NLP

EPK

04/02/2022

## Introduction

Understanding the performance of a hotel’s hospitality by its cleanliness, timely response to guest concerns, and other characteristics are measured by visitor reviews as well as their overall experience in the hotel. As text analytics is the process of extracting high-quality information from text, it plays a critical role in establishing the factors that influence determining the best and worst evaluations from consumers by analysing text review data with the help of corresponding reviews.

## Text Analysis

### Text Pre-Processing

Pre-processing the data, such as cleaning and extracting the correct text data for the topic modelling process, is required because the text in the dataset is in an unstructured format. The topic model aids in the analysis of the best and worst hotel reviews by utilising significant factors in all topics. The study started with the installation of required packages and a review of the dataset’s structure, which aids in understanding the data class type.

## 'data.frame': 1000 obs. of 2 variables:  
## $ Review.score: int 4 5 2 4 5 4 4 5 3 5 ...  
## $ Text.1 : Factor w/ 10000 levels "　ロンドンの高級ホテルとしては、比較的手軽な価格だと思います。 予約した部屋はLOFTでしたが、LOFT BALCONYにアップ"| \_\_truncated\_\_,..: 9718 1643 9578 655 1666 8470 3905 30 6877 3295 ...

#### Review Criteria

The Hotel.csv dataset has two variables: review score and review text. The variable review score’s class type is integer, however it must be converted to factor because it is used as a scale factor of satisfaction, which is a qualitative variable. This modification makes it easier to determine whether a review is positive or negative.The dataset contains a large amount of text data, and from that dataset, 1000 reviews are picked at random using the seed value, and with those reviews, the text written in English language must be extracted using the textcat function.

The review score was used as a criterion for selecting excellent and negative reviews. The dataframe contains English reviews that have been filtered using the review score function, which are used with the dplyr function “filter.” The review score of ‘4’ and ‘5’ is classified as excellent/positive, whereas the score of “1” and “2” is classified as bad/negative.

#### Text Cleaning

The first step is to convert the review text to UTF-8 format so that any data outliers that can’t be pre-processed with tm\_map or other text mining methods can be recognised for tidying the data. The next step in the analysis is text pre-processing, in which the good and bad reviews should be converted to vector format using the corpus function, and the document strings can be Lemmatized using the tm map function for cleaning and finding the frequency of terms occurring in text for each topic model using the document matrix function.

With the use of lemmatization, words like “booking,” “booked,” and “book” that are linked to room checking can be shortened to “book.” The document term matrix has a parameter named control, which is used to tidy data using the stopwords, remove\_white\_space, remove\_punctuation, and to\_lower functions.

## [1] "Excelelnt Review Document Term Matrix"

## <<DocumentTermMatrix (documents: 594, terms: 4108)>>  
## Non-/sparse entries: 30844/2409308  
## Sparsity : 99%  
## Maximal term length: 17  
## Weighting : term frequency (tf)

## [1] "Bad Review Document Term Matrix"

## <<DocumentTermMatrix (documents: 66, terms: 1525)>>  
## Non-/sparse entries: 3972/96678  
## Sparsity : 96%  
## Maximal term length: 15  
## Weighting : term frequency (tf)

#### Word Cloud

The next step is to determine the frequency of each word in the text document using the apply function, which returns the number of words in each document when passed the argument sum. To find the frequency of each word in the full document, convert the dtmdocs of excellent and bad reviews to matrix form using the as.matrix function and use the colsums function to find the frequency of each word in the entire text which helps to do the visulaisation for the frequent terms in excellent and bad review.

## hotel room good stay staff   
## 1020 1003 740 636 495

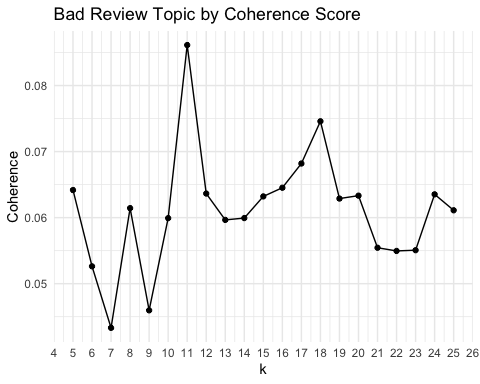
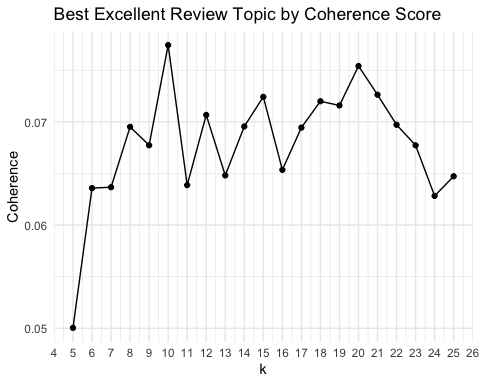
## room hotel stay good night   
## 147 113 67 58 51

 The frequent words in the excellent review, according to the wordcloud, are hotel, room, good, stay, and staff, all of which are positive words obtained from the excellent review. According to the wordcloud, the most common words in the negative review are room, hotel,and stay, all of which are evident in the bad review.

### Topic Modelling

Topic models are machine-learning techniques that use massive collections of documents to find hidden or latent theme structures (i.e. subjects). No prior labelling or annotation is required because the latent theme structures emerge automatically from the statistical features of the documents. Thematic frameworks can then be utilised to automatically categorise or summarise documents on a scale that would be hard to accomplish manually (Syed and Spruit, 2017). The most common method for topic modelling is latent Dirichlet allocation, which is used to extract the major factors of good and negative reviews that can be analysed for topic selection.

#### LDA

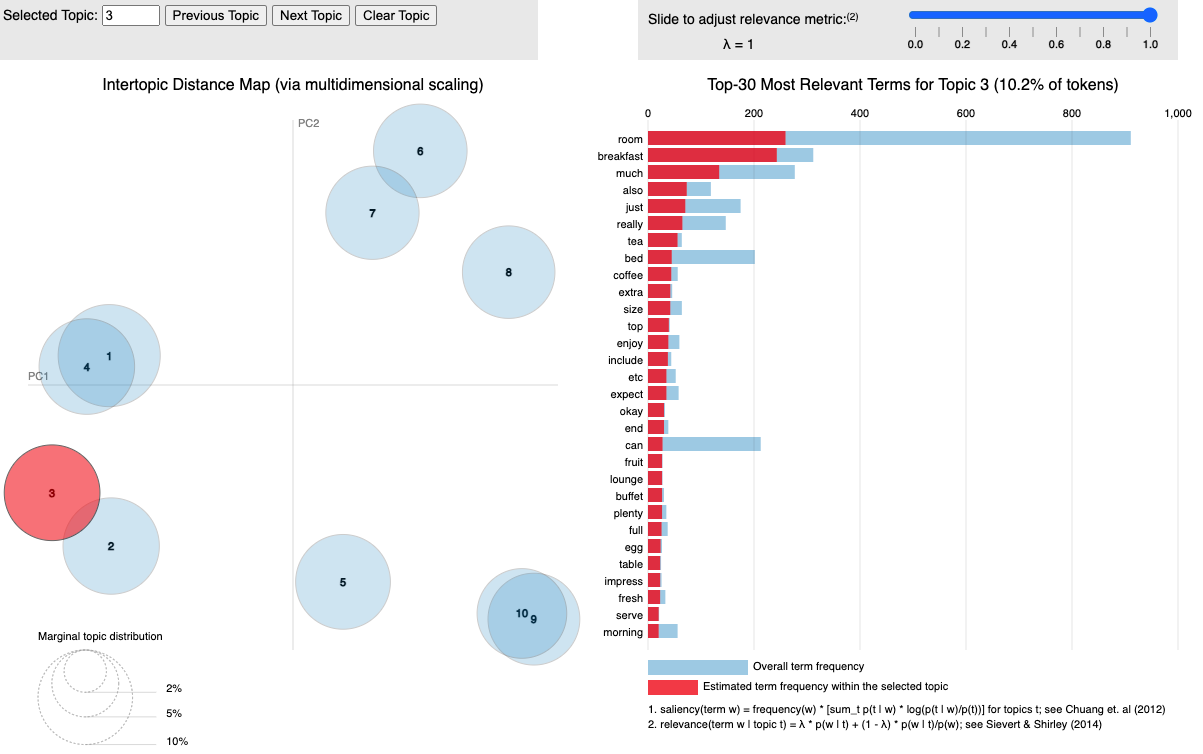
LDA aids in the generation of the probability of each word falling under the topic, which aids in the identification of the corpus hidden structure by using Gibbs method. The first step in calculating the coherence score is to approximate the posterior distribution, which allows to assign various probabilities to each word in the topic distribution. The words that occur the most frequently are the ones with the highest probability.This aids in determining the ‘k’ value using the coherence score, where ‘k’ is the number of modelling topics.  The graph clearly shows that there are 10 and 11 topics chosen for topic modelling for excellent and bad reviews, respectively.

The topic modelling is now carried out with the help of the k value acquired from the measure of coherence score, which aids in training the topic model using LDA and identifies the frequently occurring terms as well as identifying the probability of each word under the model’s topics. The same procedure is used with the good and bad review document term matrix to determine the probability of the most often appearing terms in each topic, which aids in the identification of the text model’s best factors.

#### Visualisation

#### Satisfaction of the Customers

From the excellent topic modelling the best three factors are topic 3,6 and 10 respectively. The graphs are shown below,

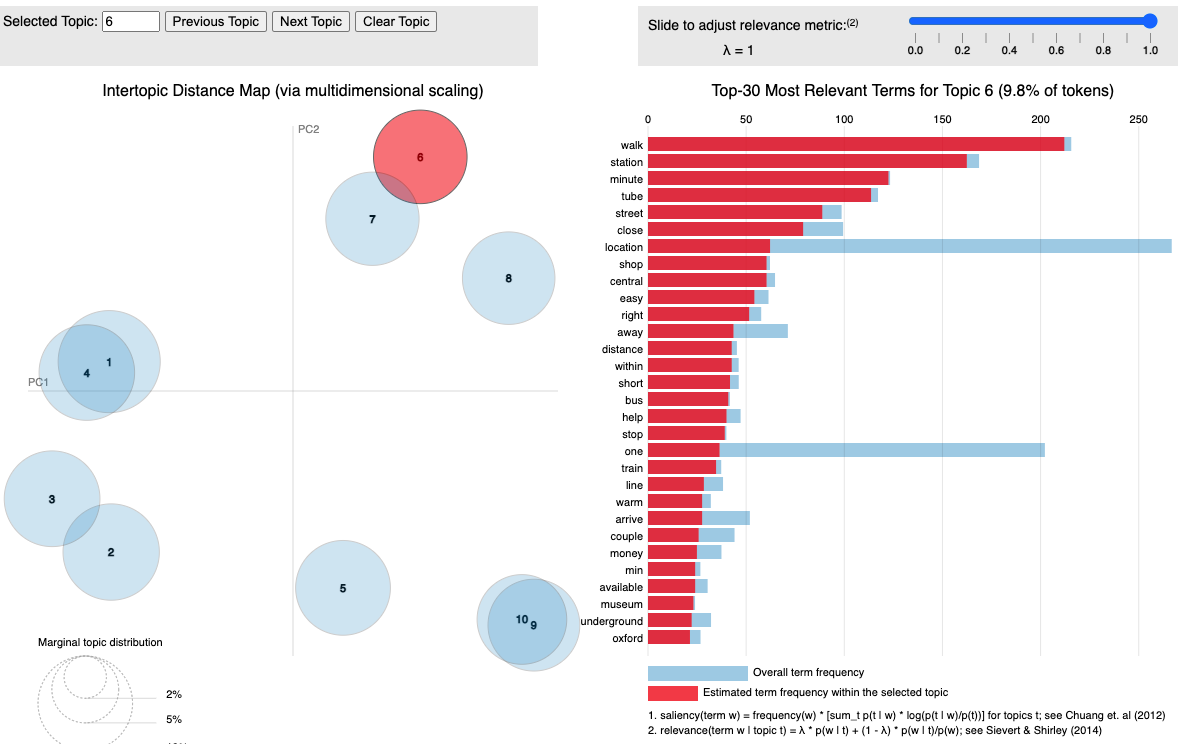


Excellent\_Topic-3

The positive ratings for topic 3 were related to the hotel’s food, which they offered with the highest quality food during breakfast time, and consumers enjoyed the buffet.

## [1] "The label for Topic 3;"

## [1] "FOOD"

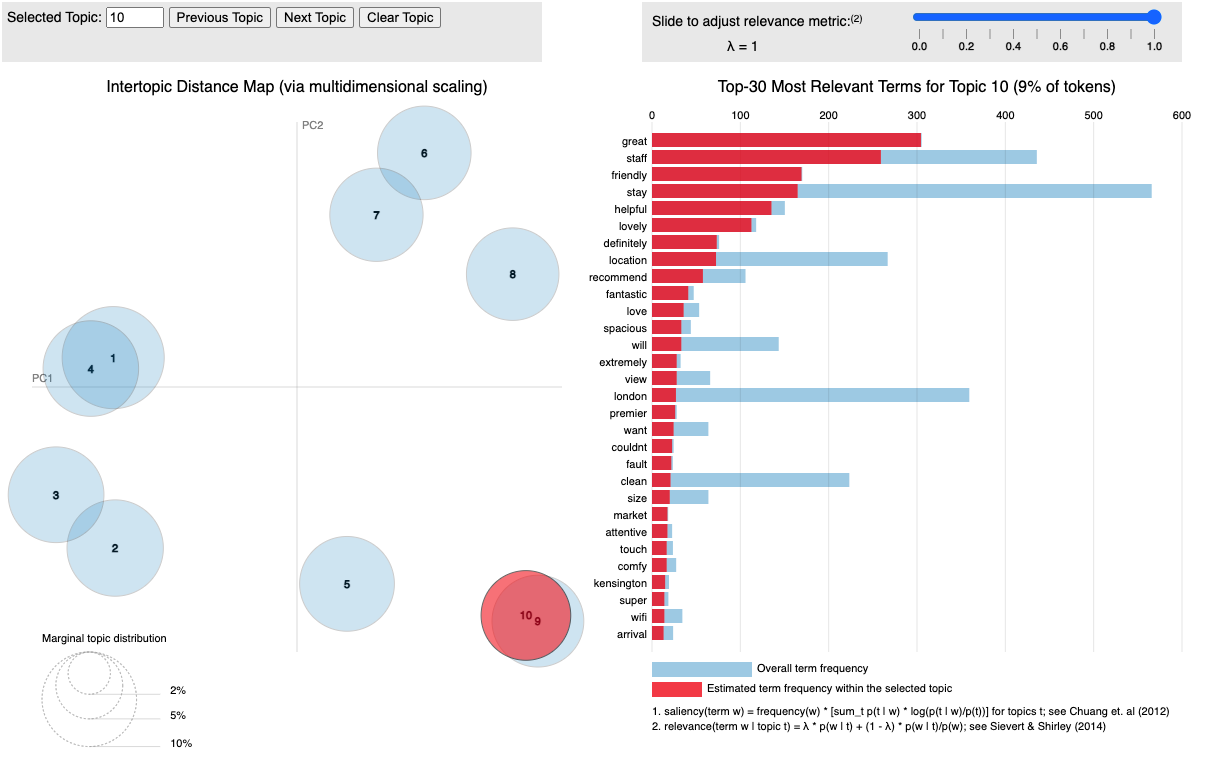


Excellent\_Topic-6

The hotel’s location, which is quite close to the main attraction and very easy to travel between with the use of an underground tube, was reviewed in topic 6.

## [1] "The label for Topic 6;"

## [1] "LOCATION"



Excellent\_Topic-10

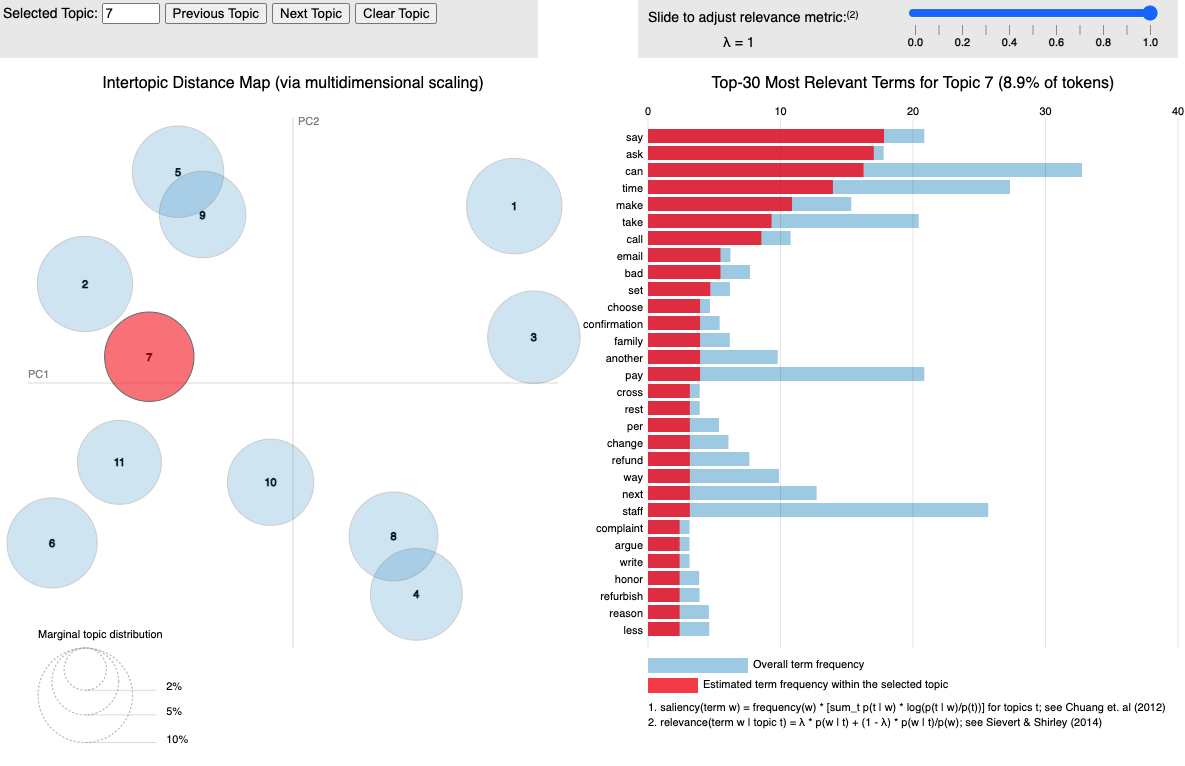
For topic 10, the hotel’s hospitality is complimented, with comments on the friendliness and attentiveness of the employees.

## [1] "The label for Topic 10;"

## [1] "HOSPITALITY"

#### Dissatisfaction of the Customers

Topics 7,9, and 11 are the three sources of dissatisfaction based on the negative review topic modelling. The graphs are shown below,

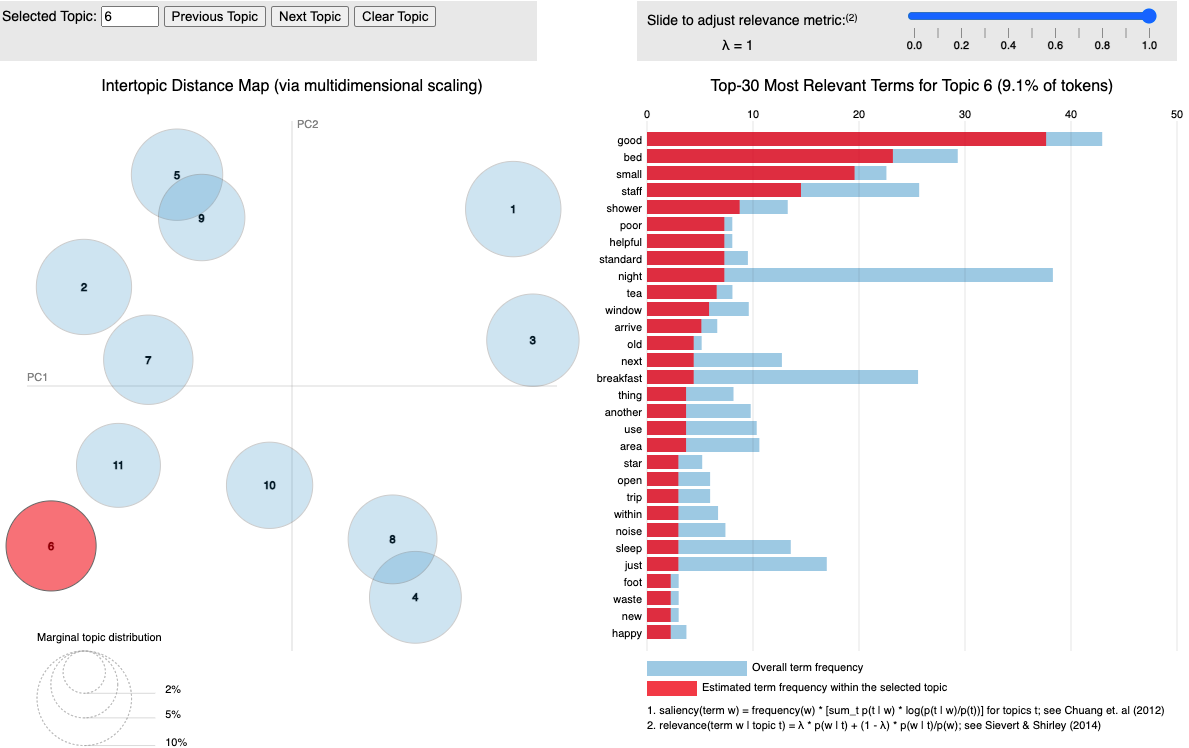


Bad\_Topic-7

The biggest source of dissatisfaction for topic 7 was the hotel’s poor customer service, as customers complained about refunds and argued over the phone conversation.

## [1] "The label for Topic 7;"

## [1] "Poor Customer Service"

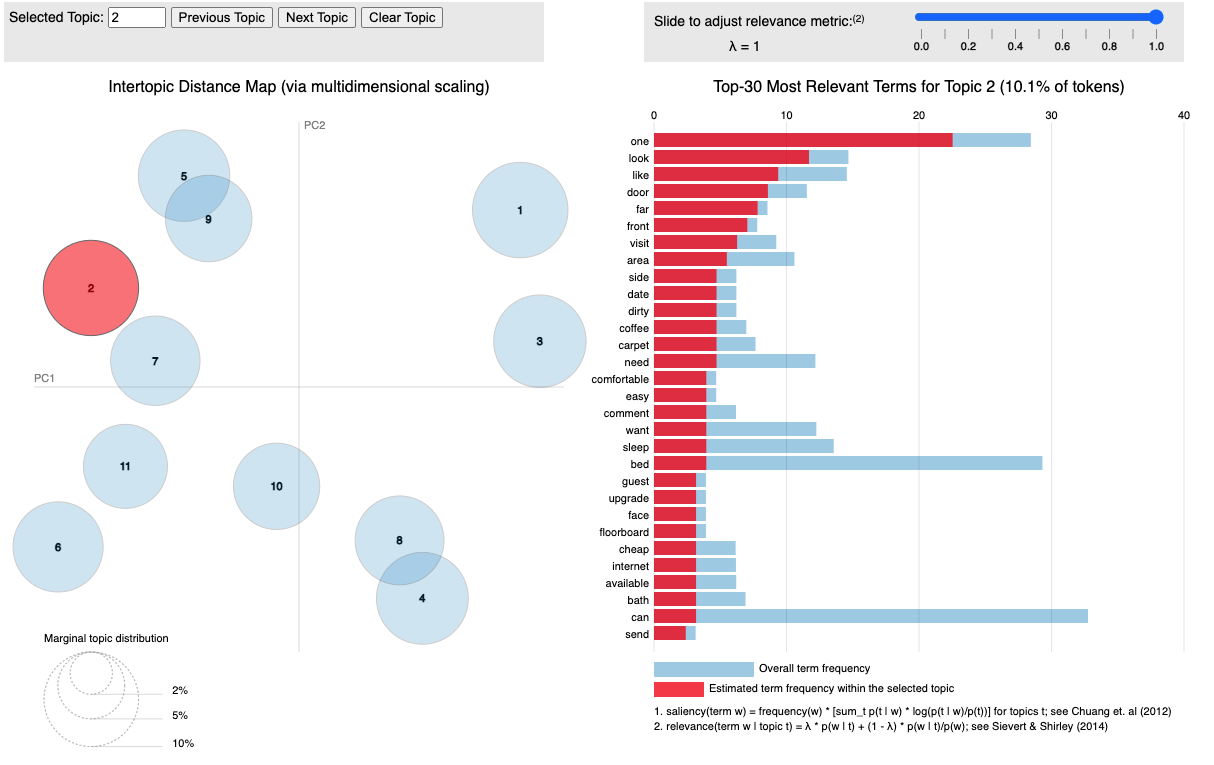


Bad\_Topic-6

The main dissatisfaction given by consumers for Topic 6 related to consumers are complaining about the bed size and shower issues.

## [1] "The label for Topic 6;"

## [1] "Bad Room Quality"

 The main dissatisfaction given by consumers for Topic 2 related to consumers are complaining about the visiting area is dirty.

## [1] "The label for Topic 2;"

## [1] "Unclean Area"

## Conclusion

The topic modelling helps to uncover satisfied and dissatisfied customer responses, providing insight for hotels to improve their systems if customers leave negative reviews. However, according to the analysis, the majority of customers are satisfied with the hotel, with approximately 75% of excellent reviews and only 10% of negative reviews.

One of the main drawback saw in the bad reviews, there is a mix of good reviews send by customers which was observed in topic modelling.