PART 2

INTRODUCTION

In the dataset "HotelsData.csv" the given data is a huge collection of customer reviews for different hotels and the ratings are from 1 to 5 which is a high satisfaction rating. These recapitulative reviews are usually believed to be the best sources of information regarding customers' experiences and opinions about hotel services and facilities. A subset of 2,000 reviews has been randomly picked from the dataset for this study. The aim is to use topic modelling to classify and understand the main topics in positive and negative customer feedback. Such structured methods are anticipated to help the hotel management in the identification of key areas of strength and those that require improvement, thus guiding the hotel management decisions to align well with the guest expectations and deliver quality services.

The dataset contains 10,000 rows and 2 variables which contain details related to hotel review.

DATE	Character	Description
Review score	Numeric (Discrete)	Customer rating from 1 to 5
Review	Text	Customer reviews.

TABLE 3 DATA DICTIONARY OF HOTEL REVIEW

DATA UNDERSTANDING AND PRE-PROCESSING

In the process of analysing the Hotel review dataset follow the following steps:

IMPORT DATA

Load the dataset from a CSV file and rename columns to "Rating" and "Review."

SAMPLING

Using the `cld3` package, identify and retain reviews written in English. Randomly sample 2,000 English reviews from the dataset for detailed analysis.

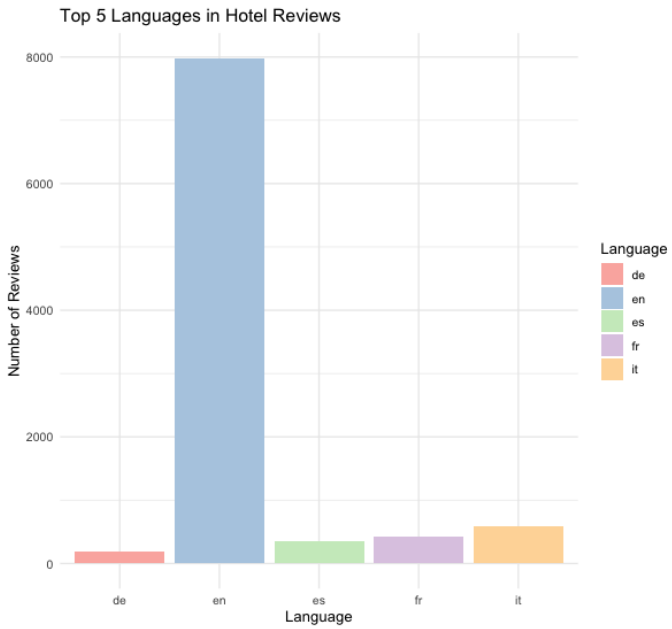

FIGURE 23 TOP 5 LANGUAGE

Cleaning And Pre-processing

Now Clean review text by removing URLs, correcting joined words, stripping special characters, converting to lowercase, removing numbers, filtering out common stopwords, replacing accented characters, eliminating extra whitespace, and lemmatisation. Applying the `tm` package to apply the defined text cleaning procedures systematically.

Subset Data

Categorize reviews into positive (ratings ≥ 4) and negative (ratings ≤ 3) based on their ratings for focused analysis. Then create a corpus for both positive and negative reviews.

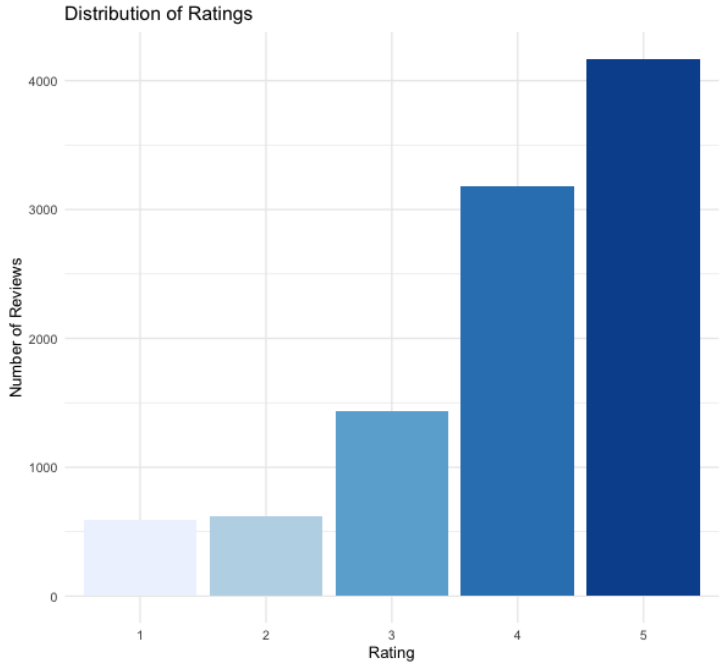

FIGURE 24 DISTRIBUTION OF RATING

Document-Term Matrices (DTMs)

The document-term matrices (DTM) of both positive and negative reviews are generated from customer review data. Initially, it begins with sparsity (99%) matrices reduced to 95% by eliminating infrequent terms. This adjustment will limit the analysis to the most common terms.

FIGURE 25 POSITIVE AND NEGATIVE WORD CLOUD

TOPIC MODELLING: LATENT DIRICHLET ALLOCATION

Latent Dirichlet Allocation (LDA) is a statistical model that generates probability distributions for discrete data, specifically for collections of text corpora. Within the context of hotel evaluations, Latent Dirichlet Allocation (LDA) seeks to represent each review as a limited combination of topics and each topic as an endless combination of words. This approach facilitates the identification of thematic patterns throughout the reviews. (Blei et al., 2003)

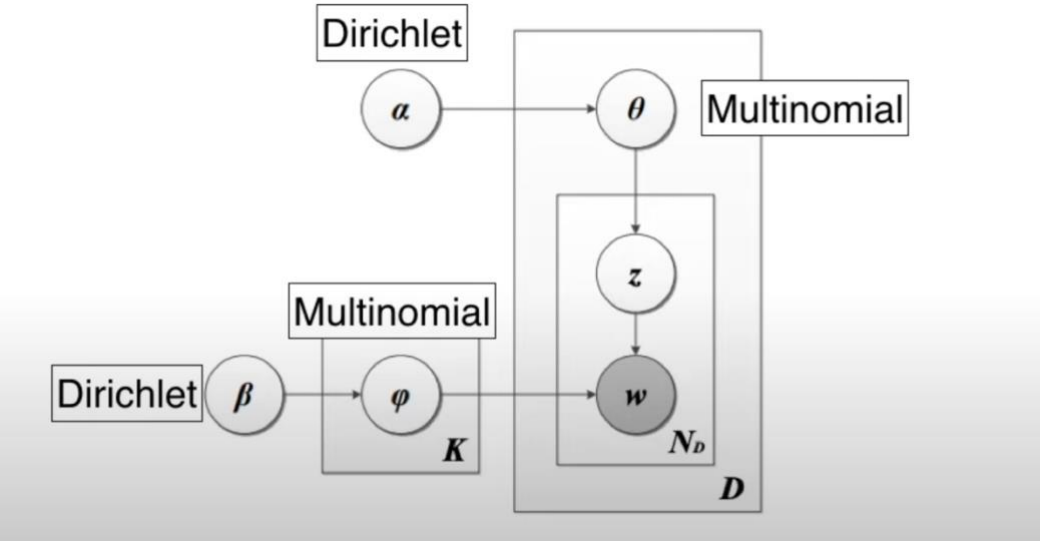

FIGURE 26 LDA MODEL (BLEI ET AL., 2003)

PROCESS OF LDA:

- Each **topic z** is a distribution over words, represented as ϕ_z where $\phi_{z,w}$ gives the p(w|z): probability of word w in topic z. The topics follow a Dirichlet distribution parameterized by β.
- Each **document d (review)** is a distribution over topics, represented as θ_d where $\theta_{d,z}$ gives the p(z|d): probability of topic z in document d. The documents follow a Dirichlet distribution parameterized by α.

Mathematical Formulation:

1 Random Initialization: Initially, each word w in each document d is assigned to a random topic z. This forms the basis for calculating the matrices ϕ (word-topic distributions) and θ (document topic distributions).

2 Iterative Refinement: For each word w in each document d, LDA adjusts the topic assignment based on the conditional distribution:

$$p(z_i = k \mid \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{k,-i}^{(w)} + \beta_w}{n_k^{(\cdot)} + W\beta} \times \frac{n_{k,-i}^{(d)} + \alpha_k}{n_d^{()} + K\alpha}$$

Here, $n_{k,-i}^{(w)}$ is the number of times word w is assigned to topic k, excluding the current instance. $n_k^{(\cdot)}$ is the total number of words assigned to topic k, and W is the vocabulary size. $n_{k,-i}^{(d)}$ is the number of times a word in document d is assigned to topic k, excluding the current instance. $n_d^{(\cdot)}$is the total number of words in document d, and K is the number of topics. (Blei et al., 2003).

OPTIMAL NUMBER OF TOPICS

Finding the optimal number of topics using the *FindTopicsNumber* function from the *ldatuning* package, which evaluates various topic numbers to determine the best fit based on different statistical metrics.

1. **Griffiths2004**: Measures the log-likelihood that the number of topics can explain the data. (Griffiths and Steyvers, 2004)
2. **CaoJuan2009**: Focuses on the consistency of word distributions across topics (Cao et al., 2009).
3. **Arun2010**: Assesses the matrix similarity of the topic-term distributions (Arun et al., 2010).
4. **Deveaud2014**: Evaluates the purity of topics based on their distinctiveness from each other (Romain Deveaud et al., 2014).

Based on FIGURE 27, the optimal number of topics for modelling the hotel review data is both positive and negative after minimising Arun2010 and Caojuan2009 and maximising arun2010 and Deveaud2014. We get the final output is a Gibbs-based LDA topic model for **positive reviews** with **7 topics.**

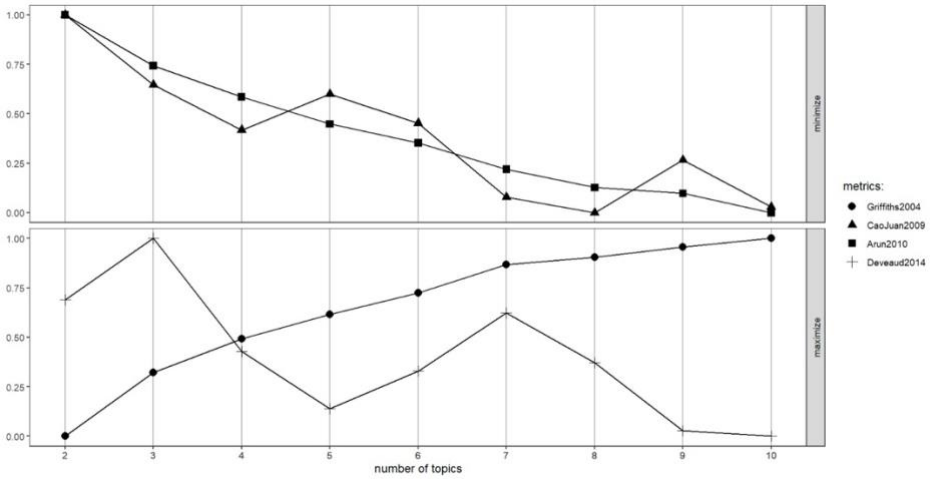

FIGURE 27 OPTIMAL TOPIC FOR POSITIVE

Similarly, the final output is a Gibbs-based LDA topic model for **negative reviews** with **7 topics.**

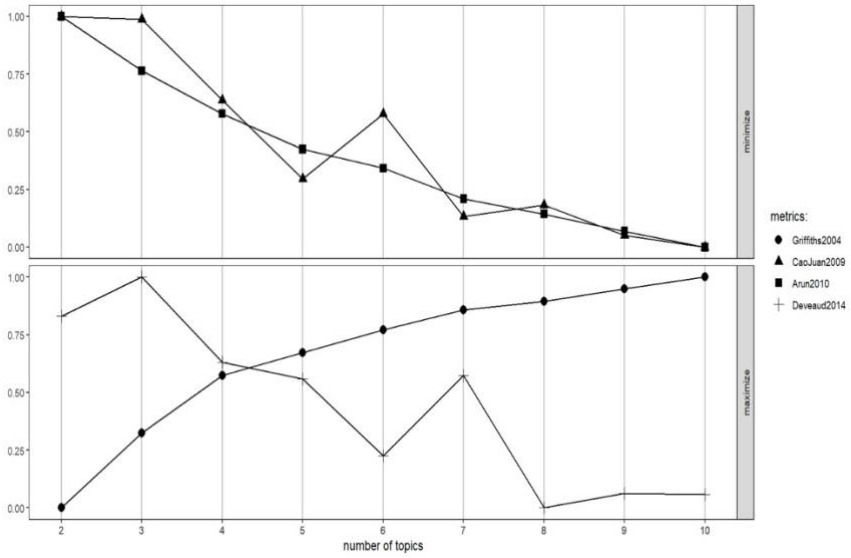

FIGURE 28 OPTIMAL TOPIC FOR NEGATIVE

We get matrices **phi** and **theta** extracted from the LDA output to represent the distribution of words across topics and the distribution of topics across documents, respectively, with top terms for each topic displayed below.

```
      Topic 1      Topic 2     Topic 3  Topic 4 Topic 5  Topic 6    Topic 7
 [1,] "walk"       "room"      "get"    "room"  "hotel"  "good"     "stay"
 [2,] "station"    "breakfast" "one"    "check" "stay"   "hotel"    "staff"
 [3,] "tube"       "bed"       "really" "even"  "staff"  "room"     "great"
 [4,] "minute"     "small"     "just"   "hotel" "lovely" "service"  "friendly"
 [5,] "hotel"      "bathroom"  "place"  "day"   "make"   "london"   "location"
 [6,] "area"       "night"     "need"   "book"  "time"   "bar"      "clean"
 [7,] "close"      "nice"      "can"    "night" "feel"   "price"    "helpful"
 [8,] "restaurant" "shower"    "also"   "give"  "will"   "nice"     "comfortable"
 [9,] "london"     "floor"     "want"   "make"  "tea"    "location" "excellent"
[10,] "park"       "large"     "didnt"  "ask"   "visit"  "food"     "recommend"
```

FIGURE 29 WORDS CORRESPONDING TO EACH TOPIC FOR POSITIVE REVIEWS

```
Topic 1    Topic 2      Topic 3 Topic 4  Topic 5    Topic 6     Topic 7
"location" "staff"      "night" "room"   "room"     "get"       "hotel"
"small"    "good"       "like"  "good"   "bed"      "check"     "stay"
"london"   "area"       "one"   "hotel"  "bathroom" "book"      "breakfast"
"good"     "also"       "just"  "make"   "shower"   "say"       "price"
"clean"    "food"       "place" "day"    "floor"    "back"      "close"
"walk"     "bar"        "stay"  "first"  "door"     "reception" "nice"
"great"    "quite"      "need"  "expect" "look"     "give"      "however"
"station"  "service"    "work"  "can"    "bad"      "arrive"    "london"
"friendly" "restaurant" "didnt" "star"   "window"   "tell"      "noise"
"tube"     "even"       "use"   "offer"  "sleep"    "pay"       "around"
```

FIGURE 30 WORDS CORRESPONDING TO EACH TOPIC FOR NEGATIVE REVIEWS

TOPIC LABELLING

Now assigning labels to **positive reviews** to "Service Quality", "Room Comfort", "Location and Accessibility", "Guest Experience", "Facilities and Services", "Booking and Management", and "Value and Dining". Then link with a positive review data frame.

Rating	Review	Topic
4	just couple day london pick hotel close need pretty ce...	Location and Accessibility
5	find hotel late room look somewhere near leicester sq...	Guest Experience
5	first stay hotel glad make change usual haunt small h...	Facilities and Services
4	just come back say night can say place firstly great lo...	Location and Accessibility
4	stay twin room annexe build room blissfully quiet cle...	Room Comfort

FIGURE 31 DATA FRAME FOR POSITIVE

Similarly assigning labels to **negative reviews** to "Transportation and Location", "Customer Service and Staff", "Sleep Quality and Noise", "Room Quality and Amenities", "Booking and Administration", "Food and Dining", and "Cost and Value". Then link with a negative review data frame.

Rating	Review	Topic
2	know friend book cheapy package expect high standa...	Transportation and Location
3	first floor room frost window view do take glitzy phot...	Customer Service and Staff
3	location great touristy do plan get around rush weeke...	Cost and Value
3	night weekend may price pay room hotel good deal h...	Booking and Administration
2	read review book expect bad need hotel near padding...	Sleep Quality and Noise

FIGURE 32 DATA FRAME FOR NEGATIVE

The extracted topic probabilities for each **positive review** demonstrate how the LDA model assigns a distribution of topics to individual reviews. Each value represents the likelihood that a given review pertains to one of the predefined topics such as "Service Quality," "Room Comfort," or "Value and Dining." For example, the first review has a 24.92% probability of being about "Location and Accessibility," indicating that this topic is a significant component of that review. This distribution enables the most prominent aspects or themes in each customer's feedback to be analysed in a more sophisticated way.

Service Quality	Room Comfort	Location and Accessibility	Guest Experience	Facilities and Services	Booking and Management	Value and Dining
0.08374384	0.1458128	0.24926108	0.11822660	0.20098522	0.12512315	0.07684729
0.18017058	0.1055437	0.08315565	0.23240938	0.14285714	0.16524520	0.09061834
0.12190476	0.1085714	0.13523810	0.13523810	0.21523810	0.12190476	0.16190476
0.10924370	0.2072829	0.21708683	0.07983193	0.14845938	0.11904762	0.11904762
0.15027829	0.3320965	0.10482375	0.13079777	0.11131725	0.07884972	0.09183673
0.07440476	0.1473214	0.14732143	0.26190476	0.12648810	0.10565476	0.13690476
0.13172542	0.1187384	0.14471243	0.10575139	0.09276438	0.19666048	0.20964750
0.18511066	0.1006036	0.21327968	0.10060362	0.10060362	0.17102616	0.12877264
0.09743252	0.1988150	0.18959842	0.23107307	0.12508229	0.09743252	0.06056616
0.20879121	0.1010989	0.11648352	0.17802198	0.13956044	0.13186813	0.12417582

FIGURE 33 PROBABILITY DISTRIBUTION OF WORDS IN POSITIVE

Similarly, the first **negative review** primarily highlights concerns about Sleep Quality and Noise (22.36%), suggesting significant dissatisfaction in this area, followed by Room Quality and Amenities (18.01%) and Food and Dining (15.30%). Less emphasis is placed on Cost and Value (7.69%), indicating minor concerns about pricing or value received.

Transportation and Location	Customer Service and Staff	Sleep Quality and Noise	Room Quality and Amenities	Booking and Administration	Food and Dining	Cost and Value
0.13664596	0.1203416	0.2236025	0.18012422	0.10947205	0.15295031	0.07686335
0.11003861	0.1505792	0.1505792	0.12355212	0.19111969	0.15057915	0.12355212
0.08869702	0.1436421	0.2919937	0.07221350	0.05572998	0.27551020	0.07221350
0.18290043	0.1450216	0.1601732	0.08441558	0.11471861	0.11471861	0.19805195
0.19447779	0.1104442	0.1776711	0.17767107	0.16926771	0.06842737	0.10204082
0.11269841	0.1015873	0.1793651	0.15714286	0.12380952	0.13492063	0.19047619
0.15654405	0.1266039	0.1505560	0.14456801	0.19846022	0.09067579	0.13259196
0.14370748	0.1079932	0.1258503	0.07227891	0.17346939	0.23894558	0.13775510
0.15138593	0.1066098	0.1364606	0.12153518	0.22601279	0.10660981	0.15138593
0.28142857	0.1014286	0.1014286	0.10142857	0.16142857	0.11142857	0.14142857

FIGURE 34 PROBABILITY DISTRIBUTION OF WORDS IN NEGATIVE

TOP 3 FACTORS

To calculate the top three topics in an LDA model, compute the mean topic probabilities across all documents from the matrix `theta` and then sort these values in descending order. Select the first three entries of the sorted list to identify the topics that are most prevalent across the corpus.

For **positive reviews**, "Service Quality" leads as the most prevalent topic at 14.41%, closely followed by "Value and Dining" at 14.39% and "Room Comfort" at 14.37%, highlighting these as key areas of customer satisfaction.

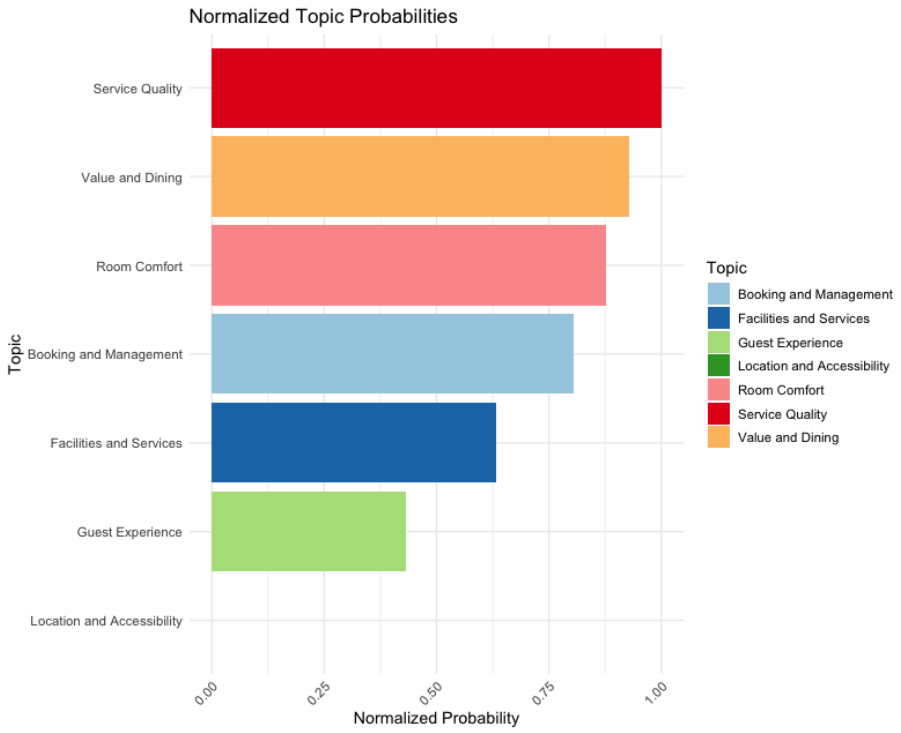

FIGURE 35 NORMALISED PROBABILITY DISTRIBUTION OF POSITIVE TOPIC

For **negative reviews**, "Food and Dining" is the most significant topic at 14.53%, slightly ahead of "Transportation and Location" at 14.52%, and "Booking and Administration" at 14.47%, indicating major areas of concern for dissatisfied customers.

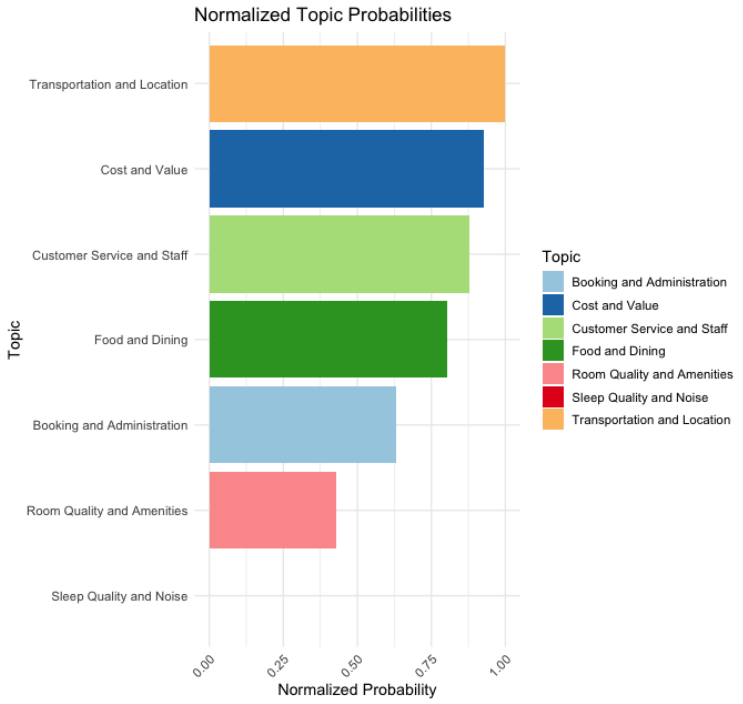

FIGURE 36 NORMALISED PROBABILITY DISTRIBUTION OF NEGATIVE TOPIC

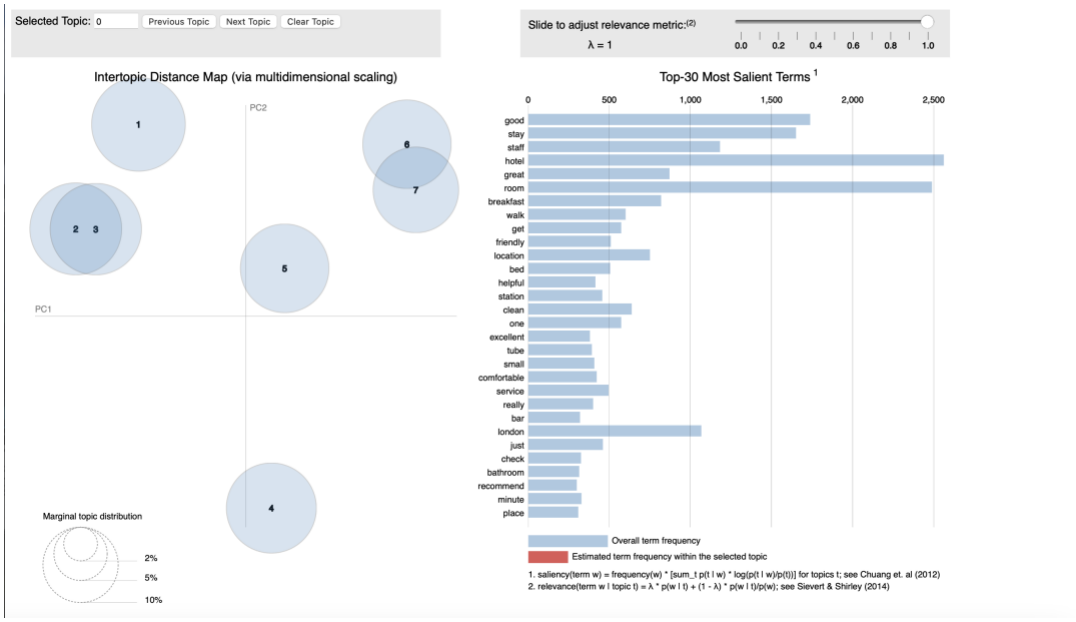

FIGURE 37 LDA VIZ FOR POSTIVE

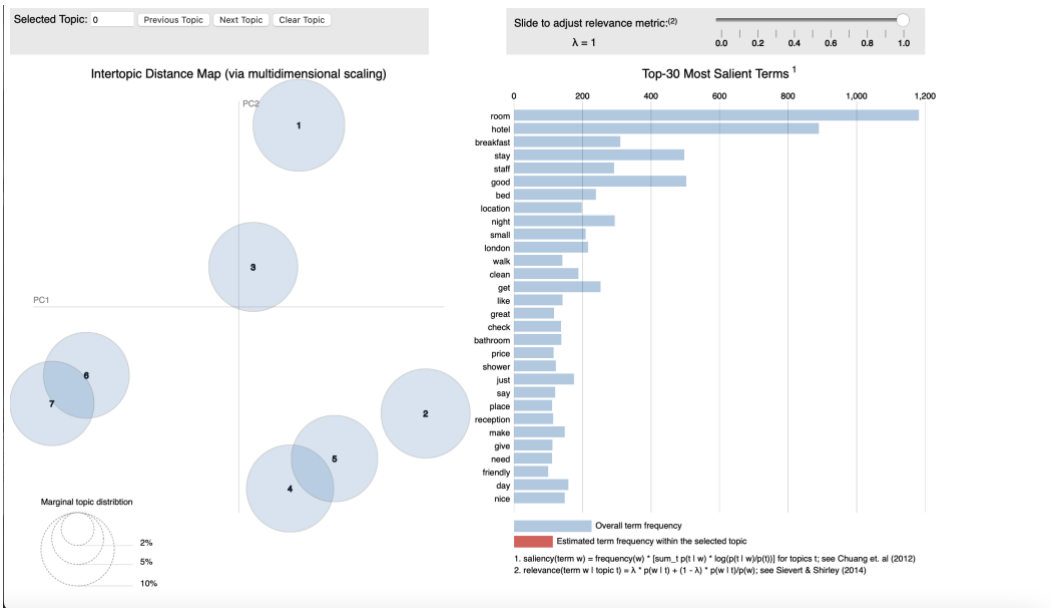

FIGURE 38 LDA VIZ FOR NEGATIVE

Conclusion

The analysis of 2,000 hotel reviews using Latent Dirichlet Allocation (LDA) was able to successfully identify the significant themes within both positive and negative feedback. The three most important satisfaction factors for positive reviews are Service Quality, Value and Dining, and Room Comfort, implying that they are areas of service that meet or exceed guest expectations. However, for negative reviews, the areas of dissatisfaction were Food and Dining, Transportation and Location, and Booking and Administration, signifying a very high level of dissatisfaction. Such information will help hotel management to focus on the areas that need to be improved and to keep the strengths to increase guest satisfaction and satisfaction levels. Consequently, the topic modelling was a successful technique for extracting the main points from customer feedback and turning them into practical recommendations.