UNIVERSITY OF CALIFORNIA
Los Angeles

Travel Guide
Using Text Mining and BERTopic

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Applied Statistics

by

Yaolan Jin

2022

ABSTRACT OF THE THESIS

Travel Guide
Using Text Mining and BERTopic

by

Yaolan Jin
Master of Science in Applied Statistics
University of California, Los Angeles, 2022
Professor Ying Nian Wu, Chair

This thesis aims to identify key topics from travel blog journals created by two bloggers Jones and handluggageonly and create interest directed recommendations to tourists. By applying Text Mining, sentiment analysis, the two writers exhibited different styles of travel guide recommendation but shares common traits in using sentimental words to lead their audiences through a trustful, exciting journey. Jones focus on providing insights in hostel rental versus handluggageonly shares his adventures on a personal level.

Both Latent Dirichlet Allocation (LDA) and BERTopic successfully categorized the journals into meaningful topics and identifies popular themes such as beach/island, music festival, historical sites, airbnb rental allocations. However, BERTopic provides additional interaction feature, which enables a powerful travel guide recommendation system that links user input to relevant documents. Whether it is historical sites, or music festivals, by associating country with the relevant documents, tourists can determine their next trip destination by referencing the firsthand travel experience provided by professional travel bloggers.

This model can efficiently capture important content with only one keyword input. It

not only save time compared to manual search but also can capture all associated themes without having to input all the keywords. By inputting "dessert", you get various topic groups identified in the documents such as "pizza", "dessert, pudding", "restaurants". The topics also reflect multiple interpretations of the word. By applying the model for both travelers and travel agents, the model can align travelers interest with travel website contents and provide feedback for a more user-focused experience. Whether the "catch" is beach, or museums, travel agents can incorporate those key attractions and create more personalized tour paths based on user preference.

The thesis of Yaolan Jin is approved.

David L. Rigby

Hongquan Xu

Ying Nian Wu, Committee Chair

University of California, Los Angeles

2022

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to Professor Ying Nian Wu and Professor David Zes for their enlightening and continuous support for guiding me to try new advanced methods on my thesis. Professor David Zes, who was not on my committee, provided valuable insights and encouraged me to enjoy each progress I made during the process. I would also like to thank Professor Hongquan Xu and Professor David Rigby for their timeless effort providing advice from structure to syntax details to further improve my report. Lastly, I am also grateful for the support I received from my family, friends, boyfriend, and classmates who always cheer me up during hard times and praise me for my accomplishments and hard work.

CHAPTER 1

Introduction

Tropical fruits, sandy beach, sound of waves of water splashing into the shore, bathing in the warm sun. What a relaxing way to enjoy our times away from the busy city life. Night clubs, fireworks, young adults laughing while heading to their favorite restaurants. Whether it is a local small family diner in a quiet street, or an expensive sashimi plate served by the most authentic chief, we would love to explore new cities, expand our horizons, and treat ourselves after all the hard work.

As exciting as the trip sounds, many hours are spent on researching on the new city/country. Some of these resources includes trip guides articles as well as bloggers who venture around the world. How can we locate the resource that tailor to our needs and preferences? How can a family with two young kids easily plan for activities and leisure beaches whereas adventurous couples or groups of friends explore the hiking trails without having to read through pages and pages of online articles?

This paper will explore applying text-mining technologies and machine learning model BERTopic to capture major topics presented in the articles and recommend sightseeing interest so families can save time and ease their mind.

Chapter 2 introduces data collection process; Chapter 3 establishes the methodology and models used in the analysis; Chapter 4 goes into details of analysis; Chapter 5 presents the results from text mining and BERTopic model; Chapter 6 concludes the study.

CHAPTER 2

Data

Two travel bloggers, Jones and handluggageonly were selected based on recommendation article "The 30 Best Travel Blogs OF 2022" [Tod22]. The two bloggers post their travel experience on their websites [Dav21, han22]. The blog journals from Asia and Europe were manually collected from the websites and saved as text files.

The two bloggers varied in styles and travel preferences. Jones would recommend hotels to stay in near the tourist attraction site with pricing details, and handluggageonly features traveling in light luggage. The different blogging content should provide a good variety of information to a wide range of travelers. Articles were randomly selected to cover a wide range of countries with a decent number articles in each country.

Mapping between each article title with its corresponding country and region and author is stored in a csv file. **R** language is used for text mining analysis, and Python is used for BERTopic topic modeling and recommendations. Each document is separated into units of sentences for further analysis.

Jones's journal counts are highly skewed towards Asia whereas Handluggageonly's journals are evenly split in both regions in Figure 2.1. Both authors have an even number with a total of 60 and 76 journals respectively. Common Asia countries for both bloggers are Vietnam, Thailand, Singapore in Figure 2.2. Unique countries for Jones are Indonesia and India; handluggageonly's unique country is Japan. Common European countries include Italy, Germany, France in Figure 2.3. Handluggageonly journals covers more variety of countries. Journals covering multiple countries are marked as either Asia, or Europe.

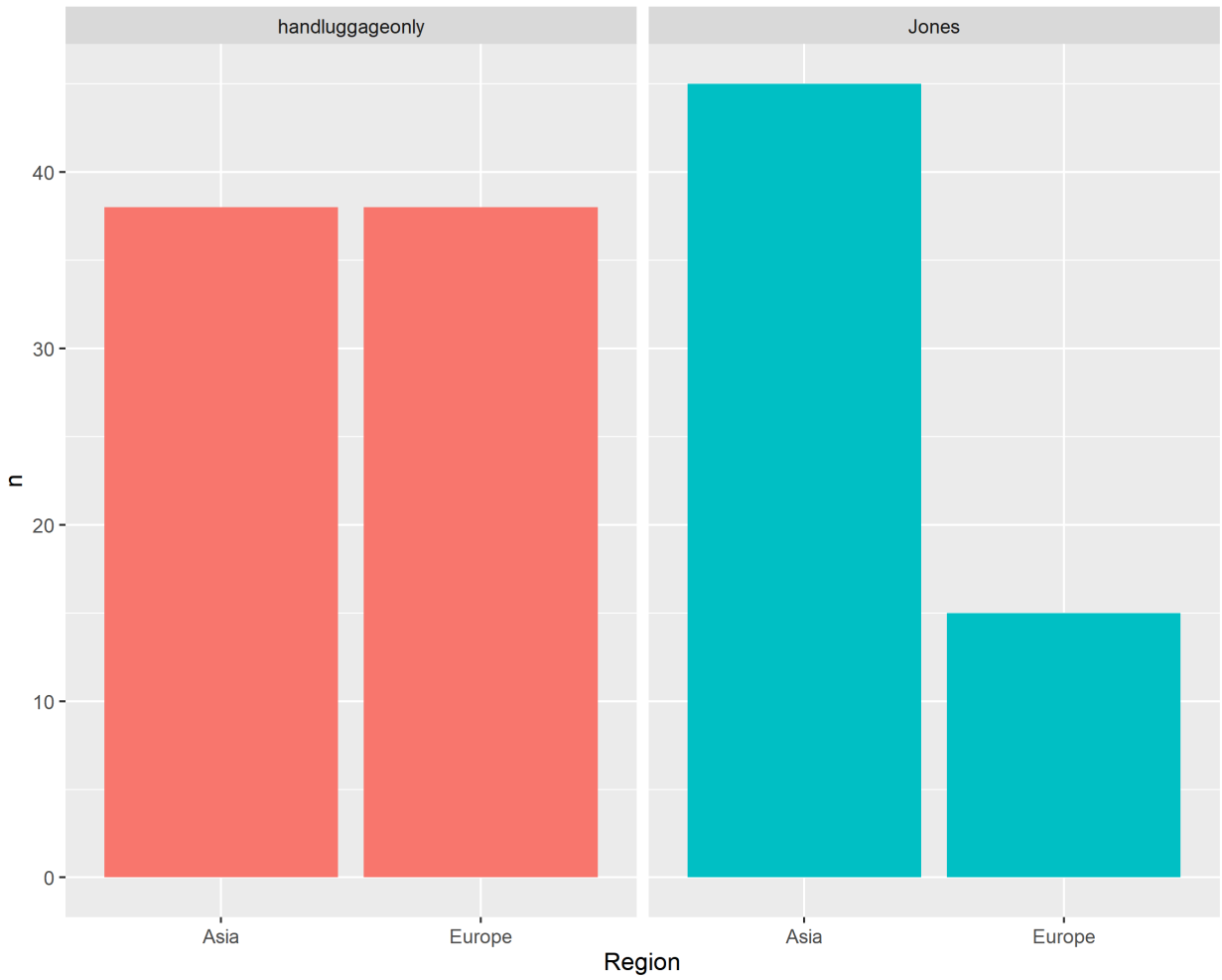

Figure 2.1: Journal Count by Region and Blogger

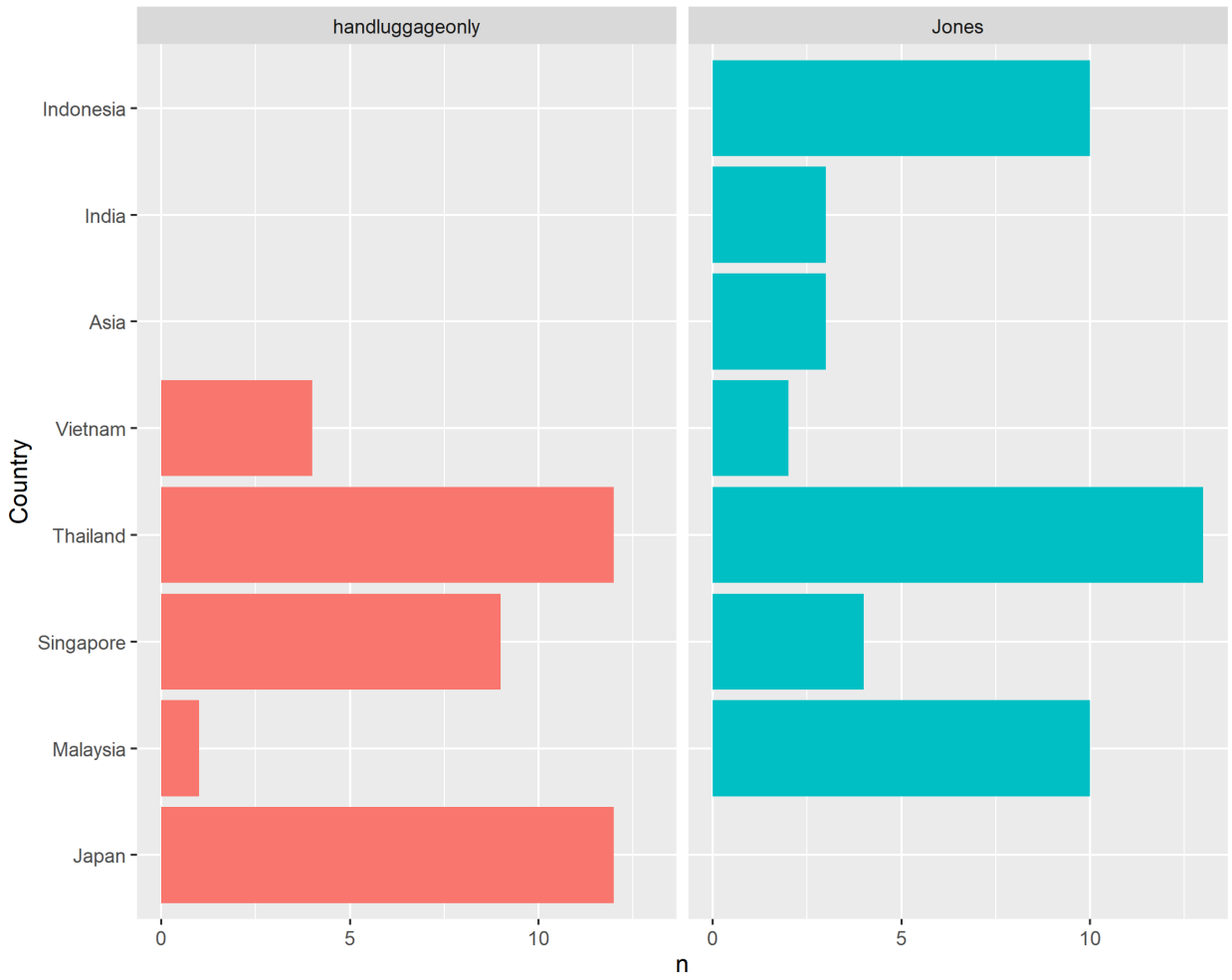

Figure 2.2: Asia Journal Count by Countries and Blogger

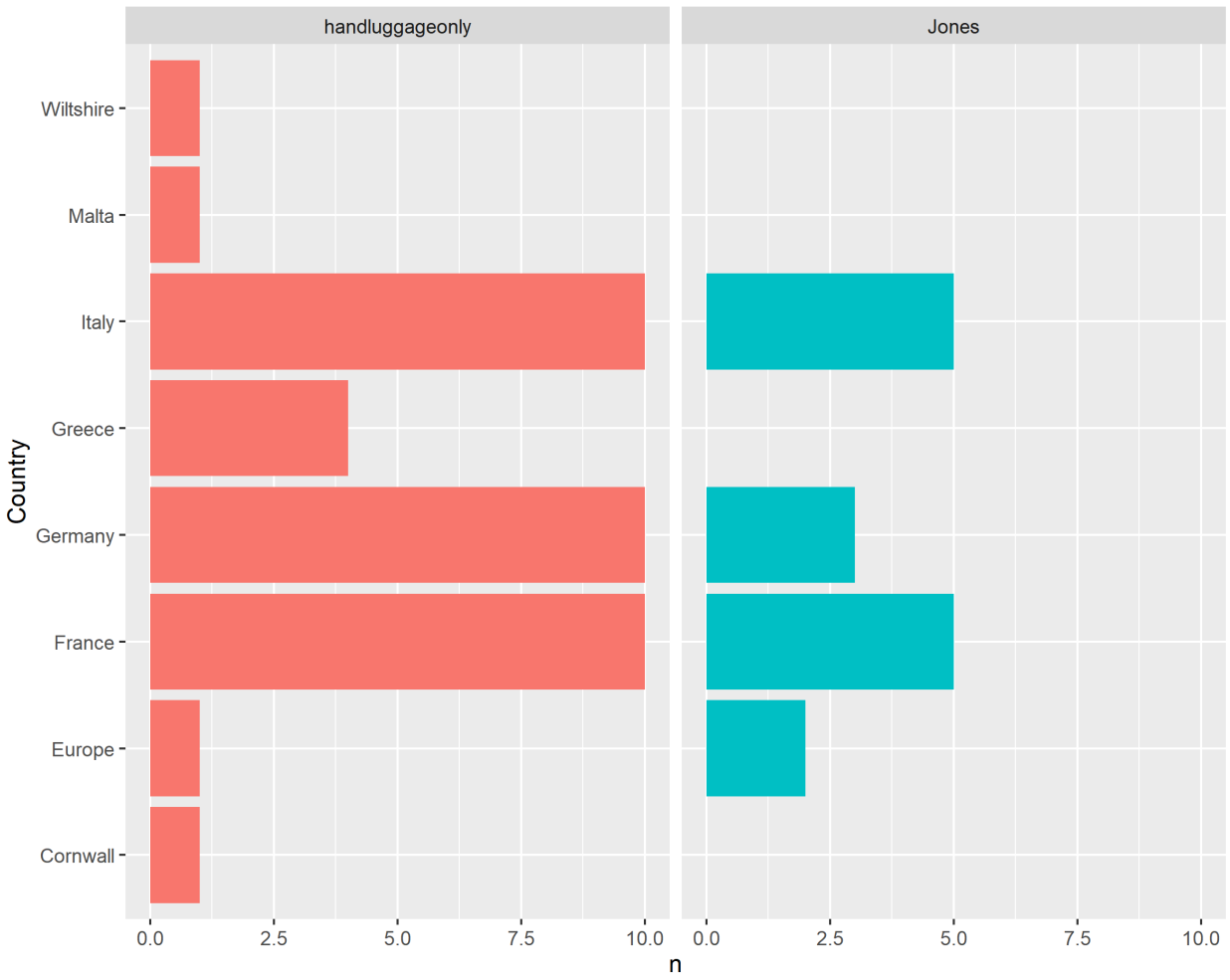

Figure 2.3: European Journal Count by Countries and Blogger

CHAPTER 3

Methodology

3.1 Text Mining Preparation

Basic text mining cleaning processes were applied as follows: transforming text into corpus with one sentence as a unit, unnest tokens, removing stop words. Removing commonly used words such as “a”, “the”, “of” could greatly reduce text size and increase information density relevant to themes of interest. After cleaning the data, region, writing style, themes were explored using text mining technologies and visualization such as word frequency plot, word cloud, topic modeling, correlation analysis.

We also want to group the journals based on traveler preferences. For example, nature lovers might be interested in national parks, wild animals and activities such as hiking, camping.

3.2 Sentiment Analysis

Sentimental analysis allows interpretation and classification of emotions within text data. It can classify words into positive or negative sentiments as well as a category of emotions such as: joy, anger, anticipation. Prebuild sentimental dictionaries such as AFINN, BING, and NRC maps words into sentimental categories and strength scores to quantitatively evaluate feelings conveyed in the messages[SR22b]. For example, NRC dictionary maps the word “abandon” as “negative, sadness, anger, fear” emotions. The NRC dictionary was used to

identify key emotions in the travel guides to understand the writing styles of bloggers and how they create their unique experience using text. Later, the sentimental words compared within each topic for similarities and differences.

3.3 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is an unsupervised topic modeling technique that create sets of words, which could be interpreted as topics. The model is based on the assumptions that all documents share the same topics, and each document contains multiple topics with a different proportion[Mal19]. Given a pre-determined number of topics, LDA will calculate the probability of a word from each topic, and then select the top words with the highest probability belonging to a topic[Lab18]. These words in the same set tends to co-occur contextually and create meaningful topics. Each document is assigned to the most relevant topic based on the chance of the words belonging to each topic. LDA is used in text mining analysis to identify themes within the journals and further extract distinctive experiences specific to each theme/topic.

3.4 UMAP

Uniform Manifold Approximation and Projection (UMAP) is a dimension reduction algorithm that is based on manifold learning techniques. It tried to reduce the dimension of data while retaining the topological data structure based on Riemannian geometry [MHM18, Mal19]. UMAP is scalable and practical in real word data given its advantage of unlimited embedding dimensions.

3.5 HDBSAN

Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSAN) is a clustering algorithm. Comparing with K-means clustering which assumes a spherical shape and equal density in clusters, HDBSAN performs better for clusters with arbitrary shapes and different densities [Ber20]. Clusters are defined as dense points surrounded by sparse points.

HDBSAN is consisted of two parts: 1. DBSAN first identified the underlying distribution of points which will be used to determine dense regions (i.e., PDF distribution regions over a certain λ threshold) 2. Hierarchy is used to represent relationship between clusters as tree structure, where the root of the tree is one cluster and more trees (i.e., dense PDF regions) are grown as the λ threshold decreases.

3.6 c-TF-IDF

Term frequency-inverse document frequency (TF-IDF) is used to find unique and important words within multiple documents by identifying terms frequent within each document and unique among all documents[SR22a]. If the word frequency is high in one document, but also occurs in every other document, such as "the", "a", then it has TF-IDF score of 0. But if it has low occurrence in other documents, then it is more representable of the document.

Term frequency ($tf_{i,j}$) is the number of occurrences of term $\{t_i\}$ in doc $\{doc_j\}$.

Inverse Document Frequency (IDF) for term t_i: $idf_i = log_2 \frac{|D|}{|\{d|t_i \in d\}|}$

Where $|\{d|t_i \in d\}|$ represents number of documents containing the word, and D is total number of documents.

TF-IDF: $tfidf = tf_{i,j} idf_i$

Class-based-TF-IDF (c-TF-IDF) is consolidating all documents in one class as one document and performing TF-IDF on class level. This would help identify unique terms among

classes.

3.7 Transformer

Transformer is a deep-learning model that utilizes self-attention mechanisms introduced by Google Brain. It is used primarily in natural language processing (NLP) and computer vison (CV). It processes the entire input rather than one word at a time to maintain positional information. The additional position knowledge allows for parallel processing, which improves training efficiency compared to recurrent neural networks (RNN).

Transformer model consists of encoding and decoding. The encoder consumes input sequence such as a sentence and creates multiple layers representing different aspects of the encodings[Ala18]. The attention matrix is learnt from a set of weight matrices: the query weights W_Q , the key weights W_K, the value weights W_V. The attention weights explain the amount of attention each token i has for another token j. Two weight matrices creates the flexibility for two tokens to pay different attention to each other. The attention is calculated as follows, normalized using SoftMax so that the sum of all attention for i is 1:

$$\mathsf{Attention}(Q, K, V) = \mathsf{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right) V$$

Where matrix $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are a collection of query, key and value vectors respectively, and d_k is the dimension of key vectors. The ijth element in $\mathbf{Q}\mathbf{K}^\top$ is the dot product between query vector q_i and the transpose of the key vector k_j where q_i is the ith row vector in matrix $\mathbf{Q}$ and k_j is the jth row vector in matrix $\mathbf{K}$.

Softmax is defined as follows so that probabilities of all possible outcomes sum to 1 and each P_c is a weighted probability:

$$P_c = \frac{e^c}{\sum_{c'} e^{c'}}$$

As shown in Figure 3.1 [VSP17], the encoder consists of multi-head attention, and feed-forward neural network. The encoder layer accepts output from its previous encoder and

generates output to feed into the feed-forward process. Each layer contains multiple attention heads, resembling different interpretation of relevance by a set of weight matrices. The decoder structure is like the encoder with additional attention to encoder outputs as well as masking self-attention to prevent leak of future information.

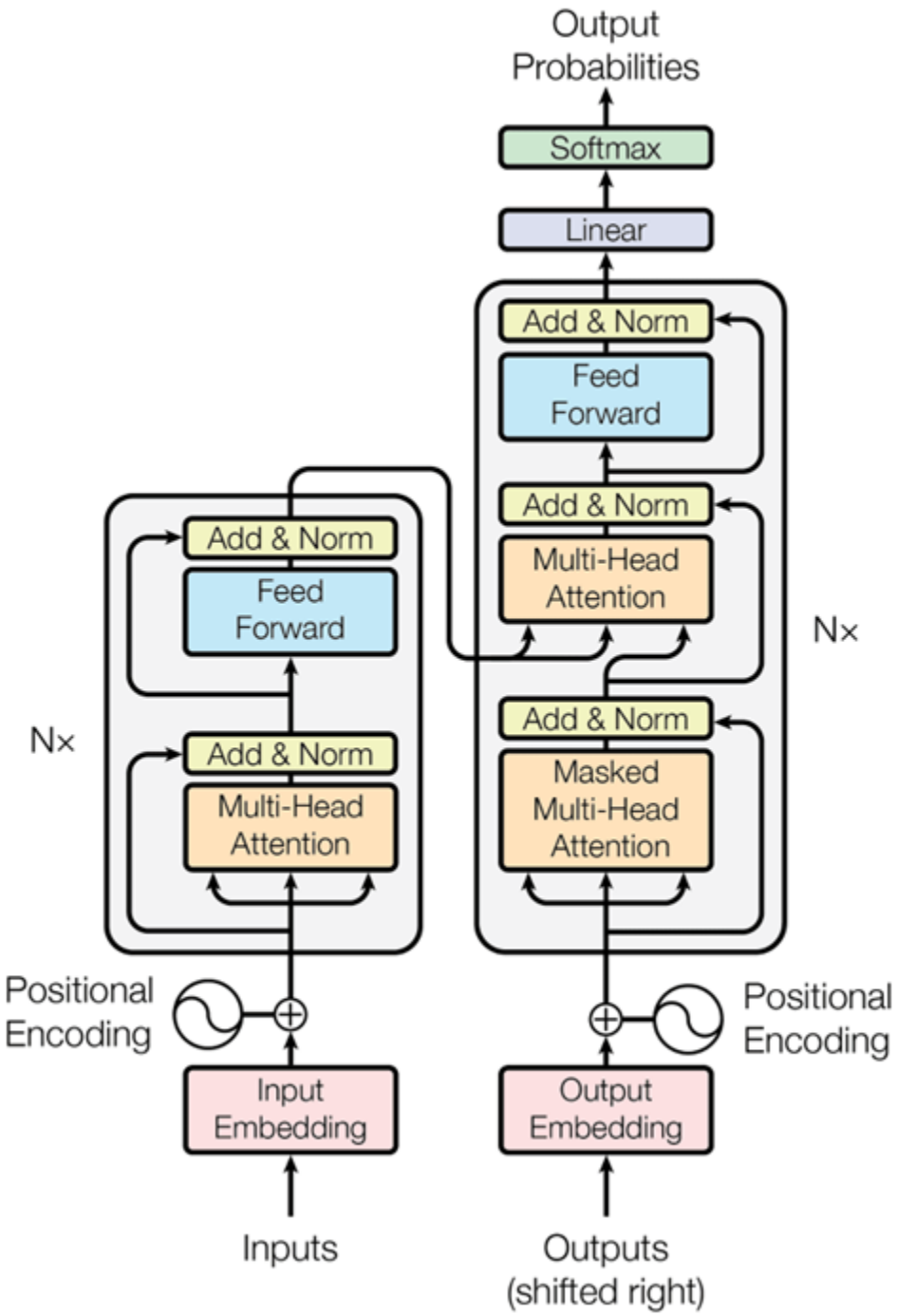

Figure 3.1: The Encoder-Decoder Structure of the Transformer Architecture Taken from "Attention Is All You Need"

3.8 BERT

Bidirectional Encoder Representations from Transformer (BERT) focuses on encoder part of Transformer model. It enhanced sentence contextual prediction power by training on two NLP tasks: Masked Language Modeling (MLM) [SLN19] and Next Sentence Prediction (NSP). BERT model is pretrained on large datasets.

To train for MLM, 15% of the words are replaced with [MASK] token and the model tries to predict masked words based on surrounding words (ie. Context) in a sentence [Hor18]. Sentence Sequence tasks pairs two sentences together with 50% sequential input and the other 50% as random pair. The BERT model tries to predict which of the two sentences are in order.

3.8.1 BERTopic

BERTopic is a topic modeling technique that aims to create easily interpretable topic clusters using c-tf-idf and BERT. It reduces BERT embedding dimensions to 5 using UMAP and create clusters using HDBSAN [Gro21a, Gro18]. Then, the model selects top 5 keywords based on class-based TF-IDF (c-TF-IDF) by identifying unique, important words in each cluster.

CHAPTER 4

Procedure

The procedure encompasses the text mining and BERTopic modeling processes. The source code and data files can be found in https://github.com/ylanJ/Thesis-BERTopic.

4.1 Text Mining

The journals were processed in R from text files into corpus. Cleaning and analysis steps were applied as follows:

Step 1: transformed corpus into unnested tokens

Step 2: remove English stop words, strip white spaces, stemming, remove punctuation

Step 3: Use word frequencies to create word clouds to compare most frequent words by bloggers

Step 4: Create topics with LDA. Identify key words and sentiments within each topic.

Step 5: Perform Tf-idf in each topic group and revealed keywords associating with location names unique to each topic.

4.2 Modeling

4.2.1 Feature Engineering

The text data were processed in Python for topic modeling. First, to clean up unnecessary context, such as picture names saved as repetitive title name, or links to hotel references, were cleaned. Useless sentences are defined by ones with repetitive first 30 characters. Majority of the removed text are trash text, with only a few lines of context, due to repetitive beginning of sentences.

4.2.2 Training and Analysis

Next, the data is fed into BERTopic model while removing English stop words and maintaining n-grams of up to 3 words. The model created around 180 topics with 5 meaningful keywords and can identify relevant topics based on specific inputs. For easier interpretability, repetitive themes such as Greek island, Cornwall beaches, Vietnam beaches are grouped into one category of beach/island. Around 90 topics remain with the most important topics containing around 43% of the data.

To graph the groups in 2 dimensions using UMAP and HDBSCAN, we extracted the embeddings using a newly defined BERTopicNEW model which returns the embedding of the BERTopic model [Gro21b]. By running the BERTopicNew model, it created slightly different topics and results than the original BERTopic model, but we assume that the majority of the groups are similar with different numbering of topics. In the next step analysis, we assigned topics using our old model so that the topic numbers and meanings are consistent throughout the analysis. The BERTopicNEW model is solely used to get embeddings of the journals as 2 dimensional vectors.

CHAPTER 5

Result and Analysis

5.1 Text Mining Data Analysis

5.1.1 Analyze journals by bloggers

In order to understand the different travel journals, we want to get to know our bloggers more closely by analyzing their writing styles and terms/sentiments they use to describe certain attractions. Do they think a hiking trip is exciting, or relaxing? Do they prefer staying at a location for long periods of time for visiting or quick short trips to cover as much grounds as possible? What themes, such as festivals, beaches, nature, museums, have they experience and recommend? Any negative association or avoid suggestions? From their journals, we can see that Jones's are tailored to travel guide and recommending hotel resources on top of major attraction remarks. On the contrary, handluggageonly's journals lead us through each footstep of their trip. We would like to capture these differences in our analysis as well as consolidate Jones's airbnb suggestions as additional resource in our tour guide.

5.1.1.1 Most common words by blogger

The most common words for handluggageonly includes name of locations such as: Thailand, Singapore, Japan, Germany; descriptive terms such as: little, beautiful; theme such as: cities, island, town; general verb: visit, explore.

For Jones, the most common words are themes such as: island, beach, festival, music,

night, Airbnb, view; verbs such as: locate, stay, review, travel; This is clearly related to the hostel reviews in Jones's journals. The symbol "-" is used a lot due to other referenced journals mentioned at the end of the post.

Jones's journals are associated with night events such as music festivals where handluggageonly explores around in the day. Both bloggers love visiting cities.

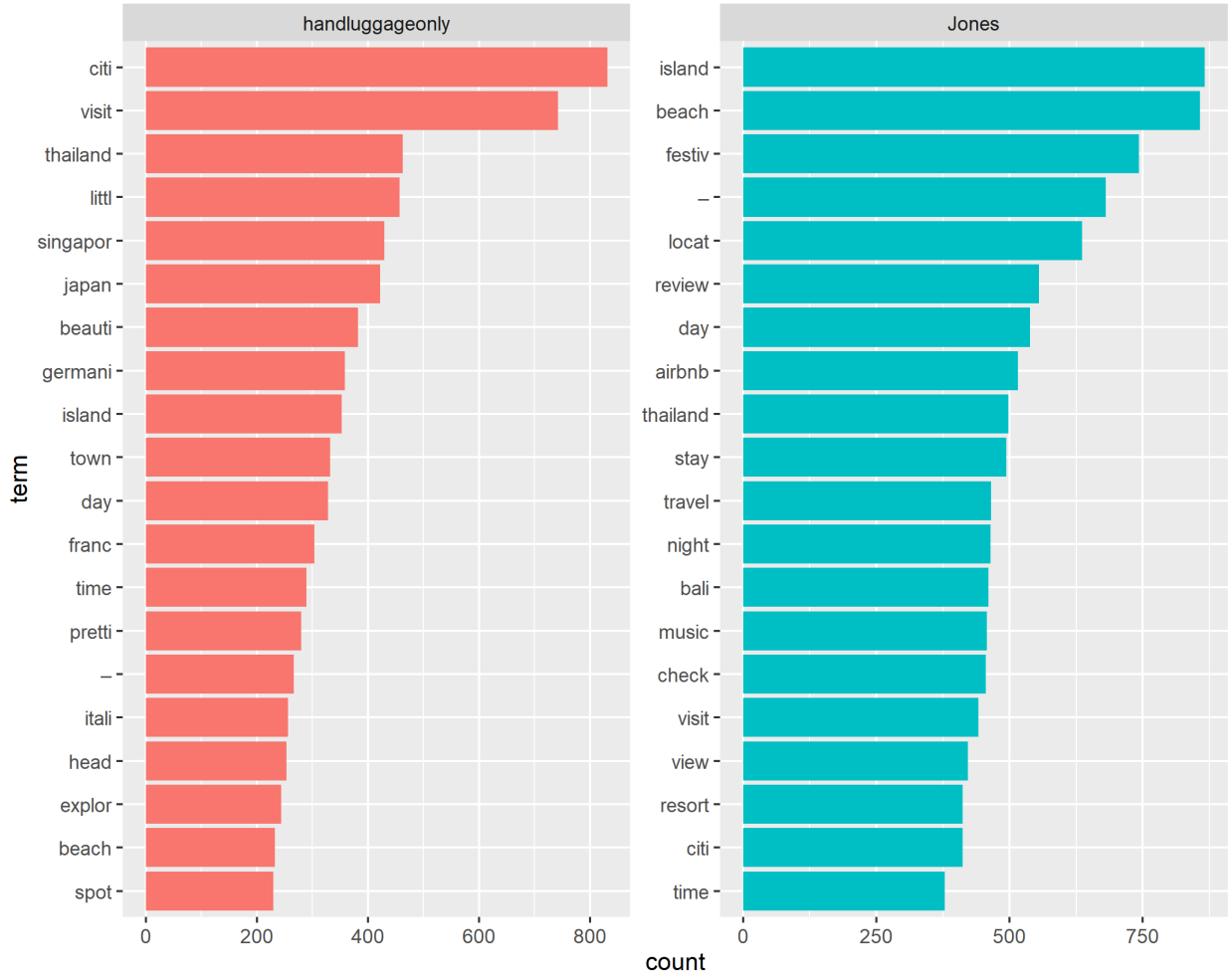

Figure 5.1: Most common words by Blogger

5.1.1.2 Comparing blogger

We plotted the words used by both authors against each other by the frequency to identify common words and differences. Some common words for both bloggers are night, island, beach, bar, drink, airport, café, main, cake, avoid, sunset, friend, absolute, accessible, advance, ahead.

The unique words for each author are further away from the diagonal line. Unique words for Jones include awesome, Airbnb, boast, ad, archipelago, atmosphere, aircondition with 0.1% chance occurring much higher than 0.01% for the other author. Unique words for handluggageonly include explore, inside, favourite, best, actual.

We can see that both bloggers had late night drinks at a bar, travelled by airport, enjoyed dessert like cake. Jones article have housing specific terms, and handluggageonly have many consuming terms such as meat, menu, bite, which creates a more personal experience.

The correlation test for words used by both bloggers are high as 0.98, meaning that the words are almost the same.

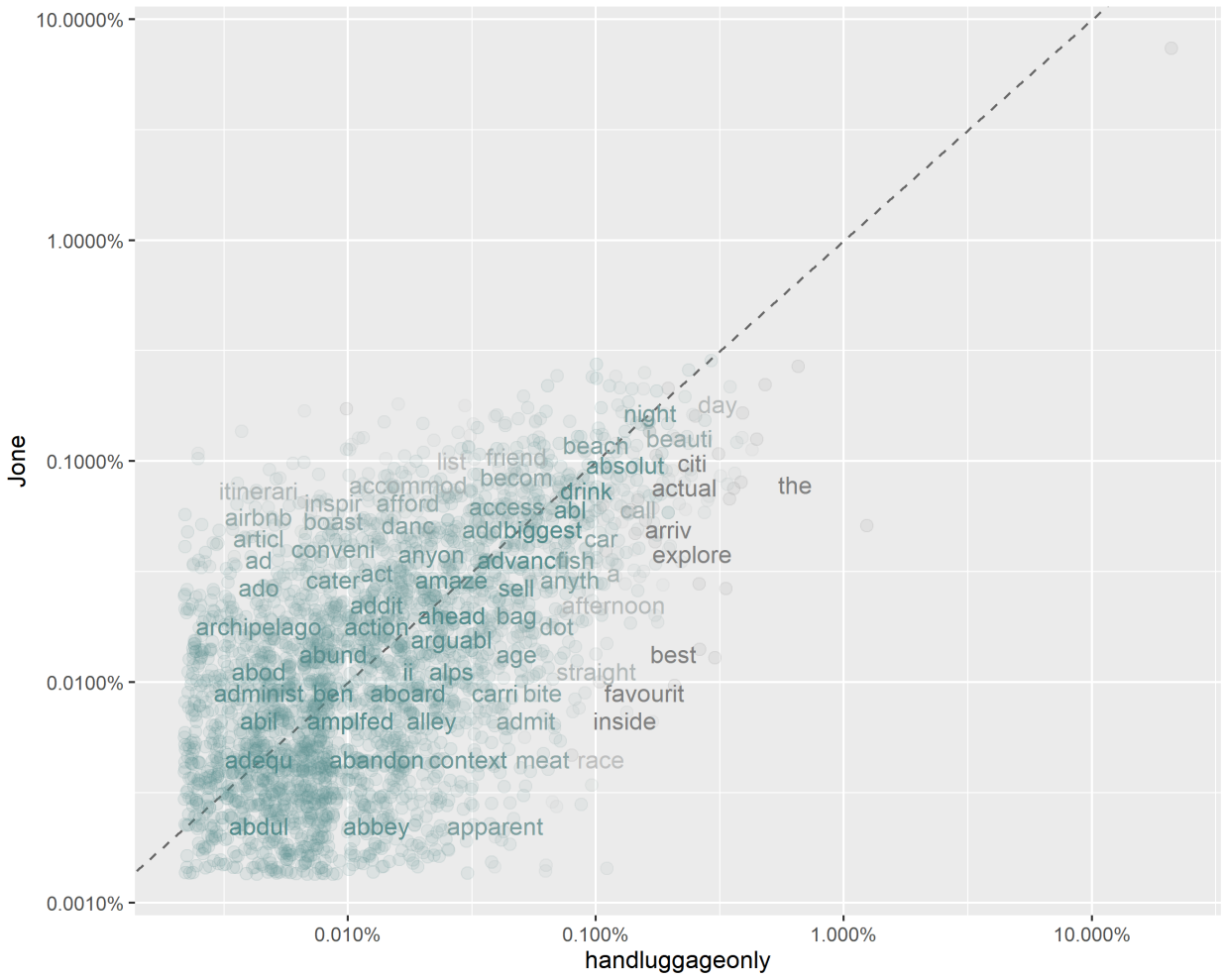

Figure 5.2: Word Comparison Blogger

5.1.1.3 Context of the words

We would like to explore further to see why some terms were mentioned.

Jones:

Afford:

The word afford was mentioned as many as in 29 journals.

As an example, in the journal "10 AMAZING THINGS TO DO ON GILI TRAWAN-

GAN", scuba-diving are recommended in Komodo National Park with affordable pricing. Snorkeling is a great affordable way to cruise around the three Fili Islands, as well as the Pandawa Beach Villas & Resort are also reasonably priced location for backpackers.

Other than recommending cheap places, Jones also use "afford" as being granted a gift by beautiful sceneries as quoted in "You'll be afforded some of the most scenic inland and oceanic views of Cote D'Azur."

Sunset:

Another popular word in Jone's journals is "sunset". It was mentioned in 40 journals. One example in "10 AMAZING & FREE THINGS TO DO IN SINGAPORE", a bridge was suggested for photographers to take sunset shots. In the same Gili Trawangan journal mentioned previously, a life style of "watching the sunset every night with a cocktail or a coconut in your hand" was used to emphasize the beauty of the view.

handluggageonly:

Best:

Handluggageonly loves to emphasize his emotions using "best". The word was mentioned in 54 of his journals ranging from "the best beaches in Thailand", "best time to visit" the beaches, "best diving and snorkelling spots" to "the best Indian and Bangladeshi restaurants" in Singapore, the best hotels in Singapore "within easy reach of some of the best sites". "The best things to do" and "the best beaches in Thailand" was repeated numerous times in the same journals.

Explore:

The word "Explore" was mentioned in 64 journals. He "explores beneath the waves" and "the summer residences" in Thailand. It is a common term for him to express the curiosity of tourists to a new city, whether it is "exploring Lake Como's gorgeous shoreline", "explore some of the museums that are perched around Como itself" in Italy to "explore the ruins of this iconic 12th Century abbey" in Germany.

Avoid:

Being a negative word, “avoid” was mentioned in 11 journals. This was used to warn travelers to “avoid peak season” travel in Kefalonia, Greece, “avoid the crowds by going early in the morning” in Maya Bay In The Phi Phi Islands, Thailand, in hindsight “avoid feeding the monkeys on the islands and allow them to forage naturally”. All the warnings were used in a positive way as someone who experienced the downside of crowds and long queues, handluggageonly provided many reminders in each instance where one should prepare for early trips or alternative solutions.

5.1.1.4 Correlation plot

We created a correlation plot by filtering words with more than 500 occurrences for both bloggers and correlation of 0.5 or more. The common words that go together are “locate, review, stay, airbnb, night, check” which are all locating words. Some other common phrases include “best thing to do”, “this also can take”, “best visit”, “I just get around”, “travel time”.

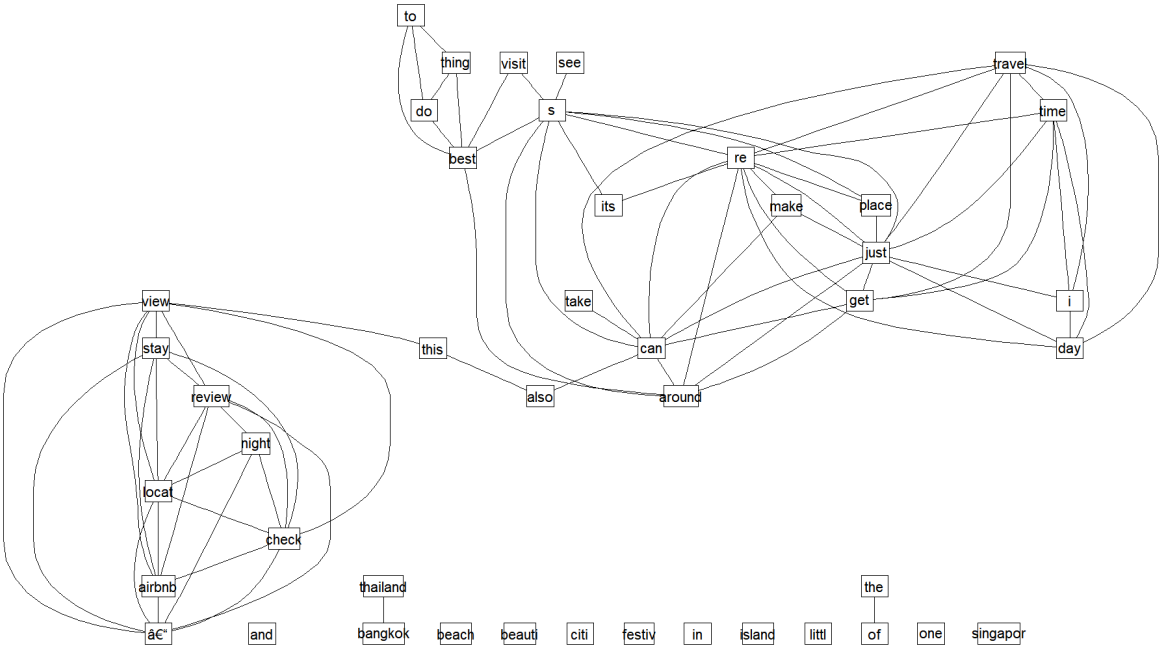

Figure 5.3: Correlation plot

5.1.2 Sentiment Analysis by blogger

5.1.2.1 Basic sentiment analysis

We want to analyze the sentiments for the journals. We joined the words with the "nrc" sentiment table [MT13] which maps sentimental words with one or more emotions. As shown in the pie chart in Figure 5.4, more than 75% of the emotions expressed are positive for all journals. The details of each sentiment for each blogger are ordered in descending order in Figure5.5. Majority of the emotions are positive including anticipation, joy, trust, and only a few negative emotions such as sadness, fear, anger, and disgust, which is expected.

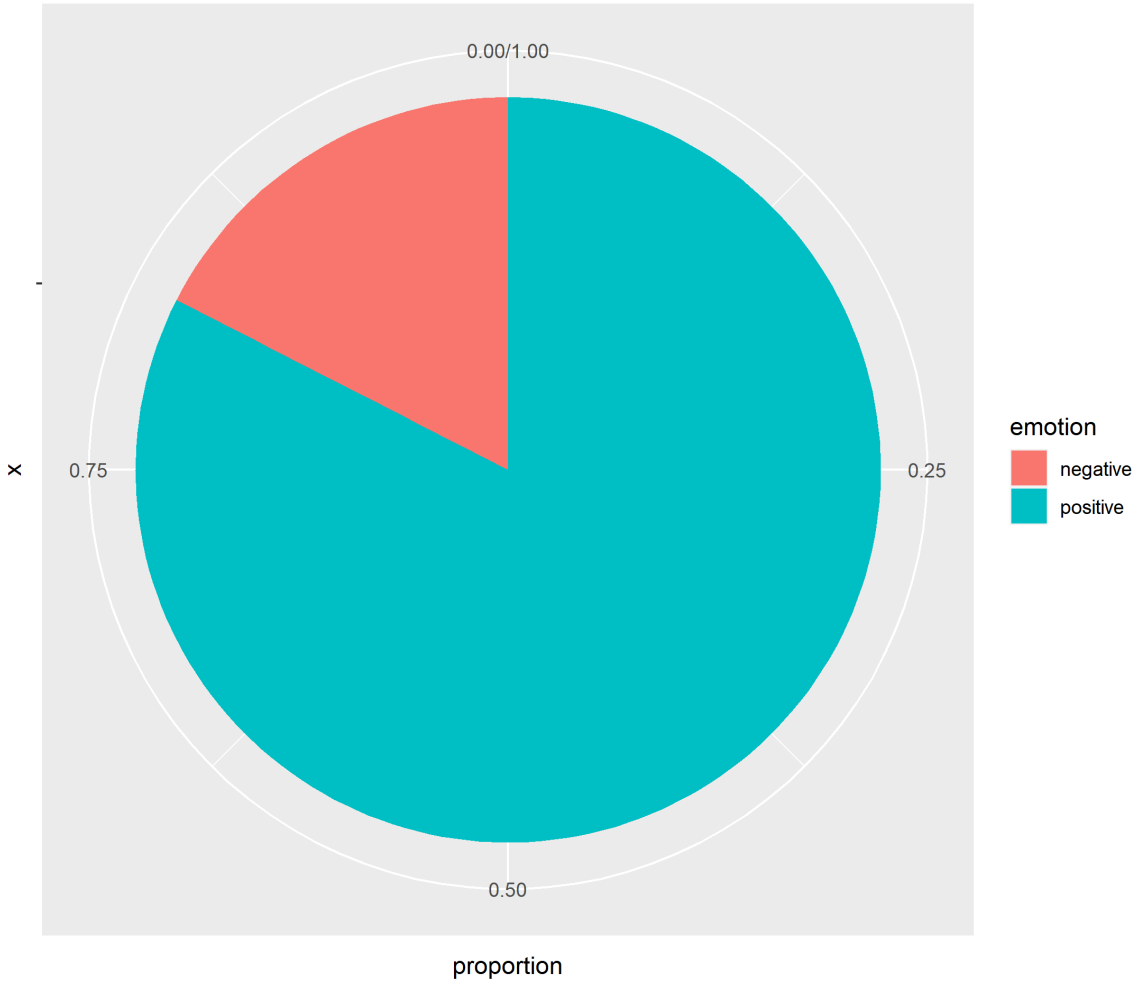

Figure 5.4: Positive vs Negative Sentiment

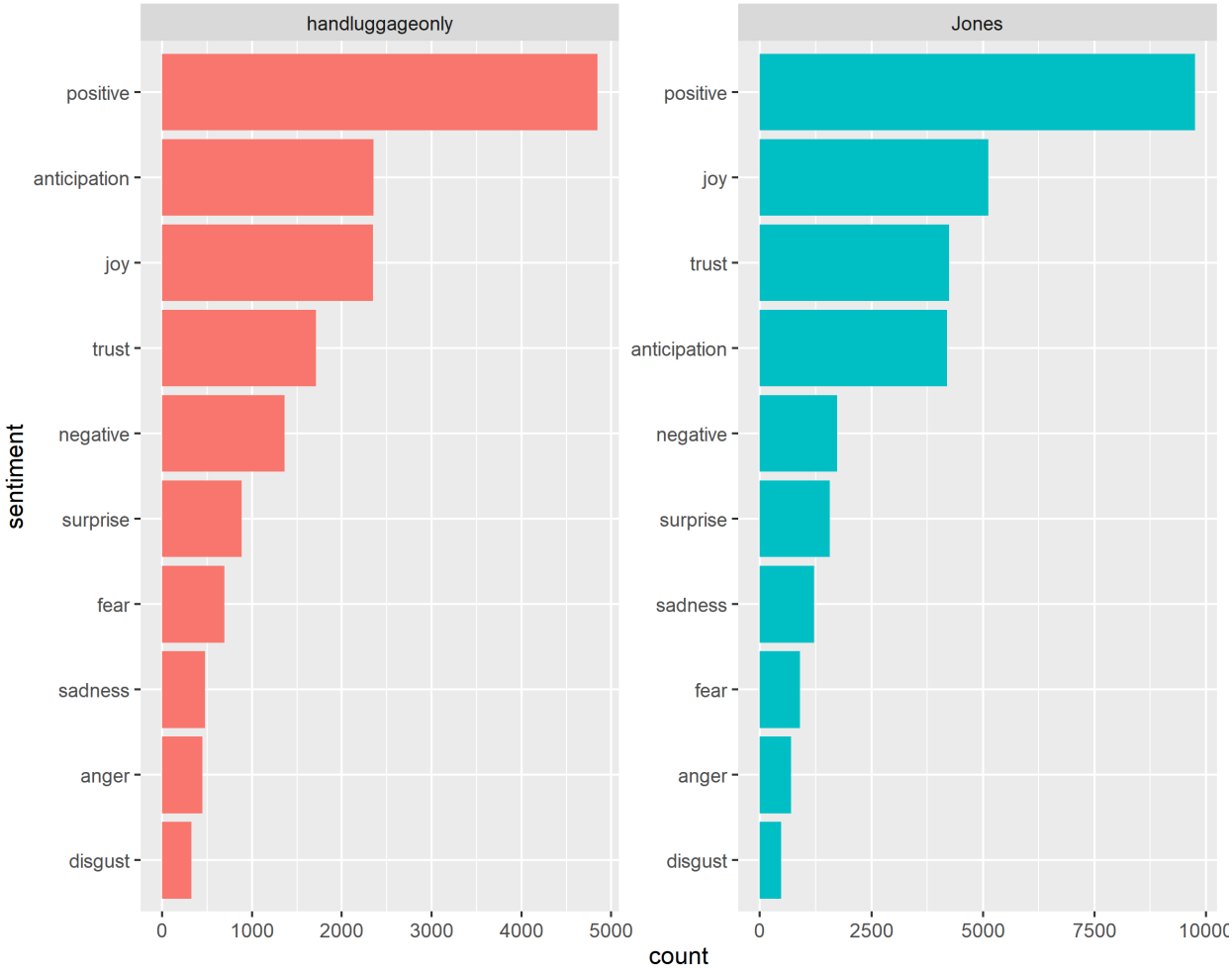

Figure 5.5: Sentiment by Blogger

5.1.2.2 Most common words for each sentiment

The most common words associated with sentiment among both bloggers are shown in Figure 5.6 and 5.7. The most common words for each sentiment are reviewed as followed.

Figure 5.6: Sentiment Wordcloud by handluggageonly

Figure 5.7: Sentiment Wordcloud by Jones

Joy:

The most common joy word among both bloggers is beach in Figure 5.8. This is a common theme since both bloggers recommended great beaches. Music the second most

frequent theme for Jones. Both loves food, have fun, and embracing nature in the gardens. Among the top 20 terms, handluggageonly's journey explores church, and search for gems of the city. Adjectives such as sweet, special, honest creates a special occasion, finding something new, sharing the fun with friends and loved ones, treating themselves with the delicious food and gorgeous views. Jones words creates a comfort, clean location to stay at, while enjoying special spa care. Activities such as swimming, attending festivals, and special honeymoon trip, all are elements of happiness.

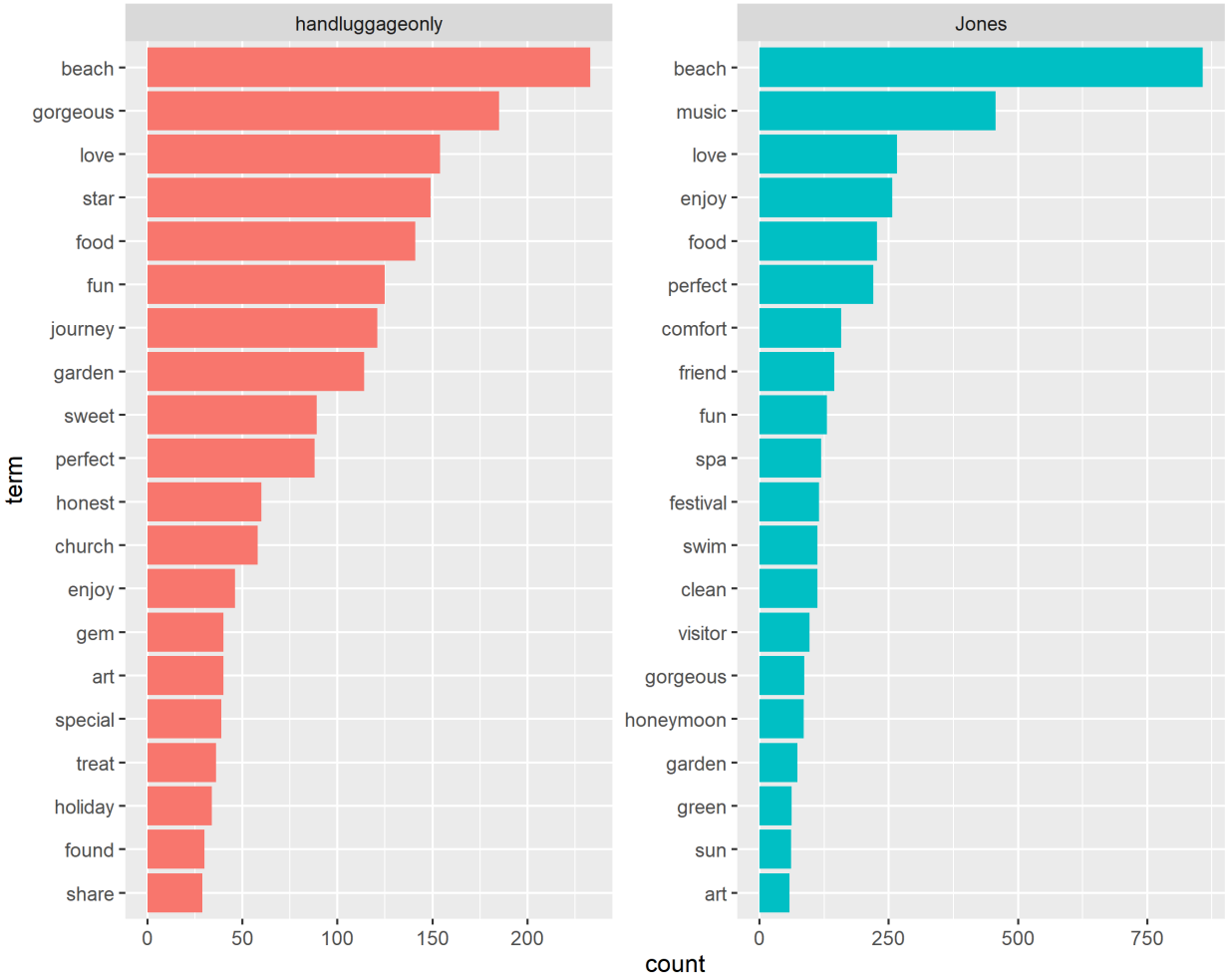

Figure 5.8: Top Joy Words by Blogger

Anticipation:

For anticipation in Figure 5.9, both bloggers use terms that illustrates an exciting experience such as: perfect, fun, top sites to visit. Words like enjoy, explore make the readers curious and amused by the new discoveries. Themes that travelers look out for: sunset, morning. Emphasize on importance of trip planning: time, expect, morning.

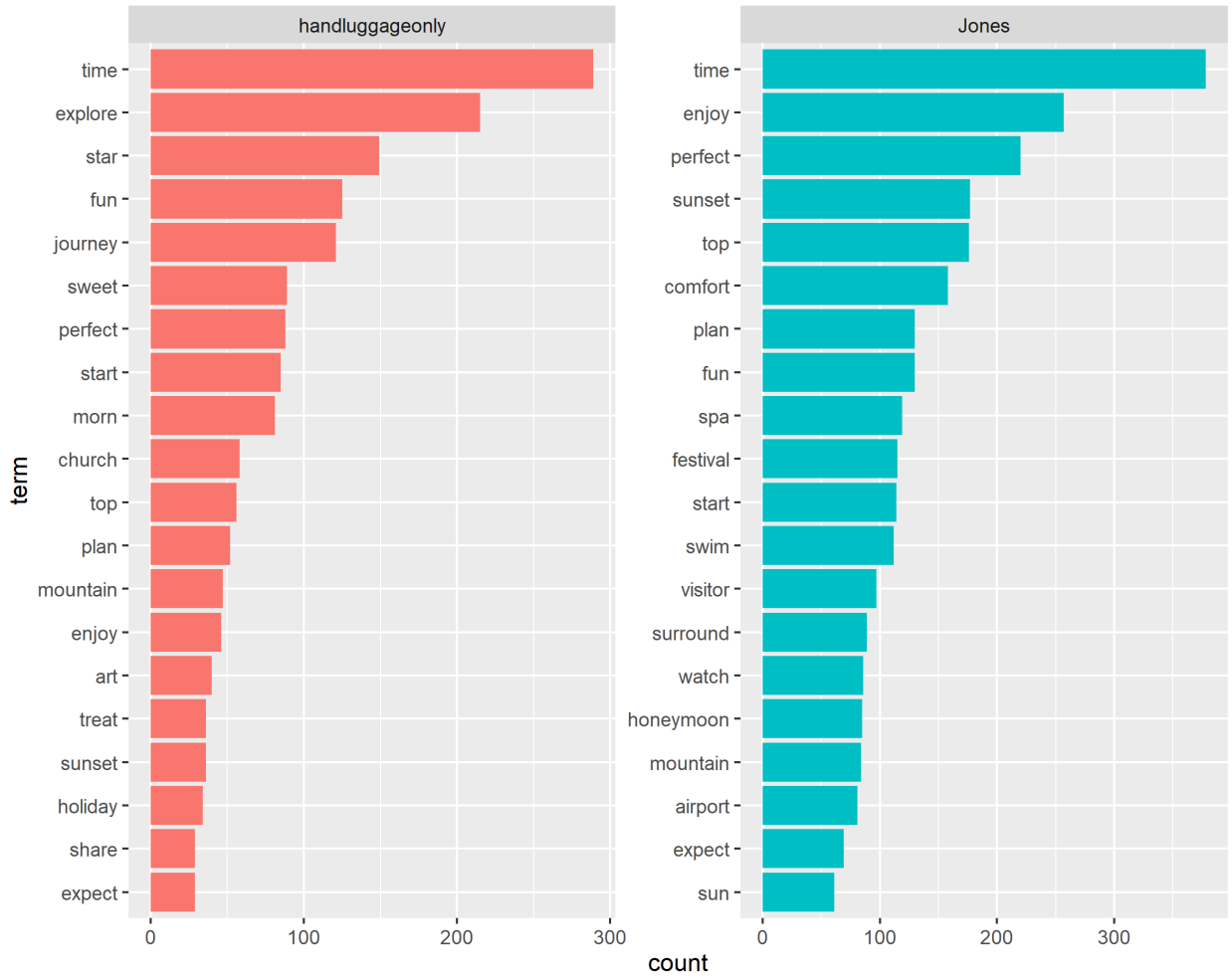

Figure 5.9: Top Anticipation Words by Blogger

Trust:

Trust as an important element in bonding with the readers, both bloggers used positive descriptions such as perfect, recommend, enjoy, to create a knowledgeable, encouraging vibe in their adventures and recommendations in Figure 5.10. They emphasize on what readers

look for: clean, comfortable spa, enjoyable food, budget-friendly planning, trust-worthy recommendations, honest opinions.

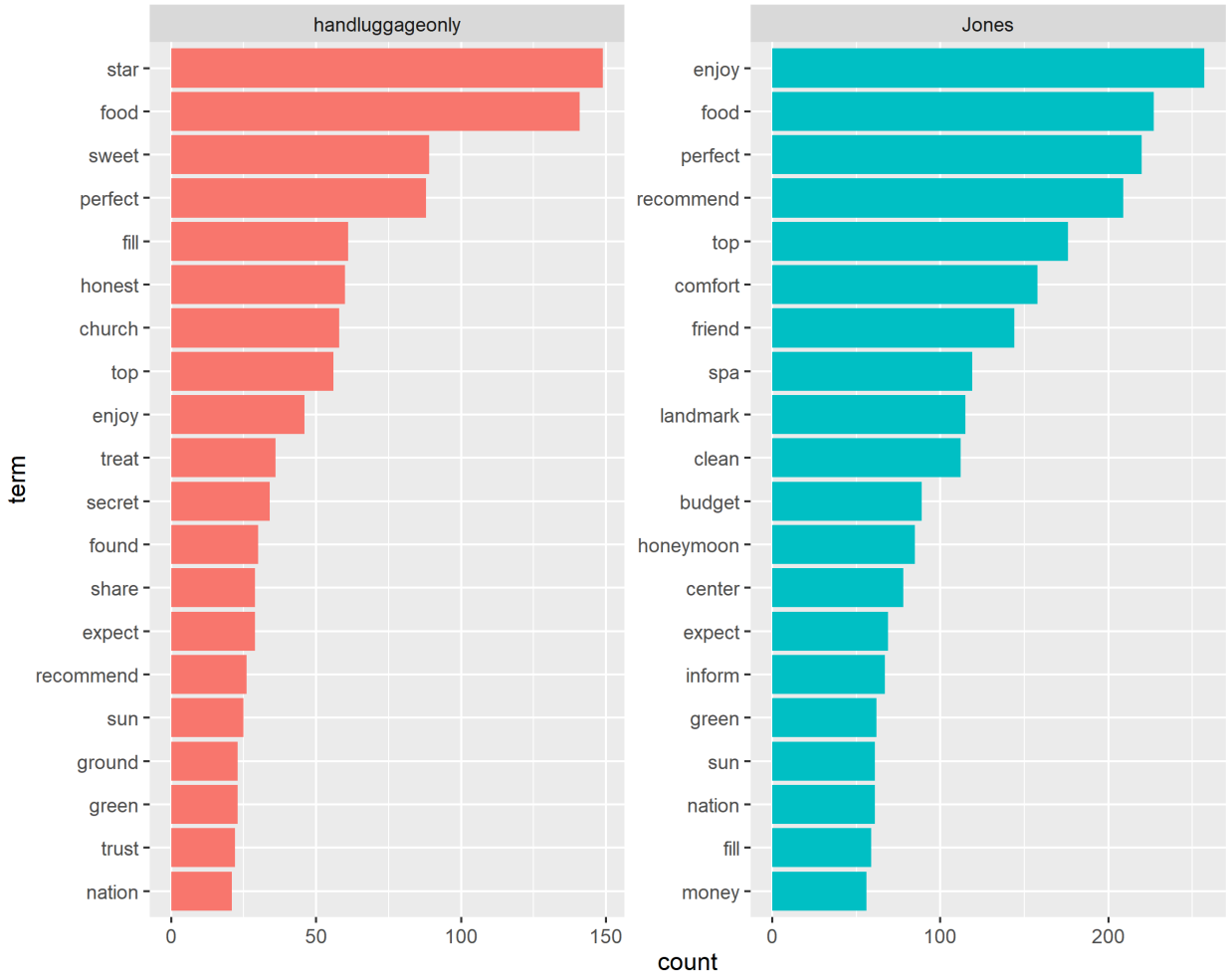

Figure 5.10: Top Trust Words by Blogger

Fear:

Fear is an element that creates a thrilling excitement that many look forward to in a new venture. Our bloggers present this element in the form of exciting events in Figure 5.11 such as swimming, exploring cliffs and coves, dancing around fire, visiting Japanese samurais, vocano visit. War and death were mentioned as one goes back in time embraced by historical sites. These potentially life-threatening factors takes one out of their comfort

zone, experiencing a thrill far different than their ordinary citizen life.

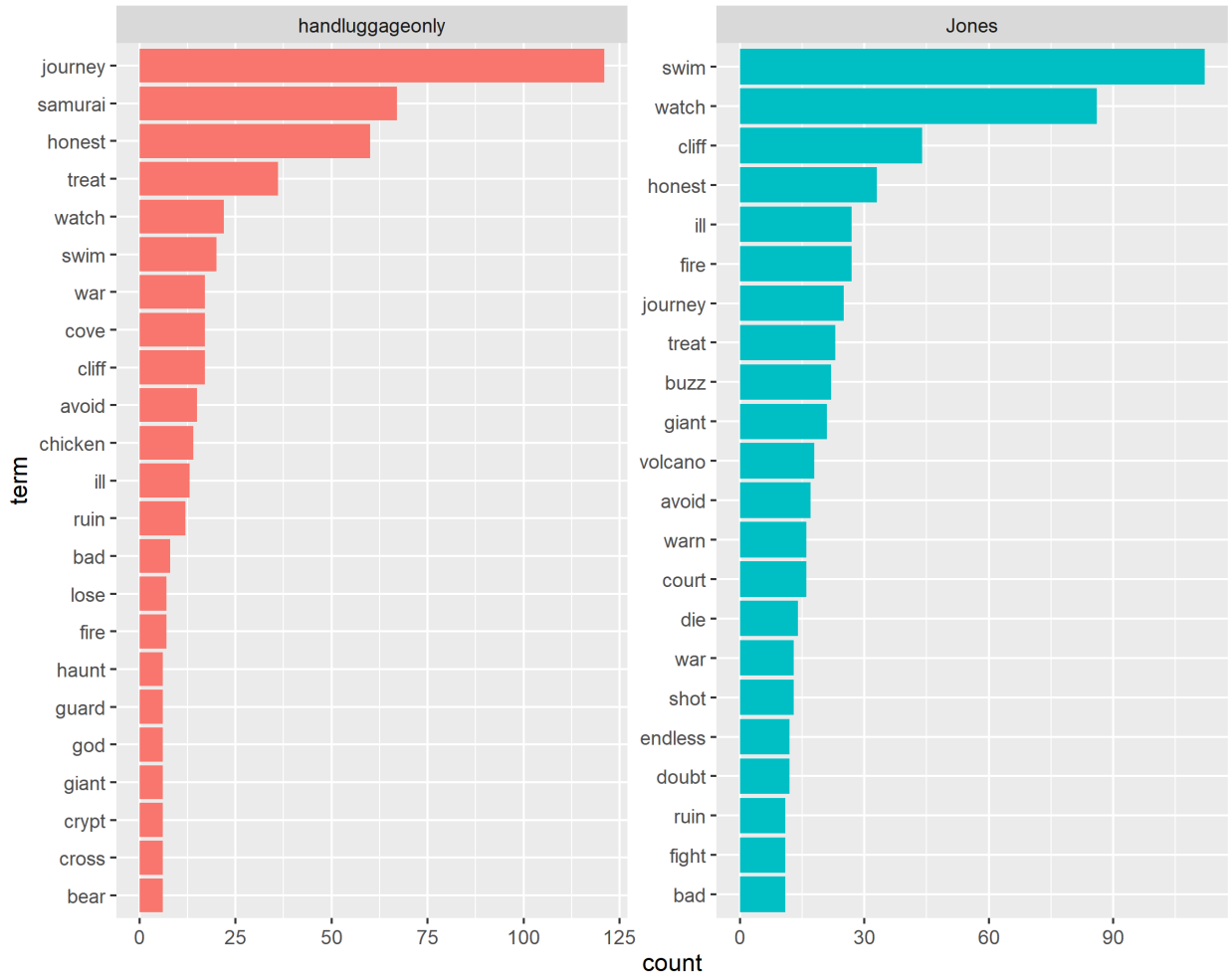

Figure 5.11: Top Fear Words by Blogger

5.1.3 LDA Topic Modeling

Next, we want to explore what types of travels experiences were suggested and can be tracked for easier user recommendations. We will use topic modeling to group the articles and see which suggestions are closely related to each other compared to others. What does the model find as significant features that differentiate one group of articles with another?

5.1.3.1 Created Four Topics by LDA

We used LDA library to create four topics based on the words in each journal. To remove common words that might occur in every journal, we calculated tf-idf which measures uniqueness and frequency for each word among all documents and removed the terms with zero tf-idf score, which are common terms that occur in every document. The LDA model resulted in the below words that are most associated with each topic.

First topic is highly related to island, beach, as well as housing reviews, and Airbnb pricing, available. It's also highly related to Malaysia and travel photos.

Second topic is related to visit cities, locations such as France, Italy, Singapore, Rome, Germany, exploring pretty, little places and beautiful forest, spot.

The third topic is related to music festival, Germany, and locations in Japan such as Kanazawa, Yokohama, Hakusan, where temples are an exclusive attraction.

Finally, the fourth topic is related to Bangkok, Thailand, Uluwatu Bali in Indonesia, where beaches are extraordinary. Market and party are also unique to this topic. Drinking in bars could be one of the main themes as well.

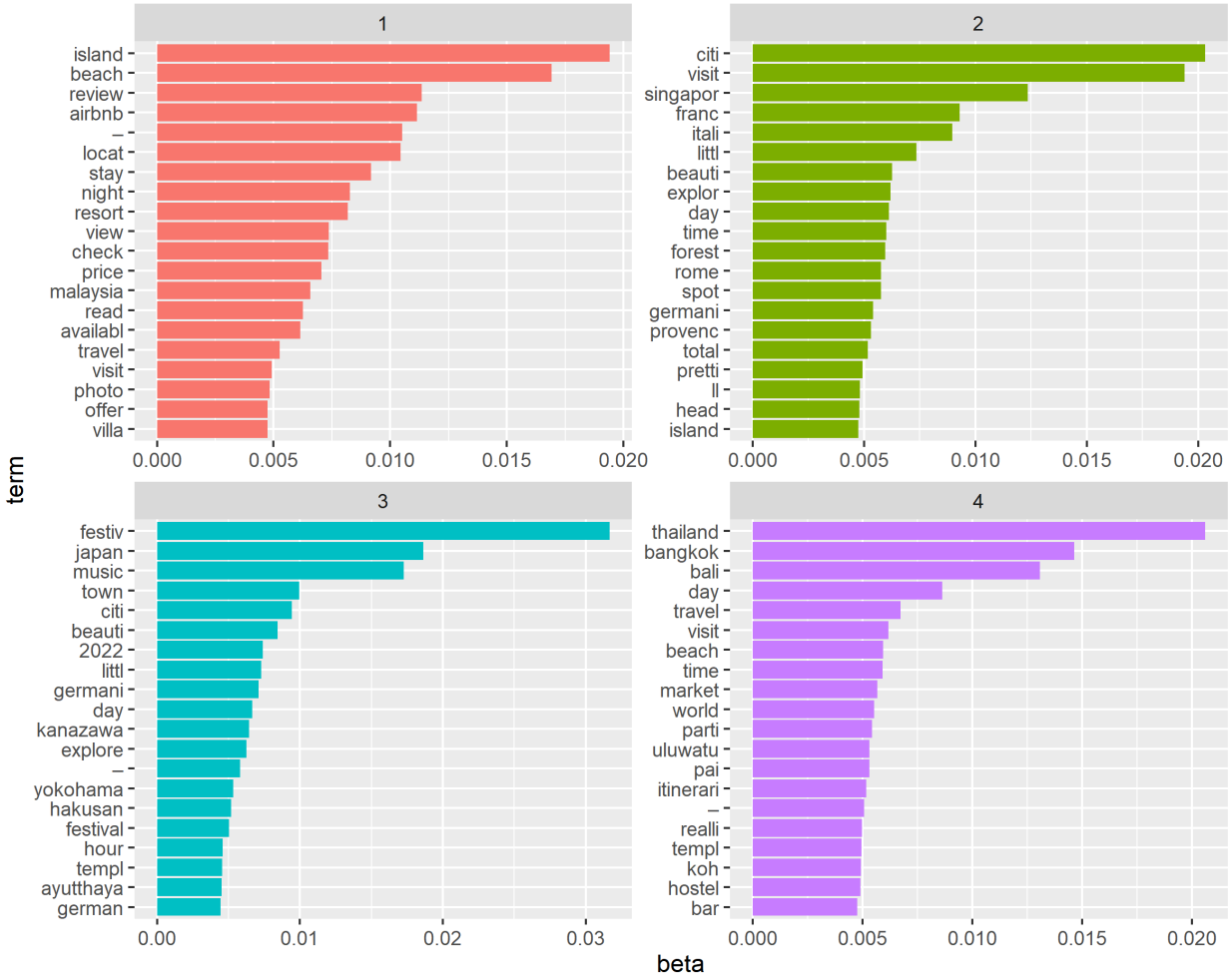

Figure 5.12: Most Common Words in each of Four Topics

5.1.3.2 Classify Journals to Topics

To confirm our interpretations of the topics, we now assign journal to its most associated topics.

Topic 1 by handluggageonly:

Journals that are categorized to first topic in Figure 5.13 are greatly related to Thailand island including ones in Phang Nga Bay and Phi Phi islands in Thailand. The model is less certain of assigning two of the journals to topic 1.

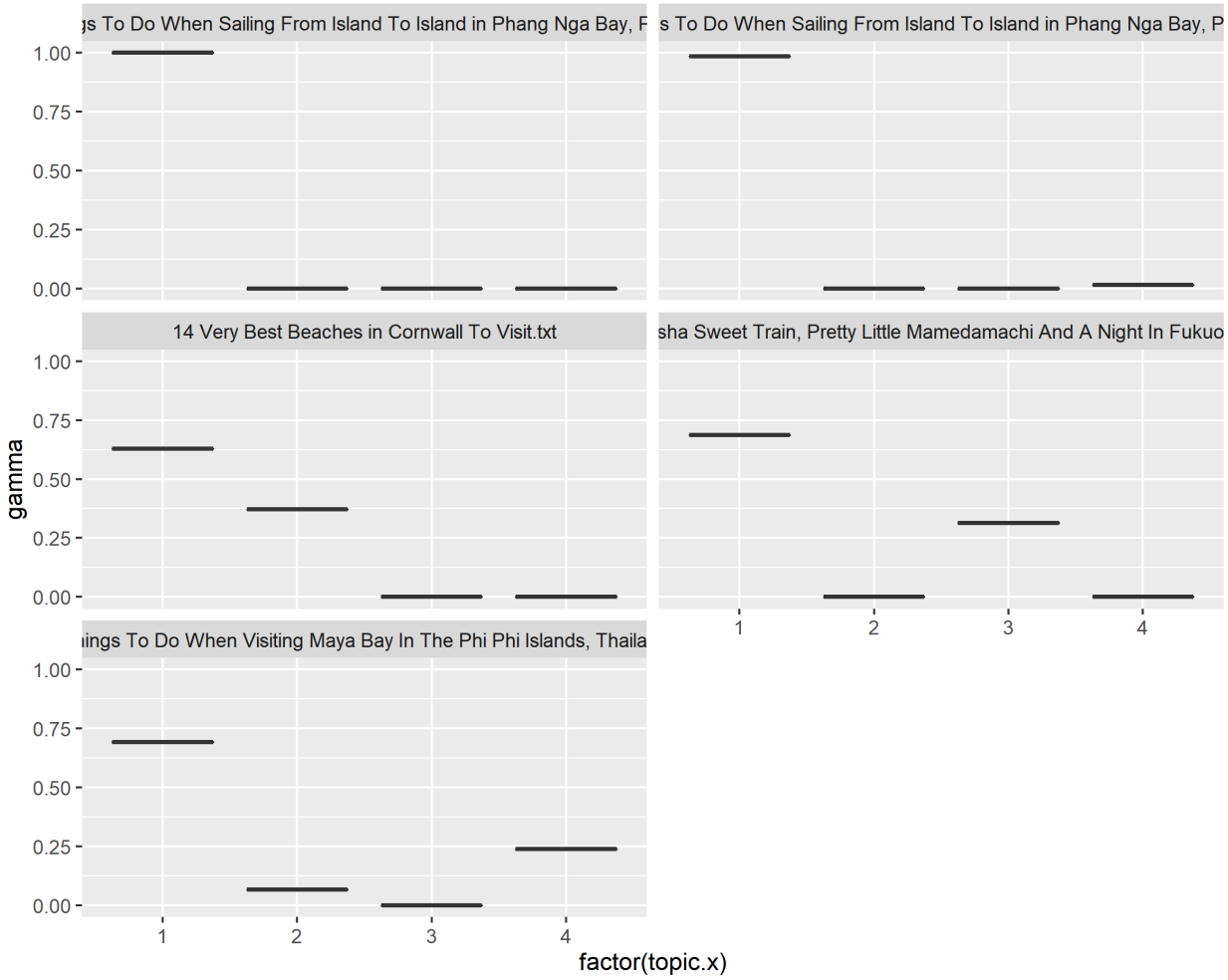

Figure 5.13: Probability of Journal categorized as Topic 1 for handluggageonly

Topic 1 by Jones:

Recommendations to the best hostels and Airbnb journals are easily categorized to topic 1. Journals in Malaysia related to islands and beaches are also in the category as expected. Shown below in Figure 5.14 are a sample of the journals.

Figure 5.14: Probability of Journal categorized as Topic 1 for Jones

Topic 2 by handluggageonly:

A subset of journals was shown in Figure 5.15, and many Avignon, France, Singapore, Germany journals falls into this topic. Dining with a view in Ho Chi Minh city is exploring the city in Vietnam.

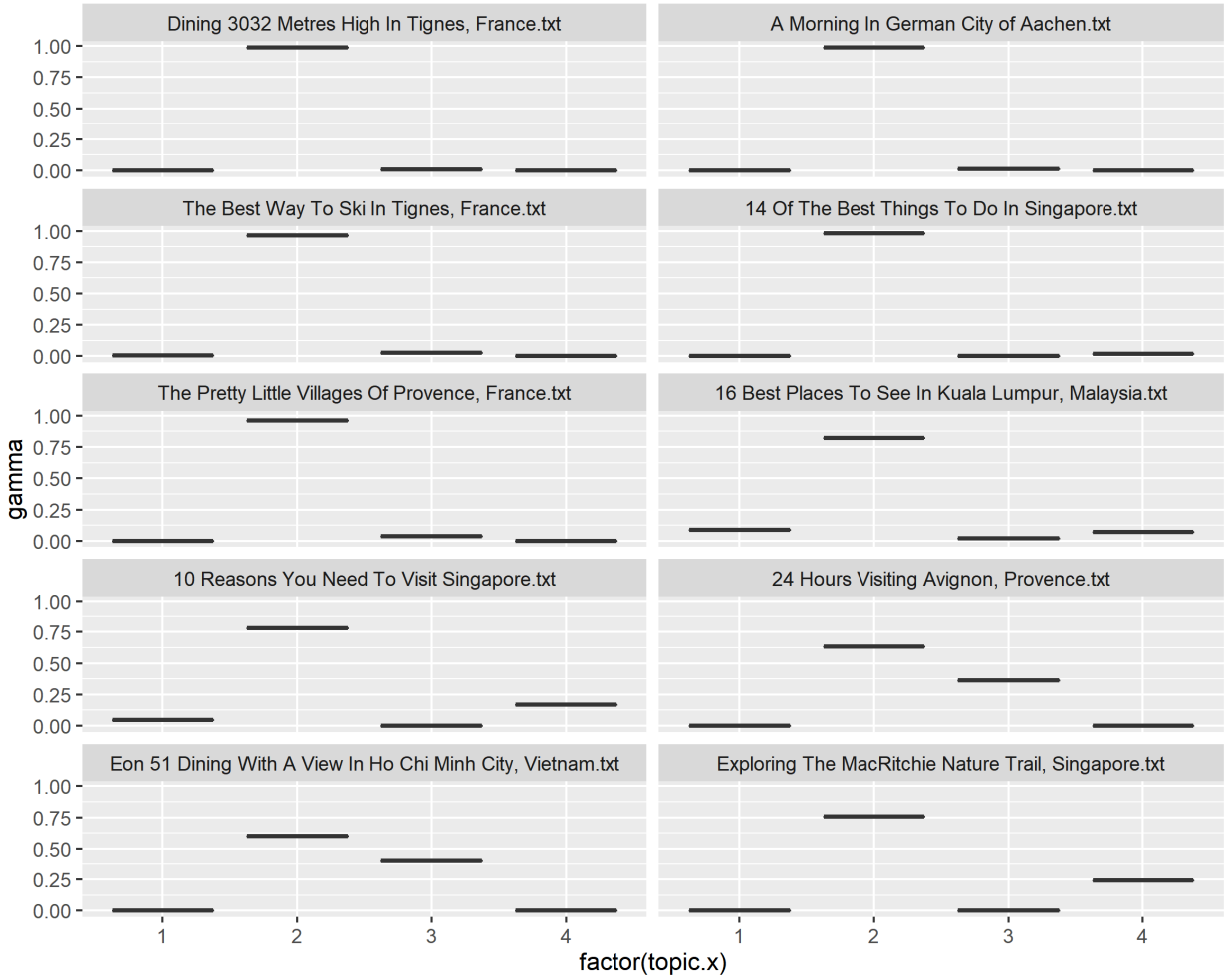

Figure 5.15: Probability of Journal categorized as Topic 2 for handluggageonly

Topic 2 by Jones:

In Figure 5.16, Jones's journals in Italy, Rome, and Singapore resides in the topic as expected.

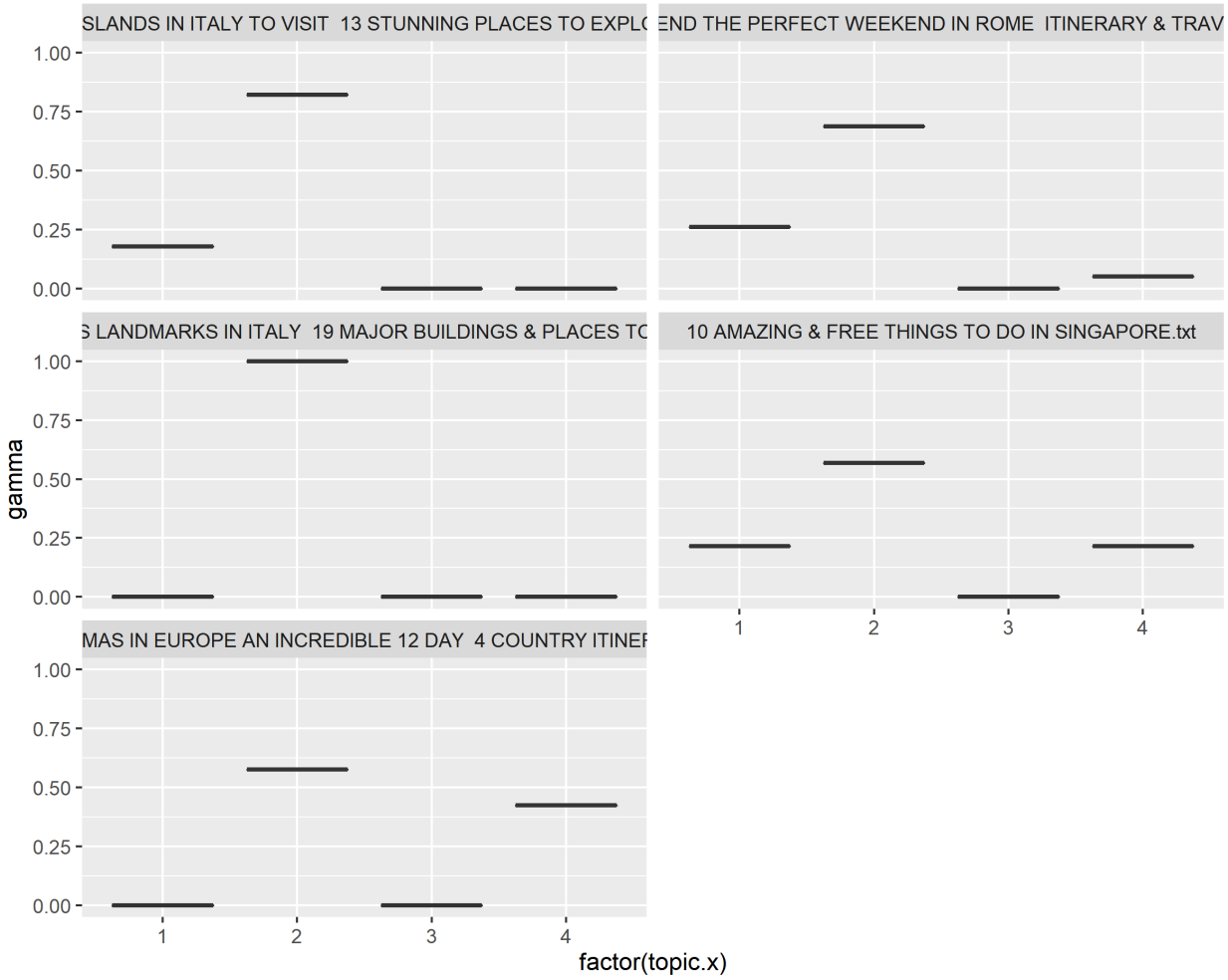

Figure 5.16: Probability of Journal categorized as Topic 2 for Jones

Topic 3 by handluggageonly:

The third topic is highlighted as temples in Japan, ancient attractions in Thailand. Notice "12 Best Hikes in Japan" weighted equally between topic 2 and topic 3 since topic 2 is related to natural sights.

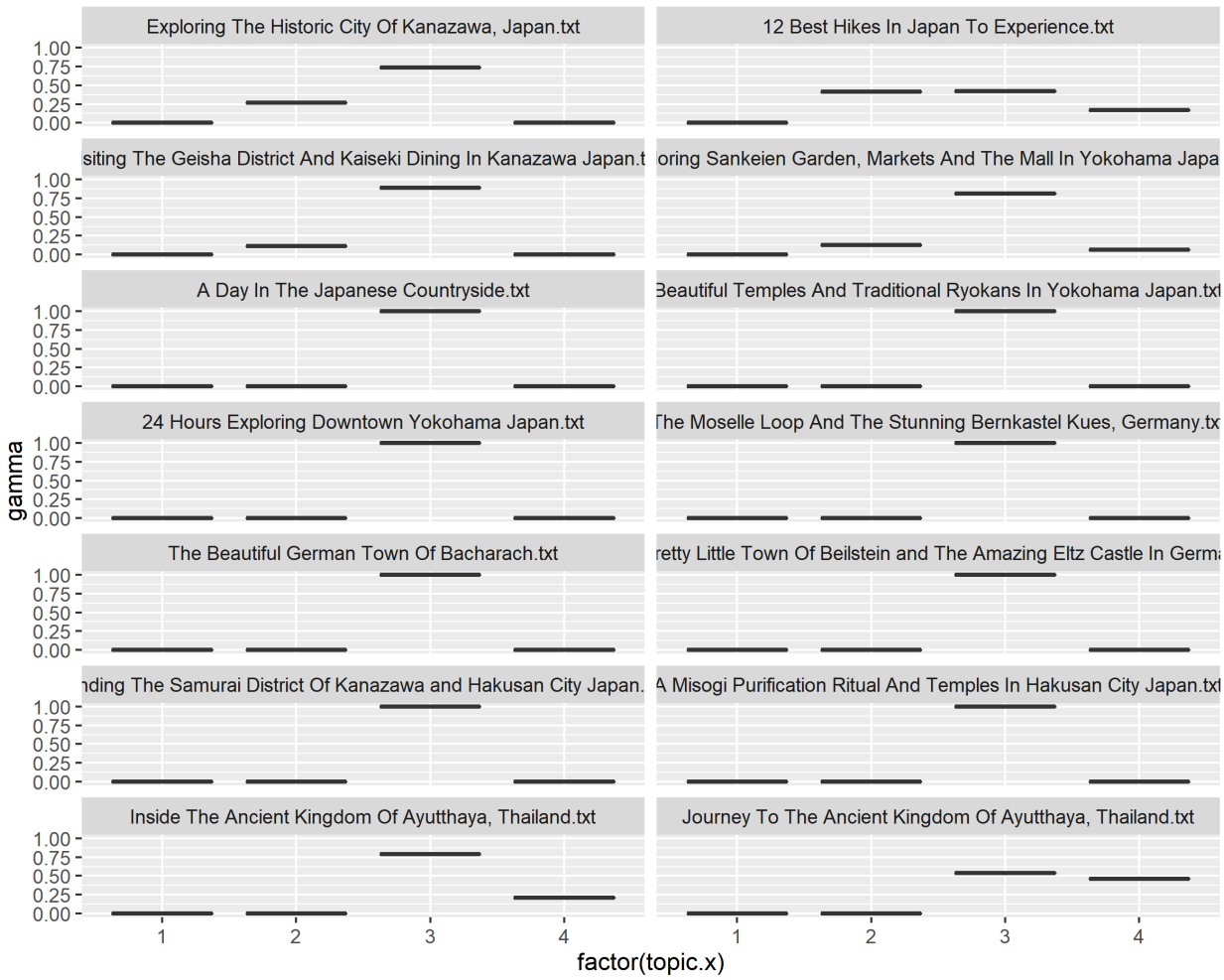

Figure 5.17: Probability of Journal categorized as Topic 3 for handluggageonly

Topic 3 by Jones:

Jones's journals contain many music festival recommendations, regardless of the location, they are all contained in this topic.

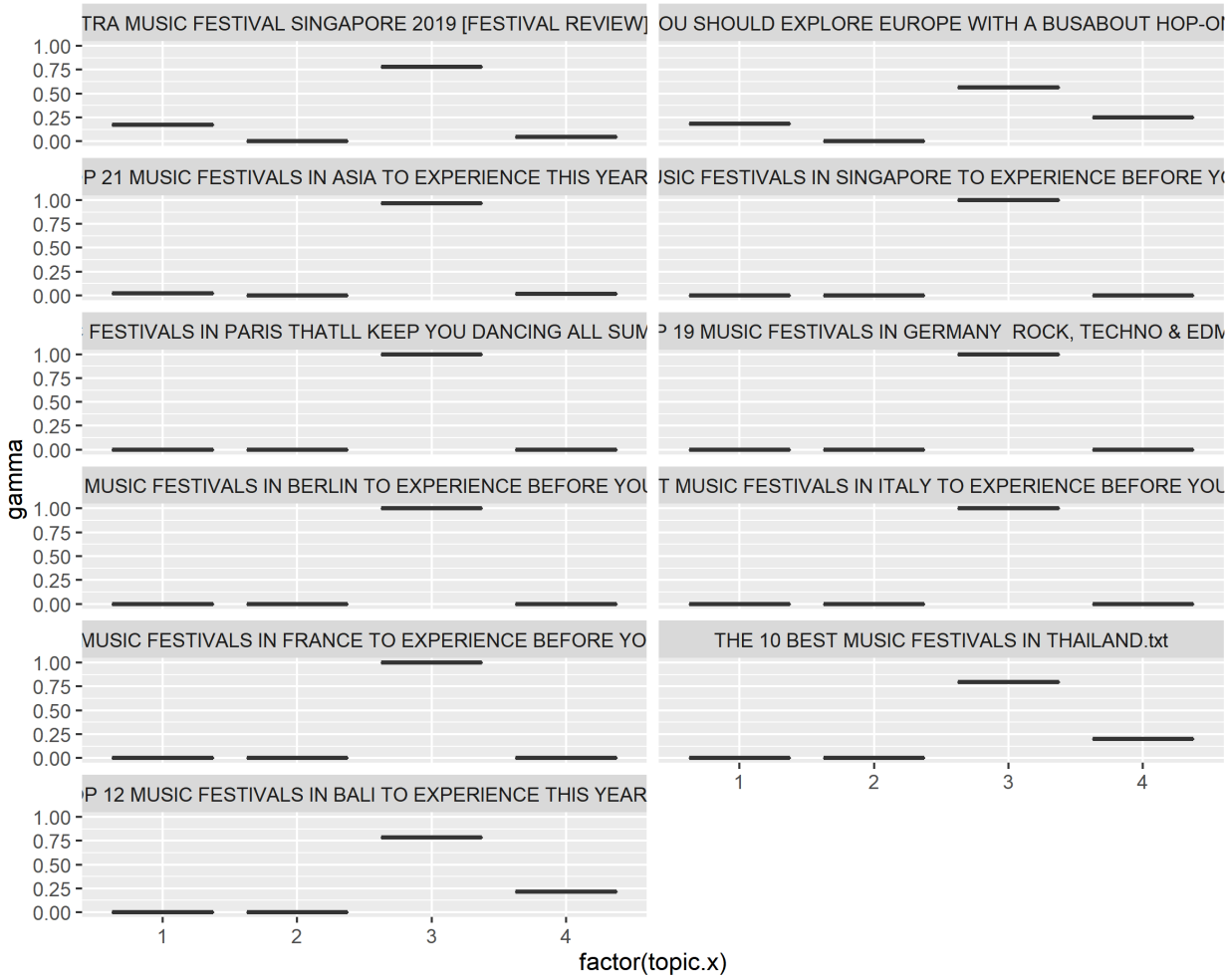

Figure 5.18: Probability of Journal categorized as Topic 3 for Jones

Topic 4 by handluggageonly:

Topic 4 is related to Bangkok, Thailand, exploring the markets where the best Thai food and dishes are found. Many luxurious trips on the Seven Star train journey are also included here where drinks are served in the bars.

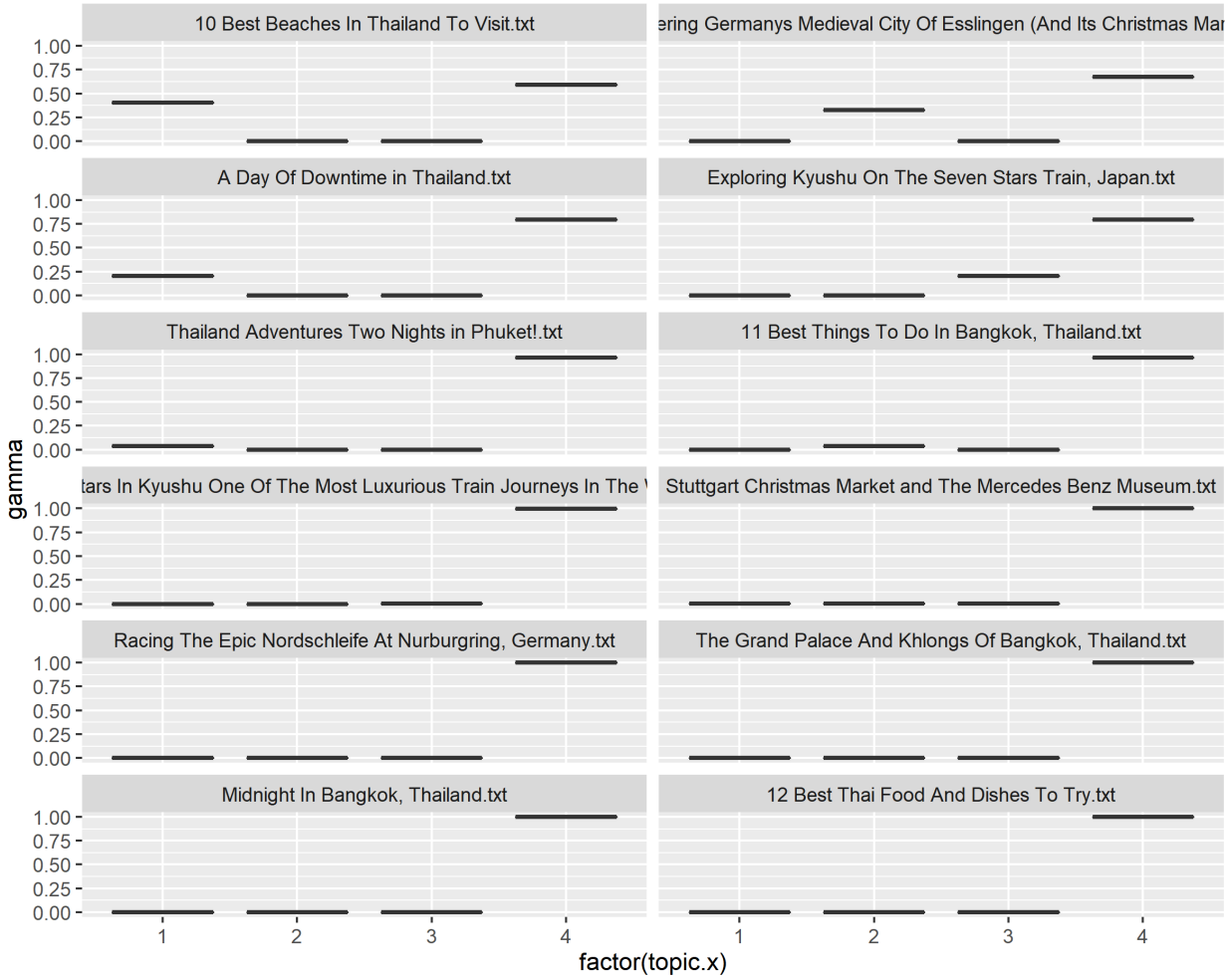

Figure 5.19: Probability of Journal categorized as Topic 4 for handluggageonly

Topic 4 by Jones:

As indicated in the fourth topic, Bangkok and Pai, Thailand as well as Uluwatu, Bali journals lays in the fourth group. "Best places to party in south-east Asia" echoes in the party theme. The model recognized that certain journals also inhibit accommodation suggestions, and the probability are split between topic 1 and 4. Similarly, historical site references journals were weighted between topic 3 and 4.

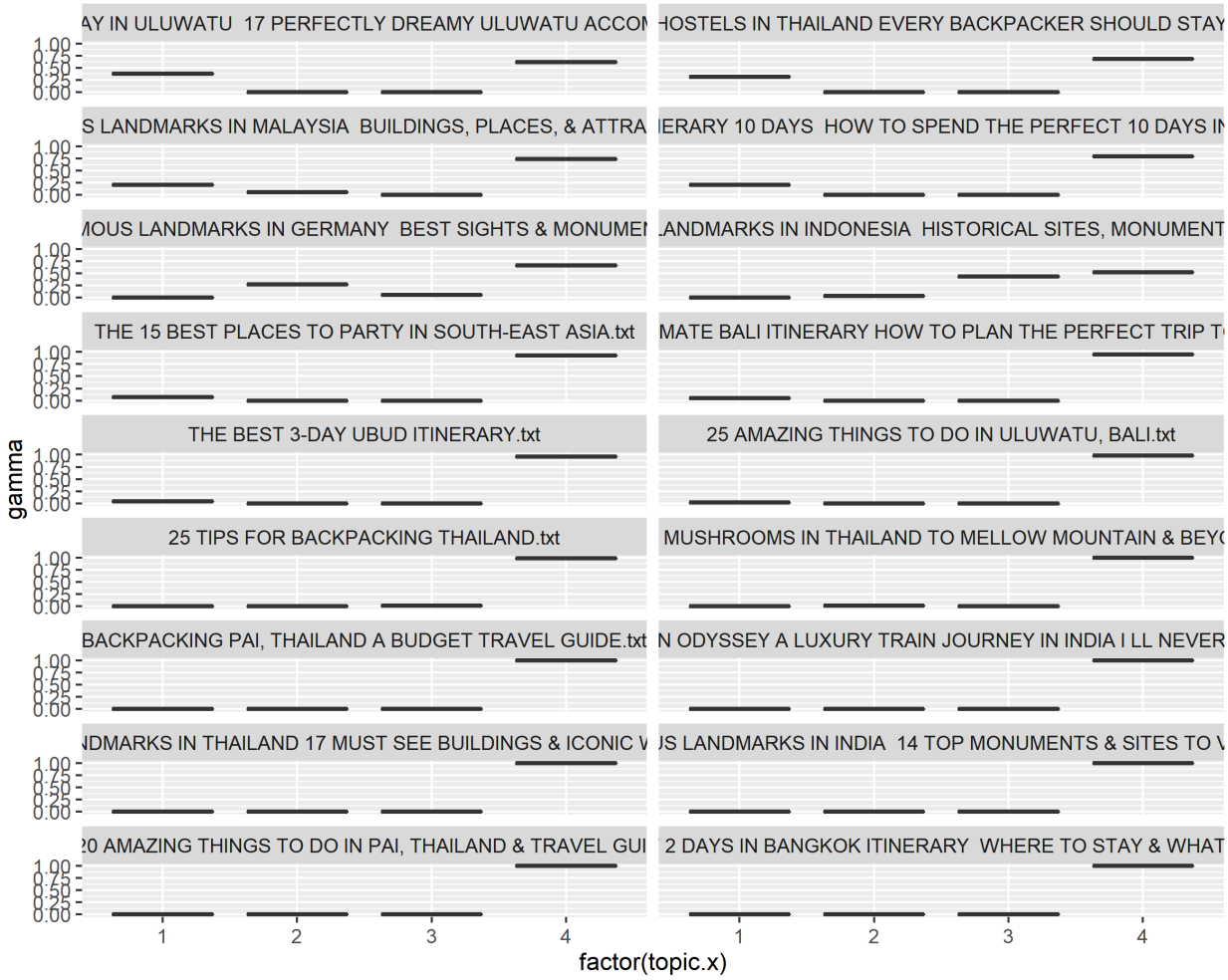

Figure 5.20: Probability of Journal categorized as Topic 4 for Jones

5.1.3.3 Random Forest to predict topics

Random Forest model was used to predict which topic the journals lie in based on the words they have. We also performed tf-tdf based on topic group to reduce number of terms used for the model. First, the data was split to 70% training (93 observations), 30% testing (43 observations). Next, we trained the data using the training dataset to build a Random Forest model. In our training set, the data are evenly spread among the 4 topics as 20% each except topic 2 is around 40%. Our testing set contains 20-30% samples from each topic.

We used the model to categorize the training dataset into 4 topics with 100% accuracy. However, our confusion matrix shows that journals from topic 2 are predicted correctly. Topic 1 journals are correct 9/13 = 69% of the time. Topic 3 and 4 have the lowest accuracy of 28% and 27%. This might be due to hostel related terms in topic 4 being categorized to topic 1. Overall, the correction rate is: (9+12+2+3)/43= 0.60.

Actual/Predit	Topic 1	Topic 2	Topic 3	Topic 4
Topic 1	9	4		
Topic 2		12		
Topic 3		5	2	
Topic 4	4	4		3

Table 5.1: Confusion matrix for topic prediction

Topic 3 associated to temples and music festivals, there were very high accuracy.

As an example, for the mis-classified topic 4 items, the below 1 to 3 and 5th journals are mis-categorized as topic 1; and journals 4 and 6 to 8 are mis-categorized as topic 2. Since the first topic is highly related to island and beaches, similarly in Uluwatu, Bali, it's reasonable that the model identifies those journals as topic 1. The fifth journal also have housing reviews, therefore, it's also highly associated with terms in topic 1.

[1] "THE ULTIMATE BALI ITINERARY HOW TO PLAN THE PERFECT TRIP TO BALI.txt"

[2] "25 AMAZING THINGS TO DO IN ULUWATU, BALI.txt"

[3] "THE BEST 3-DAY UBUD ITINERARY.txt"

[4] "Exploring Kyushu On The Seven Stars Train, Japan.txt"

[5] "WHERE TO STAY IN ULUWATU 17 PERFECTLY DREAMY ULUWATU ACCOMMODATION.txt"

[6] "Discovering Germanys Medieval City Of Esslingen (And Its Christmas Market).txt"

[7] "17 FAMOUS LANDMARKS IN INDONESIA HISTORICAL SITES, MONUMENTS & MORE.txt"

[8] "MAGIC MUSHROOMS IN THAILAND TO MELLOW MOUNTAIN & BEYOND.txt"

5.1.3.4 Unique words in each Topic

Tf-idf analysis was performed on the words in each topic to identify unique terms and further understand how we can differentiate the topics.

To explore further to see why some terms were mentioned, topic-based tf-idf was used to retain the unique terms. During collection of journals, pictures were replaced as section titles of the images, causing certain title to repeat numerous times within one article, which also boosted the if-idf rating for the terms. We could use this collection of names as a starting point for travel sites recommendations.

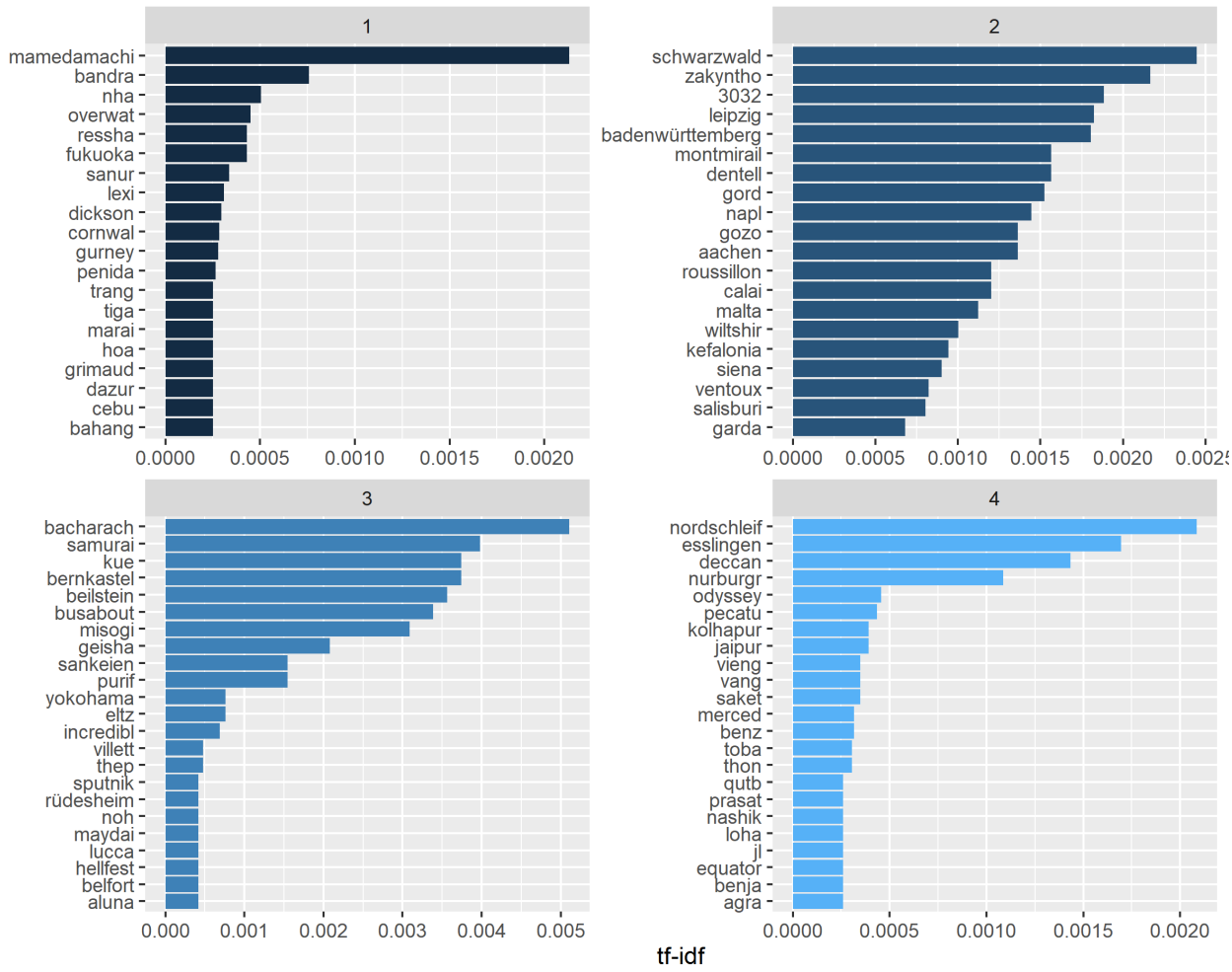

Figure 5.21: Unique words in each Topic

Topic 1:

Mamedamachi: is a location name in Japan.

Bandra: is in India in an Airbnb recommendation article.

Nha Trang: is in Vietnam

Overwater: Overwater resorts and villages in Malaysia

Topic 2:

Schwarzwald, Baden-Wurttemberg: is in Germany

Zakynthos, Greece

Topic 3: German Town Of Bacharach

Samurai District Of Kanazawa and Hakusan City – Japan

Topic 4:

Nordschleife in Germany

5.1.3.5 Sentiment by Topic

To find words associated with experience, using sentiment tables could help filter out location names and retaining adjectives or verbs with sentiments. We will do a similar analysis as sentiment of bloggers, but by topic. Since many words are associated with multiple sentiments, to avoid reviewing the same words among different sentiments, we assigned each word to only one sentiment. Next, we plotted the words by sentiment for each topic as follows.

As shown below, after filtering out duplicated sentiments, sentiments such as anger, disgust, joy, surprise have very few words remaining and common terms across all four topics. Given most of the places are visited in warm weather, hot is a very common "anger" word.

Topic 1 focuses on trip planning recommendations, therefore, words such as cheap, plan, spent, as well as housing features such as pool, breakfast, option, sunset, quiet, clean, spa, honeymoon and being in neighborhood areas are most popular.

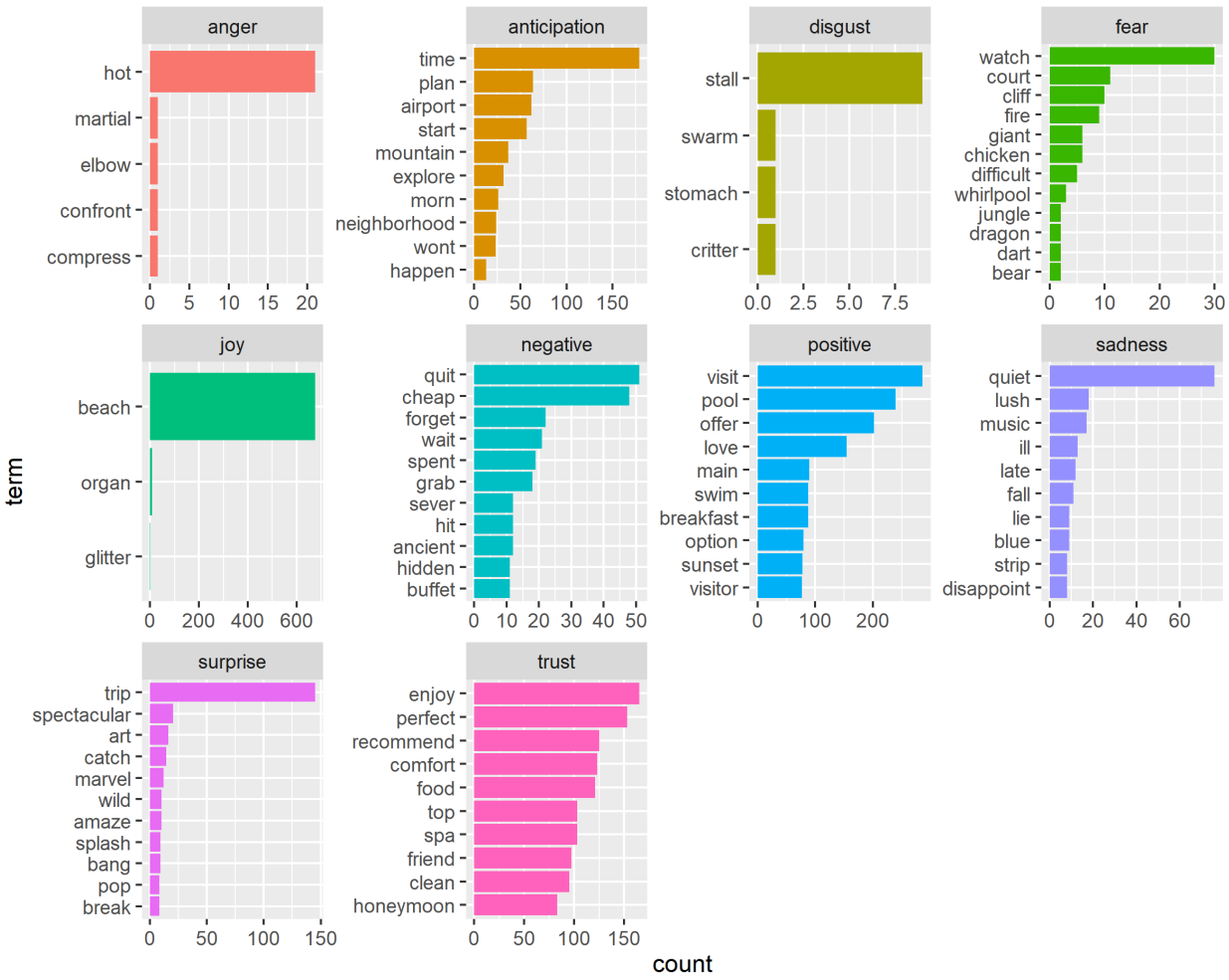

Figure 5.22: Topic 1 Top words for each Sentiment

Topic 2 focuses on exploring the cities, therefore, features such as food, church, mountain, hidden gems in the area, tomb, ruin, war, stone, tower, dragon related to ancient artifacts.

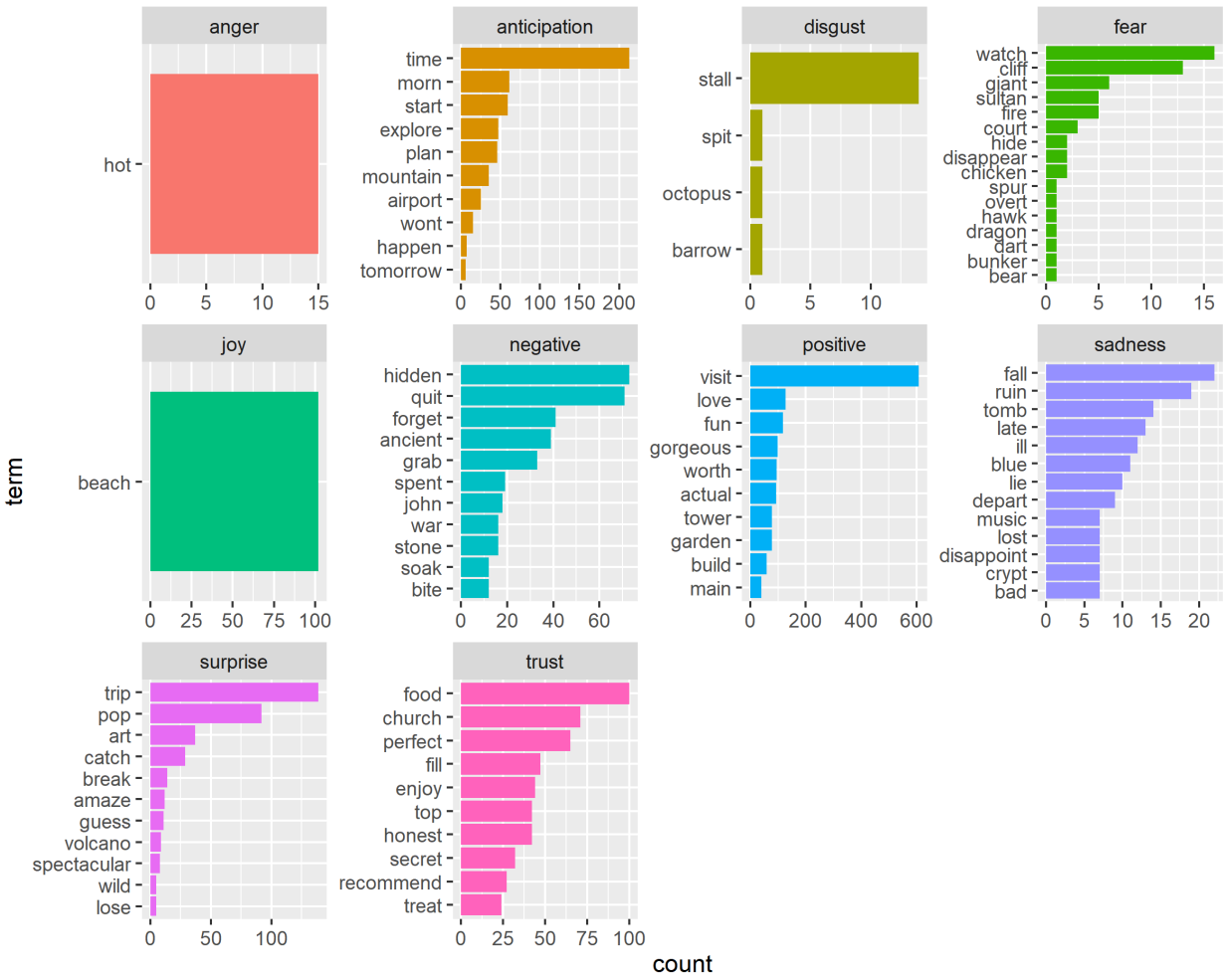

Figure 5.23: Topic 2 Top words for each Sentiment

Topic 3 concentrates on music festival, Germany, and Japanese temples. We can see music, festival, rock occurring 400 times; samurai as a Japanese warrior and temple statue, garden also occurring. Sentiments are: amaze, wild, love, gorgeous.

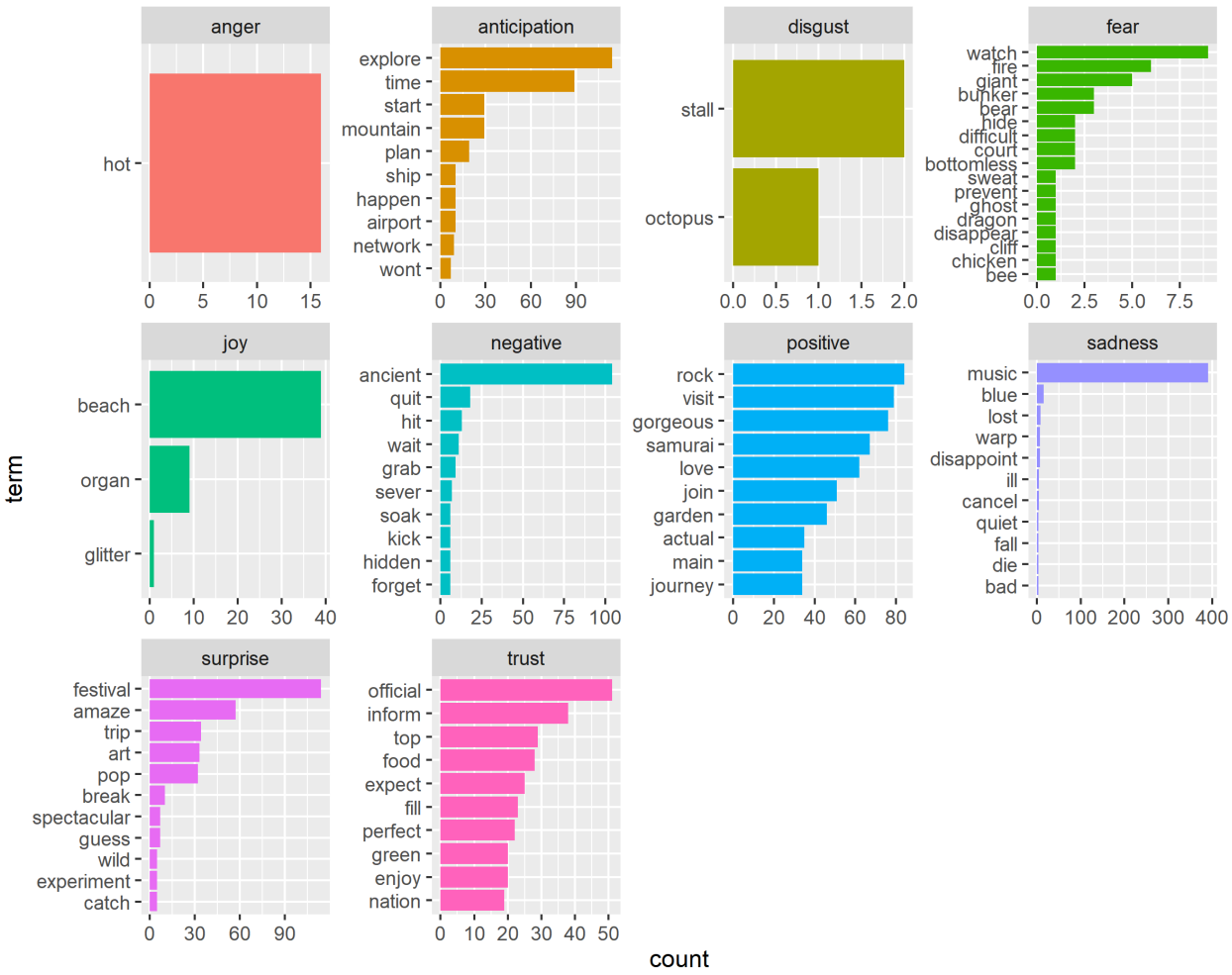

Figure 5.24: Topic 3 Top words for each Sentiment

Topic 4 is connected to Bangkok, Thailand, Uluwatu Bali, bars, markets and parties. Some association are: cheap, hidden, fee, sunset, famous, epic, pool, fall, food.

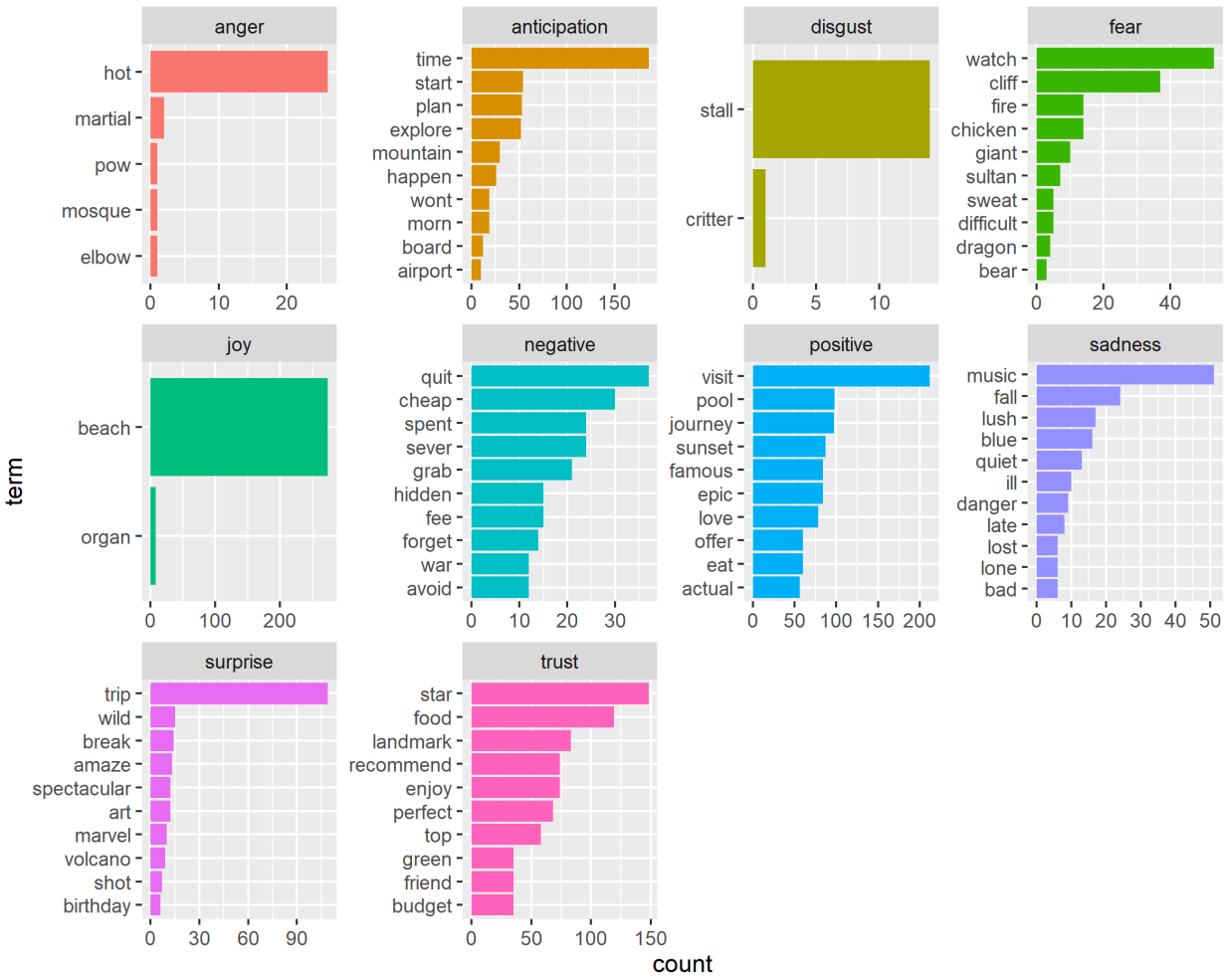

Figure 5.25: Topic 4 Top words for each Sentiment

5.1.3.6 Analyze journals by Region

To further study the main characteristics between different travel sights, we can analyze the contents by region and identify local specialty and explore how the tourist highlights are tailored towards specific visitor interest. Which regions retain historically significant buildings, museums with ancient artifacts, paintings, or nature parks, hot springs. We can associate travel sights by sentiment and common terms used for different region. Are certain areas/themes/regions pricier, busier than others?

We would first separate the journals by region, then perform analysis within each region.

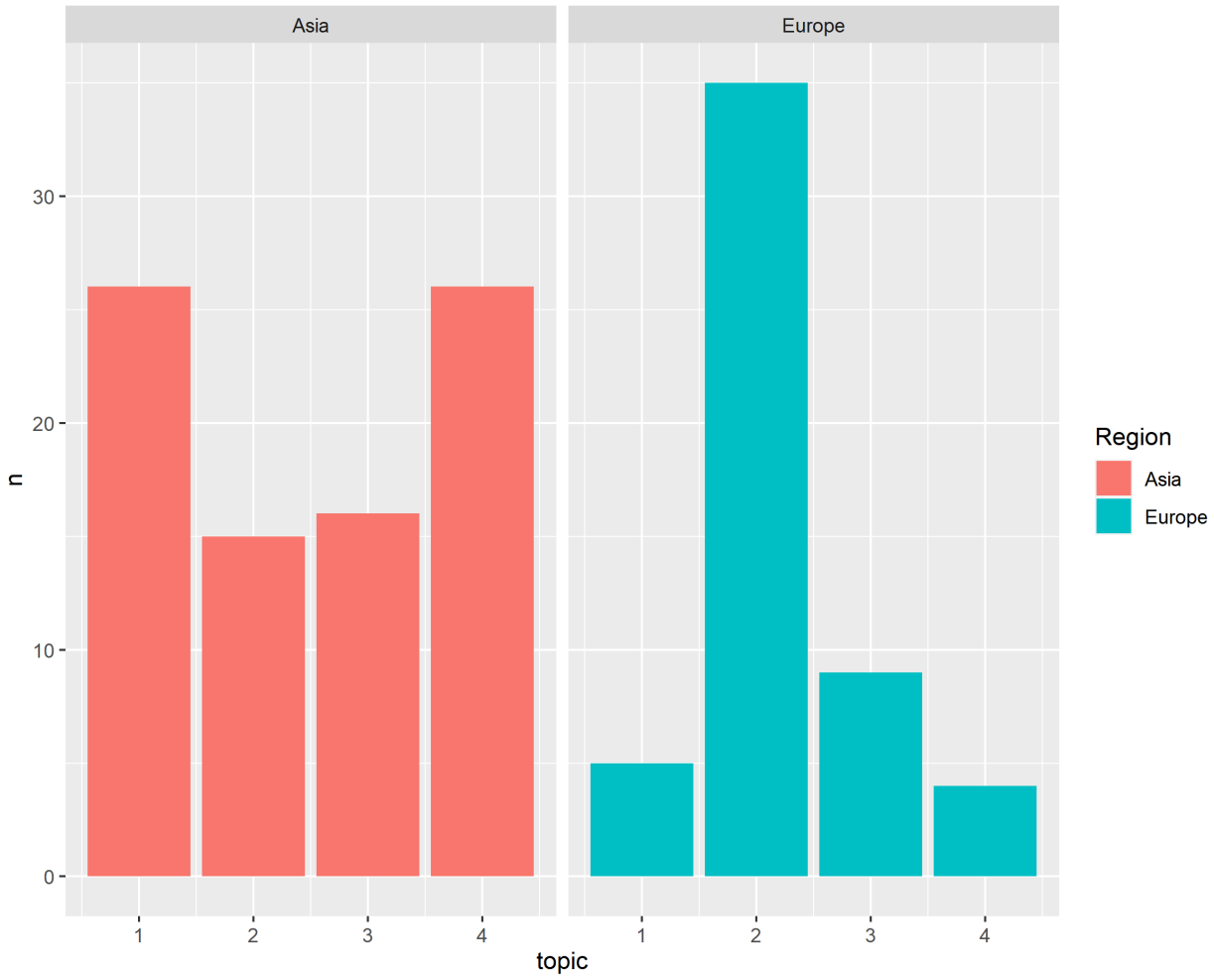

Figure 5.26: Barplot of Journal Topics by Region

We want to connect the topics with regions and identify potential patterns. Asia journals are mainly Topic 1 (island, beach, housing reviews and Malaysia) and Topic 4 (Bangkok, Bali, beaches, market, party, drinking), whereas European journals are mainly topic 2 (cities in France, Italy, Singapore, Rome, Germany, and nature). This aligns with the key locations we associate with each topic. Topic 2 locations are mainly in Europe while the others are spread within Asia.

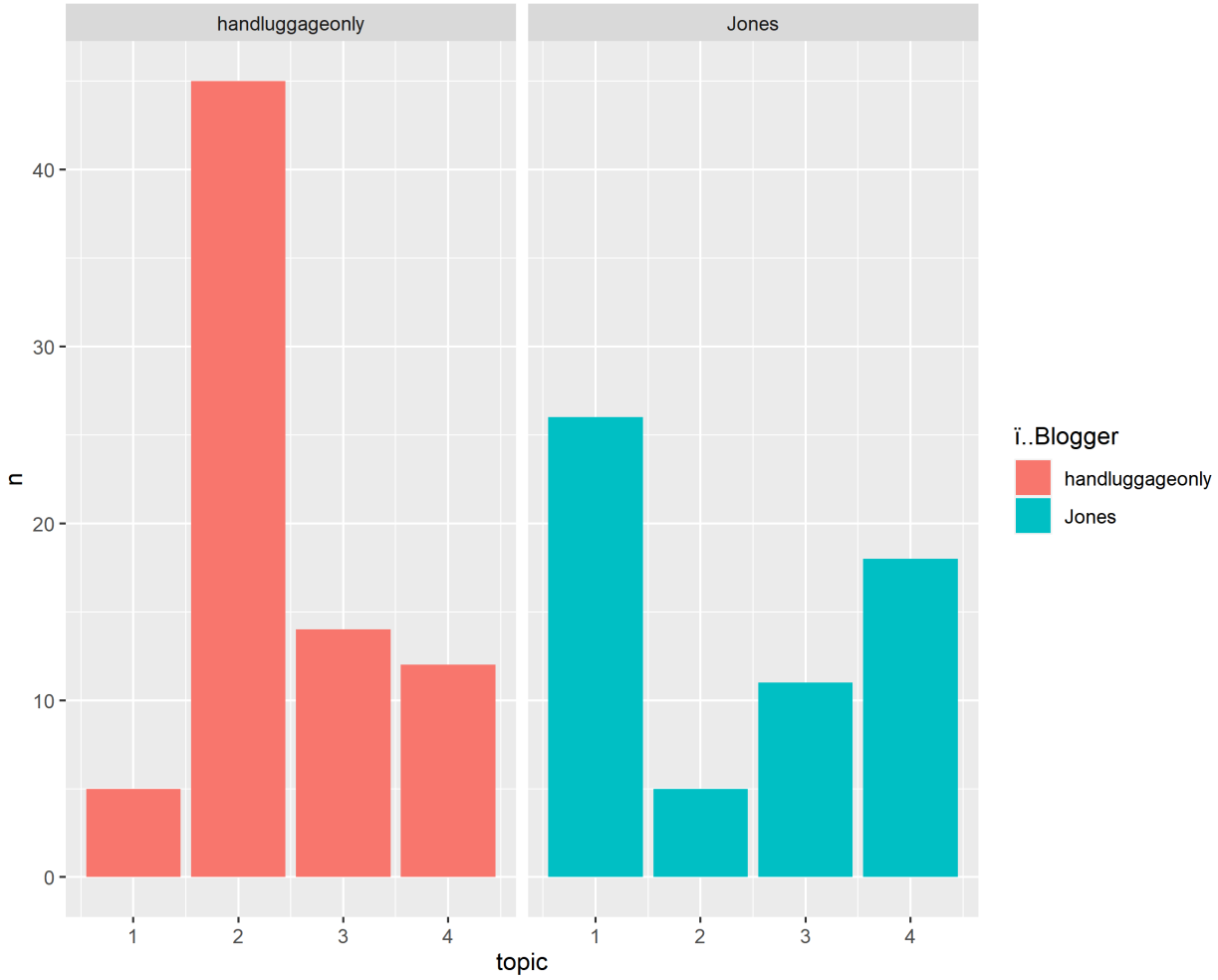

Figure 5.27: Barplot of Journal Topics by blogger

The styles/topics for both bloggers are unique. Handluggageonly's journals mainly topic 2, with exploration of nature in European counties and Singapore. Jones's journals are mainly topic 1 with housing recommendations and Uluwatu Bali beaches

Next, we review the topics by each country. We can see that one topic dominates in some countries, such as: France, Italy, Singapore is mainly topic 2; Japan mainly in topic 3, Malaysia mainly in topic 1; Thailand and Indonesia/Bali is a split between topic 1 and 4 since both topics are closely related to the islands. Therefore, we can closely associate the topics with the locations in our future analysis.

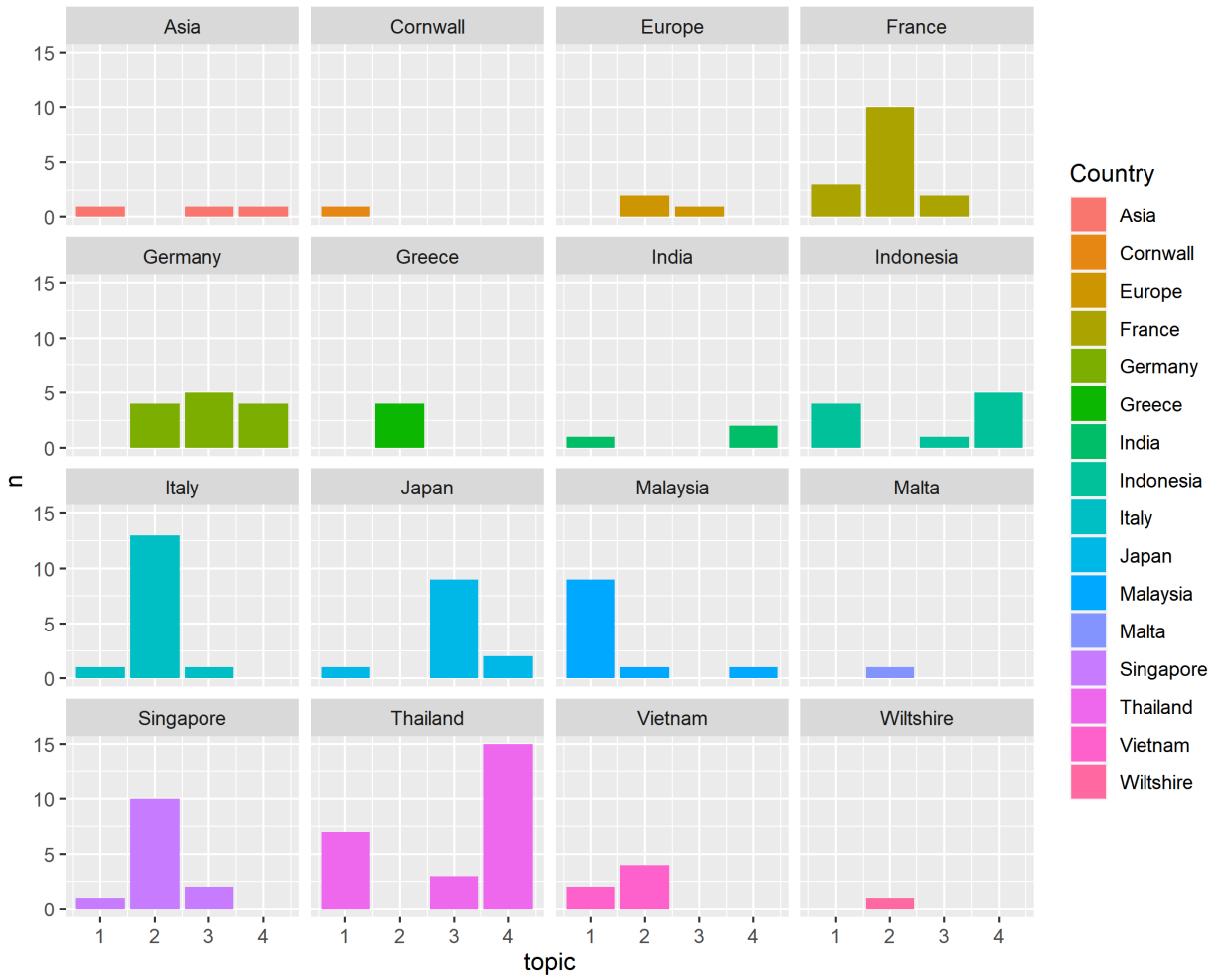

Figure 5.28: Barplot of Journal Topics by Country

5.2 BERTopic Model Analysis

5.2.1 Topic Groups

Among each topic groups, class-based tf-idf score based on uniqueness and frequency of words were used to select the most representative words, featuring the topic of the group. Table 1 shows the frequency of sentences assigned to the top 5 topics, where topic -1 includes sentences that does not belong in any of the 186 clusters created by the model.

After summarizing the topics by the themes, we arrived at 96 topics with the major themes as island/beach, Airbnb rental, festival, Thailand, and Malaysia accounting for around 43% of the data. These topics align with the findings in LDA topic modeling, with the benefit of more granular and easier to interpret groupings, whereas LDA forces certain themes to be grouped together due to restriction of total number of topics.

To visualize the clusters, BERTopic embeddings were reduced to 2 dimensions using UMAP and plotted by group. Each group consists of the original topic assignments in different colors. If the clusters in different colors are closely together, it confirms the assumption that the topics share common traits. The darkness of the colors associated with volume of data.

For example, Figure 5.29 is a 2D plot of the sentences in group island/beach. The 20 original topics listed in Table 5.2, are consolidated into one theme based on keywords "beach" and "island". In the middle of the plot, red dots cluster together with darker red representing higher density of data. On the right side of the red cluster, some violet points overlap with the light green dots indicating similarities in the two subgroups. The separation of the color clusters suggests differences in the groups even though the main theme is similar.

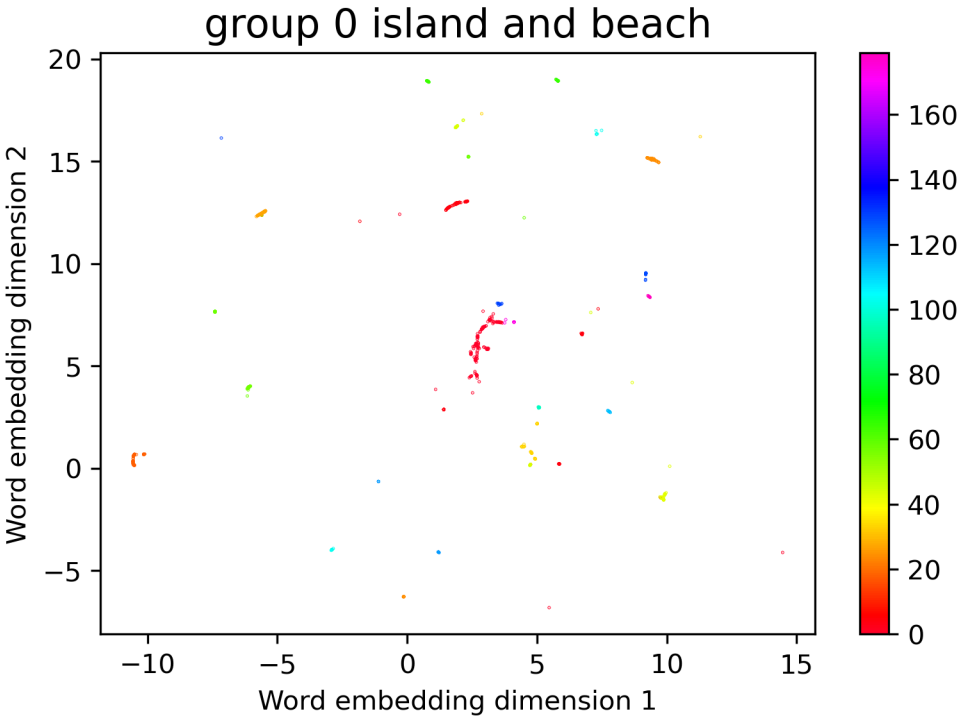

Figure 5.29: Theme Island/Beach sentence embeddings visualized in 2-dimensional space

Topic Num	Keyword 1	Keyword 2	Keyword 3	Keyword 4
4	uluwatu	bali	padang	beach
57	beach	beach 13	bai	beach 11
129	spa	resort	offers	sandy beach
179	phuket	phuket phuket	pronounced like	beach phuket
27	vietnam	beaches vietnam	beach	beaches
64	cornwall	beaches cornwall	best beaches cornwall	best beaches
113	philippines	cebu	islands	best islands
128	beaches thailand	best beaches thailand	best beaches	beaches
24	gili	trawangan	gili trawangan	gili islands
0	beach	boat	island	snorkeling
33	malaysia	redang	redang island	island
42	islands	italy	naples	island
44	kapas	pulau kapas	pulau	kapas island
65	phu	phu quoc	quoc	quoc island
97	sipadan	dive	sipadan island	kapalai
103	singapore	sentosa	beaches	sentosa island
118	island	lazarus	lazarus island	island lazarus island
174	gem island	gem island resort	gem	island resort spa
18	perhentian	perhentian islands	islands	kecil
58	kefalonia	crete	greek islands	greek

Table 5.2: Topics Keywords in "Island/Beach" Theme

Example shown below in Figure 5.30, Figure 5.31,Figure 5.32,Figure 5.33 are groups which were successfully cluster together includes festival, and groupings by country such as Koh Tao in Thailand, Singapore, and Bali.

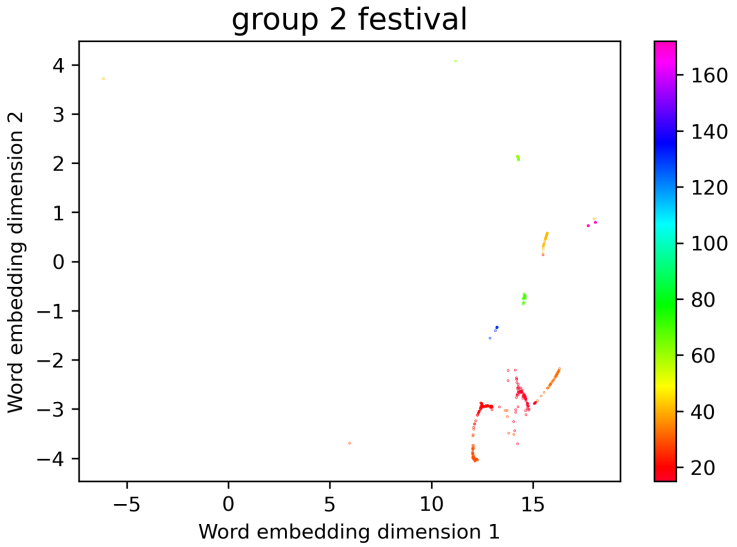

Figure 5.30: Festival Themes embeddings visualized in 2-dimensional space

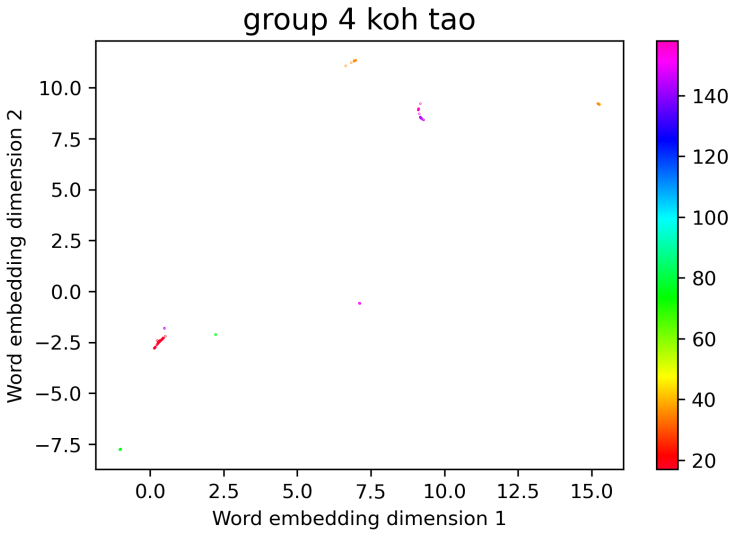

Figure 5.31: Koh Tao Themes embeddings visualized in 2-dimensional space

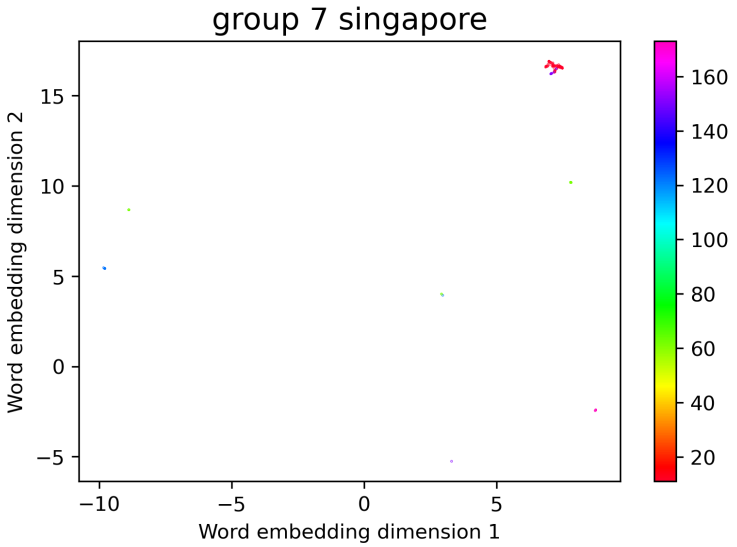

Figure 5.32: Singapore Themes embeddings visualized in 2-dimensional space

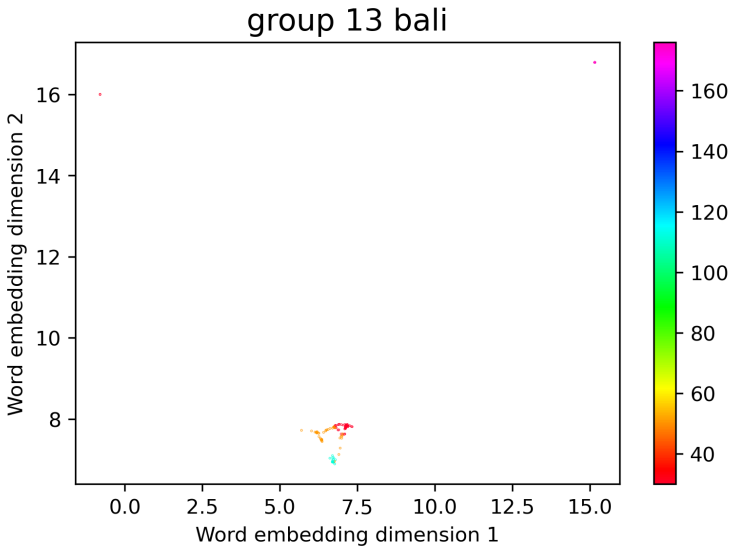

Figure 5.33: Bali Themes embeddings visualized in 2-dimensional space

5.2.2 Recommendation

The purpose of this paper is to provide travel recommendations based on user interests. Whether it is popular beaches, churches, historical sites, festivals, or food, the travel guide should provide relevant information based on the travel journals. Associating features such as themes/regions/cost with topics is key to creating tags that represent these experiences so that potential travelers can personalize their preference to select the cities and events without much hassle. For general purposes, a new audience can understand the type of experience and life style associated with a given event/region much simpler. We can also consolidate the various sources of one location for browsers so they can view multiple perspectives of a location without reading each source individually.

First, the guide will identify top 5 topics related to the keyword, then the most relevant countries based on sentences count in the topics. Highlights are used to provide travel guidance.

Example 1: Music:

The top 5 topics associated with music are listed in Table 5.3. Figure 5.34 shows most of the sentences relating to music came from travel journals in France and Italy and Asia. This result aligns with the topics 20 as music festival in Paris, topic 32 as Italy Music festivals and topic 70 as Asia music festivals, where topic 43 electronic genres are not specific to a country. The category of Asia was created due to the broader content in certain blog journals.

Topic Num	Keyword 1	Keyword 2	Keyword 3	Keyword 4
20	music	paris	festival	festivals
32	italy	festival	music	festivals
70	festivals	festivals asia	music festivals	asia
43	genre	genre electronic	electronic	techno

Table 5.3: Topics Keywords in "Music" Theme

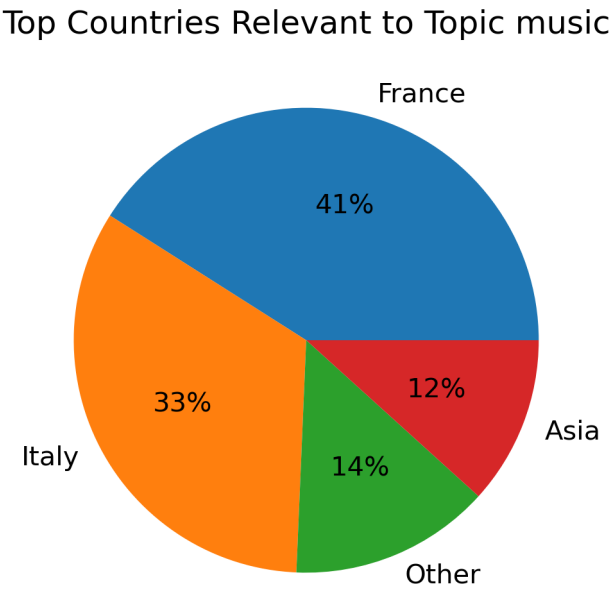

Figure 5.34: Country Revelant to Topic Music

Highlights of these topics:

Some documents selected show great reference and interesting reviews regarding music festivals in each country. Tourist attractions includes City of Lights, Paris music festival, rock summer music event in Rome, and Techno summer outdoor party in Germany.

France Highlights:

”Bonjour everyone! :ooking for the best music festivals in Paris!? Well, youve come to the right place! Paris is truly one of the worlds greatest cities, and the French capital sure knows how to put on a show to outdo the rest of us. ” ”When youre done eating croissants, taking in the many famous sights and falling in love with the City of Lights, the Paris music festival scene is your next port of call. ” ”No matter what genre youre into, you can bet your ass theres a music festival in Paris perfect for you. And with many held in stunning and unique venues (Disneyland Paris, anyone?), this is one list you have to make time for. ” ”So whether youre living in Paris, or just a tourist looking for their next international festival, this list has got you covered to make your summer trip much more memorable! ” ”œSolidays is a French music festival committed to the fight against AIDS and campaigning for equality and youth engagement, doing so through a showcasing over 80 concerts across its June weekend. Music fans can expect a diverse mix of performances “ think David Guetta, Mura Masa and Shaka Ponk “ alongside a lineup of undiscovered gems waiting to become your next favourite artist. Good days ahead for solidarity in music! ” ”œWhen it comes to considering its ecological footprint, few festivals compete with Paris We Love Green Festival. As its name suggests, the festival emphasises its dedication to its environmental responsibility, from its innovative take on eco-toilets to its scarce use of plastic to discussions and workshops on sustainable living. In addition to its progressive ethics, We Love Green reliably boasts lineups impressive enough to match its more well known European peers; previous years have seen the likes of Justice, Migos, Solange, LCD Soundsystem, Foals and Lorde. This combination of alternative thinking and billing makes We Love Green one of the highlights of the French festival scene. ” ”œHeld annually on the Plage de Torcy in the east side of Paris, Marvellous Island offers a festival experience like no other. Bringing a consistently impressive lineup of techno and house each summer, the festivals waterside location is the perfect escape from the hustle and bustle of the city. ” ”œYou know summer is on its way when Villette Sonique arrives at the end of May and the Parc de la Villette is transformed into a festival-lovers paradise. If youre into rock, electronic and experimental

sounds then youre in for a treat, with names like Marquis de Sade, Mogwai, Jon Hopkins and John Maus taking to the stage. What makes this fest even more insane is that during the day its totally free! Head to this huge 5-day extravaganza on the edge of Paris city centre to listen to epic music and discover new artists you cant believe you never knew "

Italy Highlights:

"Are you looking for the best music festivals in Italy in 2022!? Well, youve come to the right place! I used to live in Italy, and it will forever remain one of my favorite destinations in the world for travel (and now festivals)! " "In addition to literally everything else, the country shaped like a boot has also given us the gift of great music. While theres always a music festival in Italy to hit up year-round, summer is when this beautiful country really comes to life, buzzing with top performances and happy crowds. " "My list of the best Italian music festivals in 2022 will take you from the northern cities of Milan and Venice, to the historical hotspots of Rome and Florence and then deep down to the majestic coastlines of Sardinia and the Amalfi Coast. And back again. " "Rock in Roma is a major summer event dedicated to rock music, taking place from in June and July. The line-up of the 2018 edition included big-name rock, heavy metal and rap acts. The festival is based at the Ippodromo delle Capannelle venue, with several concerts taking place in other locations including Circo Massimo, Auditorium Parco della Musica and Ostia Antica. " "Post malone Live in Italy - Rock in Roma 10-7-2018 " "Firenze Rock Festival Italy 2022 " "Firenze Rocks is with no doubt one of the biggest music events of the summer in Tuscany and in Italy. Next June 2022, for the fourth time in a row, the Visarno Arena of Florence will host this great festival of rock music that in the past editions attracted thousands of fans coming from all over the country. If you are living in Italy or traveling to Tuscany next June, and you have a passion for rock music and big outdoor festivals, I would not miss Firenze Rocks! "

Germany Highlights:

"Techno is the name of the game here. While the regular spot for hard techno and EDM is the underground, FP takes the genre to the summer outdoors. It only lasts one day, so

theres no excuse to simply not make the most of the greatest techno of the summer. " "You wont be lonely, as you enjoy the range of minimal to pure techno along with 20 000 of your closest friends, of course. " "Set outdoors and essentially in the woods, the idea here is simply to engage with artists that sit outside the mainstream. That means that the appeal isnt the most famous artists or the most downloaded. Its about an unusual and passionate music lover. "

Example 2: Historical sites:

For those interested in historical sites and landmarks, the model selected 5 topics as shown in Table 5.4. Majority of the attractions resigns in India, Thailand and Indonesia with fewer in Germany, France, and Greece. The historical theme was being transferred into key features such as: landmark, palace, monument, museum, history.

Topic Num	Keyword 1	Keyword 2	Keyword 3	Keyword 4
9	india	monument	mosque	landmark
166	indonesia	trowulan	jakarta	indonesia jakarta indonesia
123	museum	exhibitions	inside	eva
88	grand palace	palace	grand	ayutthaya
180	history	palace	youre palace just	jantar mantar observatory

Table 5.4: Topics Keywords in "Historical sites" Theme

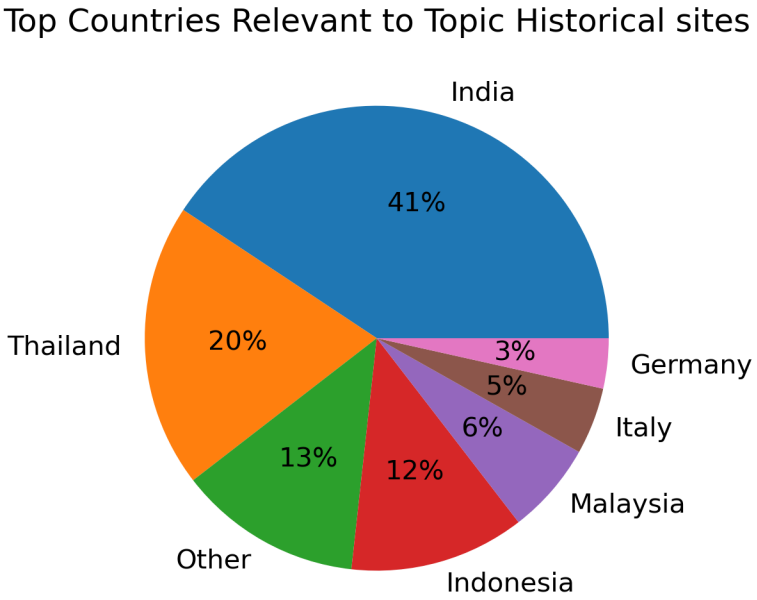

Figure 5.35: Country Revelant to Topic Historical sites

Highlights of these topics:

Indian landmark:

Taj Mahal, Agra Fort, Jahangir Palace

Indonesia:

National Monument, Trowulan Archeological sites

Italy:

Palace of Fears museum, Palazzo Lanfranchi museum, Avon Museum, Grappa Museum

India Highlights:

The Taj Mahal is an Indian landmark that needs very little introduction. This iconic landmark is situated on the right bank of the sacred Yamuna River. This stunning palace in Agra is an architectural marvel masterpiece made entirely out of white marble.

This impressive structure was built by Mughal Emperor Shah Jahan for his favorite wife,

Mumtaz Mahal, and was completed in 1648. The 42-acre (17-hectare) complex houses the tomb of Shah Jahan and his wife.

The Taj is one of the best examples of the Mughal Empires influence and became a World Heritage Site in 1983. The touching story of this iconic landmark continues to attract visitors from around the globe and it remains one of the most important sites in India.

The Taj Mahal - Landmarks in India

2) AGRA FORT

Location: Agra

The imposing Agra Fort is a sister landmark to the Taj Mahal. This famous monument served as the main residence of prominent figures in the Mughal Dynasty until the mid-1600s. It was inscribed as a World Heritage Site along with the Taj in 1983.

This 94-acre (38-hectare) complex has a semicircular formation and boasts beautiful red sandstone. The complex features several buildings, including two mosques and two palace complexes. One of the most notable palaces is the Jahangir Palace which was also constructed by Shah Jahan. The fort houses a Justice Chamber and residential complexes. One of Agras most notable sites is the intricately designed Ghaznin Gate.

3) INDIA GATE

Standing at a staggering 138-feet (42-meters) tall, the India Gate is one of the most recognizable landmarks in India. This monumental structure is a stunning sandstone arch located at the end of the Rajpath ceremonial boulevard in New Delhi.

The India Gate is a war memorial dedicated to the brave Indian soldiers who died during the Anglo-Afghan War and World War II. The arch was designed to resemble the Arc de Triomphe in Paris.

The arch has the names of thousands of soldiers inscribed on its walls. This historical monument is surrounded by manicured lawns and is a popular picnic and relaxation spot amongst locals and visitors.

Indonesia Highlights:

The country of Indonesia has an extensive history of migration, Dutch rule, Japanese occupation and finally, independence. Which brings us to one of the most treasured historical sites in Indonesia. The National Monument symbolises Indonesian independence and freedom.

The structure stands over 430 feet tall in the center of Merdeka Square, Central Jakarta. The monument is also known as the Monas and is topped with a golden flame that represents Indonesias spirit.

The Trowulan Archeological site was declared a UNESCO world heritage site one of many from Indonesia! The site is the last discoverable city remains of the Indonesian Hindu-Buddha classical age. The Trowulan site is situated in the Mojokerto Regency in East Java, which can be entered by bus.

Italy Highlights:

Once here, be sure to stop off at the Palace of Fears; a museum that houses thousands of exhibits with a really cool contemporary art collection. You can spend hours inside.

Oh, and afterwards, pop onto the roofed area (where you can freely walk) and spot the crypts and tombs.

If the weather takes a turn for the worst, pop inside to visit the museum at Palazzo Lanfranchi. Its one of the most popular art galleries and museums in the city and houses both old and new exhibitions to explore.

Example 3: Fashion:

The top 5 topics associated with fashion in Table 5.5 are: interior designs, museums and exhibitions, gorgeous things, social media and music gerne. Figure 5.36 shows fashion is related to Italy, Japan, France. The themes interpret fashion in various ways.

Topic Num	Keyword 1	Keyword 2	Keyword 3	Keyword 4
106	instagram	police	really gorgeous	lights
164	interior	floor	features	ornate
43	genre	genre electronic	electronic	techno
123	museum	exhibitions	inside	eva

Table 5.5: Topics Keywords in "Fashion" Theme

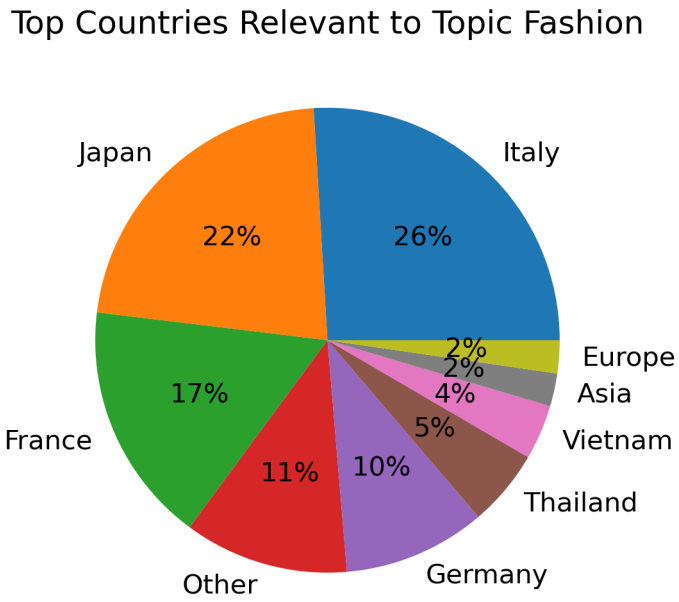

Figure 5.36: Country Revelant to Topic Fashion

Highlights of these topics:

Topic Interior Design Highlights:

Once here, be sure to keep your eyes peeled for the incredible mosaics and look down at the floors. The marble floor took hundreds of years to create and is a must-see.

Nowadays, you can head inside and see the gorgeous ceilings of the oratory.

After paying a small entrance fee, pop inside to see the anatomical theatre, Stabat Mater Lecture Hall and all the historic wall decorations that are so specific to the city. It's totally stunning.

Topic Museum Exhibits Highlights:

Once here, be sure to stop off at the Palace of Fears; a museum that houses thousands of exhibits with a really cool contemporary art collection. You can spend hours inside.

If the weather takes a turn for the worst, pop inside to visit the museum at Palazzo Lanfranchi. It's one of the most popular art galleries and museums in the city and houses both old and new exhibitions to explore.

Still, to this day, you can see her personal collections inside and even see the beautiful gardens and decorated rooms.

Example 4: Dancing:

Since dancing is related to events and festivals, as shown in Table 5.6, music festival theme is highly associated given the blog journals might only include these as dancing related events. However, it's unknown why parking theme is associated with dancing. One reason could be it being highly related with festivals. Again, France and Italy with many music festival journals are listed top in Figure 5.37.

Topic Num	Keyword 1	Keyword 2	Keyword 3	Keyword 4
21	park	car	easy	tickets
20	music	paris	festival	festivals
32	italy	festival	music	festivals
43	genre	genre electronic	electronic	techno

Table 5.6: Topics Keywords in "Dancing" Theme

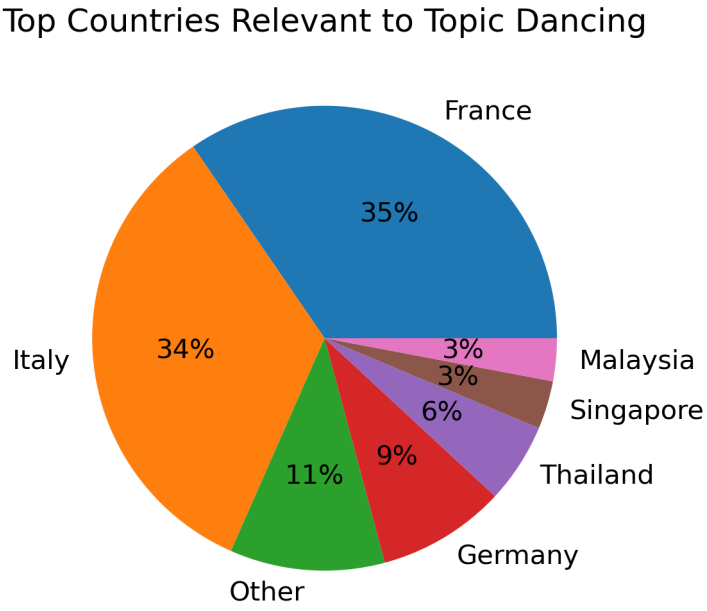

Figure 5.37: Country Revelant to Topic Dancing

Example 5: Dessert:

As a food lover, delicious food is always the drive to explore. Table 5.6 correlately associate topics such as wine, picnic food, pasta, and dessert theme. Figure 5.38 showed an even spread of association of dessert to all the countries.

Topic Num	Keyword 1	Keyword 2	Keyword 3	Keyword 4
162	wine	vineyards	wines	vineyard
66	food	picnic	stalls	food tours
160	italian	pasta	homemade	trattoria
142	menu	dinner	le	restaurant
6	lunch	dessert	menu	pudding

Table 5.7: Topics Keywords in "Dessert" Theme

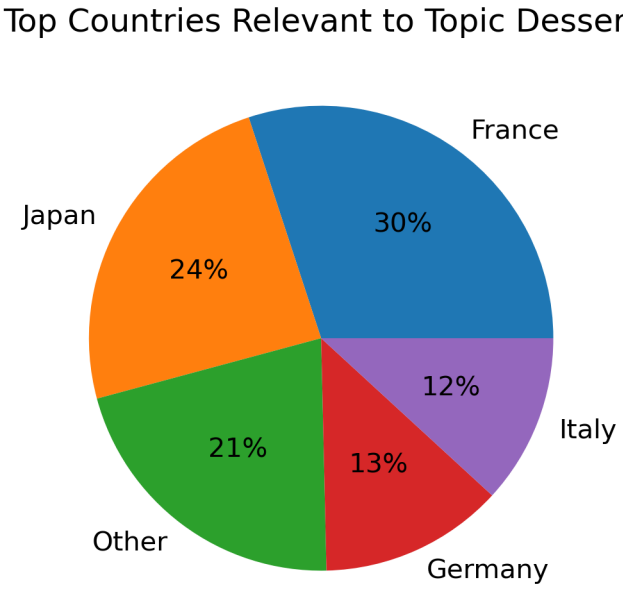

Figure 5.38: Country Revelant to Topic Dessert

CHAPTER 6

Conclusion and Discussions

Journals from two bloggers handluggageonly and Jones were analyzed in search of categorizing travel sights features to provide recommendations for future tour guides or personal trips. First, the bloggers' writing styles were compared based on word frequency and sentimental terms. Jones's journals have many hostel recommendations, money saving tips and music festivals whereas handluggageonly utilizes actions such as bite, explore giving a more personal touch. handluggageonly also gave tips to avoid crowd and busy seasons. Both enjoy food, drinks at the bar, nature, beach, and travelling through airport. Using sentimental terms, both bloggers created a trustful, joyful, exciting experience for their readers.

Next, LDA represents countries and popular themes such as rental recommendations, music festivals into 4 topics. The Random Forest model predicting journals to topics achieved an accuracy of 60%. Topic 3 and 4 struggles at 28% accuracy due to overlap of key features between topics. Applying topic-base tf-idf retained most unique terms and resulted in names of locations which could be a starting point for guides.

Lastly, BERTopic was used as an advanced machine learning technique to associate topics/themes to keywords of interest to complete our goal of creating travel recommendations. The model was able to identify related topics with words based on context. It was able to associate "cocktail" input with topics "bar, drinks" and "wine" and able to capture key groups related to any given keyword. Based on the topics, a travel guide is created with the selected documents and country.

This study showed strong promise in searching for relevant travel comments based on

topics, while interpreting the keyword in various settings. In our previous examples, the model interpretted "fashion" in different ways which could possibly trigger new interest for users. By simply manually searching for keywords within the documents, one might be able to find some sentences containing the word, but only the ones with the exact word specified, while the model is able to interpret the keyword into one or multiple themes and capture text without the requirement of containing pre-determined words as long as the context of the sentences aligns. This way, time and manual efforts can be greatly reduced in identifying useful content for websites or travel guide creation. Travelers can also search through the website using keywords to obtain recommendations of future trips. Imagine typing "hiking" into the website to find great hiking trails or "animal" for such as zoos or whale watching tours.

Some further improvements includes grouping sentences that are close within a journal with the same topic. This can capture a block of text that's discussing a specific topic and provide a more fluent and detailed explanation of the location since it's hard to understand what the object is without context.

Applying classification using BERTopic themes with unseen journals from the two bloggers could be interesting. However, this would require manual verification of the assignment.

Another key attribute to planning a trip is cost. Within the journals from Jones contains pricing information for rental properties. Analyzing the cost along with reviewer comments and features would be a great way to identify key features attractive to visitors. For example, two locations in Thailand might have similar features, but one has pool and breakfast, whereas the other provide convenient location, depending on whether visitors are willing to accept the pricing, we can learn what are important to the travellers and how agents can better price and advertise to increase profit.

To further improve the quality of themes and topics, user demographics can be considered in creating groups. Similar to key attributes for restaurants online, whether this is a family and kids friendly diner or a fancy, romantic restaurant could really alter the user's decision

depending on the setting. This idea can be implimented in group creations so new-weds can simply search for "honeymoon" for their ideal romantic trip.

REFERENCES

[Ala18] Jay Alammar. "The Illustrated Transformer.", 2018.

[Ber20] Pepe Berba. "Understanding HDBSCAN and Density-Based Clustering.", January 2020.

[Dav21] Jones Dave. "The 20 Best Beaches in Vietnam.", 2021.

[Gro18] Maarten Grootendorst. "Topic Modeling with BERT.", 2018.

[Gro21a] Maarten Grootendorst. "Interactive Topic Modeling with BERTopic.", 2021.

[Gro21b] Maarten Grootendorst. "MaartenGr/BERTopic HowTo/Feature-347.", 2021.

[han22] handluggageonly. "14 Very Best Things To Do In Zakynthos, Greece.", 2022.

[Hor18] Rani Horev. "BERT Explained: State of the art language model for NLP.", 2018.

[Lab18] Nodus Labs. "Latent Dirichlet Allocation (LDA) — How It Works.", July 2018.

[Mal19] Usman Malik. "Python for NLP: Topic Modeling.", April 2019.

[MHM18] Leland McInnes, John Healy, and James Melville. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.", 2018.

[MT13] Saif M. Mohammad and Peter D. Turney. "Crowdsourcing a Word-Emotion Association Lexicon." *Computational Intelligence*, **29**(3):436–465, 2013.

[SLN19] Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. "Pseudo-likelihood Reranking with Masked Language Models." *CoRR*, **abs/1910.14659**, 2019.

[SR22a] Julia Silge and David Robinson. "Analyzing word and document frequency: tf-idf.", 2022.

[SR22b] Julia Silge and David Robinson. "Sentiment analysis with tidy data.", 2022.

[Tod22] Smidt Todd. "The 30 Best Travel Blogs OF 2022.", 2022.

[VSP17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention Is All You Need.", 2017.