Regionality-Centric Topic Modeling Pattern Analysis: Luxury Hotels Review in Europe

Raymond Yu (cyu45@ur.rochester.edu)
Wonha Shin (wshin7@ur.rochester.edu)

I. Introduction

Booking.com, as a global platform, brings hosts and travelers from diverse nationalities and cultural backgrounds together. Understanding the nuances of guest feedback can give the hosts valuable insights, enabling them to tailor their offerings more effectively and enhance the overall guest experience. This project explores the heart of such feedback to extract patterns tied to travelers' regionality or inferred cultural backgrounds.

Europe serves as a multiracial melting pot for visitors worldwide, each bringing unique expectations, preferences, and feedback when using Booking.com. This project capitalizes on the richness of such diversity. By leveraging advanced data science techniques, we aim to decipher patterns in the comments and reviews guests leave. Such patterns can be instrumental in identifying what amenities or features resonate more with certain nationalities or even understanding if there are common pain points faced by guests from specific regions.

Using a combination of Natural Language Processing (NLP) techniques, we preprocessed and transformed the raw textual data from reviews into a structured format suitable for analysis. With the processed data, topic modeling algorithms like Gensim were employed to segment comments into distinct groups, intending to find common themes or sentiments shared by users from similar cultural or national backgrounds. The goal is to determine recurring themes or specific features consistently mentioned in conjunction with certain nationalities.

For the project, our aim is to present our motivation and rationale, highlighting why this project is necessary. We will scrutinize our datasets through an exploratory data analysis, providing a detailed, step-by-step overview of the methodologies employed. This will include a comprehensive examination of the data to understand its characteristics, limitations, and potential insights. Following this, we will delve into applying various analytical techniques, discussing each stage of our approach from data preprocessing to model implementation. Finally, we will thoroughly examine the results obtained, interpreting their implications and assessing their relevance to the project's objectives. This comprehensive approach ensures a clear understanding of our methods and findings, establishing the project's significance and contributions to the field.

II. Problem Motivation and Statement

In today’s globalized world, platforms like Booking.com or Airbnb bring hosts and travelers from various nationalities and cultural backgrounds. Each traveler, influenced by their national and cultural context, has distinct expectations and experiences when choosing their accommodation preferences. For hosts, understanding these unique preferences can be instrumental in offering a tailored and memorable experience, leading to better reviews, increased bookings, and higher guest satisfaction. Rather than manually shifting through vast amounts of reviews to discern nationality or cultural-specific feedback, it is neither feasible nor efficient.

As experienced travelers, we know that we have different considerations and priorities when choosing our accommodation according to our past experiences, values, backgrounds, cultures, and so on. Of all the attributes, nationality draws our attention the most because our countries, either Korea or Taiwan, are tourist spots in East Asia, with many foreigners visiting yearly. We are interested in what influences the foreigners who travel to Europe the most in choosing their places to stay and what kind of comments are left. When travelers visit Europe, they live in hotels or find a place via a popular housing booking platform, such as Booking.com or Airbnb. Even though their characteristics differ, they both have rating metrics and comments about the housing for analysis.

Our goal is to find a potential interesting pattern for users on Booking.com based on their nationality and reviews they had given using NLP on reviews to frequent patterns we mined from comments on different nationalities. Also we will consider two potential users of our analysis: one tailored for hosts and the other for users. For hosts, our study provides insights into the preferences and expectations of guests from various nationalities or regions. By understanding these preferences, hosts can optimize their properties for Search Engine Optimization (SEO), attracting more international travelers and potentially enhancing their ratings and earnings. For users, our analysis serves as a supplementary tool, enabling them to look beyond mere rating scores and understand the deeper sentiments expressed in previous guests' comments. We will begin our exploration by examining review datasets.

III. Dataset Analysis

1. Acquisition & Overview

Initially, we planned to employ the API or web crawler to systematically browse and extract relevant review data from Airbnb directly to gather the necessary data.; however, we could not perform that procedure due to its policy starting this year. We found a dataset from Kaggle, “515K Hotel Reviews Data in Europe,” as an alternative. The dataset was from booking.com and scraped by the author Jiashen Liu six years ago. Although the data is a bit old, we could still perform our proposed procedure and method to gain some insight and pattern first and hope to apply them to the further dataset when applicable.

The dataset contains 515K rows of data, including 1,493 luxury hotels across Europe and 17 columns. Seventeen columns describe each review and corresponding hotel. Please refer to Appendix 1 for further column details.

2. Data Analysis

(1) Discretization (Concept Hierarchy)

In our analysis, we aimed to uncover distinctive patterns based on the geographic origins of reviews, encompassing nationalities, sub-regions, and regions. However, given the dataset's extensive scope, featuring 226 nationalities, we opted for a more feasible approach by focusing on regional and sub-regional patterns.

To structure our analysis, we aligned our regional and sub-regional classifications with those defined by Google Maps as follows.

- Regions: {'Africa', 'Americas', 'Antarctica', 'Asia', 'Europe', and 'Oceania'}
- Sub-regions :{ 'Antarctica', 'Australia and New Zealand', 'Central Asia', 'Eastern Africa', 'Eastern Asia', 'Eastern Europe', 'Latin America and the Caribbean', 'Melanesia', 'Micronesia', 'Middle Africa', 'Northern Africa', 'Northern America', 'Northern Europe', 'Polynesia', 'South-eastern Asia', 'Southern Africa', 'Southern Asia', 'Southern Europe', 'Western Africa', 'Western Asia', and 'Western Europe'. }

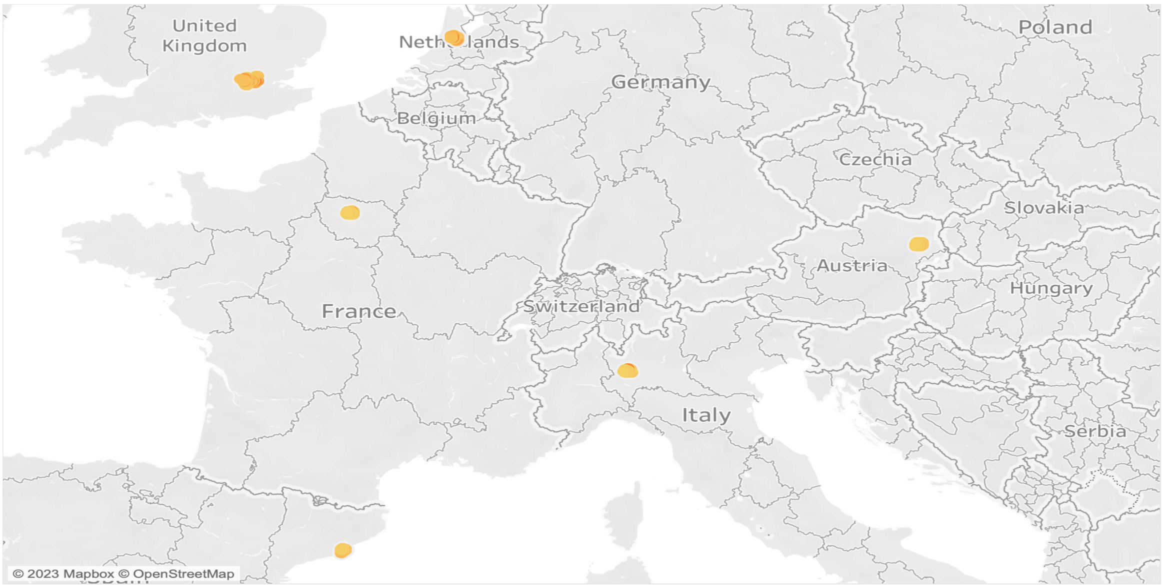
(2) Exploratory Data Analysis

After the data cleaning, we explored and summarized the columns below to better understand the data.

I. Hotel_Address, Hotel_Name

There are 1,493 hotels in the dataset, with 60K in Paris, France, 260K in London, United Kingdom, 57K in Amsterdam, Netherlands, 37K in Milano, Italy, 39K in Wien, Austria, and 60K in Barcelona, Spain, respectively.

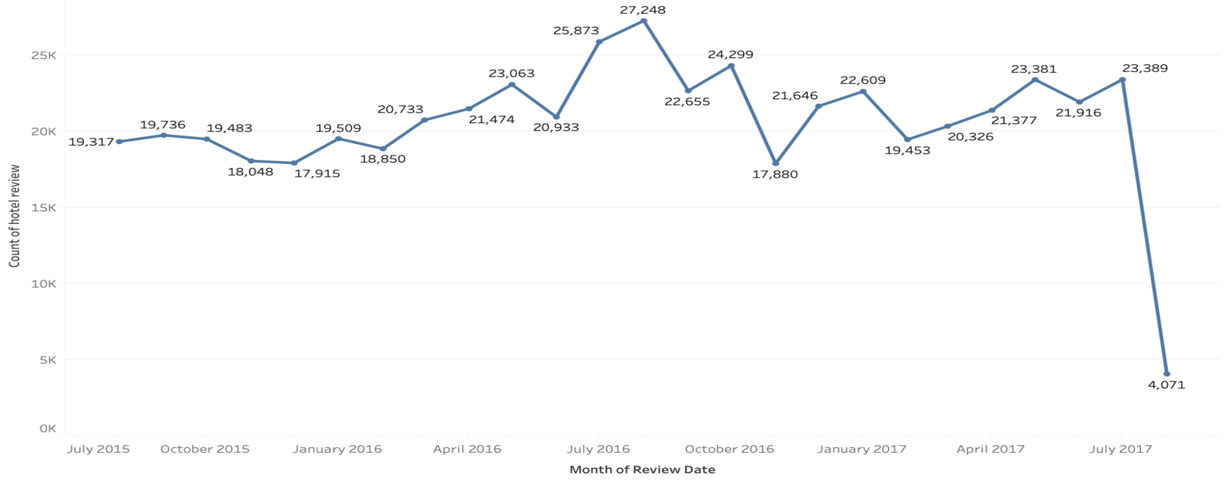
Hotel Geography



II. Review_Date

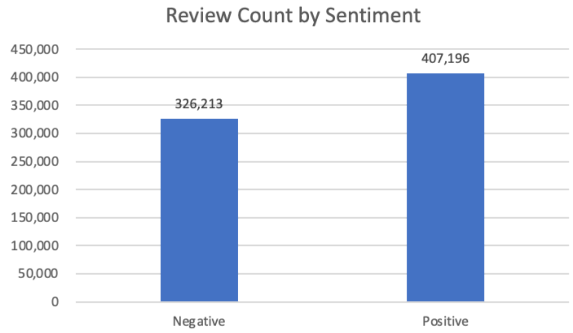
The data range spans from August 4, 2015, to August 3, 2017, encompassing two years. We could see an equivalent count of reviews with a drastic drop in August 2017, as the dataset only contained three days of data counted within August 2017.

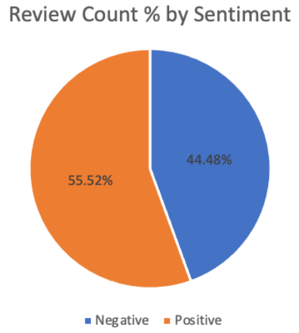
Review Count by Date



III. Counts of Negative_Review and Positive_Review

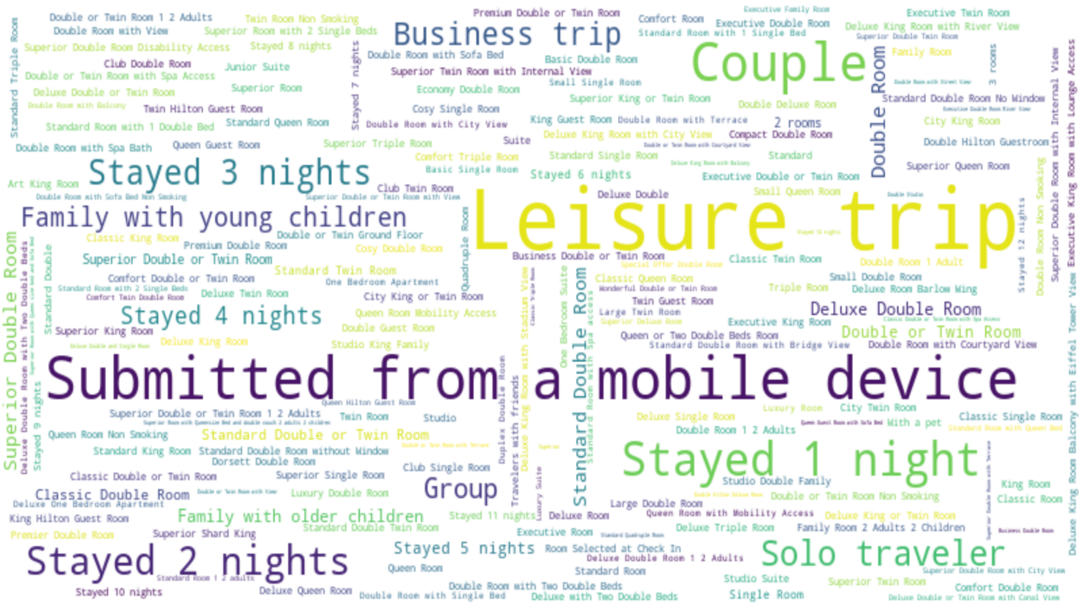
There are 326K and 407K negative and positive reviews, respectively. In proportion, negative reviews account for 44.48% of the total reviews, while the remaining 55.52% are positive reviews. So, we have a fair proportion of negative and positive reviews instead of an imbalanced ratio of reviews.





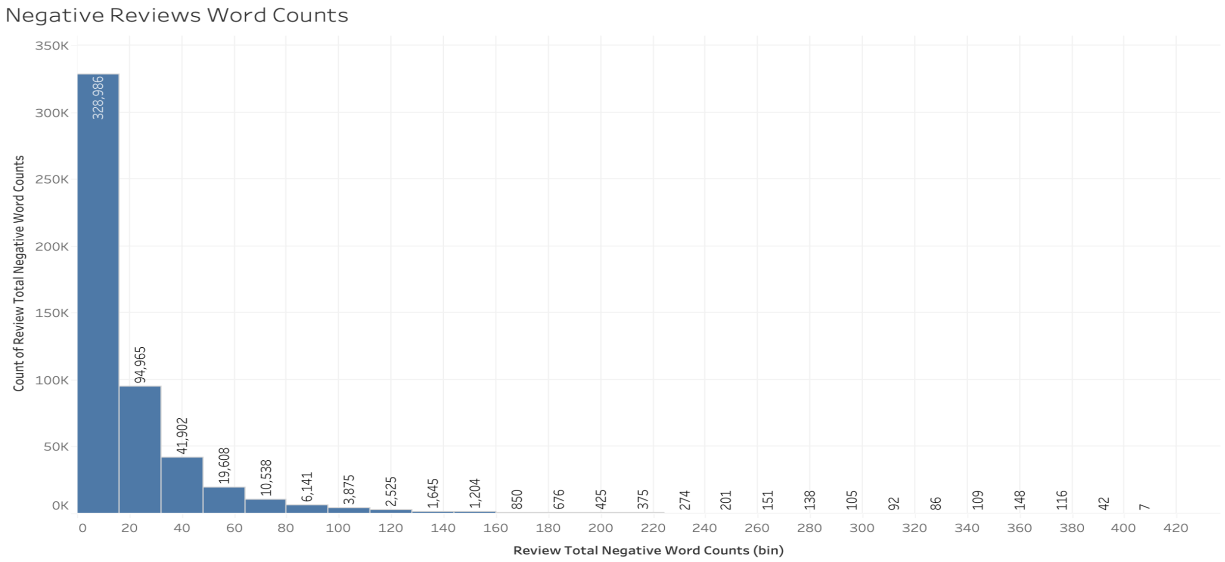
IV. Tags

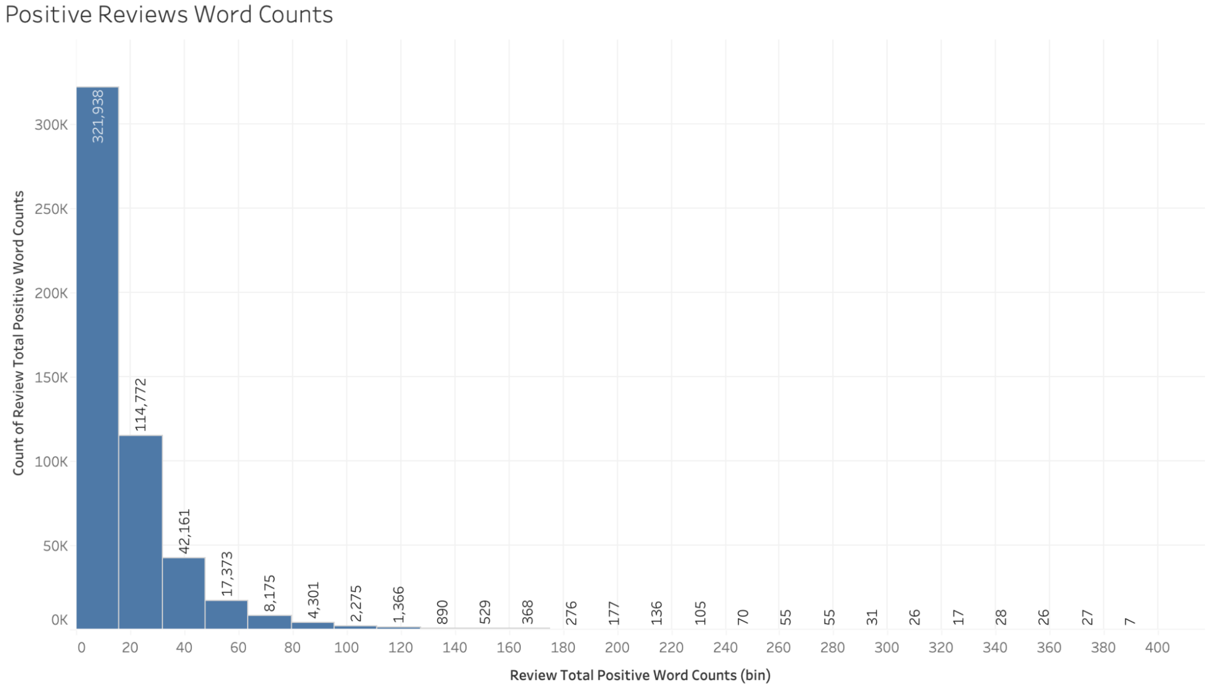
The tag column indicates the kind of tags the reviews gave the hotels. The following graph is the word cloud of tags for all cleaned datasets.



V. Review_Total_Negative_Word_Counts & Review_Total_Positive_Word_Counts

We could see that no matter whether the review is negative or positive, the word count of the reviews has the pretty same right-skewed distribution and about 90% of reviews won't exceed 40 words for their comments.





IV. Data Preprocessing

As the dataset already has been lowercase, we processed below to cleanse datasets to fit our analysis purpose.

1. Space trimming for string data type

To prevent any space value from not being considered a null value in columns whose data type is a string, we first trimmed the space in those columns and found there are 523 blank data in column “Review_Nationality.” In addition, there are 183,849 and 59 empty data in columns “Negative_Review” and “Positive_Review” with both fields blank, respectively.

2. Missing Data

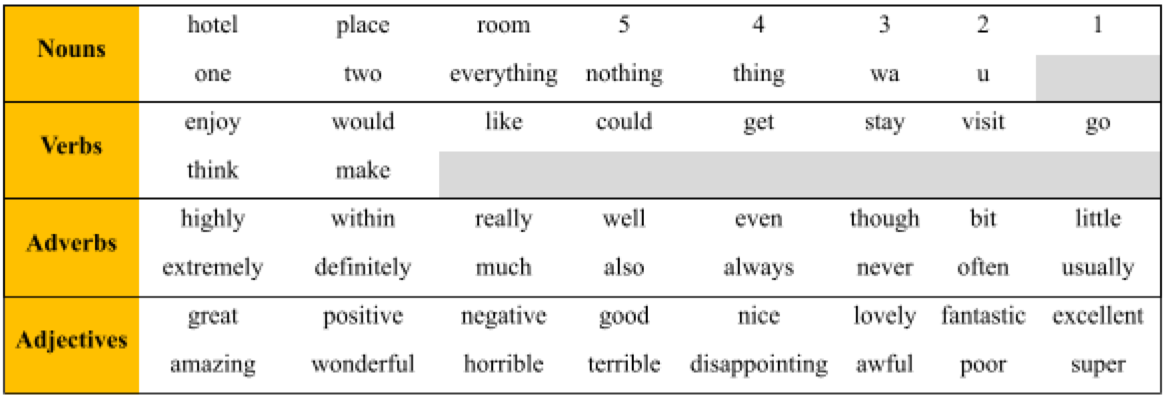
After an initial checking, besides blank data in the space trimming part, there are also 3,265 missing entries in columns “lat” and “lng.” However, as our goal is mainly about review and nationality, we decided only to drop off rows of data with column “Review_Nationality” blank, and both columns “Negative_Review” and “Positive_Review” were blank. As a result, there are 515,156 rows of data remaining, with only 582 rows of data dropped.

3. Merged Negative & Positive Reviews into One Combined Dataset

We merged the negative and positive reviews into a single, combined dataset to build a comprehensive topic modeling analysis. This unified approach allows for a more comprehensive understanding of customer preferences with the same standards.

4. Cleaning Words with Stopwords

In our analysis, we initially utilized the standard English stopword list provided by the NLTK library. Through iterative refinement involving several trial runs of our code, we identified and included additional frequently occurring words in the reviews that lacked significant meaning. These were classified into specific categories such as nouns, verbs, adverbs, and adjectives. Examples of these categories are provided below. Significantly, our customized stopword list intentionally omits certain words like "spacious", "friendly", "helpful", and "close", among others. Despite their frequent occurrence, we assessed these terms as meaningful and relevant to the context of our analysis, especially in evaluating hotel and accommodation experiences.



Nouns	hotel	place	room	5	4	3	2	1
	one	two	everything	nothing	thing	wa	u	
Verbs	enjoy	would	like	could	get	stay	visit	go
	think	make						
Adverbs	highly	within	really	well	even	though	bit	little
	extremely	definitely	much	also	always	never	often	usually
Adjectives	great	positive	negative	good	nice	lovely	fantastic	excellent
	amazing	wonderful	horrible	terrible	disappointing	awful	poor	super

<Table> A list manually added for Stopwords removal

5. Tokenization & Lemmatization

We also conduct tokenization & lemmatization of the review sets. Tokenization breaks down text into smaller units called tokens: words, phrases, symbols, or other elements. It's like slicing a sentence into individual words. Meanwhile, Lemmatization involves reducing words to their base or root form. For example, 'running', 'ran', and 'runs' would all be lemmatized to 'run'. We can convert the word to its meaningful base form through this process. Both processes were essential to our analysis.

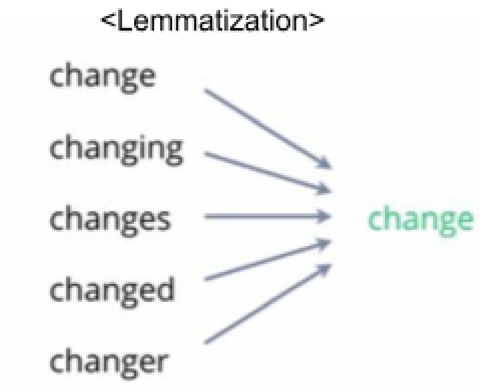


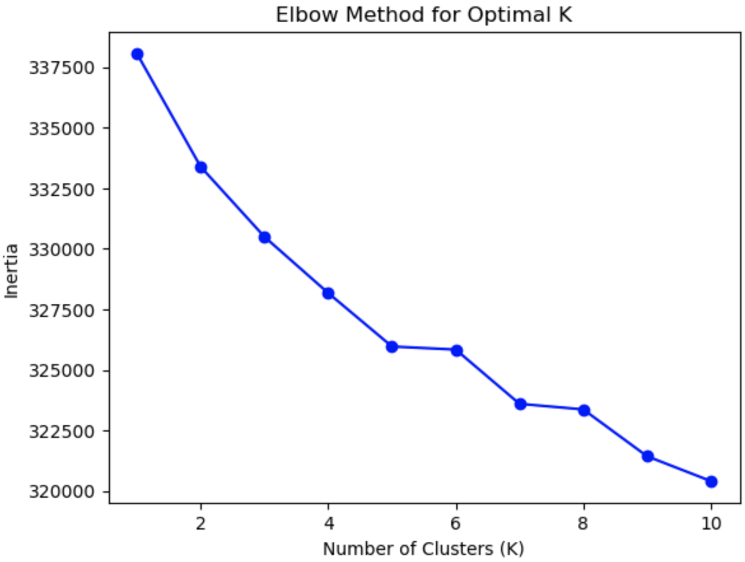
Image Source: https://www.quora.com/What-is-lemmatization-in-NLP

IV. Topic Modelings

For specific topic modeling, two prevalent methods are used: Clustering and Latent Dirichlet Allocation (LDA) modeling. K-means clustering, commonly applied for unsupervised clustering, groups data points into K clusters based on similarity. However, there may be better choices for analyzing frequent word sets in text data, especially considering different regions or sub-regions. This limitation stems from the fact that K-means does not inherently process text data and fails to capture the semantic relationships between words or phrases. Despite this, we have evaluated its performance in our context to confirm if we can draw any meaningful concepts from conducting this method.

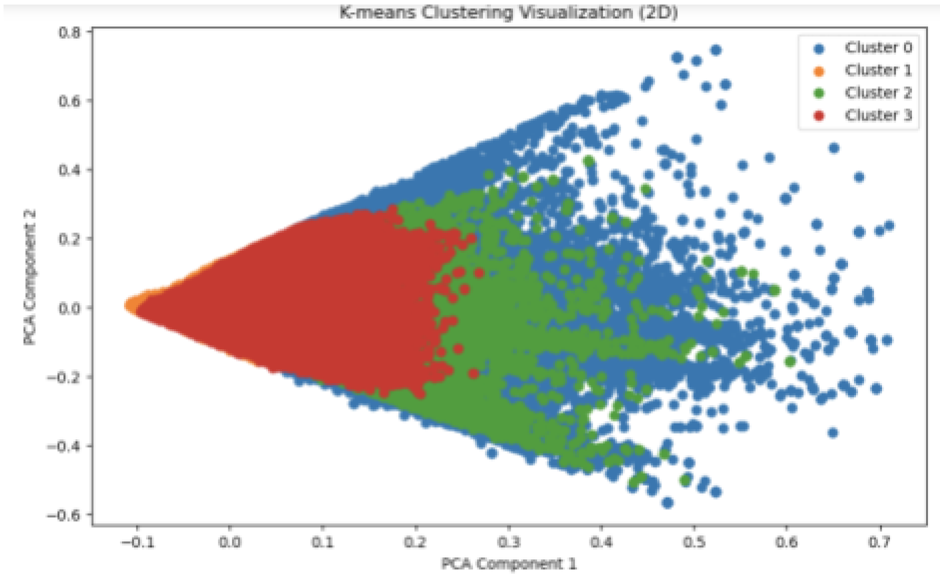
1. K Means Clustering

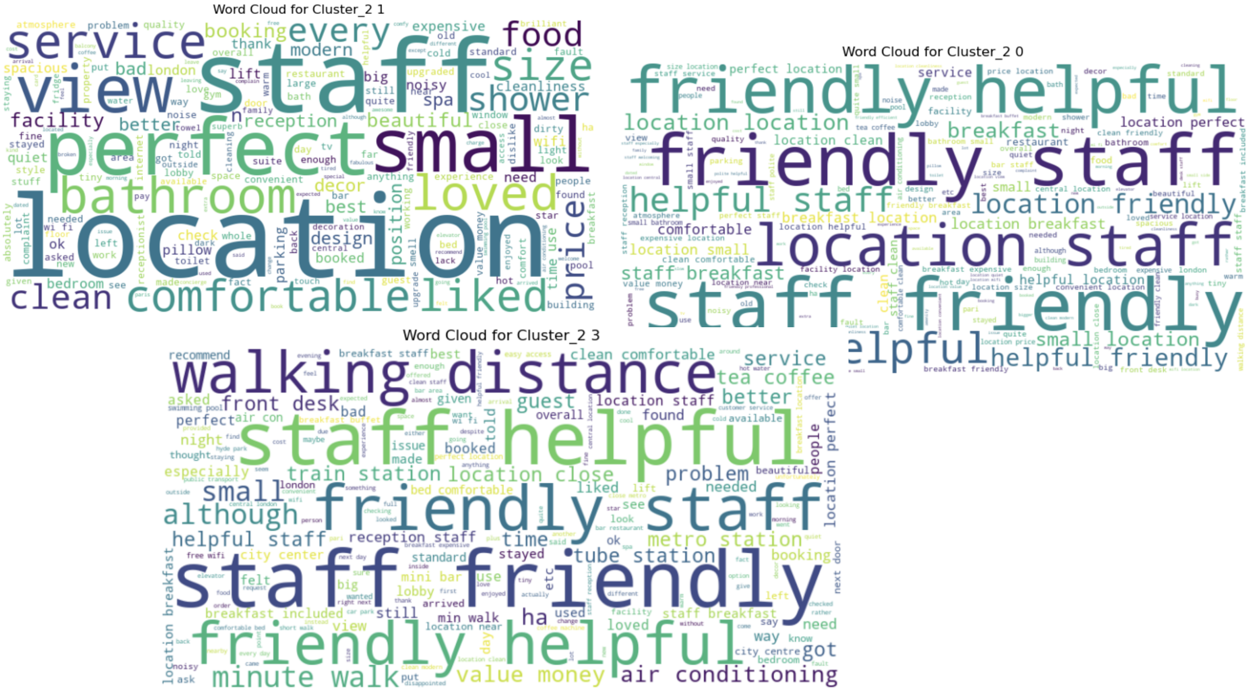
(1) the Elbow Method



Before conducting K-means clustering, we performed the Elbow Method test to determine the optimal number of clusters in K-means clustering. The Elbow Method is the method that involves plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use. We could find that the "elbow" appears to be the point after which the inertia decreases linearly. Typically, this is where the inertia begins to decrease more slowly. In our graph, there is a bend around cluster number 4, and after this point, the inertia decreases at a slower rate. Therefore, the elbow point is at k=4. Beyond this point, the gains in explained variance tend to diminish, which suggests that additional clusters do not contribute little to capturing the data structure.

(2) Results





After several trials with K-means clustering, we've confirmed that the results do not yield meaningful interpretations for our textual data. Due to the inherent limitations of clustering algorithms, the topics were represented merely as a series of words without coherent meaning. Therefore, we plan to explore alternative methods for measuring distances between feature vectors.

2. TF & IDF / NMF

(1) Concepts of TF & IDF

For building better topic models, we've tried to use Term Frequency-Inverse Document Frequency (TF-IDF) and Non-negative Matrix Factorization (NMF).

- Term Frequency (TF): A document-centric measure, quantifying the occurrence of a term within its own text. Frequent terms within a document hold high TF scores, indicating their potential topical relevance.
- Inverse Document Frequency (IDF): A corpus-level measure, assessing the uniqueness of a term across the entire collection. Terms appearing in many documents receive low IDF scores, downplaying their global importance.
- (TF-IDF)(term, document) = (TF)(term, document) * (IDF)_(term) : This score not only captures local prominence (TF) but also penalizes ubiquitous terms (low IDF), resulting in a weighted vocabulary highlighting terms that distinguish each document from its peers.

While capturing how often a term pops up in a document, this score also considers how common it is overall. Words used everywhere get less credit, ensuring the focus remains on terms that set each document apart.

(2) NMF Topic Modeling

Non-negative Matrix Factorization (NMF), as defined in [1], is a data analysis technique that decomposes a given data matrix into two smaller matrices, ensuring that none of the values in these matrices are negative. This constraint of non-negativity leads to a parts-based representation because it allows only additive, not subtractive, combinations. NMF operates on this matrix, decomposing it into two non-negative factors:

- Topic matrix (W): Each row represents a topic, a latent theme characterized by a distribution of term weights. High weights indicate terms strongly associated with that topic.
- Document matrix (H): Each column represents a document, encoded as a mixture of topic proportions. Each document exhibits varying degrees of membership in different topics.

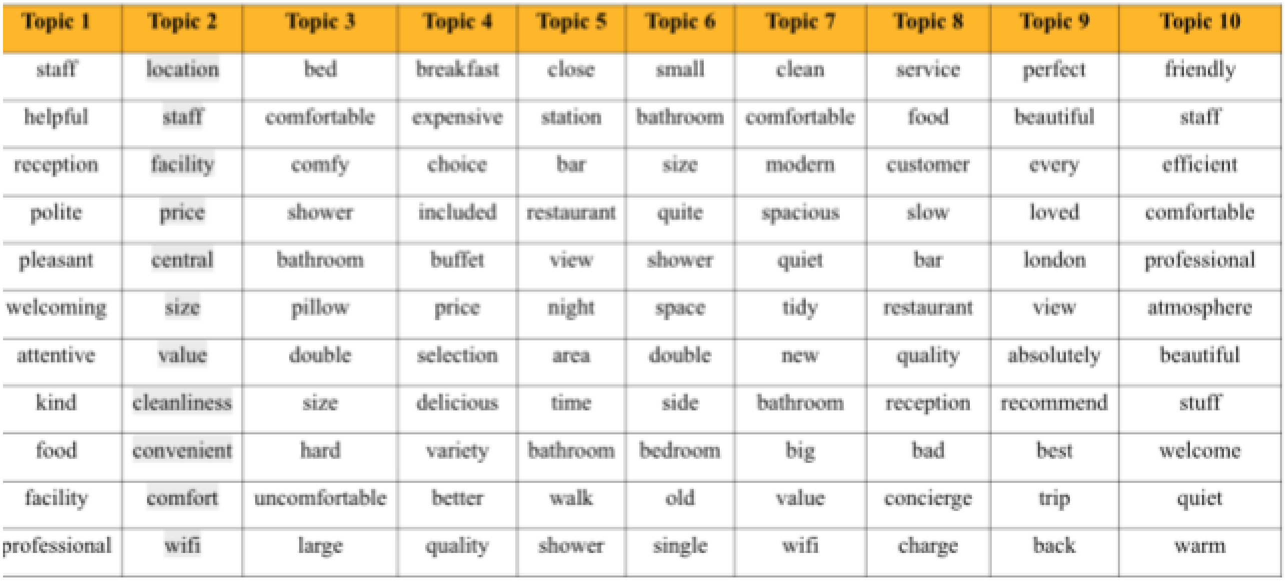
NMF essentially factors out the hidden thematic structure of the corpus, revealing topic clusters and their document affiliations.

(3) Combined Methodologies

TF-IDF pre-processes the data by providing the most informative terms for NMF to analyze. Common words, downplayed by low TF-IDF scores, do not obfuscate the thematic signal. NMF, along with this weighted vocabulary, identifies topics characterized by semantically related terms, resulting in interpretable thematic clusters. Each document's H-vector reveals its topic composition, allowing us to understand its thematic contributions and classify it based on dominant themes. In essence, TF-IDF equips NMF with a distilled dictionary to construct semantically coherent topics and decipher the thematic landscape of the document collection.

(4) Results

After performing several trials with the TF-IDF & NMF model, we were able to draw 10 topics as below.



Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
staff	location	bed	breakfast	close	small	clean	service	perfect	friendly
helpful	staff	comfortable	expensive	station	bathroom	comfortable	food	beautiful	staff
reception	facility	comfy	choice	bar	size	modern	customer	every	efficient
polite	price	shower	included	restaurant	quite	spacious	slow	loved	comfortable
pleasant	central	bathroom	buffet	view	shower	quiet	bar	london	professional
welcoming	size	pillow	price	night	space	tidy	restaurant	view	atmosphere
attentive	value	double	selection	area	double	new	quality	absolutely	beautiful
kind	cleanliness	size	delicious	time	side	bathroom	reception	recommend	stuff
food	convenient	hard	variety	bathroom	bedroom	big	bad	best	welcome
facility	comfort	uncomfortable	better	walk	old	value	concierge	trip	quiet
professional	wifi	large	quality	shower	single	wifi	charge	back	warm

<Table> Selected topics from TF-IDF & NMF

The results we observed were more refined than those obtained through K-means clustering, but there were still significant limitations in capturing the contextual nuances of words. For example, in topic 2, words like 'staff', 'facility', 'price', 'value', 'cleanliness', and 'comfort' were clustered together despite belonging to distinct thematic categories. This amalgamation of words from different topics indicated a lack of precision in the methodology. Hence, we decided to transition from this approach to Latent Dirichlet Allocation (LDA) for topic modeling, anticipating a more accurate representation of thematic structures within our data.

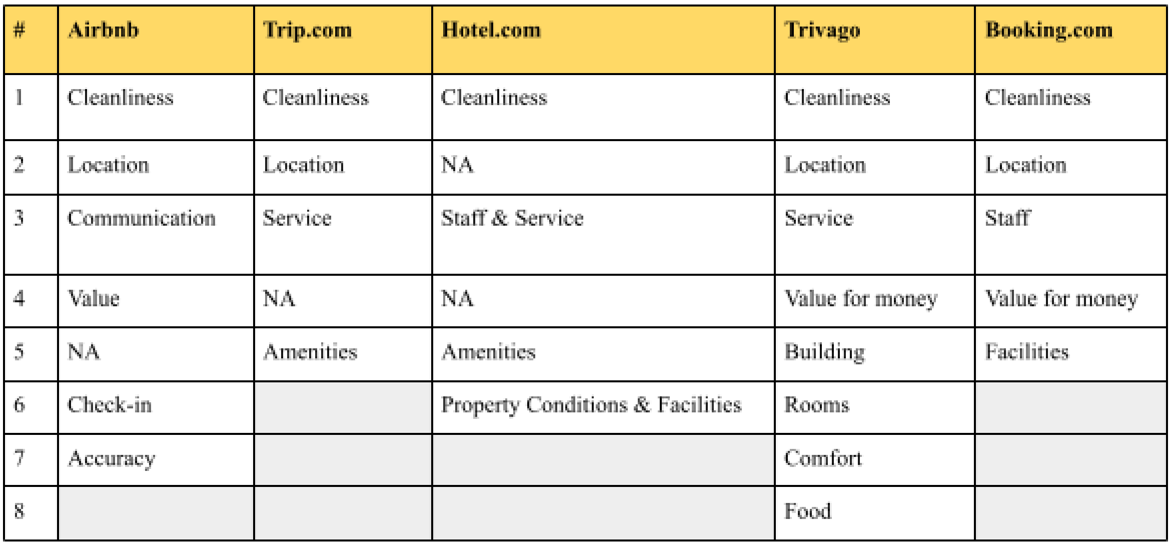
3. LDA Topic Modeling with Gensim

(1) Concepts of LDA Topic Modeling

Topic modeling, mainly through Latent Dirichlet Allocation (LDA), is a powerful data analysis technique for identifying central themes within textual data. LDA uses probabilistic modeling to analyze extensive text, determining the distribution of topics across documents and identifying keywords associated with each topic as demonstrated in [2]. This process is invaluable for extracting meaningful insights from complex data sets. In practical terms, such as when using Gensim's LDA model in Python, topic modeling is crucial for understanding varying perspectives, like regional perceptions of service aspects in customer feedback. A critical part of LDA involves setting the number of topics (k), which is determined based on domain knowledge, experimentation, or evaluation metrics like perplexity and coherence score. To find the optimal topic number of k, we started collecting topic candidates as below.

(2) Performing Topic Modeling

(i) References



#	Airbnb	Trip.com	Hotel.com	Trivago	Booking.com
1	Cleanliness	Cleanliness	Cleanliness	Cleanliness	Cleanliness
2	Location	Location	NA	Location	Location
3	Communication	Service	Staff & Service	Service	Staff
4	Value	NA	NA	Value for money	Value for money
5	NA	Amenities	Amenities	Building	Facilities
6	Check-in		Property Conditions & Facilities	Rooms	
7	Accuracy			Comfort	
8				Food	

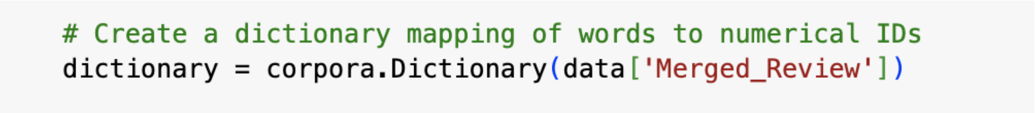
<Table> References from five popular booking websites

As we need a reference for the topics for topic modeling, we have gathered and compared the rating metrics from five popular and well-known booking platforms, including Airbnb, Booking.com, Trivago, Trip.com, and Hotel.com. The following is the comparison and correspondence with different booking platforms:

Referencing the standards that other websites use, we conducted the modeling process.

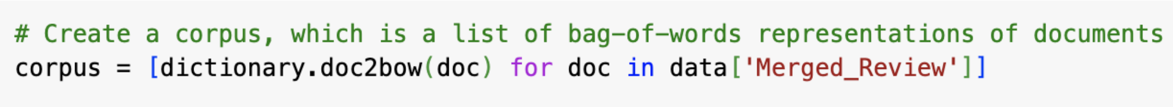
(ii) Establishing a Model with the Most Appropriate Parameters

Firstly, we must construct a dictionary of unique words to create a ‘Bag of Words.' This dictionary contains words and their unique identification numbers.



```
# Create a dictionary mapping of words to numerical IDs
dictionary = corpora.Dictionary(data['Merged_Review'])
```

Once the dictionary is created, we vectorize it by counting the frequency of each word from the dictionary appearing in the sentences. This is known as a bag of words, and this structured language data is referred to as a corpus.

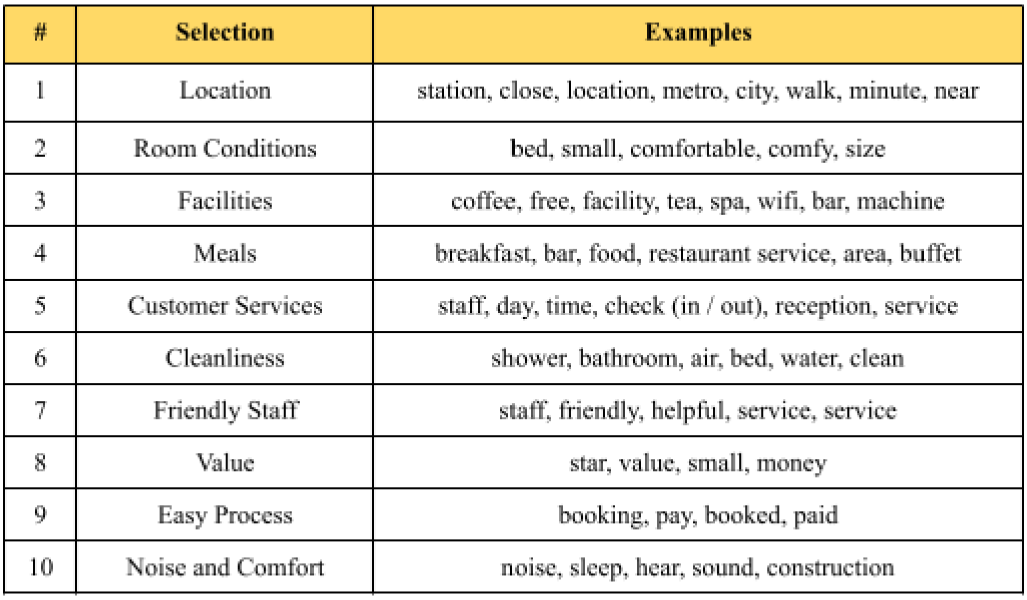


```
# Create a corpus, which is a list of bag-of-words representations of documents
corpus = [dictionary.doc2bow(doc) for doc in data['Merged_Review']]
```

In the Gensim library for topic modeling, several parameters play a crucial role in shaping the analysis, and adjusting them is critical for optimal results. Here's a detailed overview of the essential parameters:

- num_topics (Number of Topics): Determines the number of distinct topics the model will identify within the text corpus. A higher number of topics can provide more detailed thematic distinctions, while a lower number might capture broader themes. Choosing the correct number is critical for the relevance and specificity of the results.
- chunk size (Document Batch Size): Specifies the number of documents processed in a training batch. It affects the training speed and memory usage. Larger chunks might speed up training but require more memory, whereas smaller chunks can be more memory-efficient but may slow down the process.
- passes (Training Passes): Sets how often the model will pass over the entire corpus during training. Multiple passes can lead to more accurate topic models as the algorithm has more opportunities to learn and adjust the topic distributions. However, more passes also mean longer training times.
- iterations (Iterations per Document): Controls the number of iterations the model runs for each document. Higher iterations can lead to more precise topic assignments for each document, enhancing the model's accuracy. It's a balance between computational efficiency and depth of learning.

After over 50 iterations of fine-tuning each parameter and refining our corpus by removing sets of meaningless words, we achieved an optimized approach for our topic modeling analysis. This process involved meticulous updates to the corpus and the parameters influencing the model's behavior, allowing us to determine the most representative topics effectively. We strived to ensure that our final topic selection was both robust and well-suited to the nuances of our data and to achieve the highest coherence level. The coherence level measures how semantically consistent and meaningful the topics are, indicating the quality and relevance of the topics identified by the model. This iterative approach, combining corpus refinement with parameter optimization, was key to deriving the most insightful and accurate topics from our text data, ensuring that each topic not only captures the essence of the textual data but also resonates with coherent and interpretable themes. The topics we selected are below.



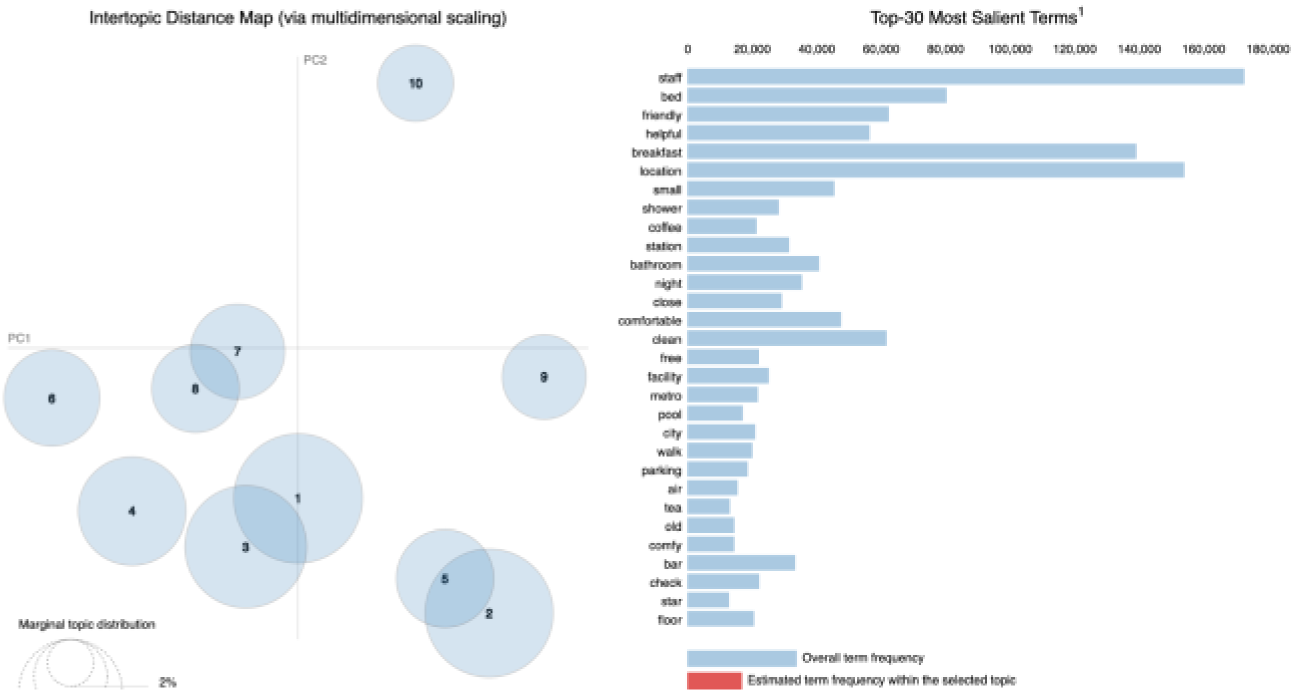
#	Selection	Examples
1	Location	station, close, location, metro, city, walk, minute, near
2	Room Conditions	bed, small, comfortable, comfy, size
3	Facilities	coffee, free, facility, tea, spa, wifi, bar, machine
4	Meals	breakfast, bar, food, restaurant service, area, buffet
5	Customer Services	staff, day, time, check (in / out), reception, service
6	Cleanliness	shower, bathroom, air, bed, water, clean
7	Friendly Staff	staff, friendly, helpful, service, service
8	Value	star, value, small, money
9	Easy Process	booking, pay, booked, paid
10	Noise and Comfort	noise, sleep, hear, sound, construction

<Table> Selected Topics by LDA Gensim modeling

(3) Topic Analysis

After conducting a topic modeling, we can utilize a visualization tool named "LDAvis" to help interpret the topics in a topic model that has been fit to a corpus of text data such as below. The method aims to assist in better interpreting the topics identified by LDA in large text corpora, making it easier for users to explore and make sense of these topics as asserted in [3].

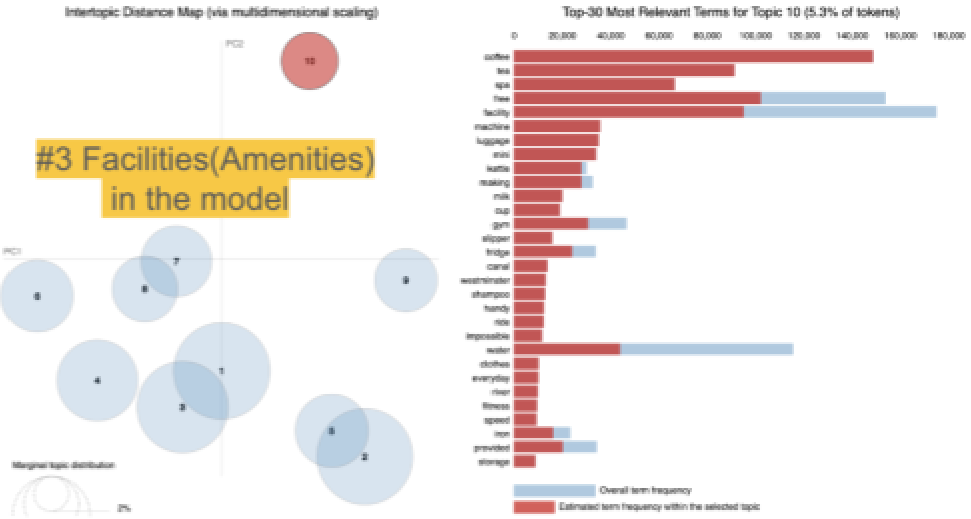
- Inter-topic Distance Map (left side): It shows the inter-topic distances in a two-dimensional plane where topics are represented as circles. The distances between topics show how similar or different the topics are. Topics that are closer together are more similar to each other than those further away. In our visualization results, topics 5 and 2 seem closely related, suggesting a thematic overlap.



<Image> The LDAvis visualization result of topic modeling

- Top-Term Bar Chart (right side): When a topic is selected, this chart shows the most relevant terms for the selected topic. Relevance is measured by a metric that balances the term frequency within a topic and the term's exclusivity to the topic. The red bar (if visible under the blue) would indicate the term's frequency within just Topic 0. It is not visible, likely due to the relevance metric being set at a level (λ=0.37) that emphasizes the term's specificity over its frequency. The relevance metric level λ indicates a balance slightly in favor of terms that are more exclusive to Topic 0. Moving the slider toward 1.0 would show more common terms while moving it towards 0 would highlight more unique terms.
- Salient Terms: The visualization also identifies and ranks terms that are interesting or have high saliency scores, combining term frequency and distinctiveness to identify terms especially characteristic of a topic.

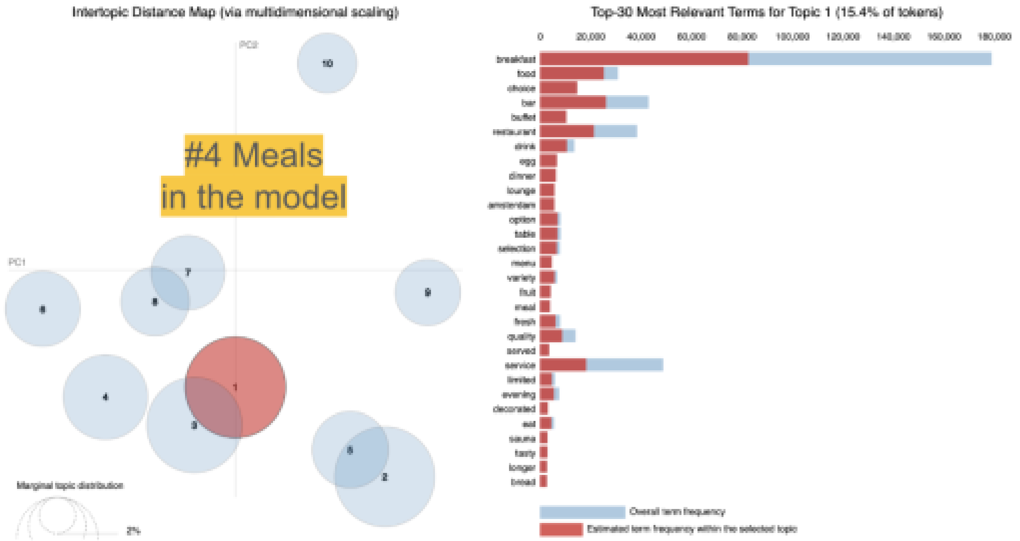
For example, when we check topic#10 below, the location of Topic 10's bubble, distant from other topics, suggests that the terms and themes in this topic are quite distinct from those in other topics.



<Image> LDA Visualization - Topic 10 (#3 in the model)

- Most Relevant Terms for Topic 10: The chart lists the most relevant terms for Topic 10. Terms like "coffee," "tea," "spa," and "free" are standard in contexts discussing hotel amenities or facilities.
- Term Relevance: The red bars represent each term's estimated frequency within Topic 10, while the blue bars represent the term's overall frequency in the corpus. We can infer that terms like "coffee" and "tea" are frequent within Topic 10 and across the entire corpus. Therefore, we named with “Facilities(Amenities)”.

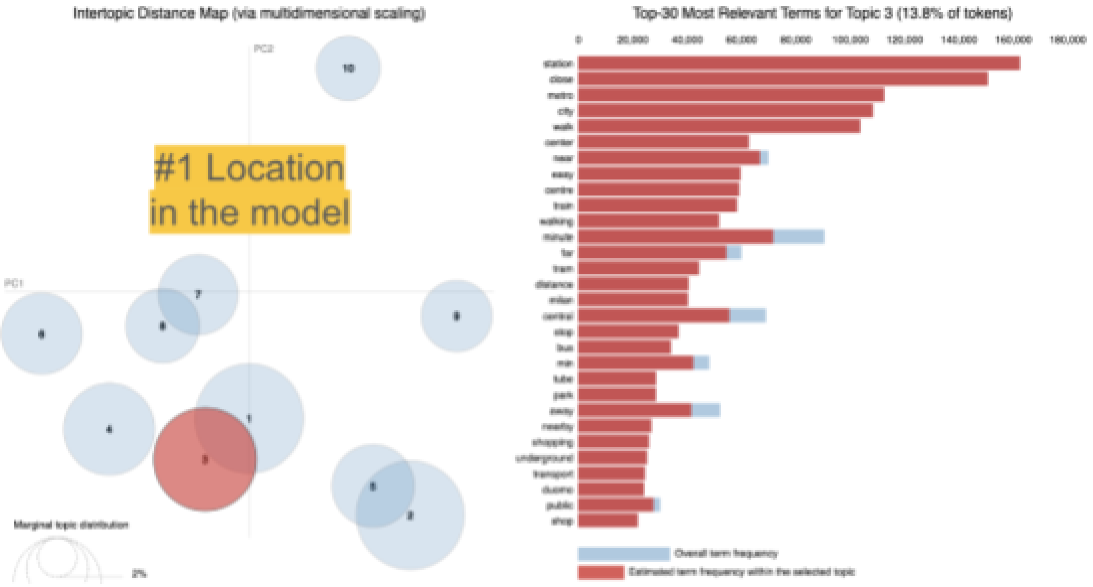
For another example of topic #1, topic one is positioned close to Topic 3, which might indicate some similarity or overlap in the themes covered by these topics.



<Image> LDA Visualization - Topic 1 (#4 in the model)

- Most Relevant Terms for Topic 1: The chart lists terms that are most relevant for Topic 1, with "breakfast" being the most prominent, followed by "food," "choice," "bar," "buffet," "restaurant," and "drink.", which allowed us name the topic as “Meals” (Dinings)
- Term Frequency: The blue bars represent the overall frequency of the terms across the entire corpus, while the red bars represent the estimated term frequency within Topic 1 expressly. The length of the bars suggests that terms such as "breakfast," "food," and "choice" are not only central to Topic 1 but are also common throughout the documents.

Meanwhile, topic 3 overlaps slightly with Topic 1, which may suggest some commonality or shared context between these topics.

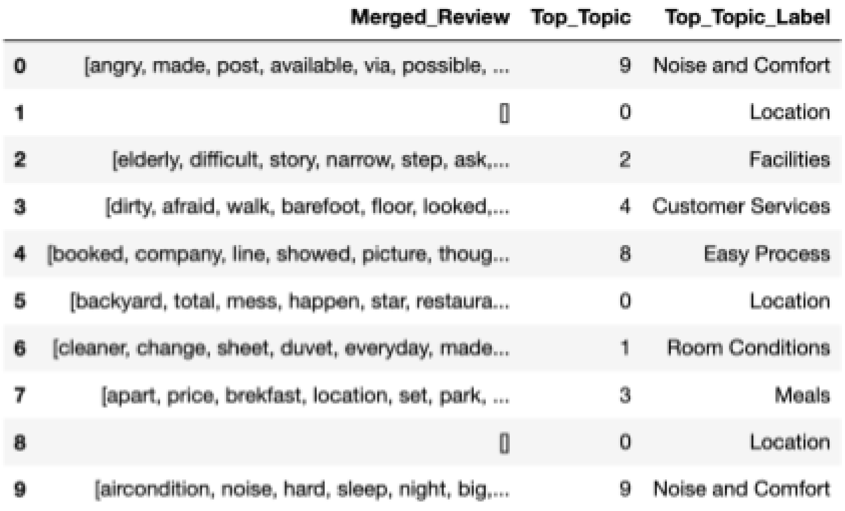


<Image> LDA Visualization - Topic 3 (#1 in the model)

- Most Relevant Terms for Topic 3: The chart lists terms that are most relevant for Topic 3, with "station," "close," "metro," and "city" being prominent, indicating a strong association with transport and urban settings.
- Term Frequency: The blue bars represent the overall frequency of the terms across the entire corpus, while the red bars expressly represent the estimated term frequency within Topic 3. The length of the bars suggests that terms like "station," "city," and "walk" are not only central to Topic 3 but are also relatively common throughout the documents.

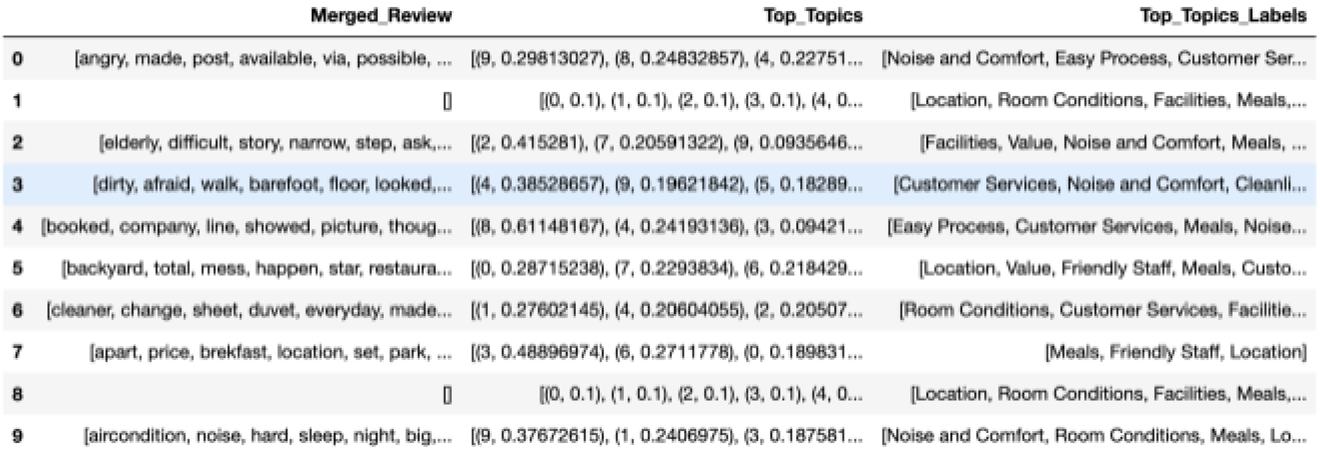
(4) Assign Topics to the Review

After establishing the topic model, we applied it to the words in each review to identify the most relevant themes. We aimed to discern the top 7 topics most prominent across the reviews, capturing the most significant and recurrent themes that could provide actionable insights into customer preferences and experiences.



	Merged_Review	Top_Topic	Top_Topic_Label
0	[angry, made, post, available, via, possible, ...	9	Noise and Comfort
1	[]	0	Location
2	[elderly, difficult, story, narrow, step, ask,...	2	Facilities
3	[dirty, afraid, walk, barefoot, floor, looked,...	4	Customer Services
4	[booked, company, line, showed, picture, thoug...	8	Easy Process
5	[backyard, total, mess, happen, star, restaura...	0	Location
6	[cleaner, change, sheet, duvet, everyday, made...	1	Room Conditions
7	[apart, price, brekfast, location, set, park, ...	3	Meals
8	[]	0	Location
9	[aircondition, noise, hard, sleep, night, big,...	9	Noise and Comfort

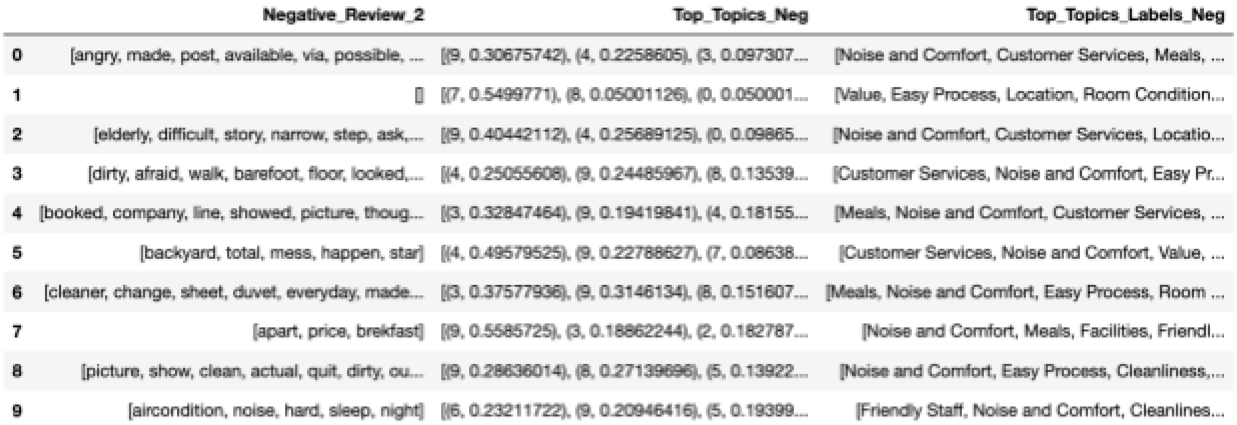
<Table> Assigned Top Topic on Merged Review



	Merged_Review	Top_Topics	Top_Topics_Labels
0	[angry, made, post, available, via, possible, ...	[(9, 0.29813027), (8, 0.24832857), (4, 0.22751...	[Noise and Comfort, Easy Process, Customer Ser...
1	[]	[(0, 0.1), (1, 0.1), (2, 0.1), (3, 0.1), (4, 0...	[Location, Room Conditions, Facilities, Meals,...
2	[elderly, difficult, story, narrow, step, ask,...	[(2, 0.415281), (7, 0.20591322), (9, 0.0935646...	[Facilities, Value, Noise and Comfort, Meals, ...
3	[dirty, afraid, walk, barefoot, floor, looked,...	[(4, 0.38528657), (9, 0.19621842), (5, 0.18289...	[Customer Services, Noise and Comfort, Cleanli...
4	[booked, company, line, showed, picture, thoug...	[(8, 0.61148167), (4, 0.24193136), (3, 0.09421...	[Easy Process, Customer Services, Meals, Noise...
5	[backyard, total, mess, happen, star, restaura...	[(0, 0.28715238), (7, 0.2293834), (6, 0.218429...	[Location, Value, Friendly Staff, Meals, Custo...
6	[cleaner, change, sheet, duvet, everyday, made...	[(1, 0.27602145), (4, 0.20604055), (2, 0.20507...	[Room Conditions, Customer Services, Facilitie...
7	[apart, price, brekfast, location, set, park, ...	[(3, 0.48896974), (6, 0.2711778), (0, 0.189831...	[Meals, Friendly Staff, Location]
8	[]	[(0, 0.1), (1, 0.1), (2, 0.1), (3, 0.1), (4, 0...	[Location, Room Conditions, Facilities, Meals,...
9	[aircondition, noise, hard, sleep, night, big,...	[(9, 0.37672615), (1, 0.2406975), (3, 0.187581...	[Noise and Comfort, Room Conditions, Meals, Lo...

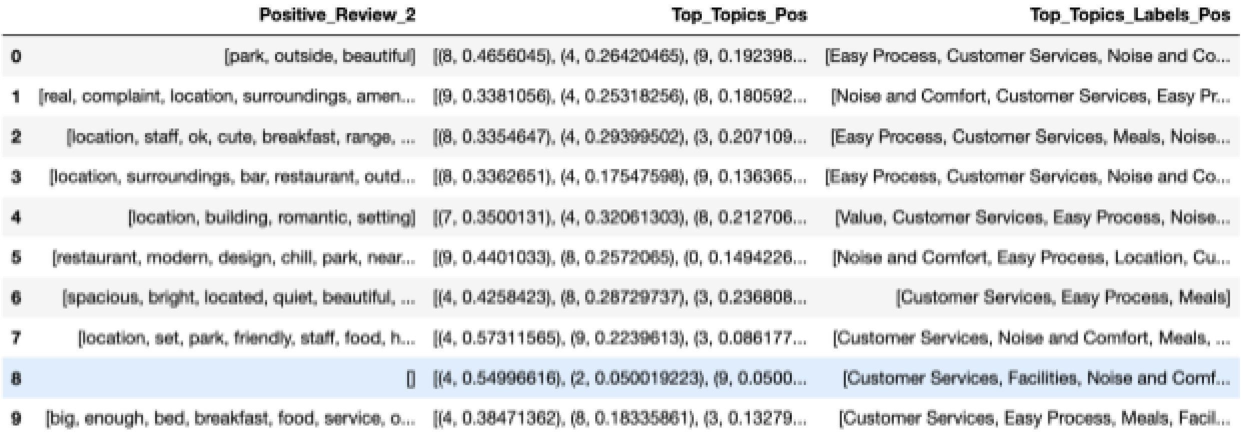
<Table> Assigned Top Topics on Merged Review

Therefore, we assigned each top 7 topics on the 'Negative Review' and 'Positive Review' respectively as follows.



	Negative_Review_2	Top_Topics_Neg	Top_Topics_Labels_Neg
0	[angry, made, post, available, via, possible, ...	[(9, 0.30675742), (4, 0.2258605), (3, 0.097307...	[Noise and Comfort, Customer Services, Meals, ...
1	[]	[(7, 0.5499771), (8, 0.05001126), (0, 0.050001...	[Value, Easy Process, Location, Room Condition...
2	[elderly, difficult, story, narrow, step, ask,...	[(9, 0.40442112), (4, 0.25689125), (0, 0.09865...	[Noise and Comfort, Customer Services, Locatio...
3	[dirty, afraid, walk, barefoot, floor, looked,...	[(4, 0.25055608), (9, 0.24485967), (8, 0.13539...	[Customer Services, Noise and Comfort, Easy Pr...
4	[booked, company, line, showed, picture, thoug...	[(3, 0.32847464), (9, 0.19419841), (4, 0.18155...	[Meals, Noise and Comfort, Customer Services, ...
5	[backyard, total, mess, happen, star]	[(4, 0.49579525), (9, 0.22788627), (7, 0.08638...	[Customer Services, Noise and Comfort, Value, ...
6	[cleaner, change, sheet, duvet, everyday, made...	[(3, 0.37577936), (9, 0.3146134), (8, 0.151607...	[Meals, Noise and Comfort, Easy Process, Room ...
7	[apart, price, brekfast]	[(9, 0.5585725), (3, 0.18862244), (2, 0.182787...	[Noise and Comfort, Meals, Facilities, Friendl...
8	[picture, show, clean, actual, quit, dirty, ou...	[(9, 0.28636014), (8, 0.27139696), (5, 0.13922...	[Noise and Comfort, Easy Process, Cleanliness,...
9	[aircondition, noise, hard, sleep, night]	[(6, 0.23211722), (9, 0.20946416), (5, 0.19399...	[Friendly Staff, Noise and Comfort, Cleanlines...

<Table> Assigned Top Topics on **Negative** Review



	Positive_Review_2	Top_Topics_Pos	Top_Topics_Labels_Pos
0	[park, outside, beautiful]	[(8, 0.4656045), (4, 0.26420465), (9, 0.192398...	[Easy Process, Customer Services, Noise and Co...
1	[real, complaint, location, surroundings, amen...	[(9, 0.3381056), (4, 0.25318256), (8, 0.180592...	[Noise and Comfort, Customer Services, Easy Pr...
2	[location, staff, ok, cute, breakfast, range, ...	[(8, 0.3354647), (4, 0.29399502), (3, 0.207109...	[Easy Process, Customer Services, Meals, Noise...
3	[location, surroundings, bar, restaurant, outd...	[(8, 0.3362651), (4, 0.17547598), (9, 0.136365...	[Easy Process, Customer Services, Noise and Co...
4	[location, building, romantic, setting]	[(7, 0.3500131), (4, 0.32061303), (8, 0.212706...	[Value, Customer Services, Easy Process, Noise...
5	[restaurant, modern, design, chill, park, near...	[(9, 0.4401033), (8, 0.2572065), (0, 0.1494226...	[Noise and Comfort, Easy Process, Location, Cu...
6	[spacious, bright, located, quiet, beautiful, ...	[(4, 0.4258423), (8, 0.28729737), (3, 0.236808...	[Customer Services, Easy Process, Meals]
7	[location, set, park, friendly, staff, food, h...	[(4, 0.57311565), (9, 0.2239613), (3, 0.086177...	[Customer Services, Noise and Comfort, Meals, ...
8	[]	[(4, 0.54996616), (2, 0.050019223), (9, 0.0500...	[Customer Services, Facilities, Noise and Comf...
9	[big, enough, bed, breakfast, food, service, o...	[(4, 0.38471362), (8, 0.18335861), (3, 0.13279...	[Customer Services, Easy Process, Meals, Facil...

<Table> Assigned Top Topics on **Positive** Review

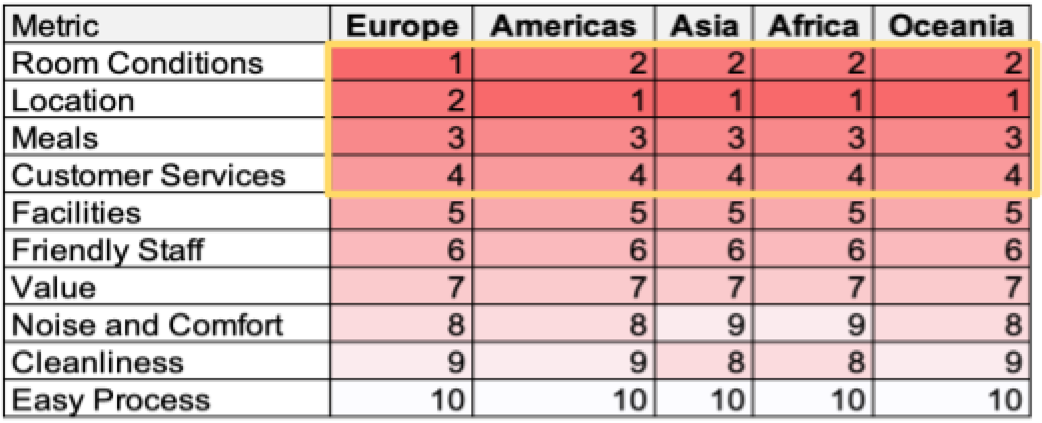
VI. Findings and Analysis

Across all review types, there is a consistent emphasis on the quality of room conditions, customer services, and the overall experience of meals and facilities. The difference in ranking between negative and positive reviews for the same region indicates areas where improvements can significantly enhance guest experiences. We assumed that analyzing the emergence of intriguing patterns based on nationality, through a combined assessment of both positive and negative aspects, would offer a more comprehensive understanding, as illustrated in the study by H. J. Kwon [4]. Cultural and regional differences can influence the emphasis on specific topics.

The overall findings of our analysis are presented in the following content.

1. Merged Review

(1) Region



Metric	Europe	Americas	Asia	Africa	Oceania
Room Conditions	1	2	2	2	2
Location	2	1	1	1	1
Meals	3	3	3	3	3
Customer Services	4	4	4	4	4
Facilities	5	5	5	5	5
Friendly Staff	6	6	6	6	6
Value	7	7	7	7	7
Noise and Comfort	8	8	9	9	8
Cleanliness	9	9	8	8	9
Easy Process	10	10	10	10	10

<Table> Ranking of Merged Review on Region

We found a similar priority of metrics between each region. However, there is a slight difference for specific metrics. The top 5 metrics for merged reviews are Room Conditions, Location, Meals, Customer Services, and Facilities, respectively. These are interpreted as follows:

A. **Room Conditions:** Consistently, room conditions hold the highest or second-highest importance across all regions. It suggests that guests have strong opinions about their living spaces, emphasizing the need for comfort, functionality, and aesthetics in rooms.

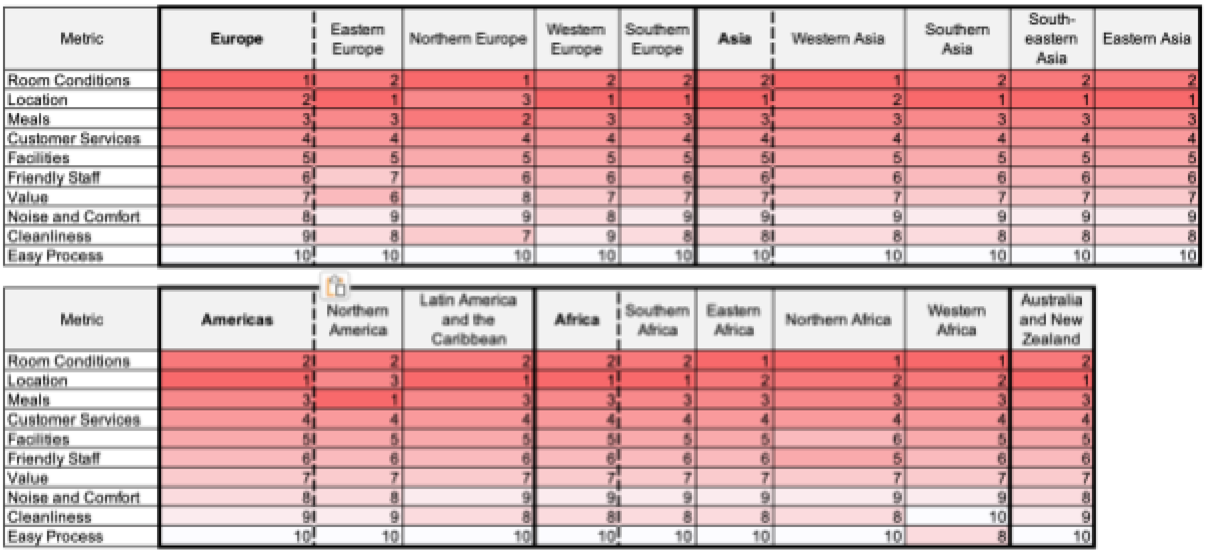
B. **Location**: Very close to room conditions in terms of priority. Location is crucial for guests. A prime location enhances the value of the hotel stay by providing convenience, access to attractions, and often a desirable view.

C. **Meals**: Food services, including the quality and variety of meals, rank high across all regions. Dining experiences are integral to guests' satisfaction, indicating that hotels should invest in high-quality culinary offerings.

D. **Customer Services**: High-quality service is a significant concern for guests, with customer service ranking prominently across all regions. The personal touch and efficiency of the staff can make a substantial difference in guest satisfaction.

E. **Facilities**: The availability and quality of facilities like spas, gyms, and pools are important to guests, though they rank slightly lower than the more immediate concerns of rooms, location, and meals.

(2) Sub-region



Metric	**Europe**	Eastern Europe	Northern Europe	Western Europe	Southern Europe	**Asia**	Western Asia	Southern Asia	South-eastern Asia	Eastern Asia
Room Conditions	1	2	1	2	2	2	1	2	2	2
Location	2	1	3	1	1	1	2	1	1	1
Meals	3	3	2	3	3	3	3	3	3	3
Customer Services	4	4	4	4	4	4	4	4	4	4
Facilities	5	5	5	5	5	5	5	5	5	5
Friendly Staff	6	7	6	6	6	6	6	6	6	6
Value	7	6	8	7	7	7	7	7	7	7
Noise and Comfort	8	9	9	8	9	9	9	9	9	9
Cleanliness	9	8	7	9	8	8	8	8	8	8
Easy Process	10	10	10	10	10	10	10	10	10	10

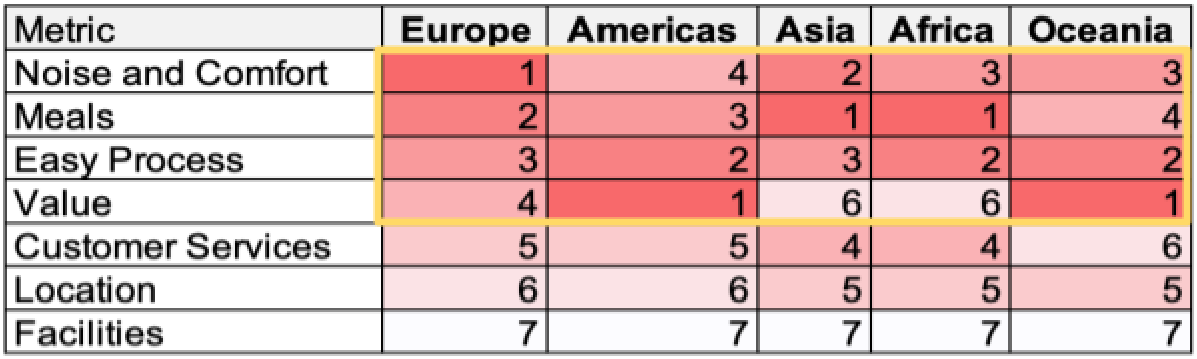
Metric	**Americas**	Northern America	Latin America and the Caribbean	**Africa**	Southern Africa	Eastern Africa	Northern Africa	Western Africa	Australia and New Zealand
Room Conditions	2	2	2	2	2	1	1	1	2
Location	1	3	1	1	1	2	2	2	1
Meals	3	1	3	3	3	3	3	3	3
Customer Services	4	4	4	4	4	4	4	4	4
Facilities	5	5	5	5	5	5	6	5	5
Friendly Staff	6	6	6	6	6	6	5	6	6
Value	7	7	7	7	7	7	7	7	7
Noise and Comfort	8	8	9	9	9	9	9	9	8
Cleanliness	9	9	8	8	8	8	8	10	9
Easy Process	10	10	10	10	10	10	10	8	10

<Table> Ranking of Merged Review on Sub-Region

Furthermore, we would like to see whether there is any difference if we view it from the sub-regional perspective as above. We found that sub-regions did influence how they rank their priority compared to the original region metric. Some ranks did differ from the regional perspective in a few rank sequences. For example, Northern America ranks meals as their priority, while America ranks location first. Each subregion in Europe and Asia also differs from each other for the first three rankings.

2. Negative Review

(1) Region



Metric	Europe	Americas	Asia	Africa	Oceania
Noise and Comfort	1	4	2	3	3
Meals	2	3	1	1	4
Easy Process	3	2	3	2	2
Value	4	1	6	6	1
Customer Services	5	5	4	4	6
Location	6	6	5	5	5
Facilities	7	7	7	7	7

<Table> Ranking of Negative Review on Region

After reviewing the merged reviews, we applied a negative review. Surprisingly, it's pretty different from merged reviews in that the top 3 are noise and comfort, meals, and easy process, which are critical factors that lead to negative reviews. For example, issues related to noise levels and overall comfort significantly impact the negative experiences of guests in these regions, and efficiency in services and perceived value for money are critical factors affecting guest satisfaction.

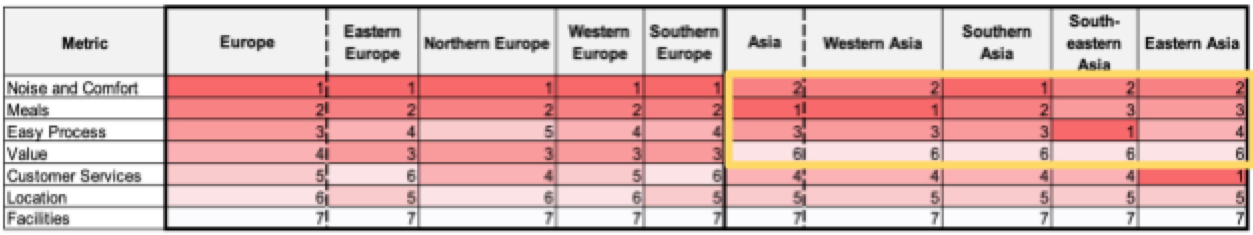
They are pretty different from each region in ranking their top 4 metrics. Each region has quite a priority for the top 4 metrics compared to the merged review. America ranked value as their top 1, which indicates they are concerned about how their money is spent. If the money is not worthy, it leads them to give negative reviews. However, Asians and Africans rank meals as their top 1, which indicates meals are a top indicator for them to leave a negative review. Following is our interpretation of the top 3 metrics:

A. **Noise and Comfort**: This is the top concern in negative reviews for Europe, indicating that inadequate soundproofing or uncomfortable furnishings can significantly affect the guest experience.

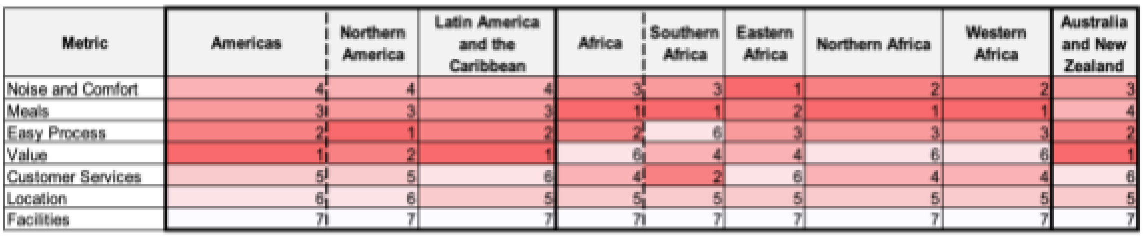
B. **Easy Process:** The prominence of this topic in negative reviews suggests that operational inefficiencies are a common pain point for guests. Streamlining these processes could lead to significant improvements in guest satisfaction.

C. **Value**: In negative reviews, value ranks higher than merged reviews, highlighting that guests who feel they aren't getting their money's worth are likely to leave a negative review.

(2) Sub-region



Metric	Europe	Eastern Europe	Northern Europe	Western Europe	Southern Europe	Asia	Western Asia	Southern Asia	South-eastern Asia	Eastern Asia
Noise and Comfort	1	1	1	1	1	2	2	1	2	2
Meals	2	2	2	2	2	1	1	2	3	3
Easy Process	3	4	5	4	4	3	3	3	1	4
Value	4	3	3	3	3	6	6	6	6	6
Customer Services	5	6	4	5	6	4	4	4	4	1
Location	6	5	6	6	5	5	5	5	5	5
Facilities	7	7	7	7	7	7	7	7	7	7



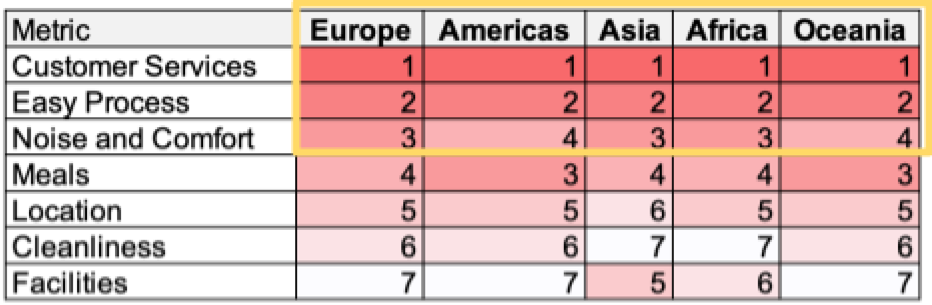
Metric	Americas	Northern America	Latin America and the Caribbean	Africa	Southern Africa	Eastern Africa	Northern Africa	Western Africa	Australia and New Zealand
Noise and Comfort	4	4	4	3	3	1	2	2	3
Meals	3	3	3	1	1	2	1	1	4
Easy Process	2	1	2	2	6	3	3	3	2
Value	1	2	1	6	4	4	6	6	1
Customer Services	5	5	6	4	2	6	4	4	6
Location	6	6	5	5	5	5	5	5	5
Facilities	7	7	7	7	7	7	7	7	7

<Table> Ranking of Negative Review on Sub-Region

There is not too much difference in the top 2 ranking for the sub-region compared to their respective region; however, the difference happened in the middle ranking within each subregion. For instance, Eastern Asians value customer service as their negative indicator the most compared with meals as to whole Asians, which is quite fair that Eastern Asia has its reputation for good service, Japan and Korea, for example. If the hotel's service doesn’t meet their expectation, it leads to negative reviews for a great chance.

3. Positive Review

(1) Region



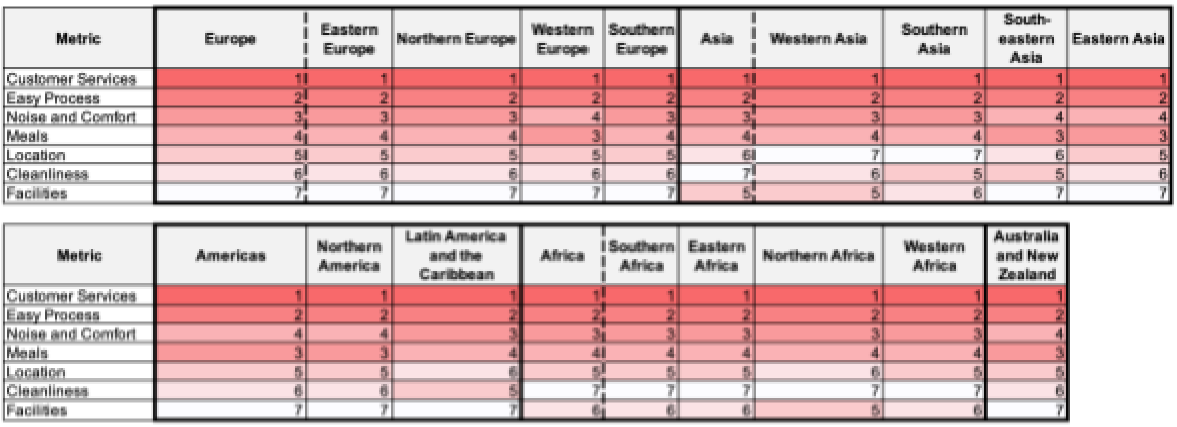
Metric	Europe	Americas	Asia	Africa	Oceania
Customer Services	1	1	1	1	1
Easy Process	2	2	2	2	2
Noise and Comfort	3	4	3	3	4
Meals	4	3	4	4	3
Location	5	5	6	5	5
Cleanliness	6	6	7	7	6
Facilities	7	7	5	6	7

<Table> Ranking of Positive Review on Region

The Top 3 critical factors that lead to positive reviews are customer service, easy process, and noise and comfort. Customer Services is the top-ranked topic in positive reviews for all regions, highlighting the crucial role of service quality in creating positive guest experiences. Furthermore, the significance of easy processes and Noise and Comfort are frequently mentioned in positive reviews across all regions, emphasizing that efficient processes and comfortable, quiet environments contribute to satisfaction.

In addition, Location is less regularly a top-ranking topic in positive reviews, suggesting that while necessary, it might not be the primary driver of positive experiences. Compared with the merged and negative reviews. Also, cleanliness is more prominent in positive reviews than negative reviews, indicating that they significantly contribute to a positive experience when they are well-maintained. The importance of 'Noise and Comfort' in negative and positive reviews underlines the significant impact of the physical environment on guest experiences.

(2) Subregion



Metric	Europe	Eastern Europe	Northern Europe	Western Europe	Southern Europe	Asia	Western Asia	Southern Asia	South-eastern Asia	Eastern Asia
Customer Services	1	1	1	1	1	1	1	1	1	1
Easy Process	2	2	2	2	2	2	2	2	2	2
Noise and Comfort	3	3	3	4	3	3	3	3	4	4
Meals	4	4	4	3	4	4	4	4	3	3
Location	5	5	5	5	5	6	7	7	6	5
Cleanliness	6	6	6	6	6	7	6	5	5	6
Facilities	7	7	7	7	7	5	5	6	7	7

Metric	Americas	Northern America	Latin America and the Caribbean	Africa	Southern Africa	Eastern Africa	Northern Africa	Western Africa	Australia and New Zealand
Customer Services	1	1	1	1	1	1	1	1	1
Easy Process	2	2	2	2	2	2	2	2	2
Noise and Comfort	4	4	3	3	3	3	3	3	4
Meals	3	3	4	4	4	4	4	4	3
Location	5	5	6	5	5	5	6	5	5
Cleanliness	6	6	5	7	7	7	7	7	6
Facilities	7	7	7	6	6	6	5	6	7

<Table> Ranking of Positive Review on Sub-Region

There are certain consistencies across regions for some metrics, which might point towards global trends or common issues that are being well addressed or need improvement across the board. For example, 'Easy Process' tends to have a score of '2' in many regions, suggesting a generally favorable perception of this aspect.

Some regions may excel in particular areas. For instance, 'Noise and Comfort' scores a '2' in Eastern Europe, Northern Europe, and Western Europe, suggesting that these regions have higher satisfaction in that metric compared to others where the score is '3' or '4'.

The metric 'Facilities' consistently scores lower across most regions, often receiving the highest number '7', which could indicate a universal area for improvement in customer satisfaction related to facilities.

There may be cultural or regional factors influencing the scores. For example, the 'Meals' metric has varying scores across different regions, which could reflect regional culinary preferences or standards. Meanwhile, Australia and New Zealand seem to perform well in 'Customer Services' and 'Easy Process', both scoring '1', but have room for improvement in 'Facilities', scoring '7'.

To draw more precise conclusions, one would need additional context about what the scores represent and how they were calculated. Also, the nature of the service or product being rated, as well as the demographic information about the customers providing these ratings, would be important to understand the underlying factors influencing these scores.

VII. Conclusion & Application

Our analysis reveals a discernible preference hierarchy based on region and subregion. This insight can be instrumental in aiding hotels to enhance their services, facilities, and other rating aspects, ultimately fostering an enriched customer experience. It isn’t solely beneficial to the hotel but advantageous for potential visitors seeking accommodation.

For the visitors, our reliance often rests on scores and metrics provided by past guests while choosing hotels. However, these ratings frequently diverge from our actual experiences. Determining the most fitting standards sometimes necessitates a deeper dive into reviews to ascertain their alignment with our specific preferences. As these ratings have limitations, historical comment analysis provides an added layer of assurance. By segregating comments into positive and negative categories initially, employing topic modeling unveils the top three attributes highlighted by previous guests. Comparing these outcomes with corresponding metric ratings enhances our confidence to meet our expectations.

For the hotels, categorizing customer reviews into positive and negative feedback yields invaluable insights. This practice helps identify weaknesses highlighted in negative thoughts, allowing hotels to pinpoint specific aspects customers dislike and make necessary improvements. Conversely, positive reviews illuminate aspects customers appreciate, enabling hotels to uphold these strengths. This strategic approach curtails the inclination to pursue higher ratings at the expense of valuable insights from reviews. Balancing both positive and negative feedback empowers hoteliers to fine-tune their offerings. Moreover, analyzing feedback from diverse regions or countries yields a nuanced understanding of preferences specific to varied demographics. This understanding enables hotels to tailor their services, addressing specific concerns that hold significance for travelers from different regions.

For future improvements, the aim is to transcend incomplete datasets by incorporating various types of data, specific hotel categories, individual establishments, particular timeframes, and distinct geographical regions. In addition, this could also be applied to a recommendation system for travelers to choose the most suitable hotels based on their regions, subregions, or nationalities. This approach allows for tailored analyses based on the available data, enabling a more precise understanding of the characteristics and preferences of diverse demographics in different areas or countries.

VIII. References

[1] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, vol. 401, no. 6755, pp. 788-791, 1999.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," Journal of Machine Learning Research, vol. 3, pp. 993–1022, Jan. 2003.

[3] C. Sievert and K. Shirley, "LDAvis: A method for visualizing and interpreting topics," in Proc. Workshop on Interactive Language Learning, Visualization, and Interfaces, Baltimore, MD, USA, 2014, pp. 63–70.

[4] H. J. Kwon, H. J. Ban, J. K. Jun, and H. S. Kim, "Topic modeling and sentiment analysis of online review for airlines," Information, vol. 12, no. 2, p. 78, 2021.

[5] TheDevastator, "Boston Airbnb Reviews: A Comprehensive Overview," Kaggle, Oct. 27, 2023. [Online]. Available: https://www.kaggle.com/datasets/thedevastator/boston-airbnb-reviews-a-comprehensive-overview. [Accessed: Oct. 27, 2023].

[6] H. Dedhia, "Airbnb Reviews Sentiment Analysis and Prediction," Kaggle, Oct. 27, 2023. [Online]. Available: https://www.kaggle.com/code/heeraldedhia/airbnb-reviews-sentiment-analysis-and-prediction. [Accessed: Oct. 27, 2023].

[7] Department of Statistics & Data Science, Carnegie Mellon University, "Statistical Analysis of Airbnb Listings in Major European Cities," Apr. 27, 2023. [Online]. Available: https://www.stat.cmu.edu/capstoneresearch/315files_s23/team5.html. [Accessed: Oct. 27, 2023].

[8] Dahyour, "Airbnb Data Set (Dirty Data)," Kaggle, Oct. 27, 2023. [Online]. Available: https://www.kaggle.com/datasets/dahyour/air-bnb-data-set-dirty-data. [Accessed: Oct. 27, 2023].

[9] "Determinants of Airbnb Prices in European Cities," Zenodo, Jan. 27, 2023. [Online]. Available: https://zenodo.org/records/4446043. [Accessed: Oct. 27, 2023].

Appendix

Appendix1 - Column Description

Column Name	Column Description
Hotel_Address	Address of hotel.
Review_Date	Date when the reviewer posted the corresponding review.
Average_Score	The average score of the hotel is calculated based on the latest comment in the last year.
Hotel_Name	Name of Hotel
Reviewer_Nationality	Nationality of Reviewer
Negative_Review	Negative Review the reviewer gave to the hotel. If the reviewer does not give a negative review, then it should be 'No Negative'
Review_Total_Negative_Word_Counts	Total number of words in the negative review.
Positive_Review	Positive review the reviewer gave to the hotel. If the reviewer does not give a negative review, then it should be 'No Positive'
Review_Total_Positive_Word_Counts	Total number of words in the positive review.
Reviewer_Score	Score the reviewer has given to the hotel, based on his/her experience
Total_Number_of_Reviews_Reviewer_Has_Given	Number of Reviews the reviewers have given in the past.
Total_Number_of_Reviews	Total number of valid reviews the hotel has.
Tags	Tags the reviewer gave the hotel.
days_since_review	Duration between the review date and scrape date.
Additional_Number_of_Scoring	There are also some guests who just scored on the service rather than a review. This number indicates how

	many valid scores without review there.
lat	Latitude of the hotel
lng	longitude of the hotel