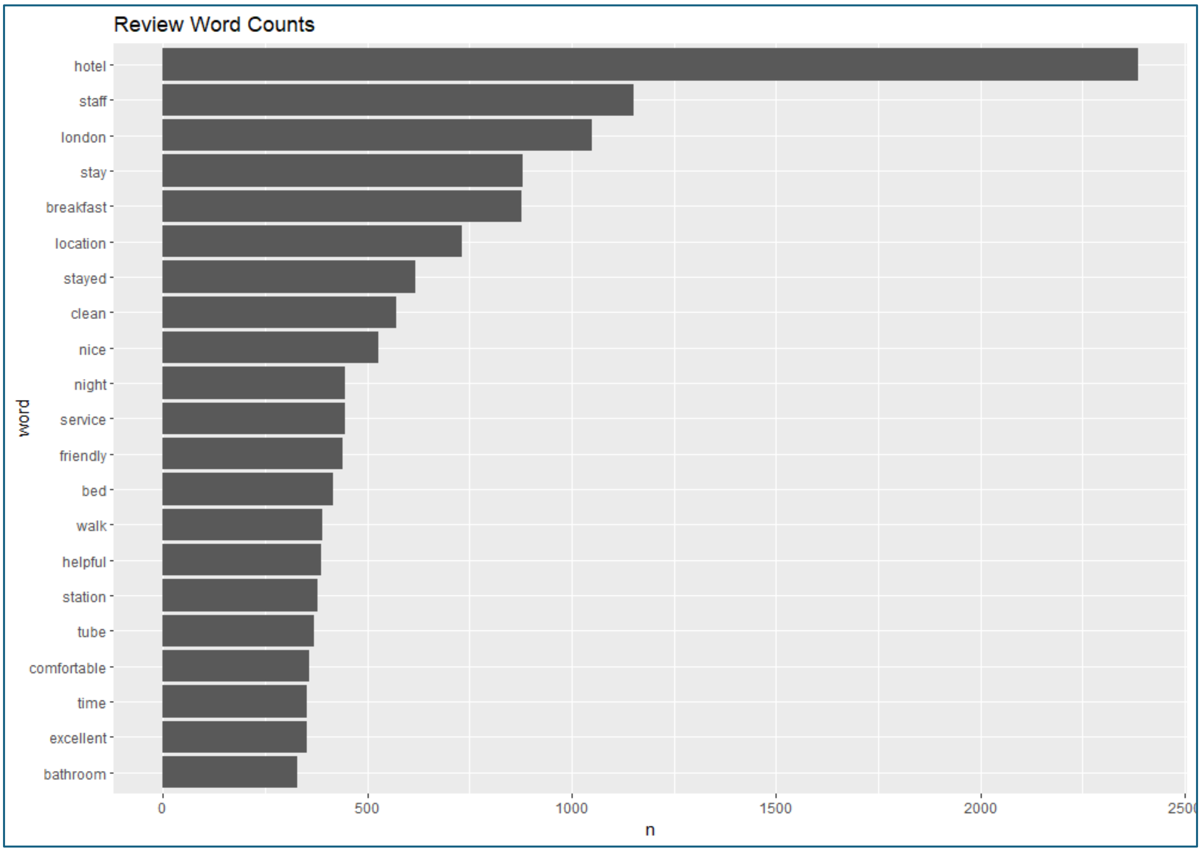
Now we must filter the tokens and remove the stop words and custom stopwords for a few numbers to do our analysis appropriately.

Removing STOPWORDS:

STEP 1: Remove stop words in the dataset like a, an, the etc.

STEP 2: Remove numbers using custom stop words like numbers.

As we can observe the word like the, in and unnecessary words have been eliminated and words like **hotel, staff** have become prominent.

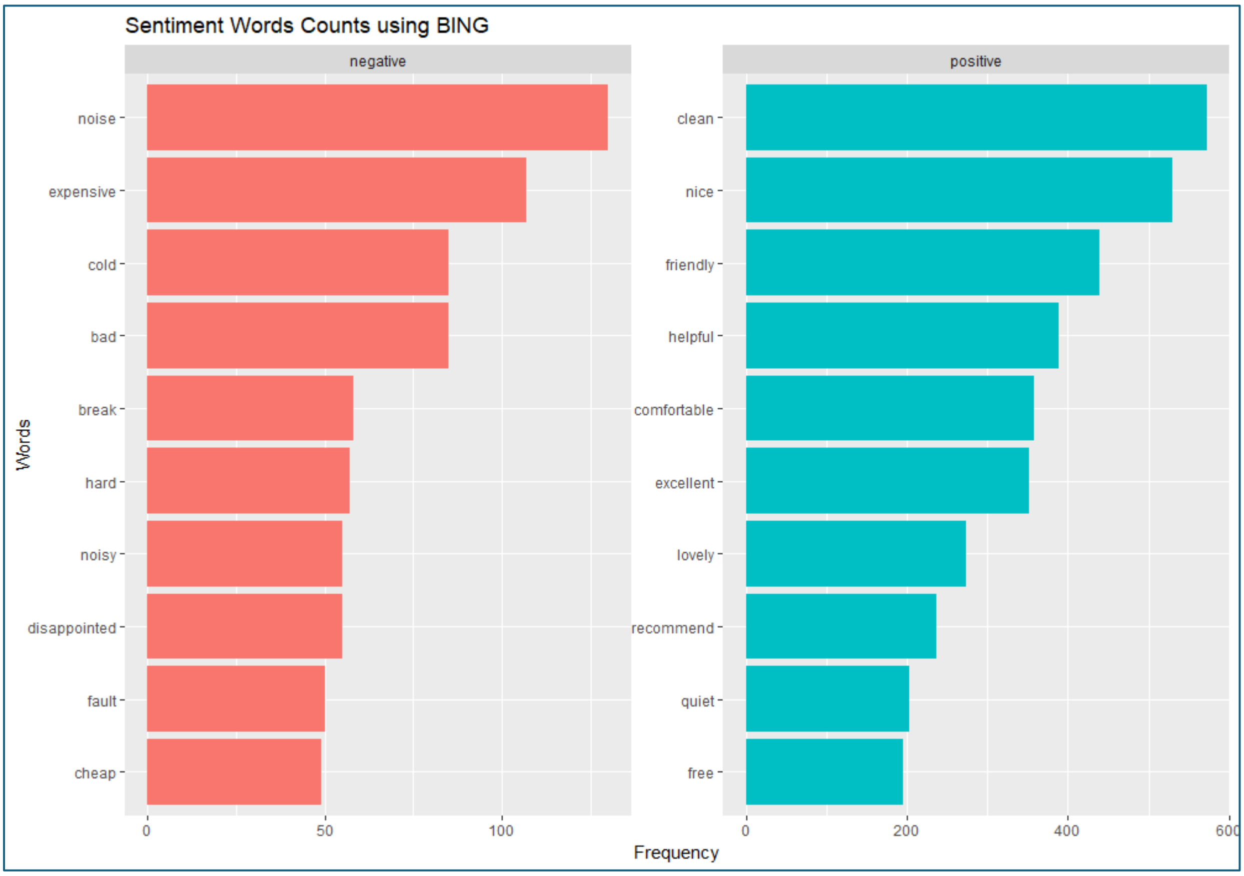


3. Sentiment Analysis

Method: “**BING**”

```
> # Total no. of data set available
> count(sentiment_review)
      n
1 12605
> # Types of emotions and there freq.
> sentiment_review %>% count(sentiment)
  sentiment    n
1  negative 3191
2  positive 9414
```

There are as many as **12605** sentiments in total in the data. Out of which **3191 are negative** and **9414 are positive**. “Noise”, “expensive”, “cold” are top negative sentiments and “clean”, “nice” “friendly” are top positive words.

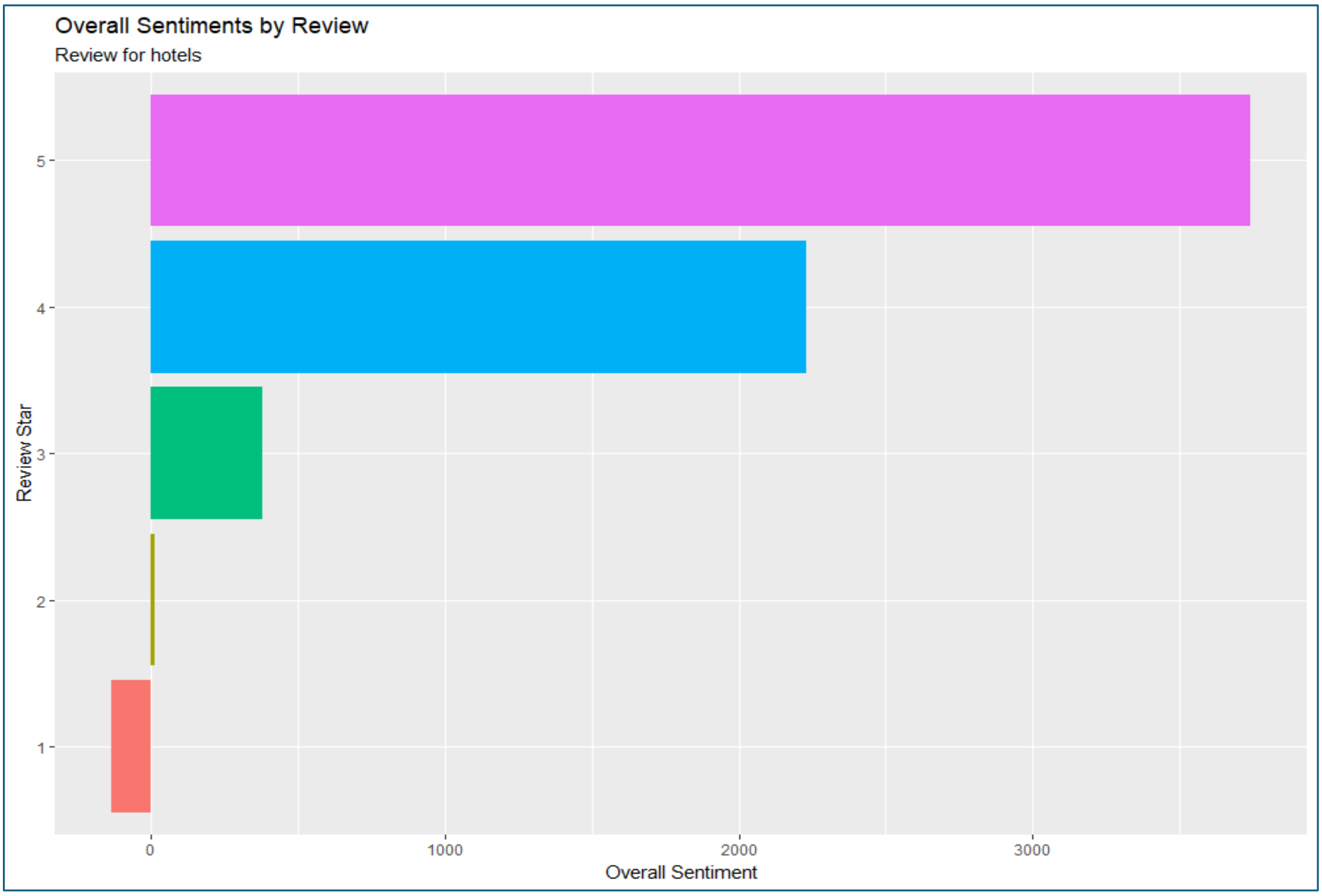


3.1 BING METHOD:

Overall Sentiment using **BING Method** *(Cheng et al., 2014.:* Using pivot wider function we arranged all negative and positive words used in each review by score.

Score	Negative	Positive	Overall Sentiment
1	389	255	-134
2	316	327	11
3	600	978	378
4	949	3177	2228
5	937	4677	3740

As mentioned here we have a graph plotted below describing each review with overall sentiment. 1-star review have a negative overall sentiment with 2-star being almost neutral but still enough data points to capture negative mood towards experience at hotel. With 77% and 83% of positive sentiments star-4 & star-5 captures positive experience. 3-Star reviews are a bit inclined towards positive sentiment but still accounts almost 40 % data covering negative sentiments exhibiting a neutral state. Using 3-star reviews we can validate positive and negative experiences while eliminating both biases from our results while topic modelling for positive and negative reviews.



Positivity Bias: individuals tend to use evaluatively positive words more frequently than evaluatively negative. *(Baumeister et al., 2001).*

4. Topic Modelling And Corpus Formation:

Topic modelling using **(LDA) algorithm**, a natural language processing technique, was employed to identify the underlying topics or themes present in the positive and negative review corpora. *(Im et al., 2019).*

Sentiment analysis using the BING method was performed and now we can use this to categorize reviews as positive (score > 4), negative (score < 2), or neutral (score = 3).

```
# Separate Samples based on scores
positive_data <- Fsample[Fsample$Score %in% c(4, 5), ]
negative_data <- Fsample[Fsample$Score %in% c(1, 2), ]
neutral_data <- Fsample[Fsample$Score == 3, ]
```

For Analysis we will be using **Positive and Negative corpus.** Further validation can be observed under appendix section using **Neutral corpus.**

4.1 Corpus Creation:

```
# Create sentiment labels
Fsample$Sentiment <- ifelse(Fsample$Score >= 4, "positive",
                            ifelse(Fsample$Score <= 2, "negative", "neutral"))

# Separate corpora
positive_corpus <- Fsample[Fsample$Sentiment == "positive", "Review"]
negative_corpus <- Fsample[Fsample$Sentiment == "negative", "Review"]
neutral_corpus <- Fsample[Fsample$Sentiment == "neutral", "Review"]
```

```
> summary(positive_data)
     Score          Review
 Min.   :4.000   Length:1171
 1st Qu.:4.000   Class :character
 Median :5.000   Mode  :character
 Mean   :4.581
 3rd Qu.:5.000
 Max.   :5.000
```

```
> summary(negative_data)
     Score        Review
 Min.   :1.0   Length:148
 1st Qu.:1.0   Class :character
 Median :1.5   Mode  :character
 Mean   :1.5
 3rd Qu.:2.0
 Max.   :2.0
```

Data Summary:

As we can observe there are as many as 1171 entries in positive data with a mean value of 4.581. For negative data there are just 148 observations with a mean of 1.5.

4.2 Positive Corpus:

Here are the top 20 terms in positive corpus with most frequency.

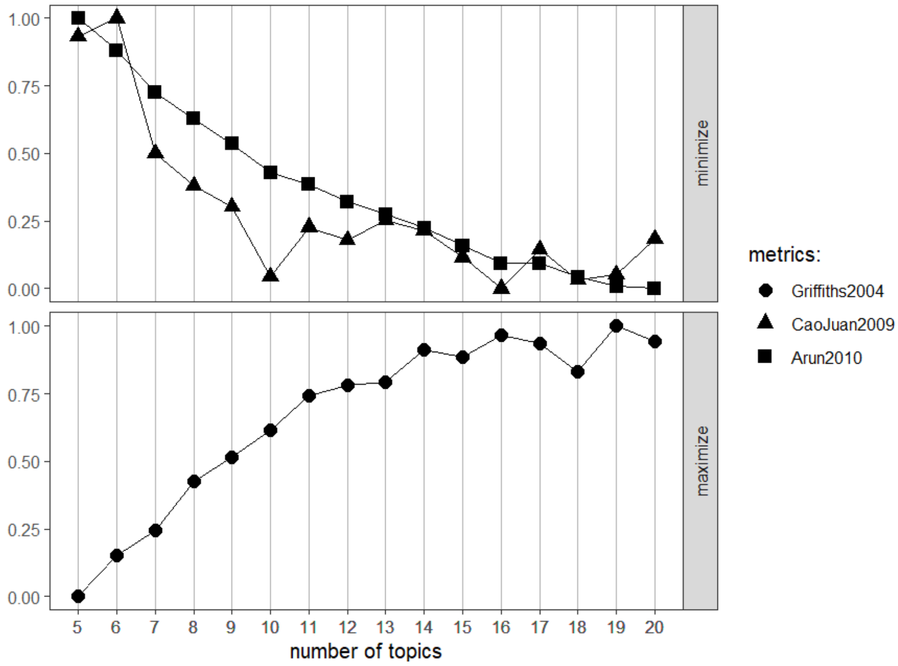
```
> p_frequency[1:20]
    hotel      room     staff    london      good      stay breakfast     great  location
     1820      1542       920       872       712       690       665       653       575
   stayed     rooms     clean      nice      well       one  friendly   service      just
      482       481       447       414       372       368       367       353       342
     also   helpful
      336       326
```

Let’s look at the words that are rarest across the corpus using inverse document frequency.

```
> head(p_freq)
         sort.colSums.as.matrix.p_idf....decreasing...TRUE.
good                                               19.00779
great                                              18.18655
london                                             16.15625
location                                           16.00595
stay                                               15.87280
rooms                                              15.80102
```

- This word cloud consists of top 100 terms with most frequency.
- Hotel, room, staff, London, good are the most popular terms in the corpus.
- Words like good, great, London etc. are the words that have received highest IDF score i.e. these words are rarest across the corpus.
- “good”, “great”, “London” has high IDF scores, the occurrence of these word is relatively low across the corpus.

Positive Corpus – Topic Modelling:



Selection Criteria:

To choose optimised number of topics is to maximize Griffiths2004 and minimize CaoJuan2009 and Arun2010.

In positive Corpus, the number seems to be at **19**.

Positive Corpus							
S. No.	**Topic 1**	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**	**Topic 6**	**Topic 7**
1	"just"	"room"	"clean"	"bar"	"hotel"	"really"	"breakfast"
2	"much"	"check"	"staff"	"food"	"stay"	"nice"	"free"
3	"time"	"day"	"will"	"restaurant"	"service"	"area"	"coffee"
4	"even"	"reception"	"perfect"	"well"	"well"	"got"	"tea"
5	"back"	"arrived"	"every"	"excellent"	"found"	"didnt"	"wifi"
6	"like"	"early"	"next"	"service"	"also"	"though"	"hot"
7	"everything"	"went"	"room"	"staff"	"extremely"	"bit"	"full"
8	"days"	"way"	"breakfast"	"lovely"	"wife"	"hotel"	"facilities"
9	"last"	"wasnt"	"visit"	"evening"	"thing"	"around"	"english"
10	"booked"	"asked"	"inn"	"drinks"	"star"	"night"	"included"

Positive Corpus							
S. No.	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14
1	"get"	"lovely"	"london"	"tube"	"great"	"room"	"staff"
2	"room"	"feel"	"also"	"restaurants"	"stay"	"one"	"friendly"
3	"place"	"beautiful"	"modern"	"close"	"location"	"night"	"helpful"
4	"nice"	"made"	"big"	"walking"	"hotel"	"stayed"	"hotel"
5	"right"	"special"	"access"	"quiet"	"recommend"	"nights"	"clean"
6	"people"	"welcome"	"etc"	"within"	"stayed"	"stay"	"excellent"
7	"find"	"make"	"tower"	"rooms"	"will"	"first"	"stay"
8	"price"	"amazing"	"easy"	"distance"	"london"	"two"	"polite"
9	"business"	"birthday"	"next"	"station"	"definitely"	"bed"	"pleasant"
10	"desk"	"book"	"excellent"	"located"	"fantastic"	"garden"	"welcoming"

Positive Corpus	
Topic 1	Overall Experience and Stay Duration
Topic 2	Check-in and Room Experience
Topic 3	Cleanliness and Hospitality
Topic 4	Dining and Beverage Experience
Topic 5	Overall Hotel Experience
Topic 6	Location and Atmosphere
Topic 7	Free Amenities and Services
Topic 8	Accommodation, people and Value
Topic 9	Persoanlized Experience and Hospitality
Topic 10	Convenient Location and Modern Amenities
Topic 11	Traveler's Comfort and Convenience
Topic 12	Highly Recommended Stay in London
Topic 13	Revisiting Experience
Topic 14	Warm Hospitality Experience
Topic 15	Convenient Urban Hub - Travelling Ease
Topic 16	Convenient Location
Topic 17	High-Value Comfort
Topic 18	Room Features
Topic 19	Top-Rated Stays

Inferring to words, these are the labels for each topic that comprises the overall sense of the topic.

Positive Corpus					
S. No.	Topic 15	Topic 16	Topic 17	Topic 18	Topic 19
1	"hotel"	"walk"	"good"	"room"	"best"
2	"london"	"street"	"rooms"	"bathroom"	"hotels"
3	"station"	"away"	"location"	"bed"	"like"
4	"can"	"minutes"	"clean"	"small"	"stayed"
5	"many"	"park"	"comfortable"	"floor"	"new"
6	"city"	"tube"	"value"	"shower"	"little"
7	"also"	"minute"	"great"	"large"	"problem"
8	"bus"	"station"	"money"	"double"	"reviews"
9	"train"	"road"	"breakfast"	"noise"	"ive"
10	"paddington"	"location"	"quality"	"use"	"now"

5. Negative Corpus:

Here are the top 20 terms in negative corpus with the most frequency.

```
> ne_frequency[1:20]
     room     hotel breakfast     staff      stay      good     night    london     rooms
      332       263        87        85        79        74        72        71        70
      get    stayed       bed       one      just reception     small      also      time
       68        64        62        59        53        52        51        49        49
     like  location
       49        47
```

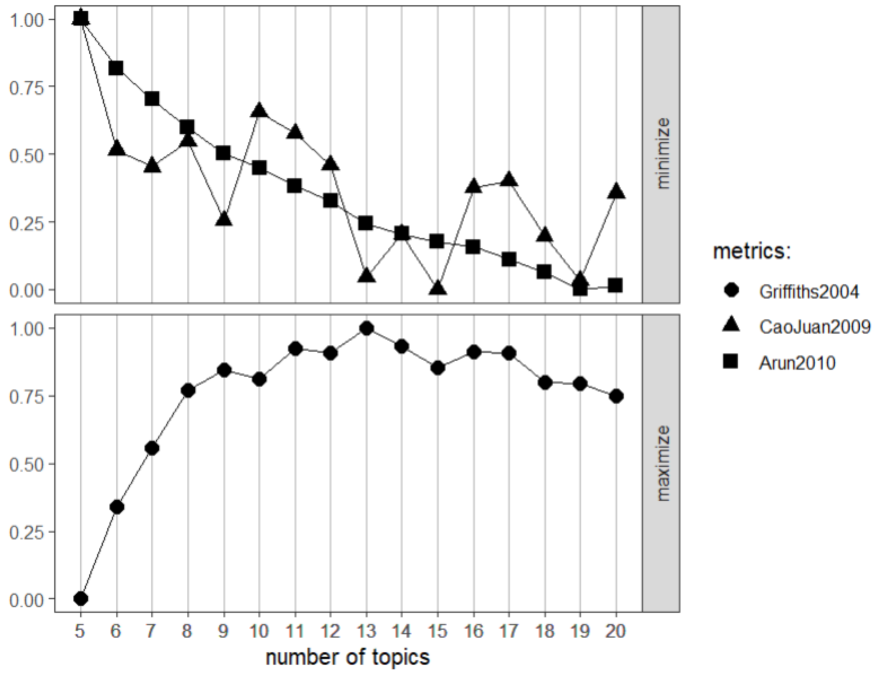
Let's look at the words that are rarest across the corpus using **inverse document frequency**.

```
> head(ne_freq)
          sort.colSums.as.matrix.ne_idf....decreasing...TRUE.
small                                                 2.067049
good                                                  1.829940
stay                                                  1.821113
clean                                                 1.782988
breakfast                                             1.726343
rooms                                                 1.692803
```

```
> ne_frequency
        room         hotel     breakfast         staff          stay 
         332           263            87            85            79 
        good         night        london         rooms           get 
          74            72            71            70            68 
      stayed           bed           one          just     reception 
          64            62            59            53            52 
       small          also          time          like      location 
          51            49            49            49            47 
```

- This word cloud consists of top 40 words.
- Hotel, room, staff, breakfast are the most popular terms in the corpus.
- Words like small, stay, clean etc. are the words that have received highest IDF score i.e. these words are rarest across the corpus.
- As small, stay, clean are the rarest across the sample we can infer that people weren't happy with the stay and found hotels untidy and small.

Negative Corpus – Topic Modelling:



Using Selection Criteria: In the Negative Corpus optimum number of topics seems to be at 13.

Negative Corpus						
S. No.	**Topic 1**	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**	**Topic 6**
1	"get"	"room"	"small"	"stayed"	"breakfast"	"rooms"
2	"day"	"hotel"	"stay"	"air"	"good"	"one"
3	"booked"	"two"	"well"	"desk"	"clean"	"staff"
4	"find"	"said"	"price"	"walk"	"location"	"stay"
5	"location"	"free"	"star"	"station"	"friendly"	"tiny"
6	"great"	"breakfast"	"found"	"nothing"	"get"	"however"
7	"many"	"nights"	"working"	"rooms"	"staff"	"helpful"
8	"long"	"will"	"place"	"stay"	"nice"	"bed"
9	"pleasant"	"bed"	"morning"	"restaurant"	"close"	"will"
10	"week"	"children"	"let"	"tube"	"english"	"great"

Negative Corpus							
S. No.	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**	**Topic 11**	**Topic 12**	**Topic 13**
1	"room"	"london"	"room"	"shower"	"hotel"	"room"	"room"
2	"like"	"hotel"	"night"	"reception"	"just"	"people"	"back"
3	"even"	"wifi"	"also"	"hot"	"didnt"	"front"	"told"
4	"bathroom"	"early"	"first"	"staff"	"really"	"bathroom	"floor"
5	"made"	"station"	"asked"	"water"	"night"	"wasnt"	"arrived"
6	"bed"	"area"	"paid"	"dirty"	"dont"	"can"	"never"
7	"noise"	"stayed"	"time"	"time"	"took"	"location"	"check"
8	"better"	"toilet"	"next"	"one"	"wanted"	"even"	"another"
9	"money"	"street"	"open"	"bags"	"window"	"bit"	"manager"
10	"double"	"hotels"	"extra"	"though"	"say"	"london"	"service"

Assumptions: In sentiment analysis, positive words in negative reviews can occur due to contextual negation, sarcasm, comparative statements, the reviewer's experience, and neutral statements. So, we assume in negative corpus words are said in a negative context. For e.g. Hotel was not **"clean",** Ambience was not **"Pleasant",** Staff was not **"friendly".**

Disguised There was no friendly customer service

Sarcasm They must be easily pleased.

Negative Corpus	
Topic 1	Frustrations with booking
Topic 2	Room and Service Concerns
Topic 3	Inconvenient Stay and Overpriced
Topic 4	Possible Inconveniences
Topic 5	Possible disappointments with hotel service and food
Topic 6	Mixed feelings of experience - Room Features
Topic 7	Uncomfortable Room Experience - bathroom
Topic 8	Poor London Hotel Stay
Topic 9	Possible Frutrations with payment
Topic 10	Poor Hotel Services - Frustrations with shower
Topic 11	Unsatisfactory Stay
Topic 12	Disappointing Experience
Topic 13	Possible Frutrations with reception services

Inferring to each topic carefully these are the labels for each topic that comprises the overall sense of the topic.

6. Data with Topic Tables

Positive Corpus:

	index	Score	Review	topics.ne_ldaOut.
1	1	2	We stayed at this hotel for a few nights in October. The loca...	4
2	2	1	I spent only one night there because my flight had been del...	1
3	3	1	This place is a cheap rip of the staff are ignorant and arroga...	7
4	4	2	The hotel is very nice, unfortunately the beds are well worn ...	11
5	5	1	The room was small, with only a single bed and although no...	6
6	6	1	Hotel is fine as a budget accommodation hotel. Clean, secur...	6
7	7	2	We spent three nights in The President Hotel in mid-Novem...	7
8	8	2	For the price I suppose it was ok, but we should have paid ...	6
9	9	1	It was ok for the price we paid, good central location, room ...	2
10	10	1	Having stayed here on a number of occasions I was disappo...	2
11	11	2	Stayed for 2 nights when we came to see Billy Connolly at th...	1
12	12	1	Well another day another bad hotel!! The only good thing a...	13
13	13	2	Stayed at this hotel on the 12th August with my wife and 2 ...	4
14	14	1	Me and hubby and friends (also a couple) stayed 5 nights in...	5
15	15	1	Don't bother, for what you get you can stay in a 2* hotel. Ni...	11
16	16	1	AVOID! AVOID! AVOID! AVOID! AVOID! AVOID! AVOID! AVOI...	11
17	17	2	The hotel was in a great location, only a few minutes walk fr...	4

Negative Corpus:

	index	Score	Review	Sentiment	topics.p_ldaOut.
1	1	5	We arrived a bit early for our check-in. Instead of having to ...	positive	2
2	2	5	This hotel is still owned by the Goring family, and it has all t...	positive	5
3	3	5	This is a 3 star hotel and it does exactly what it says on the t...	positive	16
4	4	4	Located next to the London Eye, this is a great location in th...	positive	10
5	5	5	One nights stay as it was recommended by a friend. As soon...	positive	5
6	6	4	Really enjoyed my stay here. Although we were a bit unsure ...	positive	6
7	7	5	Very nicely furnished rooms and in a great central London l...	positive	7
8	8	4	Visited this establishment on a week end break with my part...	positive	4
9	9	4	I went to the Ham Yard Hotel with a friend in august for Afte...	positive	7
10	10	5	We stayed for a long weekend the weekend after Easter (du...	positive	3
11	11	4	Not sure what the reviewer before me is really on about. Sta...	positive	1
12	12	4	just back from my yearly visit to the Strand....the location is ...	positive	15
13	13	4	We travel a lot for business but as a family also. We are very...	positive	10
14	14	5	I am famous in my company for being the hardest person t...	positive	9
15	15	4	Absolutely dreadful from uninterested check in staff to unco...	positive	18
16	16	5	Hotel was within walking distance of trafalgar square and ot...	positive	4
17	17	5	This hotel is ideally placed towards the centre of London. It i...	positive	18

7. Top 3 Factors:

Probabilities of Topic in a document for Positive Corpus:

Filter

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19
1	0.04678363	0.10738969	0.05688464	0.02658161	0.04678363	0.06698565	0.02658161	0.08718767	0.03668262	0.08718767	0.05688464	0.04678363	0.06698565	0.04678363	0.03668262	0.066985646	0.02658161	0.03668262	0.02658161
2	0.05156472	0.04480797	0.04480797	0.09886202	0.11913229	0.06507824	0.03129445	0.05156472	0.04480797	0.02453770	0.04480797	0.03805121	0.05832148	0.03129445	0.03805121	0.038051209	0.08534851	0.05156472	0.03805121
3	0.07602339	0.02373581	0.01719986	0.04987960	0.05641555	0.05641555	0.03027176	0.04334365	0.07602339	0.02373581	0.06948744	0.02373581	0.08255934	0.02373581	0.06295150	0.213278294	0.03680771	0.01719986	0.01719986
4	0.02712968	0.02712968	0.07867607	0.03743896	0.02712968	0.02712968	0.02712968	0.03743896	0.02712968	0.13022246	0.06836679	0.08898535	0.04774824	0.04774824	0.05805751	0.119913185	0.03743896	0.03743896	0.04774824
5	0.06258890	0.04907539	0.06258890	0.04907539	0.07610242	0.04907539	0.03556188	0.04907539	0.07610242	0.07610242	0.03556188	0.07610242	0.04907539	0.06258890	0.03556188	0.035561878	0.04907539	0.03556188	0.03556188
6	0.01963865	0.13904163	0.04948940	0.02710134	0.07187745	0.16142969	0.05695208	0.04948940	0.04948940	0.02710134	0.05695208	0.05695208	0.02710134	0.01963865	0.03456402	0.027101335	0.04202671	0.04948940	0.03456402
7	0.04655870	0.03373819	0.04655870	0.04655870	0.04655870	0.04655870	0.07219973	0.04655870	0.07219973	0.03373819	0.04655870	0.07219973	0.03373819	0.07219973	0.04655870	0.059379217	0.07219973	0.05937922	0.04655870
8	0.03604903	0.03604903	0.07714492	0.10454218	0.03604903	0.07714492	0.03604903	0.03604903	0.06344629	0.04974766	0.03604903	0.09084355	0.04974766	0.03604903	0.04974766	0.036049027	0.04974766	0.04974766	0.04974766
9	0.03759398	0.05187970	0.03759398	0.05187970	0.03759398	0.08045113	0.10902256	0.05187970	0.05187970	0.03759398	0.03759398	0.09473684	0.03759398	0.06616541	0.05187970	0.051879699	0.03759398	0.03759398	0.03759398
10	0.03831006	0.04511278	0.16756176	0.03831006	0.02470462	0.04511278	0.07232367	0.06552095	0.05191550	0.03150734	0.03150734	0.03831006	0.03831006	0.01790190	0.02470462	0.051915503	0.11313999	0.03150734	0.07232367
11	0.11431805	0.03904924	0.06055461	0.04980192	0.03904924	0.04980192	0.11431805	0.07130730	0.03904924	0.04980192	0.03904924	0.03904924	0.03904924	0.03904924	0.03904924	0.028296548	0.06055461	0.02829655	0.06055461
12	0.06336565	0.09626039	0.08310249	0.04362881	0.03047091	0.02389197	0.02389197	0.04362881	0.02389197	0.05020776	0.02389197	0.04362881	0.08968144	0.03047091	0.10941828	0.069944598	0.01731302	0.02389197	0.10941828
13	0.05688464	0.02658161	0.06698565	0.05688464	0.02658161	0.04678363	0.03668262	0.07708666	0.03668262	0.12759171	0.03668262	0.02658161	0.03668262	0.03668262	0.07708666	0.036682616	0.11749070	0.02658161	0.04678363
14	0.04111842	0.04111842	0.05674342	0.04111842	0.07236842	0.05674342	0.04111842	0.04111842	0.08799342	0.07236842	0.04111842	0.05674342	0.04111842	0.04111842	0.08799342	0.041118421	0.04111842	0.05674342	0.04111842
15	0.07409300	0.05467552	0.03525805	0.04496679	0.03525805	0.05467552	0.05467552	0.05467552	0.02554931	0.03525805	0.02554931	0.03525805	0.03525805	0.04496679	0.03525805	0.025549310	0.03525805	0.22943281	0.06438426
16	0.04126794	0.04126794	0.02990431	0.09808612	0.02990431	0.04126794	0.09808612	0.05263158	0.02990431	0.05263158	0.08672249	0.04126794	0.05263158	0.07535885	0.06399522	0.029904306	0.05263158	0.05263158	0.02990431
17	0.03356217	0.09877956	0.04805492	0.04805492	0.04805492	0.04080854	0.04080854	0.04080854	0.02631579	0.03356217	0.02631579	0.04080854	0.01906941	0.05530130	0.06254767	0.077040427	0.07704043	0.13501144	0.04805492

Showing 1 to 18 of 1,171 entries, 19 total columns

Probabilities of Topic in a document for Negative Corpus:

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13
1	0.064312736	0.06431274	0.05611602	0.13808323	0.13808323	0.05611602	0.04791929	0.08070618	0.06431274	0.08890290	0.08070618	0.06431274	0.05611602
2	0.153846154	0.04995005	0.04995005	0.04995005	0.08891109	0.08891109	0.04995005	0.06293706	0.12787213	0.06293706	0.10189810	0.04995005	0.06293706
3	0.085884989	0.04705004	0.11501120	0.04705004	0.04705004	0.06646751	0.13442868	0.08588499	0.07617625	0.08588499	0.04705004	0.08588499	0.07617625
4	0.074686941	0.08631485	0.06887299	0.05724508	0.09794275	0.02817531	0.06887299	0.05143113	0.08050089	0.12119857	0.15608229	0.06305903	0.04561717
5	0.088578089	0.07342657	0.07342657	0.05827506	0.07342657	0.10372960	0.07342657	0.05827506	0.08857809	0.08857809	0.07342657	0.05827506	0.08857809
6	0.082340195	0.05417118	0.09642470	0.09642470	0.08234020	0.12459372	0.05417118	0.08234020	0.06825569	0.06825569	0.05417118	0.06825569	0.06825569
7	0.088902900	0.05611602	0.04791929	0.03972257	0.14627995	0.09709962	0.16267339	0.04791929	0.07250946	0.03152585	0.06431274	0.08070618	0.06431274
8	0.060096154	0.07572115	0.07572115	0.06009615	0.06009615	0.12259615	0.07572115	0.06009615	0.10697115	0.07572115	0.07572115	0.09134615	0.06009615
9	0.081196581	0.10897436	0.10897436	0.06730769	0.06730769	0.05341880	0.05341880	0.06730769	0.09508547	0.06730769	0.08119658	0.06730769	0.08119658
10	0.037860577	0.15504808	0.07692308	0.10036058	0.07692308	0.06129808	0.07692308	0.05348558	0.06129808	0.06129808	0.06911058	0.10036058	0.06911058
11	0.126890204	0.10124918	0.05851414	0.07560815	0.08415516	0.05851414	0.07560815	0.05851414	0.10124918	0.11834320	0.04142012	0.04142012	0.05851414
12	0.062694531	0.05113384	0.04535349	0.07425522	0.09159627	0.09737661	0.06847488	0.06269453	0.06269453	0.06847488	0.08581592	0.08581592	0.14361939
13	0.065934066	0.07433743	0.04072398	0.17517776	0.05753070	0.04072398	0.05753070	0.13316096	0.04912734	0.09954751	0.09114415	0.05753070	0.05753070
14	0.048951049	0.06915307	0.05905206	0.06915307	0.14996115	0.03885004	0.06915307	0.07925408	0.06915307	0.10955711	0.08935509	0.07925408	0.06915307
15	0.068255688	0.08234020	0.09642470	0.08234020	0.08234020	0.05417118	0.08234020	0.08234020	0.06825569	0.06825569	0.11050921	0.05417118	0.06825569
16	0.055677656	0.05567766	0.08424908	0.08424908	0.06520147	0.09377289	0.04615385	0.04615385	0.09377289	0.11282051	0.12234432	0.06520147	0.07472527
17	0.076923077	0.05445117	0.07692308	0.12186690	0.06568712	0.06568712	0.09939499	0.06568712	0.04321521	0.05445117	0.11063094	0.08815903	0.07692308

Showing 1 to 18 of 148 entries, 13 total columns

Top 3 Positive Topics | By Probability	
Topic 13	Revisiting Experience
Topic 17	High-Value Comfort
Topic 18	Room Features
Top 3 Negative Topics | By Probability	
Topic 9	Possible Frutrations with payment
Topic 11	Unsatisfactory Stay
Topic 13	Possible Frutrations with reception services

```
> top3_positive
       V17        V13        V18 
0.05460408 0.05403291 0.05400815 
> top3_negative
       V13        V11         V9 
0.08163416 0.08154554 0.08146433 
```

Having Sum of all the topics' probabilities we found out that in Positive Corpus **Topics 13, 17 & 18** and in Negative Corpus **Topics 9,11 & 13** are the top 3 topics influencing **consumers' satisfaction and dissatisfaction respectively.**

8. Analysis Using Clustering Method via Json's Plot:

Positive Corpus

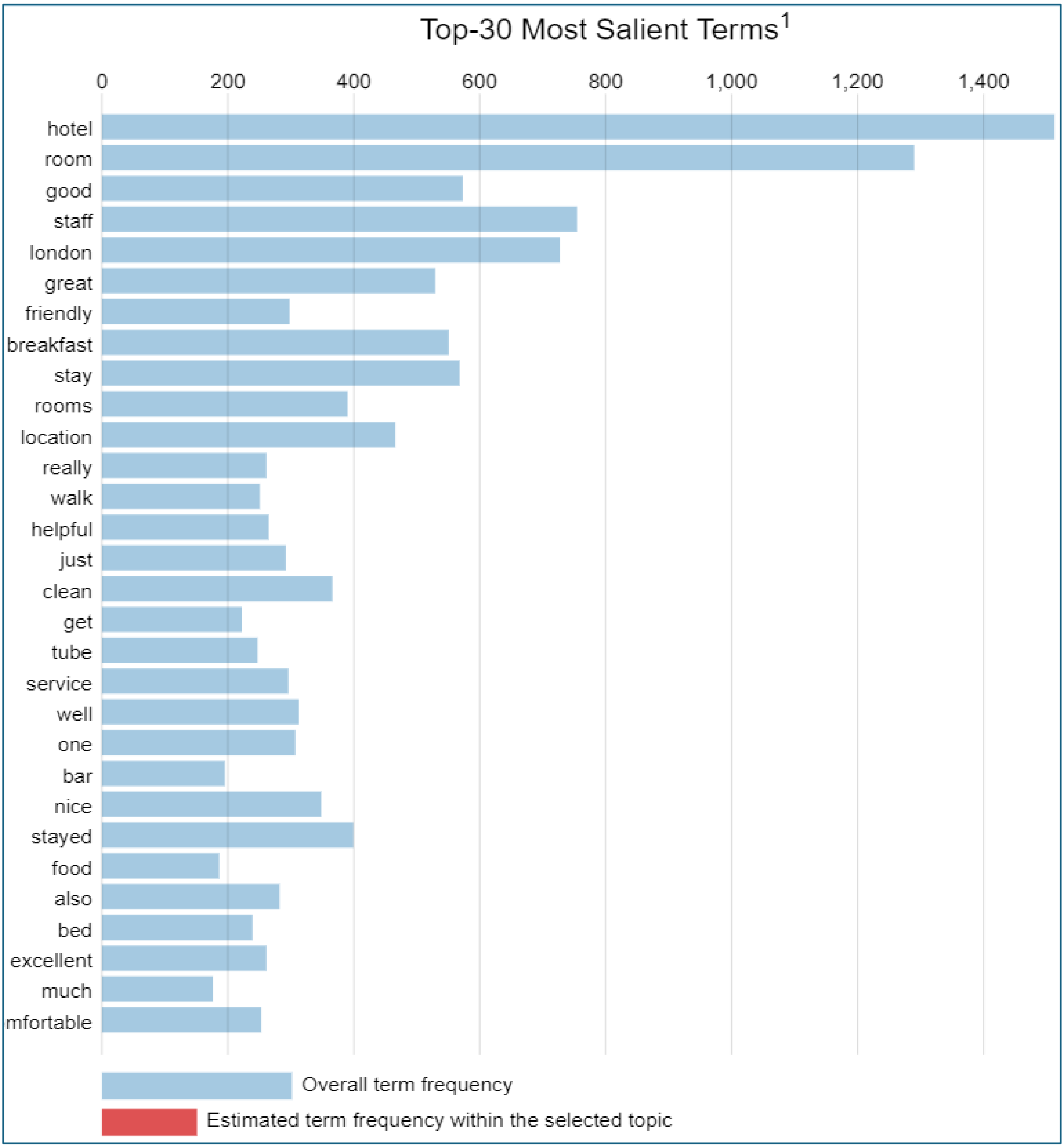
Topics 1, 2, 4 and 10 are clustered together and majorly talks about Experience oriented with hotel, rooms, dining, and city.

Topics 7, 5, 8,12 and 14 are clustered together under purple cloud and it represents the reason why people recommended and liked their stay which includes free breakfast and hot food, great Wi-Fi, fantastic services etc.

Topics 13, 18,19, 6 and 11 are clustered under blue cloud and mostly talks in detail about room features & technicalities and hotels and also a few problems they faced.

Topics 17, 16, 15, 3 and 9 are clustered under green cloud and is talking about experience upon arrival and nearby travel conveniences.

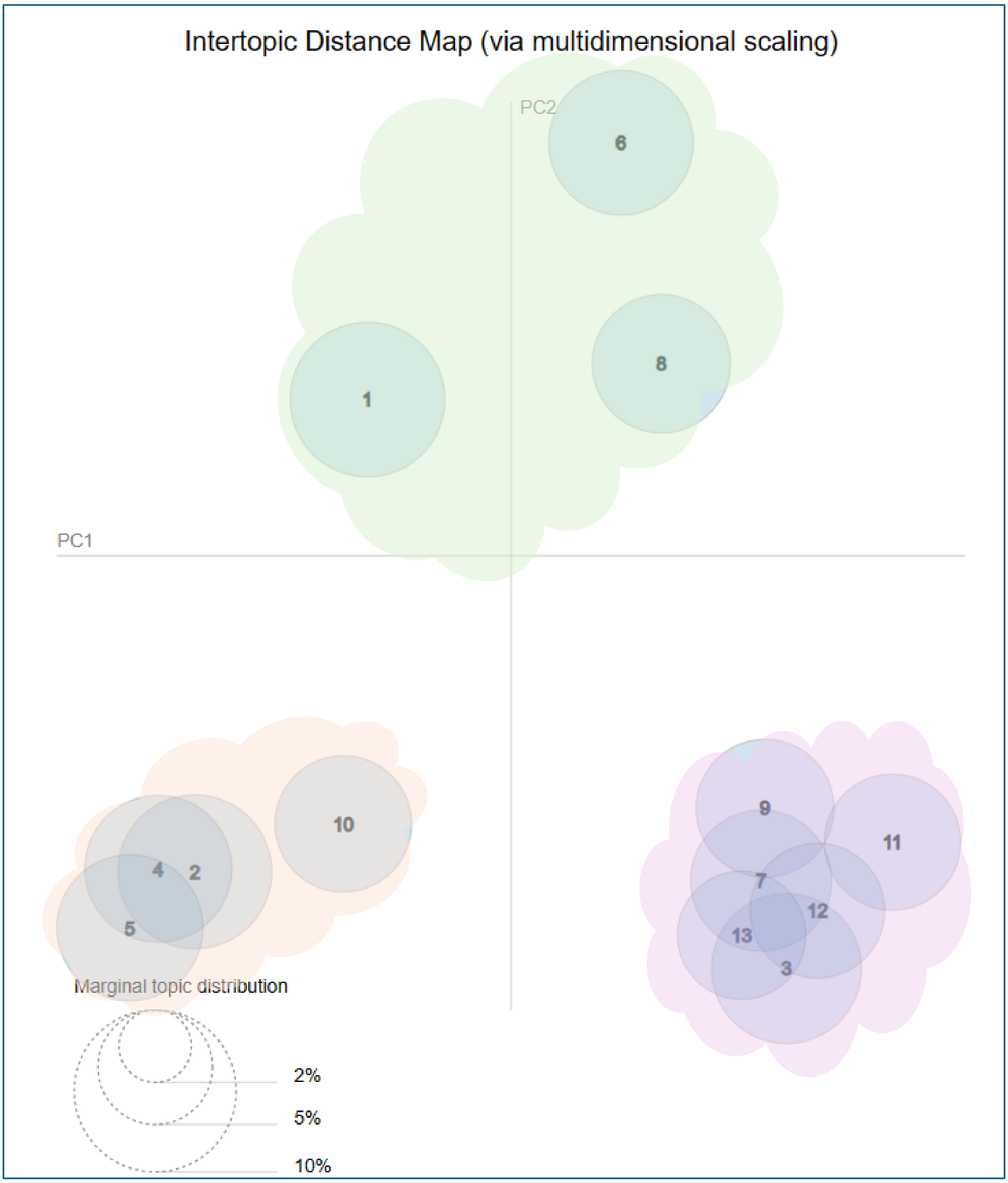
Top 30 Terms in Positive Corpus after Topic Modelling:



Positive Review Topics: By clustering we here try to maximize the common concerns and ideas sorted by highest probabilities in any topic given.

1. **Room Quality and Amenities:** This topic encompassed terms such as "room", "clean," "comfortable," "bed," and "spacious" indicating that customers highly valued the overall quality and comfort of their accommodation.
2. **Location and Convenience:** Terms like "location," "central," "walking distance," and "accessible" highlighted the importance of a hotel's strategic location and proximity to popular attractions or public transportation.
3. **Staff and Service:** The topic related to staff and service included terms such as "friendly," "helpful," "attentive," and "professional," emphasizing the significance of excellent customer service and interactions with hotel staff.

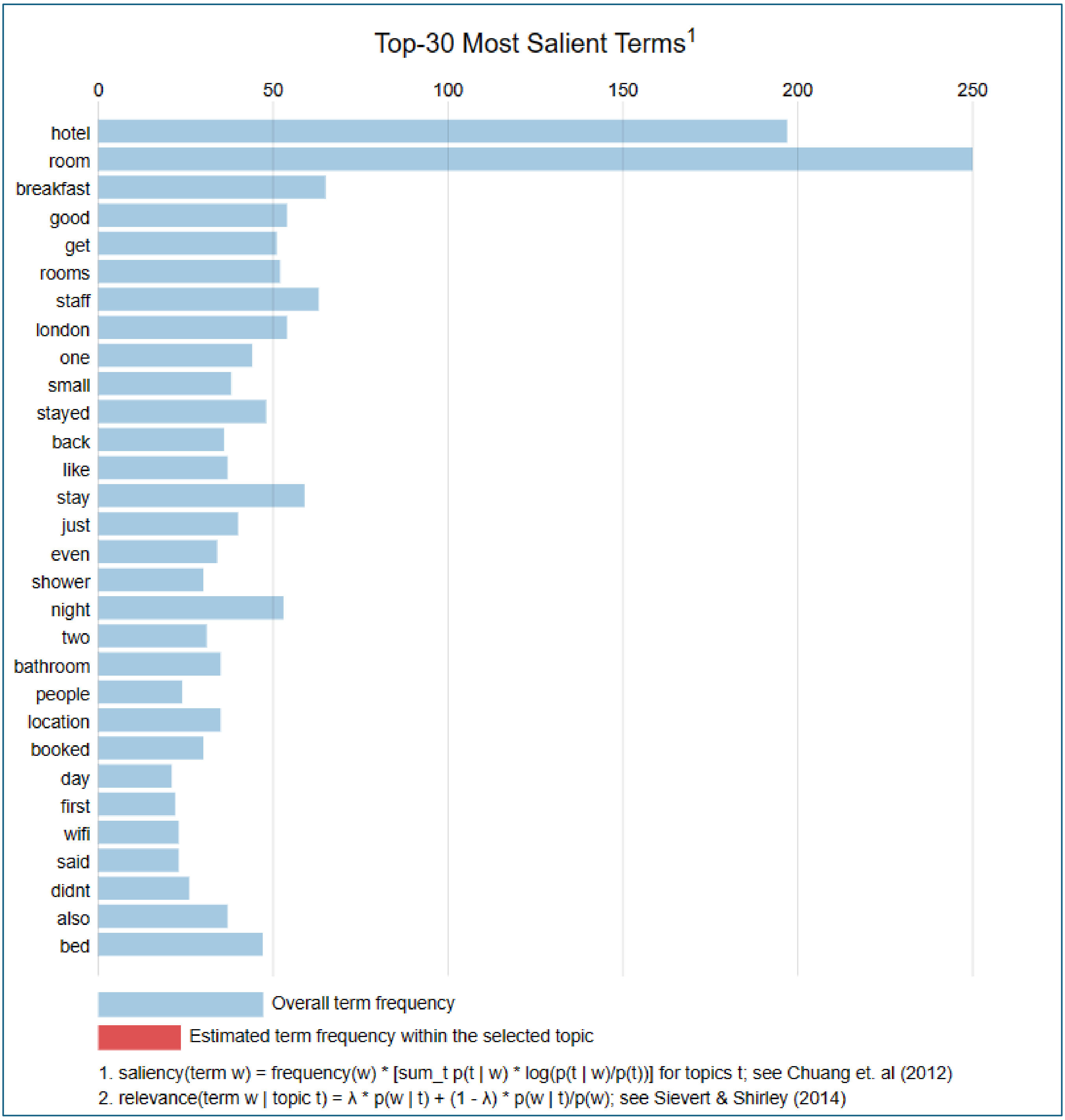
4. Analysis Using Clustering Method via Json's Plot: Negative Corpus



Topics 1, 8, and 6 are clustered together in green cloud and majorly talks about possible frustrations with stay and bookings such as "tiny", "however", "early".

Topics 4, 5, 2 and 10 are clustered together under orange cloud and it represents the reason why people might not have liked hotel upon arrival. These include frustrations related "walk", "nothing", "dirty", "water", "Friendly", "staff", "clean", "bathroom".

Topics 3, 7, 9, 11, 12 and 13 are clustered under purple cloud and mostly talks in detail about disliked room features and hotels (overpriced/noise/ extra-money/ small). This also represents the feeling of not being attended for their problems.



Negative Review Topics:

1. **Room and Amenities Issues:** This topic covered terms such as "dirty," "outdated," "noisy," and "uncomfortable," indicating that poor room conditions, cleanliness, and lack of soundproofing were major sources of dissatisfaction.
2. **Staff and Service Complaints**: Terms inferring "rude," "unhelpful," "not-friendly," and "slow & unresponsive service" highlighted instances where customers experienced unsatisfactory interactions with hotel staff or received poor service.
3. **Value for Money:** The topic related to value for money included terms such as "overpriced," "expensive," "not worth," and "poor value," suggesting that customers felt they did not receive adequate value for the price they paid.

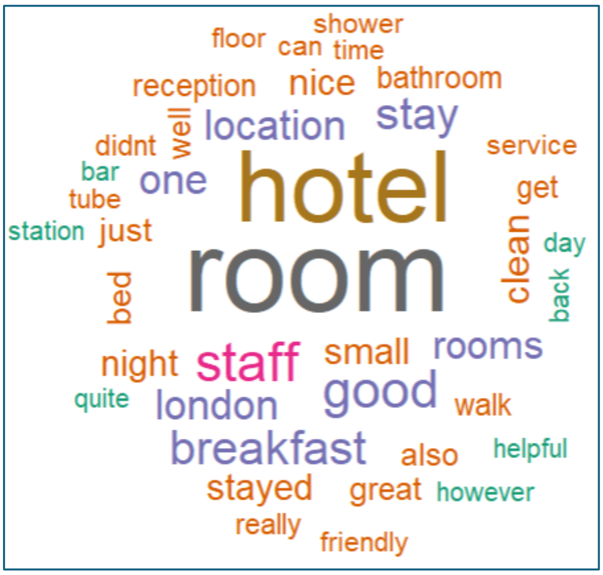
Validation and Justification Using Neutral Corpus:

Appendices:

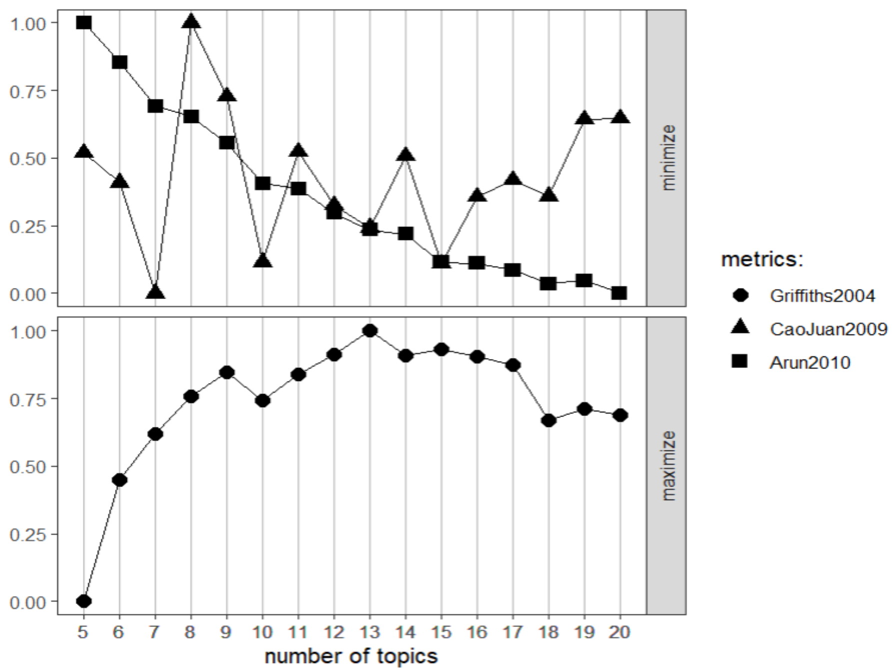
1. **Different Methods:** The na_interpolation() estimated value is the mean between the value before the missing data and the value after. The na_ma() instead uses a moving average approach. The na_kalman() either employs a structural time series fitted by maximum likelihood (default) or employs an ARIMA model.
2. **Method Selection Process:**
    - p: represents the relationship between the current observation and its past values.
    - d: the number of times differencing was applied to the time series data to make it stationary.
    - q: captures the relationship between the current observation and the residual errors from past observations.
3. **The BING method,** as proposed by , introduces an efficient approach for objectness estimation using binarized normed gradients (Cheng et al., 2014). In the realm of topic modeling, Latent Dirichlet Allocation (LDA) plays a crucial role, as highlighted in studies like 's bibliometric analysis, emphasizing its importance in big data analysis (Garg & Rangra, 2022). Additionally, discuss the iterative process of LDA for multilevel classification, showcasing its utility in classifying text data into various levels (Bhutada et al., 2014). These references collectively underscore the significance of both the BING method and LDA in their respective domains, showcasing their practical applications and contributions to computational tasks.
4. **Latent Dirichlet Allocation (LDA)** is a widely used topic modeling algorithm in natural language processing. It functions by analyzing a corpus of documents and identifying clusters of words, known as "topics," that frequently co-occur within the documents (Im et al., 2019). LDA groups documents into latent topics based on a distinct Dirichlet distribution, where each topic is represented as a distribution of terms occurring in the document set (Im et al., 2019). This algorithm is highly prevalent in text analysis due to its ability to automatically detect hidden themes within large text collections (Agarwal et al., 2020).
5. **Positivity Bias:** As Research has consistently shown a linguistic positivity bias, where individuals tend to use evaluatively positive words more frequently than evaluatively negative words (Baumeister et al., 2001). This bias has been observed in both written and spoken English (Augustine et al., 2011). Studies have demonstrated that the human-perceived positivity of commonly used English words displays a clear positive bias (Kloumann et al., 2012). This certainly strengthens the effect of negative reviews to identify the factors attributing towards negative consumers' behaviour and eventually eliminating them to boost positive experience among customers. This certainly strengthens the effect of negative reviews to identify the factors attributing towards negative consumers' behaviour and eventually eliminating them to boost positive experience among customers. In our case there are fewer negative comments, but people tend to write negative comments when they are actually frustrated, so these reviews serve as quality data for analysis.
6. **Assumptions for negative corpus:** In sentiment analysis, positive words in negative reviews can occur due to contextual negation, sarcasm, comparative statements, the reviewer's experience, and neutral statements. So, we assume in negative corpus words are said in a negative context. For e.g. Hotel was not **"clean",** Ambience was not **"Pleasant",** Staff was not **"nice".**

- Negative Reviews: **Disguised** There was no friendly customer service
- **Sarcasm** They must be easily pleased.

7. **Information on Corpuses:** As we can observe there are as many as 1171 entries in positive data with a mean value of 4.581. For negative data there are just 148 observations with a mean of 1.5 whereas the neutral data have more observations than negative data with 197 entries.
8. **Neutral Corpus:**



- This word cloud consists of top 40 words.
- Hotel, room, staff, breakfast are the most popular terms in the corpus.
- Words like "breakfast, good, nice, rooms" etc. are the words that have received highest IDF score i.e. these words are rarest across the corpus.
- As "good", "nice" are the rarest across the documents we may infer that there were less people who wrote "good" and "nice" in their reviews.

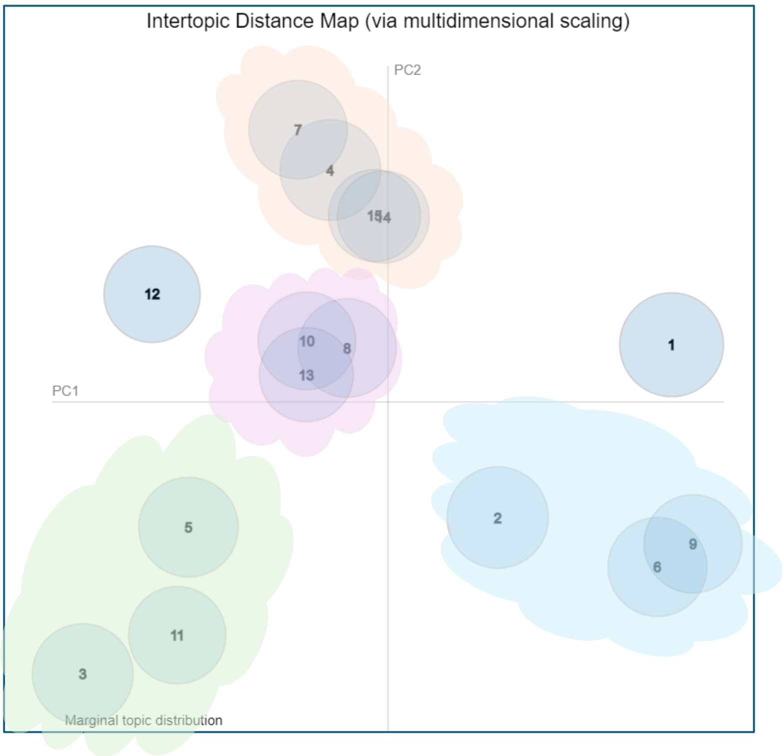


Using Selection Criteria: In the Negative Corpus optimum number of topics seems to be at 15.

Neutral Corpus								
S.no.	**Topic 1**	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**	**Topic 6**	**Topic 7**	**Topic 8**
1	"good"	"room"	"walk"	"well"	"nice"	"rooms"	"just"	"staff"
2	"much"	"great"	"nights"	"tube"	"location"	"london"	"bed"	"helpful"
3	"check"	"breakfast"	"street"	"minutes"	"friendly"	"however"	"shower"	"location"
4	"little"	"price"	"stayed"	"one"	"will"	"two"	"small"	"service"
5	"bathroom"	"use"	"window"	"get"	"also"	"view"	"hot"	"need"
6	"get"	"long"	"door"	"hotel"	"stayed"	"hotel"	"space"	"hotels"
7	"found"	"fine"	"clean"	"times"	"located"	"booked"	"breakfast"	"road"
8	"several"	"quite"	"day"	"end"	"outside"	"excellent"	"left"	"great"
9	"small"	"given"	"sleep"	"wait"	"stay"	"never"	"room"	"spacious"
10	"area"	"new"	"first"	"good"	"dont"	"easy"	"first"	"bathroom"

Neutral Corpus							
S.no.	**Topic 9**	**Topic 10**	**Topic 11**	**Topic 12**	**Topic 13**	**Topic 14**	**Topic 15**
1	"hotel"	"room"	"stayed"	"hotel"	"food"	"room"	"really"
2	"one"	"stay"	"bathroom"	"london"	"bar"	"reception	"didnt"
3	"breakfast"	"floor"	"wifi"	"clean"	"said"	"also"	"time"
4	"night"	"area"	"close"	"small"	"told"	"comfortal	"enough"
5	"restaurant"	"water"	"way"	"station"	"nice"	"went"	"even"
6	"around"	"came"	"warm"	"bit"	"service"	"can"	"though"
7	"tea"	"like"	"desk"	"location"	"mins"	"back"	"still"
8	"like"	"lovely"	"think"	"good"	"didnt"	"got"	"dont"
9	"central"	"looking"	"best"	"stay"	"minute"	"wasnt"	"things"
10	"day"	"great"	"many"	"across"	"although"	"noise"	"etc"

Neutral Corpus	
Topic 1	Average Experience - Mixed Experience
Topic 2	Comfortable Stay - Stay Experience
Topic 3	Convenient Location and Clean Accommodation
Topic 4	Convenient Access and Transportation
Topic 5	Positive Experience with Location and Service
Topic 6	Booking Ease
Topic 7	Basic Amenities and Room Size
Topic 8	Warm Hospitality and Convenient Location
Topic 9	Central Hotel with Dining Options
Topic 10	Comfortable Room and Convenient Location
Topic 11	Convenient Amenities and Thoughtful Service
Topic 12	Central London Stay with Clean but Small Rooms
Topic 13	Standard Dining and Beverage Experience
Topic 14	General Accommodation and Facilities Assessment
Topic 15	Neutral Experience and Expectations



Topics 7, 4, 14 and 15 are clustered together in orange cloud and majorly talks about overall experience. It can be validated that people did like the stay and found it comfortable, but rooms were small.

Topics 2, 6 and 9 are clustered together under blue cloud and it represents the reason why people might have had an average experience with mentioning "quite", "however" and "fine" but not great but loved the ease of booking hotels.

Topics 3, 5 and 11 are clustered under green cloud and talks about friendly location and amenities.

Topics 8, 10 and 13 under purple cloud talks about in hotel & food services and room features like spacious, lovely, great, floor, nice etc.