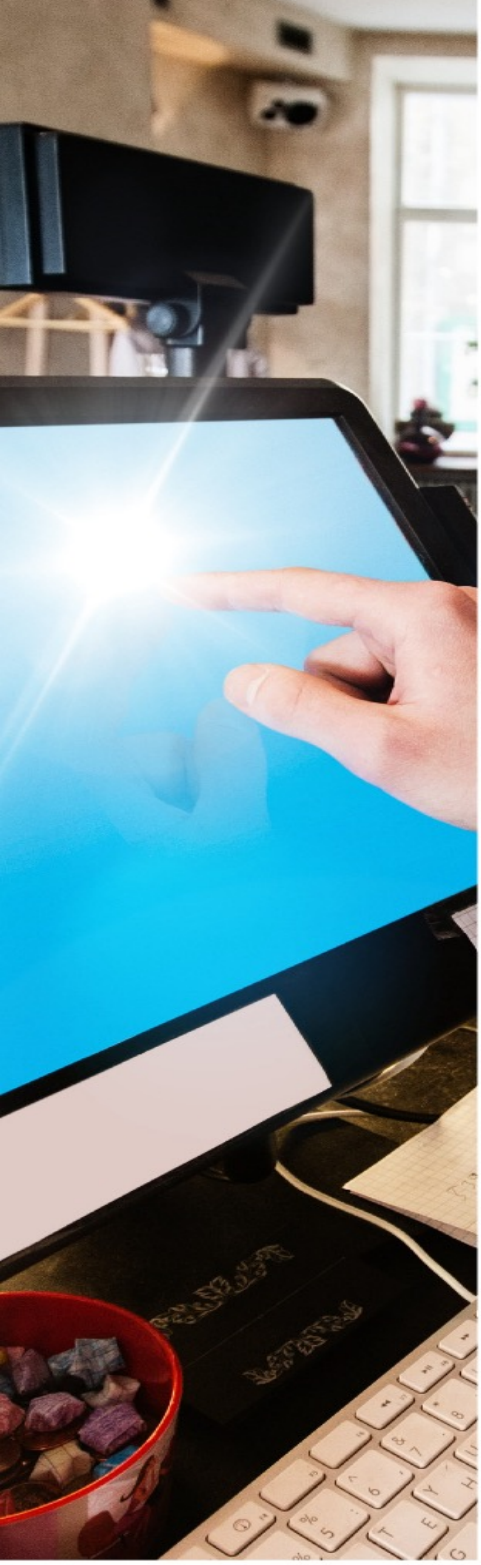

Publisher

UNIVERSITY OF SOUTH FLORIDA M3 CENTER

Advances in Hospitality and Tourism Information Technology

Editors

DR. CIHAN COBANOGLU, DR. SEDEN DOGAN,
DR. KATERINA BEREZINA, & DR. GALEN COLLINS

ISBN 978-1-7321275-8-6

ADVANCES IN HOSPITALITY AND TOURISM INFORMATION TECHNOLOGY

For all chapters, please visit: https://www.m3center.org/ahtit

ISBN 978-1-7321275-8-6

This imprint is published by University of South Florida M3 Publishing

The registered company address is: 8350 N Tamiami Trail, Sarasota, FL 34243 USA

https://digitalcommons.usf.edu/m3publishing/vol18/iss9781732127586/7
DOI: https://www.doi.org/10.5038/9781732127586

Chapter

Analytics in Hospitality and Tourism: Online Travel Reviews

Estela Marine-Roig

University of Lleida, Catalonia, Spain

SUMMARY

User-generated content, shared with other users through social media, has increased considerably in the previous decade. In particular, the content generated by travelers, mainly online travel reviews (OTRs), has grown dramatically. This abundant recorded information has served as a basis for conducting numerous researches on big data and social media analytics. Reviewers share their OTRs on travel-related websites including peer-to-peer (P2P) accommodation platforms and online travel agencies (OTAs). The aim of this chapter is to offer an overview of the state of the art of hospitality and tourism analytics based on OTRs, and explore the possibilities of gaining insight, through OTRs, about perceived image and visitor preferences. In the context of hospitality prior to the Covid-19 pandemic, empirical substantiation is obtained by crossing paratextual data from hotels, registered on TripAdvisor and in three OTAs indexed in the Nasdaq 100 (Booking, Expedia and Ctrip), located in five tourist cities (Barcelona, Cape Town, Los Angeles, Singapore, and Sydney). Regarding the content analysis of OTRs text, although there are numerous publications on Airbnb (the main P2P lodging platform), research on the influence of Airbnb OTRs on destination image construction is scarce. Therefore, the content of the Airbnb OTRs of these five cities is explored in search of patterns and metrics that allow us to measure the image perceived and transmitted by visitors.

Recommended Citation: Marine-Roig, E. (2021). Analytics in hospitality and tourism: Online travel reviews. In C. Cobanoglu, S. Dogan, K. Berezina, & G. Collins (Eds.), *Advances in Hospitality and Tourism Information Technology* (pp. 1–27). USF M3 Publishing. https://www.doi.org/10.5038/9781732127586

Learning objectives

At the end of this chapter, the student will be able to answer the following questions related to online travel reviews (OTRs):

- Is the number of OTRs proportional to the number of hotel rooms?
- Do hotels occupy a similar position in the popularity ranking of travel-related websites?
- Is the average score obtained by hotels on travel-related websites equivalent?
- Is there a relationship between hotel class (stars) and popularity or rating?
- Considering the content of OTRs on P2P accommodation, are there significant differences in tourist destination images between cities?

https://digitalcommons.usf.edu/m3publishing/vol18/iss9781732127586/7
DOI: https://www.doi.org/10.5038/9781732127586

Introduction

In the field of information and communications technologies (ICT), one of the main features of the 21st century is the profusion of user-generated content (UGC) shared among peers through social media (Park & Lee, 2021). Several authors have shown that many users read the online comments of other users and consider them when purchasing goods and services (M. S. Lin, Liang, Xue, Pan, & Schroeder, 2021). This is a paradigm shift in research because surveys or interviews are no longer essential for gathering the opinions of users and consumers (Volo, 2018, 2020). However, the large volume of available data forces researchers to employ computer-aided analytical processes (analytics). Subsequently, the scientific literature addresses social media analytics (Rathore, Kar, & Ilavarasan, 2017) and big-data analytics (Liang & Liu, 2018). Likewise, in the field of hospitality and tourism information technologies (HTIT), there is a great deal of information generated by travelers (TGC: traveler-generated content) and, in particular, by guests, diners, and tourists (Y. R. Li, Lin, Tsai, & Wang, 2015; Marine-Roig, 2019), both in textual and visual formats (Lojo, Li, & Xu, 2020; Mak, 2017). TGC impacts on tourism destination image (TDI) formation and subsequently on the overall satisfaction of tourists (Lam, Ismail, & Lee, 2020). Thus, the purpose of this chapter is to demonstrate some methods for extracting useful knowledge from the textual and paratextual elements of TGC. To facilitate learning, we will apply these methods of TGC analytics to a case study on five tourist cities: Barcelona (Catalonia), Cape Town (Western Cape), Los Angeles (California), Singapore (Singapore), and Sydney (New South Wales). The source of information will consist of more than 3.5 million online travel reviews (OTR) on accommodations in those cities posted to five travel-related web platforms: TripAdvisor, Booking, Expedia, Ctrip, and AirBnB. The data were collected just before the start of the Covid-19 pandemic, so they can serve as a reference in future studies on the recovery of the hospitality industry.

Traveler-Generated Content (TGC)

Today, many travelers share their opinions on their sightseeing adventures. The dissemination of their comments has evolved from word-of-mouth marketing (WoM) to electronic word-of-mouth (eWoM) communication through social media (Baka, 2016). TGCs' main sources are travel blogs and online travel reviews shared on travel-related websites. Another way to collect opinions from travelers is through their interactions (comments/replies, likes/favorites, and shares/retweets) with the official social media of destination management organizations (DMO) on Facebook, Twitter, and Instagram accounts (de las Heras-Pedrosa, Millan-Celis, Iglesias-Sánchez, & Jambrino-Maldonado, 2020; Huertas & Marine-Roig, 2016).

Destination Image Analytics Through TGC

Scholars have studied destination image since the 1960s (Batista Sánchez, Serrano Leyva, & Pérez Ricardo, 2020; Chon, 1990), but it the 1990s that some researchers (Baloglu & McCleary, 1999; Echtner & Ritchie, 1991; Gartner, 1993) built solid theoretical and methodological bases to define and measure TDIs. Among the many definitions of TDI

is one that stands out for its popularity (Crompton, 1979): An image may be defined as the sum of beliefs, ideas, and impressions that a person has of a destination (p. 18), and another that stands out for its scientific rigor (Lai & Li, 2016): TDI can be defined as follows: a voluntary, multisensory, primarily picture-like, qualia-arousing, conscious, and quasi-perceptual mental (i.e., private, nonspatial, and intentional) experience held by tourists about a destination (p. 1074).

Based on numerous previous studies, Marine-Roig (2019) proposed a comprehensive TDI model from a holistic perspective. In this proposal, synthesized in Figures 1 and 2, TDI is a global and complex concept perceived as a whole different from the sum of its parts (gestalt). Figure 1 shows the TDI formation circle, and Figure 2 shows three interrelated aspects that facilitate the methodological analysis of TDIs.

At the ends of the hermeneutical circle in Figure 1, there are two constructs that are fed back: the image projected by agents of the destination, and the image perceived by visitors and prospective tourists. According to the Gartner (1993) model, there are three types of agents or sources of TDI formation: induced sources that depend on destination managers, organic sources that reflect the opinion of visitors, and autonomous sources that are usually independent of the previous two. Prior personal experiences are segregated from organic sources because they enjoy maximum credibility for repeat visitors. Currently, the TGC disseminated through eWoM on social media is the most popular source of consultation and consideration among travelers (Ferrer-Rosell & Marine-Roig, 2020; Marine-Roig & Ferrer-Rosell, 2018).

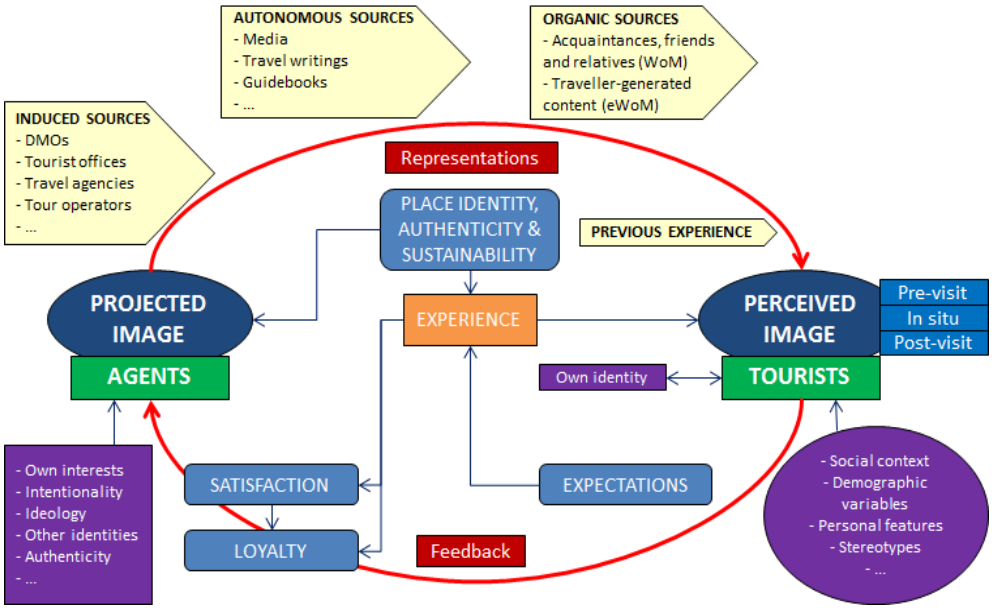

Figure 1. Circle of Destination Image Construction from a Holistic Perspective (Marine-Roig, 2019)

Classic destination image models, such as Lynch's pioneering work (Lynch, 1960), focused on the perception of urban environments. To analyze the image of cities, later authors considered three components or aspects of images: cognitive-affective-conative (Rapoport, 1977) or designative-appraisive-prescriptive (Pocock & Hudson, 1978)

https://digitalcommons.usf.edu/m3publishing/vol18/iss9781732127586/7
DOI: https://www.doi.org/10.5038/9781732127586

models. Marine-Roig (2019) later expanded a tripartite model to suit TGC as a TDI formation source (Figure 2). Due to the large volume of TGC on hospitality that influences the online destination image formation, she included in the descriptive aspect facilities to accommodate tourism-related services. That is, in addition to the physical characteristics (structure and shape) of the observed tourism resources, visitors consider the features or amenities of hotels, restaurants, transport, etc. As Lynch (1960) said, the observer's mental picture is relatively abstract when identifying a structure as a restaurant.

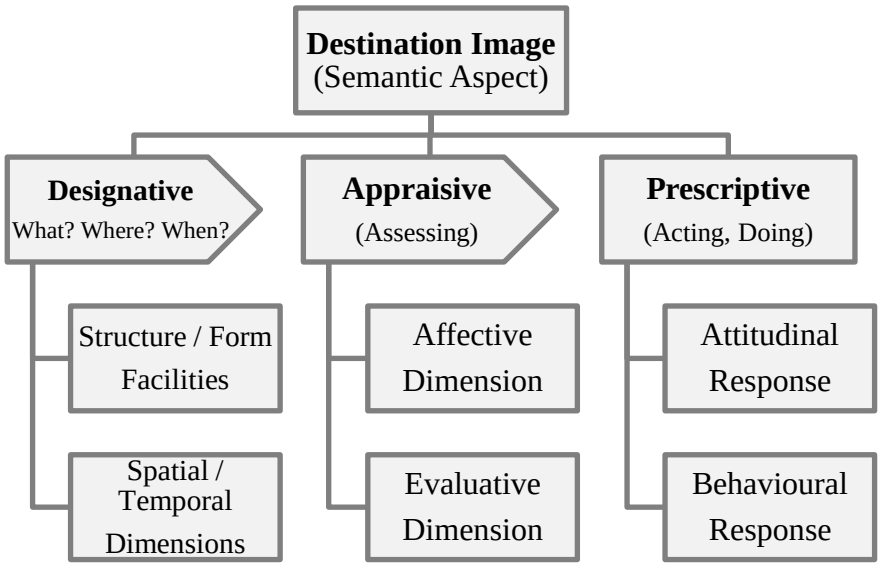

Figure 2. Tourism Destination Image Aspects (Marine-Roig, 2019)

Online Travel Review (OTR)

Since the pioneering study conducted at the Laboratory for Intelligent Systems in Tourism and led by Ulrike Gretzel (Gretzel & Yoo, 2008), OTRs have increased dramatically. For example, currently, in one of the cities analyzed in this study (Barcelona), there is an attraction (Basilica of the Sagrada Familia) that has almost 164,000 OTRs and 119,500 photos hosted on TripAdvisor.

Travel Reviews as a Data Source for Research

Several authors of studies on systematic literature review related to OTRs demonstrated that accommodation was the most researched sector in the field of tourism and hospitality (Table 1). In contrast, studies on perceived overall TDI were scarce. More than half the researchers used textual (Xiang, Du, Ma, & Fan, 2018), visual (Giglio, Pantano, Bilotta, & Melewar, 2019), or paratextual (Marine-Roig, 2017b) information retrieved from TripAdvisor OTRs (Kwok, Xie, & Richards, 2017). The most studied topics were (Hlee, Lee, & Koo, 2018) sales performance, usefulness, credibility and trust, expectation and purchasing intention, patterns, consumer satisfaction, decision making, managerial response, and motivation to read OTRs.

Table 1. Literature Review on OTRs per Tourism and Hospitality Sectors

Sector	(Schuckert, Liu, & Law, 2015)	(Kwok et al., 2017)	(Xiang, Du, Ma, & Fan, 2017)	(Hlee et al., 2018)	Average
Lodging	30	47	16	35	66.67 %
Dining	9	8	4	8	15.10 %
Tour	11	10	2	12	18.23 %

Travel Review Sources

In order to analyze the aspects of the TDI (Figure 2), it is necessary that the OTRs contain opinions of visitors about a tourist resource located in time and place as well as some type of score for evaluation. In this sense, the main source of OTRs is the online travel company TripAdvisor. In 2019, TripAdvisor served more than 15 million unique visitors daily and hosted more than 867 million reviews and comments in 28 languages on 8.7 million tourism resources (TripAdvisor, 2020). Next in importance are online travel agencies (OTA) (e.g., Booking). Today (September 2020), Booking hosts 187 million verified reviews from real guests, 25 million destination reviews from real travelers, and 14 million photos shared by real travelers (Booking, 2020). To a lesser extent, travel metasearch engines (e.g., Trivago) and web mapping services (e.g., Google Maps) also host OTRs.

Textual and Paratextual Information in Travel Reviews

An OTR's website contains structured and unstructured data that can be useful for research. A TripAdvisor review webpage, for example, provides three sources of meaningful information:

1) HTML (HyperText Markup Language) metadata (Marine-Roig, 2017a). HTML metadata is not visible to users and serves to make it easier for Internet browsers and search engines to read the essential information on the page. In this case, the three most interesting metadata for research are tagged through the Open Graph and/or App Links protocols: 'og:url', 'al:ios:url', 'og:title' and 'og:description'. The URL (Uniform Resource Locator) provides structured information about the communication protocol, server, and domain; the type of review; locality, tourist resource and review codes; and the names of the resource, locality and country. The title contains semi-structured information: on the one hand, the title written by the reviewer and, on the other hand, information added by TripAdvisor, such as locality and region. The title is also between the tags <title> and </title>. The description contains the name and the number of reviews and photos that the tourist resource has.

2) Paratextual elements (Marine-Roig, 2017b). Paratextual items consist of structured information directly related to the review, such as the name, location and score of the tourist resource, the date and language of the review, the username and country of origin of the reviewer, etc. The extraction of this data requires a computerized search and replace tool that supports regular expressions (regex).

https://digitalcommons.usf.edu/m3publishing/vol18/iss9781732127586/7
DOI: https://www.doi.org/10.5038/9781732127586

3) Text and media (Marine-Roig, 2019). The text only has the structure of the grammar rules of each language. Natural language processing techniques facilitate the analysis of textual content. For the massive analysis of other means of expression, it is necessary to have computer tools based on artificial intelligence that are not within the scope of this study.

Computing Tools for Travel Review Analytics

Due to the complexity of qualitative studies and the large volume of data (big data) in quantitative studies, OTR analytics require computer tools to collect and process data. Table 2 shows a sample of research applications used to analyze OTRs coming from several sources with various objectives. In this study we will use the algorithm in Box 1; a text editor such as Notepad++ on Windows, NotepadQQ on Linux, or Atom on MacOS; and a spreadsheet such as Microsoft Office Excel or Apache OpenOffice Calc. The recommended application to implement the Box 1 algorithm is KH Coder (Higuchi, 2020), a free software for quantitative content analysis and text mining, also employed for computational linguistics (natural language modeling). Currently, it supports text analysis in 13 languages on Windows, Linux, and MacOS operating systems. In addition to extracting and counting words, KH Coder analyzes word placement and co-occurrence, distinctive words for each part, document clustering and classification, and searching and counting documents using coded rules.

Table 2. Sample of Computer Tools Employed to Analyzing Online Travel Reviews

Software	Type	Research	Goal	Resource	Source	Dataset
NVivo	QDA	(Mate, Trupp, & Pratt, 2019)	Negative OTRs	Hotels	TripAdvisor	57
Leximancer	ACA	(Aitieva, Kim, & Kudaibergenov, 2021)	Destination image	Attractions	TravelBlog BlogSpot	360
CoreNLP	NLP	(Perikos et al., 2018)	Opinion mining	Hotels	Booking	1682
KH Coder	LSA	(K. Zhang & Koshijima, 2019)	Text mining	Hotels	Ctrip	14,850
Word2vec	ML	(W. Li, Zhu, Guo, Shi, & Zheng, 2018)	Sentiment lexicon	All tourist resources	Ctrip, Qunar	30,180
OpenCoDa	MSA	(Marine-Roig & Ferrer-Rosell, 2018)	Destination image gaps	Attractions	TripAdvisor	80,000
Algorithm (Box 1)	NLP	(Marine-Roig & Huertas, 2020)	Destination image	P2P lodgings	AirBnB	152,704
NLTK	NLP	(Hou, Cui, Meng, Lian, & Yu, 2019)	Opinion mining	All tourist resources	Ctrip, Tuniu	165,429
Algorithm (Box 1)	FA	(Marine-Roig, Ferrer-Rosell, Daries, & Cristobal-Fransi, 2019)	Gastronomic image	Restaurants	TripAdvisor	500,000
WEKA	ML	(P.-J. Lee, Hu, & Lu, 2018)	Helpfulness	Hotels	TripAdvisor	1,170,246
OpenNLP	NLP	(Guy, Mejer, Nus, & Raiber, 2017)	Travel tips	Attractions	TripAdvisor	3,362,296

LibSVM	SVM	(Martin-Fuentes, Fernandez, Mateu, & Marine-Roig, 2018)	Star rating	Hotels	Booking	18,710,881

***Note*.** QDA: qualitative data analysis; NLP: natural language processing; ACA: automatic content analysis; LSA: latent semantic analysis; ML: machine learning; MSA: multivariate statistical analysis; FA: frequency analysis; SVM: support vector machine.

Case Study: Accommodations in Five Tourist Cities

The five cities selected to apply the proposed methodology have comparable common characteristics: they are coastal cities or near the coast; they are prominent tourist destinations, and they have implemented peer-to-peer lodging in concurrence with the hotel industry. The main difference between the chosen cities is that they are located on diverse continents, namely, Europe (Barcelona) (Gutiérrez, García-Palomares, Romanillos, & Salas-Olmedo, 2017), Africa (Cape Town) (Visser, Erasmus, & Miller, 2017), North America (Los Angeles) (D. Lee, 2016), Asia (Singapore) (Koh & King, 2017), and Australia (Sydney) (Alizadeh, Farid, & Sarkar, 2018). Table 3 shows the existing accommodations in the five cities. TripAdvisor includes hotels, motels, hostels, B&Bs, and inns; AirBnB includes the rental types entire home, private room, and shared room available more than one month per year.

Table 3. Accommodations as of January 2020

Web Host	Barcelona	Cape Town	Los Angeles	Singapore	Sydney
TripAdvisor	523	121	378	189	201
AirBnB	16474	19033	29130	6090	19863

Materials and Methods

The implementation of methods to analyze the case study is divided into two parts:

A) *TripAdvisor and three OTAs' accommodation OTRs*. Quantitative analysis of the paratextual items that accompany the guest OTRs of a sample of hotels in the five cities, hosted on TripAdvisor, Booking, Expedia, and Ctrip. Scores on hotel attributes allow the deduction of guest satisfaction (Gunasekar & Sudhakar, 2019).

B) *AirBnB guest OTRs*. Measurement of the TDI of the five cities based on all guest OTRs published in English on the AirBnB platform during 2018 and 2019 (1,774,691 OTRs). This TDI is of special interest because the host-guest interaction influences the visitors' perception of their experiences (Shi, Gursoy, & Chen, 2019). The aim of this measurement is to compare the online images of cities through the metrics derived from the model in Figure 2.

Data Collection

OTR collection requires a web copier program that allows setting download filters (Marine-Roig & Anton Clavé, 2016), such as HTTrack Website Copier or Cyotek

https://digitalcommons.usf.edu/m3publishing/vol18/iss9781732127586/7
DOI: https://www.doi.org/10.5038/9781732127586

WebCopy. It is essential to study the structure of the website and configure the filters properly. For example, the TripAdvisor server dynamically manages more than 1 billion HTML pages, and each TripAdvisor OTR webpage contains more than 500 HTML links. Any copier configuration error can crash the local computer.

A) *TripAdvisor and three OTAs' accommodation OTRs*. The download process is divided into several phases: first, downloading the web pages of all the properties in each city registered in the Hotels and Places to stay section of TripAdvisor; second, extraction of the main paratextual items of each property (i.e., class, # of rooms, # of OTRs, and score); third, ranking from highest to lowest the properties by number of reviews; fourth, searching and downloading on the other platforms of the most popular properties on TripAdvisor; and fifth, extraction of the paratextual items of these properties.

B) *AirBnB guest OTRs*. Due to the difficulty of downloading data directly from the AirBnB platform, it is constructive to obtain this information from InsideAirbnb (Murray, 2020), a prestigious website, independent of AirBnB, which periodically publishes detailed data on AirBnB activity in numerous cities and regions around the world. Table 4 shows the number of OTRs available in each city.

Table 4. AirBnB Guest Reviews at January 2020

Period and Language	Barcelona	Cape Town	Los Angeles	Singapore	Sydney
All times and languages	749,047	335,339	1,358,535	108,271	628,517
2018-2019 (English)	293,175	196,982	830,807	58,561	395,166

Natural Language Processing (NLP)

Once the data is downloaded, it is necessary to extract, clean, arrange, and store the significant textual and paratextual data. The data storage files are in CSV (plain-text, comma-separated values) format because it allows text processing with a parser such as the one in Box 1, and it is compatible with text editors, database systems, and spreadsheets. The character code chosen is UTF-8 (Unicode transformation format-8 bit) because, thanks to its variable-width encoding (one to four bytes), it supports all valid characters in standard Unicode (universal coded character set).

Recognition of the text language of OTRs is complicated because there are some very short and also bilingual comments (Marine-Roig & Huertas, 2020). This study used the Java Lingua library (Stahl, 2020) based on an n-gram probabilistic model (1-gram to 5-gram) that supported more than 70 languages. To improve accuracy, the algorithm made two detections: the first with the from all spoken languages option, detected 67 languages in the dataset; and the second, with the from languages Catalan, Chinese, Dutch, English, French, German, Italian, Japanese, Korean, Portuguese, Russian, Spanish option, detected the 12 most common languages and classified the others as unknown.

To perform the frequency analysis in languages such as French, German, and Italian, languages that contain many endings, it is necessary to reduce the words to their roots

using NLP stemming or lemmatization techniques based on, for example, the SnowBall library for Java (Porter, 2021) and multilingual resources at UniNE (UniNe, 2020).

Content Analysis

Content analysis is a set of research techniques used to transfer unstructured information to a data matrix suitable for statistical analysis. In case study A (sample of hotels), the data collected is numerical and, therefore, can be analyzed statistically to answer the questions posed in this research. In contrast, in case study B (AirBnB guest OTRs), reviewers give feedback on accommodations and other experiences at the destination. Each OTR contains structured data such as dates and scores, as well as non-structured text.

Researchers agree that the terms mentioned most often are those that arouse the greatest interest to writers. Therefore, the analysis of textual content consists of counting and categorizing key terms. Counting consists of listing the key terms and calculating the percentage they represent in relation to the total number of words, including stop words. Categorization consists of grouping by theme the univocal key terms with similar meanings or connotations (Weber, 1990) based on some theory (a priori coding), a preliminary examination of the data (emergent coding), or a combination of both encodings (Marine-Roig & Huertas, 2020). In the case of big-data analytics with data sets of several million words, it is almost impossible to achieve exhaustive categories, but it is easier to control the categories being mutually exclusive. The metric used to make comparisons is the percentage of key terms belonging to the category, that is, the sum of the percentages obtained for each key term in the counting phase.

Counting Paratextual Items

Once the downloaded data is arranged and stored in CSV files, we can import it into a spreadsheet to perform statistical calculations using, for example, a Microsoft Office Excel pivot table.

Counting Key Terms

According to the definition of terms, it is necessary to prepare a list of key terms with two or more consecutive words that have their own meanings (e.g., los angeles, definitely come back, not book it, not miss it). To generate the frequency table, the algorithm in Box 1 performs the following steps: 1) converts the text to lower case; 2) counts and eliminates compound terms that have priority over other terms (stop words and terms with a single word) (in case of overlap of two compound terms, priority is given to the one that is first on the list, for example, not book it comes before book it); 3) divides the text into words by considering any character that is not a letter as a word separator; 4) eliminates stop words; 5) reduces words to their root or main parts through a stemming or lemmatization process based on NLP techniques; and 6) returns a frequency table with three columns: in the first column all the unique terms are listed, in the second the amount

https://digitalcommons.usf.edu/m3publishing/vol18/iss9781732127586/7
DOI: https://www.doi.org/10.5038/9781732127586

of each term in integers, and in the third the calculation of the percentage that the term represents in relation to the total words processed.

Box 1. Simplified Algorithm Used for Quantitative Content Analysis

```
Load compWord; // group of two or more words (list of)
Load stopWord; // non-significant word (list of)
Load text;     // text to analyze quantitatively
Load nonLetter; // word delimiter (non-letter characters)
New  words;    // words in text (list of)
New  stems;    // root or main part of word (list of)
New  result;   // table to store frequency of key terms

text := text.toLowerCase();            // lowercase text
for each compWord do
{
  if exists compWord in text then
  {
    Count occurrences of compWord in text;
    Add compWord to result with its frequency;
    Delete occurrences of compWord in text;
  }
}
words := tokenize(text, nonLetter);    // list of tokens
for each stopWord do
{
  Delete occurrences of stopWord in words;
}
stems := stemming(words);              // list of stems
for each stem in stems do
{
  if exists stem in result increase its frequency;
  else  add stem to result with frequency 1;
}
Print result;
```

In the case of counting and categorizing terms in English, the stemming process is not necessary because this language has very few grammatical inflections, and it is easy to include in the categories all the most frequent forms that appear in OTRs.

Sentiment Analysis

With regard to the affective dimension and the appraisal aspect of the image (Figure 2), sentiment analysis, also known as opinion mining, is intended to deduce the feelings and moods of reviewers based on the positive, negative, or neutral polarity of the OTR key terms. Sentiment analysis, in addition to being a useful method for measuring the affective dimension of the perceived image, also contributes to the analysis of reviewers' satisfaction and loyalty (Figure 1). The polarity of sentiment expressed in OTRs is directly related to overall customer satisfaction (Zhao, Xu, & Wang, 2019). The doctrine (H. Zhang, Fu, Cai, & Lu, 2014) considers attitudinal loyalty to be the intention to recommend and behavioral loyalty the intention to visit or revisit. Regarding OTRs, recommendations and visits do not refer to intentions but to facts confirmed by the reviewers' own accounts.

Since the textual data are not structured, it is necessary to construct categories to measure the affective, attitudinal, and behavioral dimensions. Based on the combination of both coding methods seen above and considering the particles that change polarity, the terms selected for the categories (i.e., adjectives, nouns, verbs, and phrases) have positive or negative polarity in the vast majority of cases within the context of the OTRs. The algorithm in Box 1 manages the overlap between terms to avoid counting a term as positive and negative at the same time (e.g., nice, not so nice), but does not control for irony (e.g., a nice way to).

Affective dimension categories:

- Positive feelings and moods (*feel+*): e.g., great, dream come true, chill-out, happy
- Negative feelings and moods (*feel-*): e.g., dirty, illegal, off-putting, disappointed

Attitudinal response categories:

- Positive recommendations (*recom+*): e.g., recommend, recommended, book it, look no further
- Negative recommendations and warnings (*recom-*): e.g., avoid, not recommend, not stay here, be careful

Behavioral response categories:

- Positive behavior (*behav+*): back next time, return in the future, stay there again, would not stay anywhere else
- Negative behavior (*behav-*): not stay here again, not stay there again, will not be back, would not return

In relation to the evaluative dimension of the image (Figure 2), guests scores of accommodation features (Table 5) may be different depending on the platform hosting the OTR; however, all platforms agree to rate the cleanliness of the accommodations. In the Staff/Service column, AirBnB guests value the relationship with the host under the concepts of check-in and communication, and in the Other column, they value the accuracy of the offer, i.e., whether the features of the accommodation advertised on the portal correspond to reality.

Table 5. Accommodation Features and Amenities Scored by Reviewers

	Staff/Service	Cleanliness	Comfort	Location	Value for Money	Other
AirBnB	X	X		X	X	X
Booking		X	X	X	X	X
Ctrip	X	X		X		X
Expedia	X	X	X			
TripAdvisor	X	X		X	X	

Findings and Discussion

The results are divided into two parts as recorded in the Materials and Methods section. A part of the discussion remains pending for the readers, who have the opportunity to solve the exercises and problems presented in the appendix.

https://digitalcommons.usf.edu/m3publishing/vol18/iss9781732127586/7
DOI: https://www.doi.org/10.5038/9781732127586

Guest Reviews on TripAdvisor and Three Online Travel Agencies

Table 6 shows the six most popular hotels on TripAdvisor in each of the five cities by number of OTRs. The first three columns contain the hotel data common to the four travel-related companies: main airport code (BCN: Barcelona; CPT: Cape Town; LAX: Los Angeles; SIN: Singapore; and SYD: Sydney), hotel class (star rating from 1* to 5*), and number of rooms. The other columns show the number of OTRs and the average score given by the reviewers, considering that Booking scores from 1 to 10 and the other platforms from 1 to 5.

Table 6. Scores From the Six Most Popular Hotels on TripAdvisor

			TripAdvisor		Expedia		Booking		Ctrip	
Airport	**Class**	**Rooms**	**OTRs**	**Score**	**OTRs**	**Score**	**OTRs**	**Score**	**OTRs**	**Score**
BCN	5*	473	7688	4.5	996	4.3	2612	8.5	95	4.3
BCN	4*	169	5615	4.5	1101	4.7	1417	9.3	21	4.8
BCN	5*	483	5554	4.5	991	4.5	1194	8.5	94	4.6
BCN	3*	108	5249	4.5	993	4.6	811	9.0	31	4.7
BCN	4*	167	5113	4.5	9416	4.4	846	8.7	23	4.4
BCN	3*	105	5072	4.5	985	4.6	771	9.0	47	4.7
CPT	5*	329	3022	4.5	475	4.6	1244	9.0	24	4.7
CPT	4*	537	2745	4.0	179	4.3	280	8.2	8	4.9
CPT	5*	483	2672	4.5	746	4.6	320	8.8	21	4.8
CPT	5*	120	2506	5.0	135	4.8	86	9.4	10	4.7
CPT	5*	176	2462	4.5	999	4.7	1282	9.0	17	4.3
CPT	4*	394	2328	4.5	125	4.4	180	8.5	16	4.8
LAX	4*	1234	6139	3.5	10462	3.9	3713	7.8	2368	4.4
LAX	4*	495	5043	4.0	6404	4.3	1593	8.6	439	4.6
LAX	4*	628	4799	4.5	5689	4.4	1479	8.7	77	4.6
LAX	4*	453	4390	4.5	2134	4.4	487	8.6	51	4.6
LAX	4*	747	4044	4.0	7187	4.0	1550	8.3	595	4.4
LAX	3*	257	3842	4.5	6189	4.4	581	8.7	35	4.3
SIN	5*	2561	28897	4.5	1000	4.4	23196	9.0	8616	4.8
SIN	5*	1077	11949	4.0	1270	4.2	3934	8.4	825	4.6
SIN	5*	790	10587	4.5	1289	4.6	5513	9.0	2769	4.8
SIN	5*	1252	10114	4.0	5993	4.3	3274	9.0	2813	4.8
SIN	5*	575	7719	4.0	4708	4.3	2146	8.4	833	4.6
SIN	5*	778	7718	4.5	2274	4.6	1485	9.0	1305	4.8
SYD	5*	415	11131	4.5	4209	4.5	4058	8.9	306	4.7
SYD	5*	430	7902	4.5	1029	4.5	6998	8.9	856	4.7
SYD	5*	564	7894	4.5	4212	4.5	6086	8.8	1058	4.7
SYD	5*	531	7814	4.5	4517	4.6	2710	9.1	910	4.7
SYD	4*	382	7798	4.5	4414	4.5	3846	8.9	416	4.6
SYD	3.5*	413	6815	3.5	1770	3.8	3237	7.9	37	4.6

Hotel class: Star ratings indicate the general level of features and amenities to expect

Looking superficially at Table 6, it turns out that the largest hotel (SIN, 5*, 2,561 rooms) has only 1,000 OTRs on Expedia, but it has the maximum number of reviews on the other platforms. The highest-rated hotel (CPT, 5*, 120 rooms) has the highest rating on three platforms; however, on Ctrip it does not reach the highest rating. Also, hotels located in Europe and Africa have very few reviews on Ctrip.

Destination Image Analytics Through Guest Reviews on Airbnb

Table 7 shows the 20 most frequent keywords as a percentage of the total words in the OTRs regarding each city. Bearing in mind that the most frequent terms are those that arouse the greatest interest to reviewers, it turns out that the most important topics related to the design aspect are experience in accommodations (e.g., apartment, house, stay, or room) and the location of the accommodations (e.g., place, location, and city name). Other terms (e.g., close, metro, mrt, and station) can also be related to location. Regarding the

affective dimension, the main themes are cleanliness (i.e., clean) and the relationship with the host (i.e., host). In addition, all qualifying adjectives have positive polarity (e.g., great, nice, and good). Regarding the prescriptive aspect, the attitudinal response (i.e., recommend) stands out.

Table 7. Most Frequent Key Terms in Percentages

Barcelona		Cape Town		Los Angeles		Singapore		Sydney	
great	1.033	great	1.126	place	1.221	place	1.241	great	1.329
apartment	0.890	place	1.049	great	1.196	great	0.889	place	1.063
location	0.832	stay	0.822	stay	0.730	stay	0.724	location	0.970
place	0.800	location	0.653	location	0.653	location	0.708	stay	0.793
stay	0.573	apartment	0.648	clean	0.621	clean	0.654	clean	0.610
clean	0.522	host	0.481	nice	0.463	good	0.603	apartment	0.610
nice	0.499	clean	0.464	host	0.449	room	0.503	host	0.487
barcelona	0.489	recommend	0.371	house	0.355	nice	0.497	nice	0.389
host	0.464	cape town	0.364	comfortable	0.322	host	0.486	close	0.377
good	0.405	nice	0.359	recommend	0.318	mrt	0.439	good	0.373
room	0.363	beautiful	0.353	home	0.303	apartment	0.426	sydney	0.366
recommend	0.339	amazing	0.351	perfect	0.287	singapore	0.400	recommend	0.365
really	0.300	lovely	0.340	space	0.282	close	0.274	comfortable	0.327
close	0.298	perfect	0.339	room	0.277	really	0.267	lovely	0.320
metro	0.295	really	0.303	definitely	0.274	easy	0.261	easy	0.319
perfect	0.283	home	0.302	los angeles	0.274	recommend	0.254	perfect	0.293
comfortable	0.222	view	0.299	good	0.261	station	0.234	house	0.291
easy	0.218	house	0.280	really	0.261	check	0.233	really	0.287
walk	0.211	good	0.275	easy	0.256	walk	0.231	walk	0.269
time	0.211	comfortable	0.272	close	0.242	just	0.227	home	0.233

Note. los angeles = LA + Los Angeles; mrt = MRT + Mass Rapid Transit
Source: AirBnB guest OTRs in English posted during 2018 and 2019 (85 million words)

Table 8 shows the results of the sentiment analysis. The first row represents the evaluative dimension as an average of the overall score. AirBnB accommodations in all cities have a score of more than 90%, very high ratings consistent with previous research (Bulchand-Gidumal & Melián-González, 2020; Marine-Roig, 2021b). In terms of the affective dimension, terms with positive polarity represent more than 5% of all words. For example, Los Angeles English AirBnB OTRs posted during 2018 and 2019 contain 2.7 million positive key terms.

The top-rated city is Cape Town as it has the highest positive ratings and negative minimums, except that it has an intermediate score on positive behavior. In addition, the best-rated hotel in Table 6 is located in Cape Town.

Table 8. Sentiment Analysis Results in Percentages

Group	Barcelona	Cape Town	Los Angeles	Singapore	Sydney
Avg. score	90.98	94.70	94.39	90.61	93.37
Feel-	0.33273	0.19306	0.28871	0.30047	0.22715
Feel+	5.96106	7.58604	6.81942	5.77276	7.22038
Recom-	0.03789	0.02957	0.03279	0.03010	0.02975
Recom+	0.55609	0.67502	0.50372	0.45769	0.61699
Behav-	0.00091	0.00046	0.00096	0.00073	0.00073
Behav+	0.12657	0.17086	0.25700	0.13873	0.21896

Concluding Remarks

This research shows that you can extract from hospitality online travel reviews useful knowledge for destination marketing and management organizations. The opinions freely

https://digitalcommons.usf.edu/m3publishing/vol18/iss9781732127586/7
DOI: https://www.doi.org/10.5038/9781732127586

expressed by many thousands or millions of visitors (big data) facilitate the improvement, promotion, and distribution of tourist resources, and they aid in co-creating new tourist experiences and improving existing ones (Lalicic, Marine-Roig, Ferrer-Rosell, & Martin-Fuentes, 2021; M. P. Lin, Marine-Roig, & Llonch-Molina, 2021). In applying the proposed methodology to analyze a sample of 3.5 million OTRs on accommodations in five cities (Barcelona, Cape Town, Los Angeles, Singapore, and Sydney), it was found that Cape Town stood out for being the city valued best by the reviewers. These results may be of interest to the DMOs and other stakeholders of these destinations.

This all-encompassing model for measuring tourism destination images (Figure 1 and Figure 2) (Marine-Roig, 2019)—using narratives, opinions, and ratings shared on social media and based on visitors' experiences in traveling, sightseeing, entertaining, shopping, lodging, and dining in tourist destinations (Marine-Roig & Huertas, 2020)—is useful for analytics in different spatial areas (neighborhood, city, country, region, etc.) and times (years, months, seasonality, temporal trends, etc.). In addition, OTRs can be segregated by language, country of origin of the reviewer, and other features listed in their profiles (Marine-Roig, 2017b).

The main limitation of methodologies based on UGC big data is the classification and categorization of key terms. For example, the number of words (e.g., 85 million in Table 7) makes it difficult to create exhaustive categories because there may be dialects, slang, irony, ambiguous words, misspellings, etc. Another limitation for similar reasons is the less than 100% accuracy of natural language processing such as that executed by language recognition and translation programs. Content analysis succeeds or fails according to its categories (Berelson, 1952). Therefore, the main challenge in research based on TGC big data is to refine the classification of key terms into categories and check their accuracy against other data sources such as OTRs on attractions, transport, restaurants, and other tourism-related resources and activities.

Key Terms and Definitions

- *Traveler-generated content (TGC)* (Marine-Roig & Huertas, 2020): Narratives, opinions, pictures, audiovisual files, and ratings shared on social media and based on visitors' experiences traveling, sightseeing, entertaining, shopping, lodging, and dining in a tourist destination.
- *Regular expression (regex)*: Regex is a sequence of characters that defines a search-and-replace pattern in plain-text documents. Regular language is useful for extracting web page data from online travel reviews. The three recommended text editors (Notepad++, NotepadQQ, and Atom) support regular expressions.
- *Term*: Minimum unit of analysis consisting of a word (e.g., great, stay) or a group of consecutive words with their own meaning (e.g., Cape Town, not stay here).
- *Key term*: Term that can be significant in relation to any of the categories (e.g., nouns, adjectives, verbs).
- *Stop word*: Word that is not meaningful for content analysis like most determiners, conjunctions, prepositions, pronouns, and adverbs.
- *Comma-separated values (CSV)* (Marine-Roig & Huertas, 2020): Plain-text file used to store a data table. Each line represents a record, and a record is composed of fields containing a piece of information (e.g., code, date, text). Fields are separated by commas or semicolons. CSV files are compatible with any text editor, spreadsheet, or database system.

https://digitalcommons.usf.edu/m3publishing/vol18/iss9781732127586/7
DOI: https://www.doi.org/10.5038/9781732127586

Discussion Questions (10)

Please try to argue your answers based on the information available in this chapter and the related bibliography.

A) TripAdvisor and three OTAs' accommodation OTRs (Table 6). Starting from the most popular hotels in each city on TripAdvisor by number of reviews, several questions arise:
1) Is the number of reviews proportional to the number of rooms?
2) Do hotels occupy the same position in the popularity rankings of TripAdvisor and the three OTAs?
3) Are the average scores obtained by hotels on TripAdvisor and in the three OTAs equivalent?
4) Is there a relationship between the class of the hotel (star rating) and its popularity or score?
5) Are there significant differences between cities?
6) Are there significant differences between travel agencies in relation to cities?

B) AirBnB guest OTRs.
7) Why does the close relationship between host and guest affect the lived experiences and the image perceived by visitors?
8) Do AirBnB guest accommodation reviews shared on social media contribute to the formation of TDI online?
9) According to the most frequent key terms (Table 7), what are the topics that AirBnB guests prefer to discuss?
10) According to the most frequent key terms (Table 7), are there adjectives with evident positive polarity that allow us to deduce the satisfaction of the guests? What does it mean that the adverbs really and definitely are among the most frequent words?

References

Aitieva, D., Kim, S., & Kudaibergenov, M. (2021). Destination image of Kyrgyzstan: A content analysis of travel blogs. *Journal of Quality Assurance in Hospitality & Tourism*, (in press). https://doi.org/10.1080/1528008X.2021.1964412

Alaei, A. R., Becken, S., & Stantic, B. (2019). Sentiment analysis in tourism: Capitalizing on big data. *Journal of Travel Research*, *58*(2), 175–191. https://doi.org/10.1177/0047287517747753

Alizadeh, T., Farid, R., & Sarkar, S. (2018). Towards understanding the socio-economic patterns of sharing economy in Australia: An investigation of Airbnb listings in Sydney and Melbourne metropolitan regions. *Urban Policy and Research*, *36*(4), 445–463. https://doi.org/10.1080/08111146.2018.1460269

Baka, V. (2016). The becoming of user-generated reviews: Looking at the past to understand the future of managing reputation in the travel sector. *Tourism Management*, *53*, 148–162. https://doi.org/10.1016/j.tourman.2015.09.004

Baloglu, S., & McCleary, K. W. (1999). A model of destination image formation. *Annals of Tourism Research*, *26*(4), 868–897. https://doi.org/10.1016/S0160-7383(99)00030-4

Batista Sánchez, E., Serrano Leyva, B., & Pérez Ricardo, E. del C. (2020). Bibliometric study of tourism destination image in Science Direct. *Interamerican Journal of Environtment and Tourism*, *16*(1), 97–105. https://doi.org/10.4067/s0718-235x2020000100097

Berelson, B. (1952). *Content analysis in communication research*. New York, NY: Free Press.

Booking. (2020). How was your stay? Retrieved September 20, 2020, from Booking reviews website: https://www.booking.com/reviews.html

Bulchand-Gidumal, J., & Melián-González, S. (2020). Why are ratings so high in the sharing economy? Evidence based on guest perspectives. *Current Issues in Tourism*, *23*(10), 1248–1260. https://doi.org/10.1080/13683500.2019.1602597

Chon, K.-S. (1990). The role of destination image in tourism : A review and discussion. *The Tourist Review*, *45*(2), 2–9. https://doi.org/10.1108/eb058040

Crompton, J. L. (1979). An assessment of the image of Mexico as a vacation destination and the influence of geographical location upon that image. *Journal of Travel Research*, *17*, 18–23. https://doi.org/10.1177/004728757901700404

de Borda, J. C. (1781). Mémoire sur les élections au scrutin. In B. de Fontanelle, J. J. Mairan, J. P. Grandjean de Fouchy, & J. A. Condorcet (Eds.), *Histoire de l'Académie Royale des Sciences avec les mémoires de mathématique & de physique* (pp. 657–665). Paris: Imprimerie Royal.

de las Heras-Pedrosa, C., Millan-Celis, E., Iglesias-Sánchez, P. P., & Jambrino-Maldonado, C. (2020). Importance of Social Media in the image formation of tourist destinations from the stakeholders' perspective. *Sustainability*, *12*(10), article 4092. https://doi.org/10.3390/su12104092

Echtner, C. M., & Ritchie, J. R. B. (1991). The meaning and measurement of destination image. *Journal of Tourism Studies*, *2*(2), 2–12.

Ferrer-Rosell, B., & Marine-Roig, E. (2020). Projected versus perceived destination image. *Tourism Analysis*, *25*(2–3), 227–237. https://doi.org/10.3727/108354220X15758301241747

Gartner, W. C. (1993). Image formation process. *Journal of Travel & Tourism Marketing*, *2*(2–3), 191–215. https://doi.org/10.1300/J073v02n02_12

Giglio, S., Pantano, E., Bilotta, E., & Melewar, T. C. (2019). Branding luxury hotels: Evidence from the analysis of consumers' big visual data on TripAdvisor. *Journal of Business Research*, (in press). https://doi.org/10.1016/j.jbusres.2019.10.053

Gretzel, U., & Yoo, K. H. (2008). Use and impact of online travel reviews. In P. O'Connor, W. Höpken, & U. Gretzel (Eds.), *Information and Communication Technologies in Tourism 2008* (pp. 35–46). Vienna: Springer Vienna. https://doi.org/10.1007/978-3-211-77280-5_4

Gunasekar, S., & Sudhakar, S. (2019). How user-generated judgments of hotel attributes indicate guest satisfaction. *Journal of Global Scholars of Marketing Science*, *29*(2), 180–195. https://doi.org/10.1080/21639159.2019.1577155

Gutiérrez, J., García-Palomares, J. C., Romanillos, G., & Salas-Olmedo, M. H. (2017). The eruption of Airbnb in tourist cities: Comparing spatial patterns of hotels and peer-to-peer accommodation in Barcelona. *Tourism Management*, *62*, 278–291. https://doi.org/10.1016/j.tourman.2017.05.003

Guy, I., Mejer, A., Nus, A., & Raiber, F. (2017). Extracting and ranking travel tips from user-generated reviews. *Proceedings of the 26th International Conference on World Wide Web - WWW '17*, 987–996. New York, USA: ACM Press. https://doi.org/10.1145/3038912.3052632

Higuchi, K. (2020). KH Coder: for quantitative content analysis or text mining. Retrieved September 20, 2020, from https://khcoder.net/en

Hlee, S., Lee, H., & Koo, C. (2018). Hospitality and tourism online review research: A systematic

analysis and heuristic-systematic model. *Sustainability*, *10*(4), article 1141. https://doi.org/10.3390/su10041141

Hou, Z., Cui, F., Meng, Y., Lian, T., & Yu, C. (2019). Opinion mining from online travel reviews: A comparative analysis of Chinese major OTAs using semantic association analysis. *Tourism Management*, *74*, 276–289. https://doi.org/10.1016/j.tourman.2019.03.009

Huertas, A., & Marine-Roig, E. (2016). Differential destination content communication strategies through multiple Social Media. In A. Inversini & R. Schegg (Eds.), *Information and Communication Technologies in Tourism 2016* (pp. 239–252). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-28231-2_18

Koh, E., & King, B. (2017). Accommodating the sharing revolution: a qualitative evaluation of the impact of Airbnb on Singapore's budget hotels. *Tourism Recreation Research*, *42*(4), 409–421. https://doi.org/10.1080/02508281.2017.1314413

Kwok, L., Xie, K. L., & Richards, T. (2017). Thematic framework of online review research: A systematic analysis of contemporary literature on seven major hospitality and tourism journals. *International Journal of Contemporary Hospitality Management*, *29*(1), 307–354. https://doi.org/10.1108/IJCHM-11-2015-0664

Lai, K., & Li, X. (2016). Tourism destination image: Conceptual problems and definitional solutions. *Journal of Travel Research*, *55*(8), 1065–1080. https://doi.org/10.1177/0047287515619693

Lalicic, L., Marine-Roig, E., Ferrer-Rosell, B., & Martin-Fuentes, E. (2021). Destination image analytics for tourism design: An approach through Airbnb reviews. *Annals of Tourism Research*, *86*, article 103100. https://doi.org/10.1016/j.annals.2020.103100

Lam, J. M. S., Ismail, H., & Lee, S. (2020). From desktop to destination: User-generated content platforms, co-created online experiences, destination image and satisfaction. *Journal of Destination Marketing & Management*, *18*, article 100490. https://doi.org/10.1016/j.jdmm.2020.100490

Lee, D. (2016). How Airbnb short-term rentals exacerbate Los Angeles's affordable housing crisis: Analysis and policy recommendations. *Harvard Law & Policy Review*, *10*(1), 229–253.

Lee, P.-J., Hu, Y.-H., & Lu, K.-T. (2018). Assessing the helpfulness of online hotel reviews: A classification-based approach. *Telematics and Informatics*, *35*(2), 436–445. https://doi.org/10.1016/j.tele.2018.01.001

Li, Q., Li, S., Zhang, S., Hu, J., & Hu, J. (2019). A review of text corpus-based tourism big data mining. *Applied Sciences*, *9*(16), article 3300. https://doi.org/10.3390/app9163300

Li, W., Zhu, L., Guo, K., Shi, Y., & Zheng, Y. (2018). Build a tourism-specific sentiment lexicon via Word2vec. *Annals of Data Science*, *5*(1), 1–7. https://doi.org/10.1007/s40745-017-0130-3

Li, Y. R., Lin, Y. C., Tsai, P. H., & Wang, Y. Y. (2015). Traveller-generated contents for destination image formation: Mainland China travellers to Taiwan as a case study. *Journal of Travel & Tourism Marketing*, *32*(5), 518–533. https://doi.org/10.1080/10548408.2014.918924

Liang, T.-P., & Liu, Y.-H. (2018). Research landscape of business intelligence and big data analytics: A bibliometrics study. *Expert Systems with Applications*, *111*, 2–10. https://doi.org/10.1016/j.eswa.2018.05.018

Lin, M. P., Marine-Roig, E., & Llonch-Molina, N. (2021). Gastronomic experience (co)creation: evidence from Taiwan and Catalonia. *Tourism Recreation Research*, (in press). https://doi.org/10.1080/02508281.2021.1948718

Lin, M. S., Liang, Y., Xue, J. X., Pan, B., & Schroeder, A. (2021). Destination image through social media analytics and survey method. *International Journal of Contemporary Hospitality Management*, (in press). https://doi.org/10.1108/IJCHM-08-2020-0861

Lojo, A., Li, M., & Xu, H. (2020). Online tourism destination image: components, information sources, and incongruence. *Journal of Travel & Tourism Marketing*, *37*(4), 495–509. https://doi.org/10.1080/10548408.2020.1785370

Lynch, K. (1960). *The image of the city*. Cambridge, MA: The MIT Press.

Mak, A. H. N. (2017). Online destination image: Comparing national tourism organisation's and tourists' perspectives. *Tourism Management*, *60*, 280–297. https://doi.org/10.1016/j.tourman.2016.12.012

Marine-Roig, E. (2017a). Measuring destination image through travel reviews in search engines. *Sustainability*, *9*(8), article 1425. https://doi.org/10.3390/su9081425

Marine-Roig, E. (2017b). Online travel reviews: A massive paratextual analysis. In Z. Xiang & D. R. Fesenmaier (Eds.), *Analytics in Smart Tourism design: Concepts and methods* (pp. 179–202). Heidelberg, Germany: Springer. https://doi.org/10.1007/978-3-319-44263-1_11

Marine-Roig, E. (2019). Destination image analytics through traveller-generated content. *Sustainability*, *11*(12), article 3392. https://doi.org/10.3390/su11123392

Marine-Roig, E. (2021a). Contributions on Destination image analytics through traveller-generated content. In A. Correia & S. Dolnicar (Eds.), *Women's voices in tourism research: Contributions to*

knowledge and letters to future generations (pp. 306–309). Brisbane, Australia: University of Queensland. https://doi.org/10.14264/817f87d

Marine-Roig, E. (2021b). Measuring online destination image, satisfaction, and loyalty: Evidence from Barcelona districts. *Tourism and Hospitality*, *2*(1), 62–78. https://doi.org/10.3390/tourhosp2010004

Marine-Roig, E. (2022). Content analysis of online travel reviews. In Z. Xiang, M. Fuchs, U. Gretzel, & W. Höpken (Eds.), *Handbook of e-tourism* (p. forthcoming). Cham, Switzerland: Springer.

Marine-Roig, E., & Anton Clavé, S. (2016). A detailed method for destination image analysis using user-generated content. *Information Technology & Tourism*, *15*(4), 341–364. https://doi.org/10.1007/s40558-015-0040-1

Marine-Roig, E., & Ferrer-Rosell, B. (2018). Measuring the gap between projected and perceived destination images of Catalonia using compositional analysis. *Tourism Management*, *68*, 236–249. https://doi.org/10.1016/j.tourman.2018.03.020

Marine-Roig, E., Ferrer-Rosell, B., Daries, N., & Cristobal-Fransi, E. (2019). Measuring gastronomic image online. *International Journal of Environmental Research and Public Health*, *16*(23), article 4631. https://doi.org/10.3390/ijerph16234631

Marine-Roig, E., & Huertas, A. (2020). How safety affects destination image projected through online travel reviews. *Journal of Destination Marketing & Management*, *18*, article 100469. https://doi.org/10.1016/j.jdmm.2020.100469

Martin-Fuentes, E., Fernandez, C., Mateu, C., & Marine-Roig, E. (2018). Modelling a grading scheme for peer-to-peer accommodation: Stars for Airbnb. *International Journal of Hospitality Management*, *69*, 75–83. https://doi.org/10.1016/j.ijhm.2017.10.016

Mate, M. J., Trupp, A., & Pratt, S. (2019). Managing negative online accommodation reviews: evidence from the Cook Islands. *Journal of Travel & Tourism Marketing*, *36*(5), 627–644. https://doi.org/10.1080/10548408.2019.1612823

Murray, C. (2020). Get the data. Retrieved February 20, 2020, from Inside Airbnb: Adding data to the debate website: http://insideairbnb.com

Park, D.-H., & Lee, S. (2021). UGC sharing motives and their effects on UGC sharing intention from quantitative and qualitative perspectives: Focusing on content creators in South Korea. *Sustainability*, *13*(17), article 9644. https://doi.org/10.3390/su13179644

Perikos, I., Tsirtsi, A., Kovas, K., Grivokostopoulou, F., Daramouskas, I., & Hatzilygeroudis, I. (2018). Opinion mining and visualization of online users reviews: A case study in Booking.com. *2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA)*, 1–5. IEEE. https://doi.org/10.1109/IISA.2018.8633597

Pocock, D., & Hudson, R. (1978). *Images of the urban environment*. London, UK: Macmillan.

Porter, M. F. (2021). SnowBall: A language for stemming algorithms. Retrieved March 19, 2021, from https://snowballstem.org/

Rapoport, A. (1977). *Human aspects of urban form*. Oxford, UK: Pergamon Press.

Rathore, A. K., Kar, A. K., & Ilavarasan, P. V. (2017). Social media analytics: Literature review and directions for future research. *Decision Analysis*, *14*(4), 229–249. https://doi.org/10.1287/deca.2017.0355

Schuckert, M., Liu, X., & Law, R. (2015). Hospitality and tourism online reviews: Recent trends and future directions. *Journal of Travel & Tourism Marketing*, *32*(5), 608–621. https://doi.org/10.1080/10548408.2014.933154

Shi, S., Gursoy, D., & Chen, L. (2019). Conceptualizing home-sharing lodging experience and its impact on destination image perception: A mixed method approach. *Tourism Management*, *75*, 245–256. https://doi.org/10.1016/j.tourman.2019.05.012

Stahl, P. M. (2020). Lingua: Natural language detection library. Retrieved November 20, 2020, from https://github.com/pemistahl/lingua

TripAdvisor. (2020). About us. Retrieved January 1, 2020, from TripAdvisor Media Center website: https://tripadvisor.mediaroom.com/us-about-us

UniNe. (2020). IR multilingual resources at UniNE. Retrieved March 19, 2021, from Université de Neuchâtel website: http://members.unine.ch/jacques.savoy/clef/

Visser, G., Erasmus, I., & Miller, M. (2017). Airbnb: The emergence of a new accommodation type in Cape Town, South Africa. *Tourism Review International*, *21*(2), 151–168. https://doi.org/10.3727/154427217X14912408849458

Volo, S. (2018). Tourism data sources: From official statistics to big data. In *The SAGE handbook of tourism management: Theories, concepts and disciplinary approaches to tourism* (pp. 193–201). London, UK: SAGE Publications. https://doi.org/10.4135/9781526461452.n12

Volo, S. (2020). Tourism statistics, indicators and big data: a perspective article. *Tourism Review*, *75*(1), 304–309. https://doi.org/10.1108/TR-06-2019-0262

Weber, R. (1990). *Basic content analysis* (2nd ed.). Thousand Oaks, CA: SAGE Publications, Inc. https://doi.org/10.4135/9781412983488

Xiang, Z., Du, Q., Ma, Y., & Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. *Tourism Management*, *58*, 51–65. https://doi.org/10.1016/j.tourman.2016.10.001

Xiang, Z., Du, Q., Ma, Y., & Fan, W. (2018). Assessing reliability of social media data: lessons from mining TripAdvisor hotel reviews. *Information Technology & Tourism*, *18*(1–4), 43–59. https://doi.org/10.1007/s40558-017-0098-z

Ylijoki, O., & Porras, J. (2016). Conceptualizing big data: Analysis of case studies. *Intelligent Systems in Accounting, Finance and Management*, *23*(4), 295–310. https://doi.org/10.1002/isaf.1393

Zhang, H., Fu, X., Cai, L. A., & Lu, L. (2014). Destination image and tourist loyalty: A meta-analysis. *Tourism Management*, *40*, 213–223. https://doi.org/10.1016/j.tourman.2013.06.006

Zhang, K., & Koshijima, I. (2019). Trend analysis of online travel review text mining over time. *Journal of Modelling in Management*, *15*(2), 491–508. https://doi.org/10.1108/JM2-10-2018-0178

Zhao, Y., Xu, X., & Wang, M. (2019). Predicting overall customer satisfaction: Big data evidence from hotel online textual reviews. *International Journal of Hospitality Management*, *76*, 111–121. https://doi.org/10.1016/j.ijhm.2018.03.017

Problems (15)

The sample file LA_UTF8.csv is available **(pending)**. The LA_UTF8.csv file contains the AirBnB OTRs from the city of Los Angeles posted during 2018 and 2019. Content analysis issues to solve using an advanced text editor. The following instructions are for the Notepad++ version 8 tool. Depending on the power of the available hardware, operations can run in near real time or take a few minutes. If the available hardware and/or software are obsolete, it is preferable to divide the sample file into two parts (e.g., OTRs posted during 2018 and OTRs posted during 2019) to speed up data processing. Please open the LA_UTF8.csv file with Notepad++. When solving exercises, keep in mind that there may be overlapping terms and that there are particles that change the terms' polarities. Through the menu Search - Find... or Ctrl-F and the Count operation:

1) Find the frequency of 2 terms in the Positive feelings and moods category.
2) Find the frequency of 1 term in the Negative feelings and moods category.
3) Find the frequency of 2 terms in the Positive recommendations category.
4) Find the frequency of 1 term in the Negative recommendations category.
5) Find the frequency of 2 terms in the Positive behaves category.
6) Find the frequency of 1 term in the Negative behaves category.
7) Find the frequency of 2 composite terms (e.g., Los Angeles).

Please open the website of the top 10 Los Angeles hotels in TripAdvisor.com's Best Value ranking (this rating does not have to match the hotels in Table 6). Download the web page of an OTR from each hotel with the option Save as HTML only (plain HTML file) and save them in an empty folder (MyFolder). Open all 10 HTML files with Notepad++. Run Search - Find in files. If the student masters the regular expressions, he/she can solve the exercises in thousands of OTRs at once. Otherwise, you need to go step by step.

8) To simplify the exercise, we will place a line break before and after the HTML tags. For example:
Find what: <title>, Replace with: \r\n<title>, Filters: *.htm *.html, Directory: path to \MyFolder, Search mode: Extended, and run Replace in files.
Find what: </title>, Replace with: </title>\r\n, Filters: *.htm *.html, Directory: path to \MyFolder, Search mode: Extended, and run Replace in files.
Ditto in front of <meta name description and in front of <meta property al:ios:url, as well as behind '>' and behind '/>'.

9) Find what: <title>, Filters: *.htm *.html, Directory: path to \MyFolder, Search mode: Normal, and run Find All. Ditto for the description and al:ios:url tags.
10) Extract the names of the 10 hotels.
11) Extract the codes from the 10 hotels.
12) Extract the number of reviews from each hotel.
13) Extract the number of photos from each hotel.
14) Extract the titles from the 10 OTRs.
15) Extract the codes from the 10 OTRs.

https://digitalcommons.usf.edu/m3publishing/vol18/iss9781732127586/7
DOI: https://www.doi.org/10.5038/9781732127586

Application Exercises in Excel (5)

The sample files LA_UTF8.csv and LAl_UTF8.csv are available **(pending)**. The LA_UTF8.csv file contains a sample of 10,000 AirBnB OTRs from the city of Los Angeles posted during 2018 and 2019. The LAl_UTF8.csv file contains a sample of 1,000 AirBnB listings for the city of Los Angeles in January 2020.

The following instructions correspond to the Microsoft Office Excel 2016 version.

Import the files LA_UTF8.csv and LAl_UTF8.csv into Excel
Excel - Data - Get External Data - From text - Text import wizard
Step 1: Delimited - 65001: Unicode (UTF-8) - My data has headers - Next
Step 2: Delimiters (Semicolon) - Text qualifier (none) - Finish

Count items from LA_UTF8.csv and LAl_UTF8.csv files
Excel - Insert - Tables - Pivot Table - New Worksheet

Exercise 1 (Ranking by Popularity)

Import Table 6 into Excel. Number the hotels in a city and build a combined hotel popularity ranking by number of OTRs per room, taking into account the data from TripAdvisor together with the data from the three OTAs. One method can be Borda's count function (de Borda, 1781). There is an example of an application of Borda's method in Marine-Roig (2021b).

Exercise 2 (Weighted Average Score)

Import Table 6 into Excel. Number the hotels in a city and calculate the weighted average of the score for each hotel, taking into account that Booking scores from 1 to 10, and the other platforms score from 1 to 5. Also, calculate the weighted average score for the set of six hotels in each city. Weighted score means that the weight of the score varies depending on the number of OTRs.

Exercise 3 (Scored Features)

Import the LAl_UTF8.csv file into Excel. Find out what features or amenities of the accommodation AirBnB guests score. Check if the average of the specific scores equals the overall score.

Exercise 4 (Tourist Seasonality)

Import the LA_UTF8.csv file into Excel. Find out if the influx of AirBnB guests varies by month, quarter, or season.

Exercise 5 (Language Recognition Accuracy)

Import the LA_UTF8.csv file into Excel. Select a random sample of OTRs using the Excel RAND() function. Please check what percentage of reviews are correctly classified by language. With the auto detect option, you can do the checks in:

https://translate.google.com/
https://www.bing.com/translator/
https://translate.yandex.com/

https://digitalcommons.usf.edu/m3publishing/vol18/iss9781732127586/7
DOI: https://www.doi.org/10.5038/9781732127586

Test Bank Questions (20)

1. Scholars have studied city image formation since:

    a) 1960s
    b) 1970s
    c) 1990s

2. What decade began scientific analysis of online travel reviews?

    a) 1990s
    b) 2000s
    c) 2010s

3. About how many online travel reviews are hosted on travel-related websites?

    a) One million (1,000,000)
    b) A billion (1,000,000,000)
    c) A trillion (1,000,000,000,000)

4. What is the travel-related website hosting the most online travel reviews?

    a) TripAdvisor
    b) Booking
    c) Ctrip

5. Considering online travel reviews as a data source for research, which tourism sector has been studied the most?

    a) Lodging
    b) Dining
    c) Sightseeing

6. Looking at Gartner's model on destination image formation, travel guidebooks are included in:

    a) Induced sources
    b) Autonomous sources
    c) Organic sources

7. Looking at Gartner's model on destination image formation, destination marketing and management organizations are included in:

    a) Induced sources
    b) Autonomous sources
    c) Organic sources

8. Looking at Gartner's model of destination image formation, word-of-mouth communications are included in:

    a) Induced sources
    b) Autonomous sources
    c) Organic sources

9. Looking at Marine-Roig's model of destination image formation, what aspect related to tourists is central in the hermeneutic circle?

    a) Expectations
    b) Experience
    c) Loyalty

10. In the interrelated and hierarchical destination image aspects model, the prior aspect is:

    a) Prescriptive
    b) Designative
    c) Appraisive

11. In the destination image aspects model, the evaluative dimension is included in the aspect:

    a) Prescriptive
    b) Designative
    c) Appraisive

12. In the destination image aspects model, the spatial dimension is included in the aspect:

    a) Prescriptive
    b) Designative
    c) Appraisive

13. In the destination image aspects model, the behavioral response is included in the aspect:

    a) Prescriptive
    b) Designative
    c) Appraisive

14. Narratives, opinions, pictures, audiovisual files, and ratings shared on social media and based on visitors' experiences traveling, sightseeing, entertaining, shopping, lodging, and dining in a tourist destination is a definition of:

    a) Social media
    b) User-generated content
    c) Traveler-generated content

https://digitalcommons.usf.edu/m3publishing/vol18/iss9781732127586/7
DOI: https://www.doi.org/10.5038/9781732127586

15. What is a naming of non-significant words for textual content analysis?

    a) Keyword
    b) Stop word
    c) Key term

16. Regular expressions (regex) are useful patterns for:

    a) Interpret pictures
    b) Find and replace text
    c) Digitize sounds

17. What is a peer-to-peer (P2P) lodging platform?

    a) TripAdvisor.com
    b) Booking.com
    c) AirBnB.com

18. The average overall rating given by guests to P2P accommodations is:

    a) High
    b) Average
    c) Low

19. What is the most common language in AirBnB online travel reviews?

    a) Russian
    b) English
    c) Chinese

20. Which language has the fewest grammatical inflections?

    a) English
    b) German
    c) French