Text Mining of TripAdvisor Reviews – What Attractions Need to Know

Group 7 Datourism: Akshay Chauhan (A0218900L), Jiale Mi (A0218935W), Jiaqi Cao (A0218896J), Woo Jian Sheng (A0218849M), Yijie Wen (A0218946R)

GitHub Repo Link: https://github.com/akshay7chauhan/group7_bt5153_TA/

Abstract

With the closure of international borders due to COVID-19, attractions operators in Singapore have the chance to renovate and retool their experiences with minimal disruption to existing visitor satisfaction. This study mined TripAdvisor reviews from 2016 to 2020 for insights to suggest specific aspects that attraction operators should focus on to have the greatest impact on visitor satisfaction. Using latent dirichlet allocation (LDA), 18 topics were identified from the reviews for 4 sub-categories of attractions. Sentiment analysis was done on the topics to diagnose the overall sentiment for each topic. Further analysis was also conducted at selected attractions to understand the reasons for their sentiment scores. For this study to create a greater business impact, a recommendation system was built for the Singapore Tourism Board (STB) based on the results of the topic modelling and sentiment analysis. The recommendation system uses the visitor profile as input and suggests attractions based on the profile similarity and attraction similarity. The study concludes with a brief discussion on the scope for future enhancements to this approach.

1. Introduction

Due to the COVID pandemic, Singapore's borders were largely closed in 2020 and 2021 to international leisure travelers. Nevertheless, the slump in volume of international visitors provides the tourism industry an opportunity to renovate their premises, refine their product offerings and retrain their staff, without significant impact to the onsite visitor experience. More importantly, this downtime allows the tourism industry the time to transform, so that they can prepare for the gradual recovery of international travel in the next 2-3 years' time.

1.1 Problem Statement

The objective of this study is to support the Singapore Tourism Board (STB) in guiding the transformation of the tourism industry to better address visitor's needs and expectations. This is achieved through a 2-prong approach.

(1) Topic modelling using TripAdvisor reviews: to identify areas that deliver the greatest impact to the visitors' experience, topic modelling on online reviews will be done to uncover key topics and their corresponding sentiment. Further analysis is done to identify if this is a systemic issue across the entire attraction sector, or specific to certain types of attractions.

(2) Recommendation system: Singapore was ranked 31st in a 32-city study on Timeout's 2018 Most Exciting Cities to Visit. However, this is assessed to be an awareness gap as Singapore has many world-class attractions and immersive experiences. A recommendation system using online reviews would allow visitors to easily discover relevant Places of Interests (POIs) to visit.

1.2 Project Methodology

Online reviews of POIs can reflect the characteristics of this POI. For example, if the user likes this POI very much, the user will praise the good points in the review. Conversely, if a user had a bad experience, the user would complain about it in the review. Some users like to share their experiences, and from those experiences we can uncover the key information on the aspects that attract visitors to this POI. Topic modeling on these reviews enables us to get different topics for analysis.

User ratings are an important factor to judge the popularity of a POI. Sentiments reflected in user comments are a good proxy for user ratings. Therefore, sentiment analysis is conducted after topic modelling to better understand the review ratings.

At the same time, an unsupervised machine learning method is used to establish a recommendation system which recommends revelant POIs that users may be interested in.

2. Literature

Among the academic studies on data mining of online reviews, many studies used latent dirichlet allocation (LDA) to uncover dimensions found within the text reviews. Among the papers which covered the mining of online reviews in the tourism sector, many preferred to mine hotel reviews as the number of possible dimensions were more controllable (Y. Guo et al, 2016). Only a few

studies covered tourist attractions, which is this study's topic of interest.

One of those studies used TripAdvisor's reviews on tourist attractions in Phuket, Thailand (V. Taecharungroj, 2019). The paper also developed two practical tools, the dimensional salience-valence analysis (DSVA) and the lexical salience-valence analysis (LSVA), which allowed non-practitioners to easily understand the outputs. The findings from the paper were also used to suggest actions for the Tourism Authority of Thailand. This inspired the inclusion of a recommendation model for this study as a means to bridge the gap between theory and actual implementation.

3. Dataset Description

TripAdvisor is a leading global travel review website. According to its official website, TripAdvisor had an average of more than 490 million monthly visitors and a total of more than 730 million reviews and opinions by the end of 2018. The data was web scrapped from TripAdvisor website using beautifulsoup python package and was made available for this study by STB.

As TripAdvisor had a different page structure for attractions in Sentosa, it was left out in the dataset that was scrapped by the creator of the dataset.

The dataset crawled from TripAdvisor consists of 68 attractions across Singapore. There are a total of 126,021 reviews in the dataset from January 2016 to February 2020. Each row provides information such as the date of the review, rating of the attraction, review body and the profile of the reviewer. The variables in the dataset are listed in Appendix Table 1.

4. Data Exploration

Some form of analysis of the dataset prior to running the ML models can be handy for better feature engineering and model building. Exploratory data analysis was conducted on 3 variables in the dataset – attraction type, user profile and text.

4.1 Attraction description

There are 4 attraction sub-categories in our dataset. The first sub-category is "Leisure & Recreation" and include popular attractions like Gardens by the Bay, Singapore Flyer, Singapore Botanic Gardens and Marina Bay Sands Skypark. The second sub-category is "Precinct & Street", which mainly consists of cultural precincts such as Chinatown and lifestyle precincts like Orchard Road and Singapore River. The third type is "Nature & Wildlife", which mainly includes natural landscapes like MacRitchie Reservior and the zoos. The fourth type is "Art, History & Culture", which includes museums, historical locations and temples. In total, the dataset contains 68 Singapore attractions. Gardens by the Bay is the most popular attraction, followed by Singapore Zoo, Singapore Botanic Gardens, Marina Bay Sands SkyPark and Chinatown. They all have more than 5,000 reviews. Appendix Figure 2 shows the distribution of reviews in each sub-category.

4.2 User profile

We have a diverse visitor profile group. TripAdvisor's reviewers come from 188 different countries and regions. There are 53 countries and regions with more than 100 reviews. Around 10% of the reviews in our dataset have missing regional information. According to user profiles, the largest number of tourists come from Australia, followed by the UK, Singapore, India and the US. The number of tourists' reviews from Southeast Asia and East Asia is relatively small, but this may be because the review data is only in English, and excluded reviews in Chinese, Malay and other languages.

According to TripAdvisor's classification, trip types are classified into five categories: Couples, Family, Solo, Friends, and Business. Couples have the highest number of reviews, followed by Family. This shows that in the minds of most tourists, Singapore is an ideal family vacation destination. In particular, Couples from the United Kingdom, Australia and the United States, spend their honeymoons in Singapore for an Asian experience. Singapore is such a country which blends all kinds of Asian cultures and has a similar lifestyle to Western countries. Americans and Europeans could adapt themselves quickly to local life as well and have a glance at diverse Asian cultures.

Appendix Figure 3 shows the heat map of user profiles.

4.3 Text mining

4.3.1 Text Preprocessing

Several preprocessing steps were done to prepare the data for topic modelling:

(1) Expand Contractions.

Abbreviated words are expanded into full format, and some examples are shown in Table 1

Table 1 Convert abbreviations to full words

Abbreviated Word	*Full Word*
Ain't/Aren't	Is/Are not
Can't/Couldn't	Can/Could not
Cause	Because
Could've/Might've	Could/Might have
Don't/Didn't/Doesn't	Do/Did/Does not
Haven't/Hadn't/Hasn't	Have/Had/Has not
He'd/He'll/He's	He would/will/is

How'd/How'll/How's	How do/will/is
I'll/I've/I'd/I'm	I will/have/would/am
I'd've/I'll've	I would/will have

(2) Remove accented characters using package '*unicodedata*'.

(3) Remove special characters and numbers.

(4) Characters other than English letters, ends, commas, exclamation marks, question marks, slashes, colons, semicolons, quotes and spaces are removed.

(5) Remove punctuations.

(6) Remove stop words.

(7) Remove extra white spaces.

(8) Lowercase conversion.

(9) Lemmatization with WordNetLemmatizer.

4.3.2 Word Cloud

Several word clouds were generated based on attraction sub-categories, attraction names and travel types. This allows us to get an initial view of the text data. And the examples of word clouds based on attraction types – 'Leisure & Recreation' and 'Precinct & Street' is shown below:

Figure 1 Word cloud of 'Leisure & Recreation'

Figure 1 WordCloud of 'Precinct & Street'

As is shown in the word clouds, there are similar high-frequency words in reviews from different attraction sub-categories, which will be inconvenient for us to make recommendation for users. Therefore, topic modelling is utilized in this analysis.

5. Machine Learning models

5.1 Topic Modelling

In natural language processing, latent Dirichlet allocation (LDA) is a 'generative probabilistic model for collections of discrete datasets'[1], such as text corpora. A Topic modeling using LDA is a type of unsupervised learning. The graphical explanation is presented below:

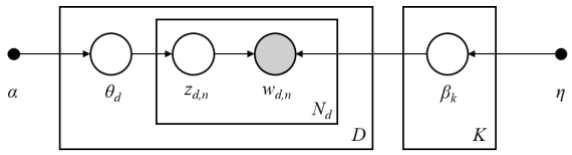

Figure 2 Graphical Topic Modeling

where

(1) The corpus is a collection of D documents.
(2) The document is a sequence of N words.
(3) There are K topics in the corpus.
(4) The boxes represent repeated sampling.

In topic modeling, it assumes that each document contains several topics, and each word belongs to one of the document's topics. LDA model that uses variational Bayes algorithm can be easily implemented from *sklearn.decomposition.*

5.2.1 Data Segmentation

We will use topic modeling method to explore latent sub-topics for each attraction type respectively, as we believe that sub-topics are distinct under different data type. For example, discussions on a zoo may not have the same focus with that of a museum. By detecting topics of each attraction sub-category, we can identify if a systemic issue exists or pertaining to certain types of attractions only.

5.2.2 Document-Word Matrix

After a standard text cleaning process referred to section 3 and 4, we create the document-word matrix as LDA model requires a numerical format of text data as its main input. There are two common methods for creating word vectors:

(1) TF-IDF Vectorizer: Returns a score with takes both word frequency and document frequency that contains the word into consideration.
(2) Count Vectorizer: Returns a count of a word

Table 2 Attraction type and number of reviews on each topic

Leisure & Recreation			*Precinct & Street*			*Nature & Wildlife*			*Art, History & Culture*		
1	Flowers and Plants	76	1	Nightlife	3889	1	Environment	11	1	Cultural Atmosphere	94
2	Service	100	2	Shop and Restaurants	11951	2	Animal Show	10	2	Artwork Quality	12659
3	Transportation and Tickets	5336	3	Cultural Atmosphere	14	3	Exploration Efficiency	7	3	View	125
4	Entertainments and Activities	185				4	Interaction with Animals	14953	4	Temple History	23
5	Indoor Environment	116				5	Nature				
6	Garden Tour	54179									
Total Reviews		60003			15909			14989			12832

Both methods can be efficiently implemented with the help of Scikit Learn, where all the heavy lifting is done by the feature extraction functionality provided for text datasets. Ideally, CountVectorizer should have performed better than TfidfVectorizer when feeding into the LDA model, as it is a model that deals with probabilities and only requires raw counts. However, after we have tried out both TfidfVectorizer and CountVectorizer, the former produced more interpretable topic results. Hence in this analysis, we proceed with the TfidfVectorizer.

5.2.3 K-Means Clustering

Instead of arbitrarily decide the number of topics we would like the model to detect, we use K-Means clustering and Silhouette Score to find the best approximate number of clustering for each attraction sub-category. The steps of choosing the best K are:

(1) For each attraction type, run K-Means clustering on K ranging from 2 to 10.

(2) Choose the K resulted in a relatively high Silhouette Score, which is an evaluation of the quality of the clustering. If two K result in similar scores, we choose the smaller one for modeling efficiency.

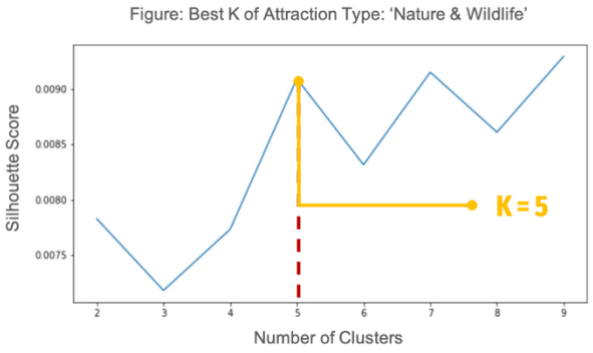

Figure 3 Optimal K with highest Silhouette Score

Figure 4 shows a sample of the best K on attraction type: 'Nature & Wildlife'. In this case, K equals to 5. Hence, the number of topics we would like to detect under this attraction sub-category is 5.

5.2.4 LDA and Modeling Results

After fitting the LDA model, we use the functionality *lda.transform* from Scikit Learn to tag each review with

the probabilities of belonging to each topic. We filter the result with the most likely topic (with the highest probability) and attach it as a new feature to the data.

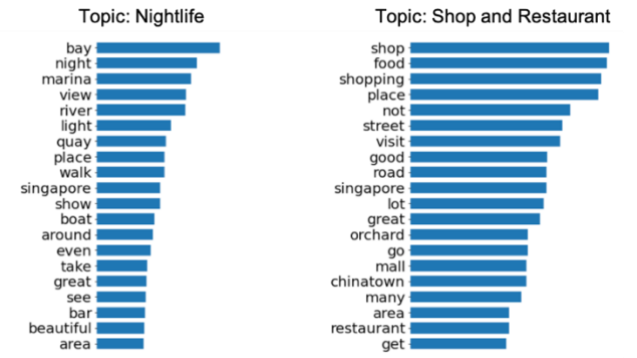

Figure 4 Sample Keywords Under Attraction Type: 'Precinct & Street'

Figure 5 shows some sample keywords for the two topics under the attraction type: 'Precinct and Street'. One researcher named each topic by looking into the key words as well as some original text bodies. For example, in the left diagram, top words are 'bay', 'night', 'light' etc., so we decide to name it as 'Nightlife'. Whereas in the right diagram, key words are related to food and shopping experiences, so we name it as 'Shop and Restaurant'. A second researcher independently named each topic too, and when the naming was different from the first researcher, the two researchers discussed and aligned with an agreed naming. This is a similar approached used by another study on mining of TripAdvisor reviews on attractions in Phuket, Thailand (V. Taecharungroj, 2019). After the naming of each topic, the data can be easily interpreted and investigated together with sentiment scores, in a hope of gaining a deeper insight of the Singapore tourism industry. A complete topic results under each attraction type and the corresponding number of reviews is shown in Table 2.

5.2 Sentiment analysis

As there is a strong bias towards 4 and 5-star ratings, the reviews ratings were recoded to balance out the distribution. Reviews with 1 to 3-star ratings were recoded as 'negative' and 4 to 5-star ratings as positive.

The Sentiment Intensity Analyser from the NLTK package was used on the original review text to calculate the intensity of the review sentiment. Each review was given a sentiment score ranging from -1 to 1.

The resulting dataset would include the original scrapped details of the attraction, the review text, the dimensions which each review is assigned to and the resulting sentiment analysis. with both identified dimensions, analysis could be done.

5.3 Recommendation System

In our approach we leverage the collaborative filtering approach of formulating a recommender system for tourist attractions in Singapore. Under normal circumstances data would be filtered by unique user ID to provide personalized recommendations but due to the lack of multiple reviews by the same user. We relied on establishing unique user profiles by the means of combining the country and trip type columns. This enabled us to recommend based on a richer model. Based on this approach we generated 633 unique user profiles.

The scores generated for the review text associated with each review was then then grouped by the profile name and the name of the attraction to gauge the mean scores. In the next step, we pivot the table to generate a 2D matrix of scores in such a manner that each index represents an attraction while each column represents a profile name. At this point we use an unsupervised Nearest Neighbor algorithm to map out the relative position of each profile based on their scores with respect to all the tourist attractions available to us.

5.3.1 Cosine Similarity and Nearest Neighbor Algorithm

We used the cosine similarity as a metric to measure how similar the profiles are on the basis of their reviews for different attractions. Mathematically, cosine similarity measures the cosine of the angle between two vectors projected in a multi-dimensional space. The cosine similarity is advantageous because even if the two similar documents are far apart by the euclidean distance (due to the size of the document), chances are they may still be oriented closer together. The smaller the angle, higher the cosine similarity.

After generating similarity scores for each of the profile and each of the attraction. We made use of the Nearest neighbour algorithm to make recommendations.

5.3.2 Recommender System Working Demonstration

Based on the above information we developed our algorithm for the recommender system which utilized this information to make recommendations based on the user profile. This recommendation system was able to make recommendations for profiles for which we had limited data available as well by means of finding similar profiles based on the available data and then using these similar profiles to make recommendations. We realized that with the existing system we were unable to make recommendation to profiles that did not exist in our data.

Here we provide the demonstration of how our recommender system is able to make recommendations for even profiles which we have limited data available. In our data the profile Philippines solo has only a limited data available in terms of the reviews and ratings for various attractions. But even with this limited data we are able to find similar profiles to Philippines solo and recommend new attractions based on the knowledge about the scores from these similar profiles.

Figure 6 demonstrates the how the recommender system actually makes the decisions to recommend attractions using the knowledge of similar profiles established by Cosine similarity.

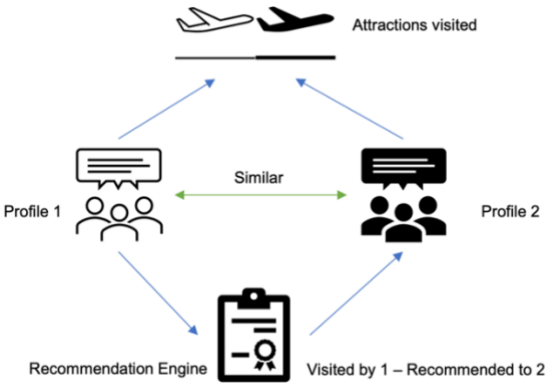

Figure 6 Recommendation system

6. Model Evaluation and Selection

As our first step as validating our approach we carried out the exercise of predicting the ratings of the reviews based on the features. Table 3 summarizes the results of the different Machine learning models that were used.

Table 3 Model evaluation

Model	*Train Accuracy*	*Test Accuracy*
Logistic Regression	0.6347	0.6293
SVM	0.6355	0.6344
Naïve Bayes (BOW)	0.78	0.75

Of all the model. Naïve Bayes performed the best and for further validation of our approach we calculated the ROC score for this model and found it to be 0.9416. It performed equally well for the ratings 1-5. This provided us with the confidence to move forward with LDA for topic modelling and expecting to get relevant topics for further analysis and the building of the recommendation system.

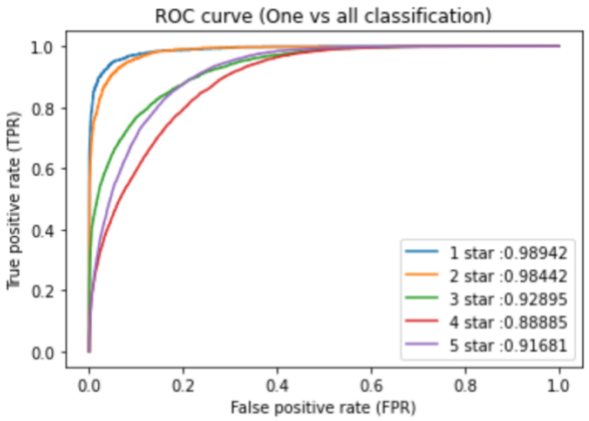

Figure 7 ROC curve of Naïve Bayes

7. Results and Discussion

7.1 Analysis of Topic Modelling and Sentiment Scores

Table 4 shows the results of the topic modelling and sentiment analysis.

Table 4 Result of Topic Modeling and Sentiment Scores

Leisure & Recreation	***Min Sentiment***	***Max Sentiment***	***Average Sentiment***
Entertainments and Activities	-0.57	0.99	0.54
Flowers and Plants	0.00	0.97	0.62
Garden Tour	-0.99	1.00	0.77
Indoor Environment	-0.78	0.98	0.53
Service	-0.96	0.98	0.58
Transportation and Tickets	-0.99	1.00	0.63
Grand Total	-0.99	1.00	0.76

Nature & Wildlife	***Min Sentiment***	***Max Sentiment***	***Average Sentiment***
Animal Show	-0.96	0.96	0.53
Environment	-0.36	0.96	0.63
Exploration Efficiency	0.20	0.90	0.65
Interaction with Animals	-0.99	1.00	0.70
Nature	0.56	0.91	0.76
Grand Total	-0.99	1.00	0.70

Art History & Culture	*Min Sentiment*	*Max Sentiment*	*Average Sentiment*
Artwork Quality	-0.99	1.00	0.67
Cultural Atmosphere	-0.50	0.98	0.71
Temple History	-0.30	0.90	0.51
View	-0.65	0.94	0.46
Grand Total	-0.99	1.00	0.67

Precinct & Street	*Min Sentiment*	*Max Sentiment*	*Average Sentiment*
Cultural Atmosphere	-0.90	0.96	0.38
Nightlife	-0.77	1.00	0.75
Shopping and Restaurant	-0.99	1.00	0.61
Grand Total	-0.99	1.00	0.65

Attractions in the Leisure & Recreation sub-category included popular POIs like Gardens by the Bay and Singapore Botanic Gardens (which is Singapore's first UNESCO world heritage site), and also POIs like the Singapore Flyer and Marina Bay Sands Skypark, which are famous for their breathtaking views of the city skyline. 6 topics were identified and labelled as 'Flowers and Plants', 'Garden Tours', 'Indoor Environment', 'Entertainment and Activities', 'Service' and 'Transportation and Ticketing'. The average sentiment scores for 'Garden Tour' and 'Flower and Plants' were 0.77 and 0.62, indicating that most reviews had positive sentiments regarding the curated flora and fauna in these POIs. However, 'Entertainment and Activities' scored 0.54, indicating that these POIs may have to relook at the activities that they offer, and consider how to improve the entertainment value.

Attractions in the Nature & Wildlife sub-category included popular POIs like the Singapore Zoo, River Safari, Night Safari and Jurong Bird Park and MacRitchie Nature Trail. 5 topics were identified and labelled as 'Animal Show', 'Interaction with animals', 'Environment', 'Exploration efficiency' and 'Nature'. The average sentiment scores for 'Nature' and 'Interactions with animals were 0.76 and 0.70, indicating that most reviews had very positive sentiments regarding the opportunity to get upclose with animals in their natural environment and habitats, something which is not common in zoos in other countries. However, 'Animal Show' scored 0.53. Upon further investigation, these were due to several reviews recounting the bad experience they had at Night Safari where they were turned away from the animal show because of the limited seating capacity.

Figure 5 Merlion Park

Attractions in the Art, History and Culture sub-category included POIs like Merlion Park, where an 8.6m Merlion stands at the edge of the waterfront, with a scenic backdrop of the Central Business District, and various museums and temples around Singapore. 4 topics were identified and labelled as 'Artwork Quality', 'Cultural Atmosphere', 'Temple History' and 'View'. The average sentiment scores for 'Cultural Atmosphere' was 0.71 and can be attributed to the colourful and immersive experiences visitors can get visiting the POIs like Buddha Tooth Relic Temple, Peranakan Museum and Sri Mariamman Temple.

Attractions in the Precinct & Street sub-category included cultural precincts like Chinatown and Little India, and lifestyle precincts like Orchard Road and Singapore River, where visitors can shop and dine. 3 topics were identified and labelled as 'Cultural Atmosphere', 'Nightlife' and 'Shopping and Restaurant'. The average sentiment scores for 'Nightlife' was a positive 0.75, debunking TimeOut's article of Singapore being a boring city. However, 'Cultural Atmosphere' had an average sentiment of 0.38. Upon further analysis, this was attributed to the inauthentic experiences at the cultural experiences at Chinatown, where one reviewer mentioned that it was a place where expensive souvenirs were sold and restaurants which were run to cater to unsuspecting tourists and not the locals.

In addition, more indepth analysis was done at the individual attraction level to understand their sentiment scores. Among the POIs which had very positive sentiment scores, the reviews mentioned specific attributes such as well-maintained facilities, immersive experiences, and world-class content. Among the POIs which had very low sentiment scores, reviews indicated the opposite; many of the POIs with low sentiment scores had poorly maintained facilities, poor crowd management, superficial experiences and limited content. The Intan, a small privately run Peranakan museum, had the highest average sentiment score of 0.84. Reviews clearly showed that it performed well on all aspects:

"*WOW thanks Alvin to view your private collection of Peranakan culture was **truly informative & inspiring**. Learning about your heritage & cultures thru your tales has given us an **insight** into how Singapore came about. I will **definitely recommend** your museum to all my mates." – New Zealand visitor travelling with friends*

Haw Par Villa, once a popular attraction in the 1960s, scored an average sentiment score of 0.49, one of the lowest among all the attractions in Singapore. Based on the online reviews, Har Par Villa did not fare well in the content and experience aspects:

"Those with an understanding of Chinese folk stories will get more out of it. Not hugely recommended for younger kids (our 6 y.o. was fairly bored and only really enjoyed feeding the terrapins). There are some juice vending machines outside but not much food." – New Zealand visitor travelling with family

7.2 Business Application

The combination of topic modelling and sentiment analysis of the TripAdvisor online reviews had demonstrated the ability to detect key topics for each sub-category of attractions and allowed a deep dive to diagnose poor performing aspects. The results are useful to STB as it provides clarity on what are the areas of improvement. STB can then create a concerted effort to encourage attraction operators to focus on these areas, usually through supports like financial grants and content co-creation.

The recommendation system provides POI recommendations based on the visitor profile. This is important as research showed that there is an awareness gap among the less popular POIs. The current model built is fully functional and can be deployed ie integrated as a search function on travel apps or travel websites.

7.3 Limitation of the Current Study

While TripAdvisor is the largest online platform globally, it under-represents the non-native English-speaking markets like North Asia and SEA countries which account for approximately 60% of the total international visitor arrival to Singapore in 2019. The approach in this study can be applied to other English and non-English travel review websites, and the performance can be compared against the results from the TripAdvisor data to further identify market nuances.

7.4 Possible Future Work

One major assumption is that consumers are largely homogenous within each profile. This means that travel considerations, prior travel experiences and preferences are not considered in the modelling. Rising incomes and growth of specific interests like sustainable tourism have also resulted in a more fragmented market in recent years. The challenge would be to identify emerging topics in the topic modelling process and take these additional considerations to refine the recommendation model.

8. References

[1] 2.5. decomposing signals in components (matrix factorization problems) —scikit-learn 0.24.1 documentation. https://scikit-learn.org/stable/modules/decomposition.html#latentdirichletallocation. (Accessed Apr 24, 2021)

[2] Yue Guo, Stuart J. Barnes, Qiong Jia. *Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation*. Tourism Management, Volume 59, Pages 467-483, 2017.

[3] Viriya Taecharungroj, Boonyanit Mathayomchan. *Analysing TripAdvisor reviews of tourist attractions in Phuket, Thailand*. Tourism Management, Volume 75, Pages 550-568, 2019.

[4] McKinsey & Company (2020, October 20). COVID-19 tourism spend recovery in numbers. Retrieved from https://www.mckinsey.com/industries/travel-logistics-and-infrastructure/our-insights/covid-19-tourism-spend-recovery-in-numbers#

[5] Nicole-Marie Ng (2018, January 31). Time Out City Index 2018: apparently Singapore is boring? Retrieved from https://www.timeout.com/singapore/news/time-out-city-index-2018-apparently-singapore-is-boring-013118

Appendix

Table 1 : Variables Description

Variable Name	*Variable Description*	*Tyoe*
REVIEW_INDEX	Unique ID for each review	number
REVIEW_DATE	Date of review	date
REVIEW_RATING	Review rating by user	number
REVIEW_TITLE	Title text of review	text
REVIEW_BODY	Body text of review	text
DATE_OF_EXPERIENCE	Date of visit	date
TRIP_TYPE	Type of trip	text
REVIEW_CRAWLED_TIME	Time of this review crawled	date
REVIEWER_NAME	Name of the reviewer	text
HOME_COUNTRY	Home country of reviewer	text
ATTRACTION_NAME	Name of attraction	text
ATTRACTION_TYPE	Type of attraction	text
ADDRESS	Address of attraction	text

Figure 2: Number of reviews by attraction

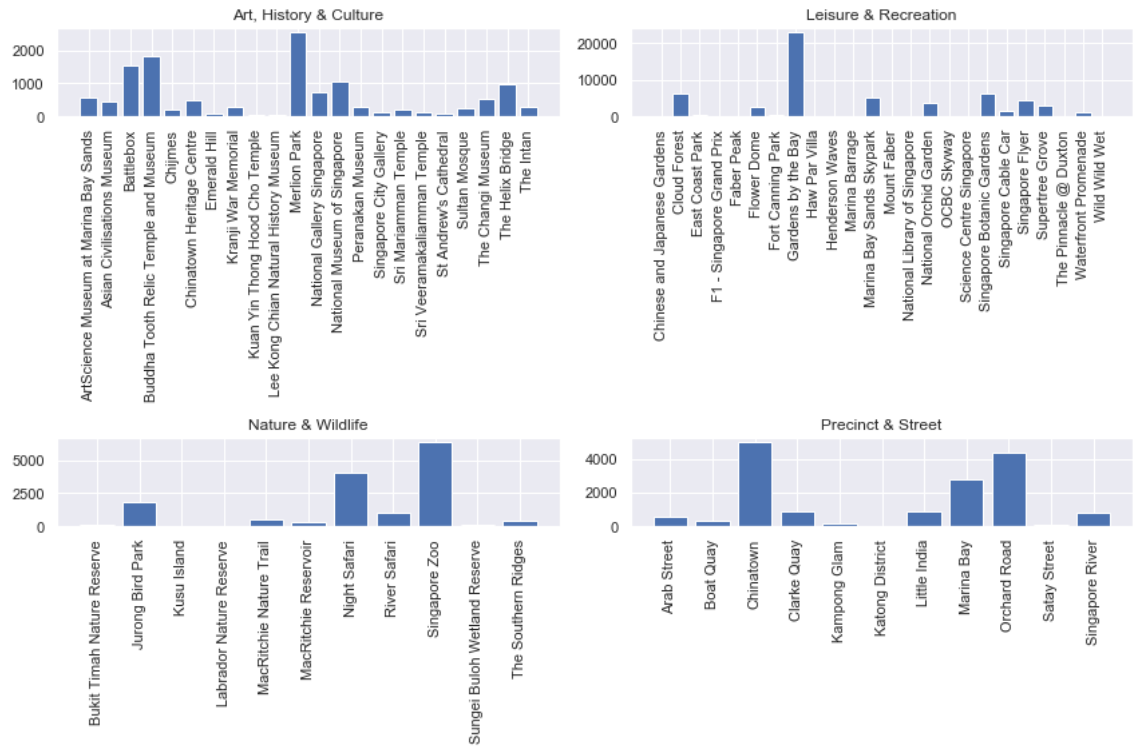

Figure 3: Heat map of user profile