A data-driven approach to exploring similarities of tourist attractions through online reviews

Grant McKenzie[a] and Benjamin Adams[b]

[a]McGill University, Montréal, Canada;
[b]University of Canterbury, Christchurch, New Zealand

PRE-PRINT

ABSTRACT
The motivation for tourists to visit a city is often driven by the uniqueness of the attractions accessible within the region. The draw to these locations varies by visitor as some travelers are interested in a single specific attraction while others prefer thematic travel. Tourists today have access to detailed experiences of other visitors to these locations in the form of user-contributed text reviews, opinions, photographs, and videos, all contributed through online tourism platforms. The data available through these platforms offer a unique opportunity to examine the similarities and difference between these attractions, their cities, and the visitors that contribute the reviews. In this work we take a data-driven approach to assessing similarity through textual analysis of user-contributed reviews, uncovering nuanced differences and similarities in the ways that reviewers write about attractions and cities.

KEYWORDS
similarity; tourist attraction; user-generated content; topic model; tripadvisor

1. Introduction

A recent Nielson survey reported that 63% of respondents who shopped or purchased in the *travel services or products* category researched the service or product online first (2016). These survey results reflect travelers' increased reliance on online platforms as part of their decision making process. Not only are consumers looking to reviews for opinions on everyday products (e.g., Amazon.com) or restaurant suggestions (e.g., Yelp), but increasingly they turn to travel review sites to help them make decisions on where to travel and what attractions[1] to visit.

Today, these review sites play a substantial role in determining where people travel and the activities that they conduct at those locations (Xiang and Gretzel 2010). Recent research has shown these sites to be incredibly persuasive, with more and more travelers trusting the opinions of contributors (Arsal, Backman, and Baldwin 2008; Sparks, Perkins, and Buckley 2013). The perceived value of these reviews is reflected in the millions of individuals producing and consuming content related to travel destinations and tourist activities. In principle, these sites empower the travel community (Lee, Law, and Murphy 2011), allowing users to bypass financially-incentivized travel agents, activity agents, and tourism bureaus. The travel review platform *Tri-*

Corresponding Author Email: grant.mckenzie@mcgill.ca

pAdvisor, for example, boasts 415 million average unique monthly visitors with over 535 million reviews and opinions contributed at the time of writing (TripAdvisor, Inc. 2017). The shear amount of data being produced and consumed through travel review platforms is staggering, and these contributions offer an unparalleled opportunity to better understand how travelers experience tourism destinations.

Furthermore, the users of these platforms are diverse. For example, although the largest percentage of TripAdvisor users come from the United States, users from around the world are represented as well. Together with the large amount of data available, a heterogeneous user base means that we can explore group-level patterns, such as differences in the travel behavior of users from different countries.

Gaining insight into the opinions and views of travelers towards the cities and attractions to which they travel is important, not just for a better understanding of inter-societal differences and similarities, but also for the tourism industry and governmental sectors related to travel and tourism. As one example, the combined direct and indirect contribution of tourism to the New Zealand economy in 2016 was 22.7 billion NZD or roughly 10% of the gross domestic product (Tourism New Zealand 2016). Knowing which international cities tourists find most similar to Auckland, or which types of attractions are most unique to New Zealand is therefore of considerable value to the New Zealand Tourism board. This information can provide a foundation on which to develop an advertising campaign, target specific demographic groups, or determine where to invest additional resources (Pan, MacLaurin, and Crotts 2007; Hays, Page, and Buhalis 2013).

Common methods of determining where to invest and advertise rely on gathering knowledge of similar cities and unique attractions through time consuming and often expensive survey methodology, asking a small subset of visitors to rate their experiences and provide feedback on their travel (Dolnicar, Laesser, and Matus 2009; Hung and Law 2011). For instance, the Ministry of Business, Innovation and Employment in New Zealand conducts international visitor surveys to understand the travel and spending patterns of visitors to New Zealand.[2] While these methods have proved successful in the past, in this article we present a complementary approach that makes use of the millions of user-contributed textual reviews. This work aims to address some of the image/information gaps found in tourism websites (Marine-Roig and Clavé 2016). Furthermore, this work demonstrates that analysis of nuanced differences in content contributed by actual travelers permits comparison of cities, attractions, and reviewers at a level that is not possible through traditional surveys or government tourism websites. Although this passively-collected, user-contributed content is not necessarily focused on answering highly-targeted questions, such as those found in visitor surveys, we show that these data are useful in exploring themes at an aggregate level to identify overall trends as well as similarities and differences between countries, cities, and attractions.

Through large-scale text-based analysis we demonstrate the significance of specific attractions or categories of attractions in defining how travelers view a city. For instance, do most travelers visit Rome, Italy for a single dominant attraction or a specific category of attractions? Similarly, which cities are most similar to Rome based on the terms, phrases and linguistic patterns of contributing reviewers? Through the investigation of term context, this work also identifies national level difference between travelers, for example how Canadians differ from Americans in their descriptions of the Eiffel tower. This provides further evidence to the commonly held assumption that locals or inhabitants view a city quite differently than visitors (Calantone et al. 1989; Urry 1992).

In this paper we make the argument that a user-contributed, data-driven approach to exploring the similarities between cities and attractions can substantially contribute to travel behavior research. We demonstrate that analysis of content contributed by actual visitors can lead to the exposure of nuanced similarities and differences between places. The primary contribution of this work is presented through the following four research questions. These questions each address similarity through a different lens, focusing on attractions, cities, reviewers and finally a combination of the three.

*RQ*1 Given different categories of attractions in cities around the world, can we determine (a) what the single most prototypical attraction is for a category and (b) what the prototypical attraction is for a city based on the linguistic patterns of contributing reviewers? These prototypes are useful in that they demonstrate the most representative single attraction, regardless of popularity.

*RQ*2 Is the identity of a city based on a common theme and broad set of attractions or it is dominated by one or two prominent attractions? Based on review statistics and linguistic similarity assessment we show that cities can be compared along two dimensions, namely the dominance of individual attractions and the dominance of attraction categories.

*RQ*3 Can similarity between cities be measured based on the popularity, categorical assignment, and textual review contributions to their attractions? We assesses the similarity of each city using a series of weighted, ranked attraction approaches. Combining a number of models, we report on which cities are closest in *attraction-space* and which are most dissimilar.

*RQ*4 Finally, is there a notable difference between travel reviewers from different countries? Provided the same cities and attractions, we identify differences in the subject matter of reviewers from different countries. In addition, we show that travel reviewers from different countries differ in the attractions they identify as similar.

The remainder of the paper is organized as follows. Existing research on related work is presented in the next section. The dataset used in this research along with an overview of the methods are given in section following. The *Representative Attractions* section introduces the idea of prototypical attractions while the *Dominant Attractions* section discusses the importance of individual attractions and categories in defining a city. The *Comparing Cities* segment provides an overview of a weighted approach to measuring city similarity and the *Comparing Reviewers* section further investigates the differences between reviewers from different countries. Finally, we discuss the implications of this research and provide conclusions and directions for future work.

2. Related Work

There has been considerable research focused on the role of online user-contributed reviews. Generally speaking the related work falls into two main categories: research related to (a) non-travel consumer behavior prediction and recommendation systems and (b) trust and credibility of online travel review sites. To the best of our knowledge, our work is unique in that it takes a data-driven approach to assessing the similarities and difference between cities and attractions based on user-contributed travel reviews.

In previous work, online travel review platforms, such as TripAdvisor, have been the basis for investigating patterns in hotel reviews and ratings. Work by (Vermeulen and

Seegers 2009) demonstrated the importance of user-contributed reviews on consumer choice comparing and contrasting positive and negative reviews of experts and non-experts. Banerjee and Chua (2016) found significant differences in rating patterns for various types of hotel travel (e.g., business, family) and independent vs. chain hotels.

Other work on user-contributed travel reviews in general has been aimed at assessing the validity of online reviews as a data source. TripAdvisor in particular has received negative press in previous years related to claims that the reviews are false or inaccurate. Research findings on this subject vary with some claiming these reports to be unfounded (O'Connor 2008) and others exposing suspicious rating behavior (Schuckert, Liu, and Law 2016). Much of this research has focused on the trustworthiness and credibility of these reviews as well as methods for assessing their credibility (Fang et al. 2016; Ayeh, Au, and Law 2013; Filieri, Alguezaui, and McLeay 2015). Still other work has shown that the content of travel reviews have a substantial influence on near or far future travel (Shin et al. 2016).

From a methodological perspective, topic modeling (Blei 2012) (detailed in the Section 3.2.1) and other related natural language processing techniques have been applied to user-generated online contributions to construct thematic search engines (Adams, McKenzie, and Gahegan 2015), trip recommender systems (Lu, Chen, and Tseng 2012; Borràs, Moreno, and Valls 2014) and to assess similarity between places (Kim, Vasardani, and Winter 2017; Preoţiuc-Pietro, Cranshaw, and Yano 2013; Adams and Raubal 2014; McKenzie et al. 2015a). The application of topic modeling to understand the experiences of travelers from long form user generated travel content, such as travel blog entries, has been used for exploratory analysis and to test theory about the phenomenology of tourism experiences (Rahmani, Gnoth, and Mather 2017). However, that work was performed with a relatively small data set of under 2,000 documents. Other studies have used topic modeling on larger data sets to identify patterns in the themes and emotions expressed in travel blog entries (Menner et al. 2016; Adams and McKenzie 2013; Ballatore and Adams 2015).

Work similar in scope to this research has utilized the presence landmarks within a region to help define the region of interest (Zhou et al. 2017), though this work has focused on the extraction of landmarks from natural language text rather than the characterization of a region or city through the topics associated with the place. Other work by Yan et al. (2017) uses word embedding to characterize regions based on the types of points of interest (e.g., nightclub) within the region. While similar, our approach focuses on the linguistic context in which an attraction or city is reference in a review rather than the specific types of attractions. Shin et al. (2017) investigated words and phrases employed by visitors to certain landmarks to develop a destination personality scale. Additional efforts in this area uncovered unique attributes of tourist destinations using online travel reviews as a source (Toral, Martínez-Torres, and Gonzalez-Rodriguez 2017).

On the topic of regional variability in linguistic patterns, current work has explored the use of search engine analytics to predict tourism behavior (Zhang et al. 2017). Additional work on geosocial *check-ins* has shown that nuanced differences in the use of the English language can be used to differentiate locale (McKenzie and Janowicz 2017) and recent work by Gao et al. (2017) demonstrates that linguistic patterns clearly differentiate often vaguely defined regions (e.g., Southern California vs. Northern California).

3. Data & Methods

In this section we give an overview of the data used in this work as well as details on the techniques and methods employed in conducting the analysis.

Table 1.: The focal cities for this research along with the number of attractions and reviews per city.

City	Country	Attr.	Reviews	City	Country	Attr.	Reviews
Athens	Greece	249	25798	Mexico City	Mexico	384	29579
Auckland	New Zealand	172	24129	Milan	Italy	711	32411
Bangkok	Thailand	440	50822	Moscow	Russia	1783	34722
Barcelona	Spain	585	52231	Mumbai	India	401	33370
Beijing	China	984	32900	Munich	Germany	236	26581
Berlin	Germany	652	56341	New Delhi	India	326	31846
Bogota	Columbia	228	11039	New York City	USA	851	90960
Cairo	Egypt	163	14987	Nice	France	149	13487
Capetown	South Africa	136	25412	Oslo	Norway	238	21349
Chiang Mai	Thailand	191	17988	Paris	France	957	84277
Christchurch	New Zealand	141	17449	Rio de Janeiro	Brazil	609	38306
Crete	Greece	482	42801	Rome	Italy	1264	77327
Edinburgh	Scotland	294	44585	Santiago	Chile	412	24538
Glasgow	Scotland	190	29747	Sao Paulo	Brazil	600	28436
Johannesburg	South Africa	108	11662	Shanghai	China	778	27721
London	England	1366	174900	St. Petersburg	Russia	1641	28263
Los Angeles	USA	380	42469	Sydney	Australia	335	48038
Madrid	Spain	632	44056	Toronto	Canada	432	40098
Manchester	England	118	22965	Vancover	Canada	237	29619
Melbourne	Australia	287	43114	Zürich	Switzerland	147	12579

3.1. *TripAdvisor: Attractions, Reviews & Reviewers*

Data on the English language version of the travel review site, *TripAdvisor* were accessed through the web platform in April of 2017. All of the attractions, descriptions, and the 1000 most recent reviews for each attraction listed on the *Top Things to Do*[3] pages for 46 cities were accessed.[4] Two major cities from 23 countries were initially selected to provide variance within a country and between countries. These cities (and countries) were chosen based on their high number of attractions and reviews. The data associated with a specific attraction consists of the name, categories, average rating out of five stars, number of reviews, and popularity ranking within the city. Of the accessed data, there are 114 unique TripAdvisor-defined categories associated with the attractions.[5]

Any city that had fewer than 33 distinct attraction categories was removed from analysis as well as any category that existed in fewer than 10 cities.[6] Additionally, any attraction category that consisted of fewer than 10,000 words when the reviews were aggregated, was removed. The purpose of this data cleaning was to ensure a robust dataset on which to construct similarity models. Restricting the cities and attractions to those with a large number of contributions limits the overall impact of any single review thus reducing some of the bias and impact of *fake* reviews.

All of this reduced the set of attraction categories to 81 and number of cities to 40. After reduction, the analysis in this work makes use of 1,695,333 unique text reviews contributed by 548,573 users to 20,409 distinct attractions within the 40 cities. A list

of these cities along with their associated country, number of attractions and total number of reviews are shown in Table 1. A *review* of an attraction consisted of a contributor identifier, contributor location (user specified), review title, review text, and date of contribution.

3.2. *Corpora & Text Analysis*

Prior to analysis, all reviews were cleaned by removing non-alphanumeric characters and stop words (e.g., the, and, etc.), reducing all text to lowercase, and stemming all words using the Porter stemming algorithm (Porter 1980). Furthermore, all references to the city in which an attraction is located were removed from the reviews (e.g., New York, NYC, etc.). Though every effort was made, it is probable that some colloquial references (e.g., The Big Apple) remained in the dataset.

After cleaning, four corpora were generated from the review data for further analysis. The first corpus was constructed by grouping all text from reviews by category of attraction, regardless of city designation, resulting in a single *bag-of-words* for each attraction category. For example, all reviews for attractions of type *Zoo*, e.g., Auckland Zoo, Sydney Zoo, Vancouver Zoo, were combined into a single document representing zoos. We will refer to this as the C_{Cat} corpus (e.g., C_{Zoos}). The second corpus was constructed by grouping all text from all reviews by city, regardless of attraction category. For example, the document representing *Paris, France* would contain reviews contributed about the Eiffel Tower, Louvre Museum, Seine River, etc. We will refer to this as the C_{City} corpus. The third corpus groups reviews by a combination of both city *and* category resulting in corpus $C_{CityCat}$, for instance, $C_{Paris\_Gardens}$ or $C_{London\_HistoricBuildings}$. The fourth corpus ($C_{Attr}$) aggregates reviews by individual attraction instance, the highest resolution with only reviews contributed about the *Eiffel Tower* being aggregated to a document, for example. Overall, these four corpora form the foundation for the topic modeling analysis discussed in the next section.

3.2.1. Topic Modeling: Latent Dirichlet Allocation.

Latent Dirichlet allocation (LDA) is a form of unsupervised topic modeling that takes a bag-of-words approach to extracting themes or topics in natural language text (Blei, Ng, and Jordan 2003). This approach uses the co-occurrence of words within a document to identify word groups that are often found together. These word groupings result in topics that represent specific themes (e.g., beach terms or words related to music). For each of our corpora, introduced in the previous section, we generated an LDA topic model that would allow us to describe each city, category, or category *in a* city as a probability distribution over the same topic space. For example, this might result in a city such as Vancouver, Canada being described as high in a topic related to *winter sports* but low in a topic made up of *historical artifact* words. These topics are generated in an unsupervised manner. The label *winter sports*, in this case is manually assigned based on observed winter sport related terms. Rome, Italy on the other hand would likely show a very different distribution for these same topics. The Mallet toolkit (McCallum 2002) was used to execute the topic modeling analysis with forty topics.

3.2.2. Word Embedding: Word2Vec.

LDA topic modeling is an approach that examines the overall linguistic pattern and co-occurrence of words both within a document and across documents in a corpus. However, LDA does not consider the context in which a term is used or the structure and organization of the terms surrounding it. Yet, as plainly put by the linguist, J. R. Firth (1957), "*You shall know a word by the company it keeps*," and so it follows for attractions and their descriptions. An alternative approach, Word2Vec (Mikolov et al. 2013), produces word embeddings in a latent factor vector space based on the linguistic context of terms. Thus, a Word2Vec approach to assessing the similarity or difference between tourist attractions focuses on the individual terms within a document, the relationship between those terms, and the terms surrounding the denoted attraction. These contextual terms provide considerable insight into the similarity of attractions. We employ this form of analysis to identify similar entities within our corpora, be they cities or tourist attractions.

3.3. *Measuring Similarity*

Jensen-Shannon distance (JSD) (Lin 1991) is a method for calculating the dissimilarity between two probability distributions of equal size. These probability distributions are the topic value outputs of the LDA models described in Section 3.2.1. The resulting value is bounded between 0 (complete similarity) and 1 (complete dissimilarity). The method is used in this work as it is well suited for symmetrical analysis and has been successfully employed in previous text-based similarity analyses (Hall, Jurafsky, and Manning 2008; Adams and McKenzie 2013).

4. Representative Attractions

In this section we explore the concept of a prototypical attraction as it relates to an attraction category or a city. In this case we refer to a prototype as the most representative or typical instance of either a category or a city. Following from research in cognitive science, we adopt a similarity-based approach to prototype theory where the prototypical instance of a category is represented as an average from a set of exemplars (Rosch 1978).

4.1. *Prototypical Category Attractions*

In addressing **RQ1**, we investigate the categories of tourist attractions that contribute to the tourism profile of a city. Specifically, this section focuses on extracting the linguistic patterns that define a category and using these patterns to identify *prototypical* attractions. In other words, what is the individual attraction that best represents a specific category?

The text corpus C_{Cat} was used to extract a set of topics across all text contributed to all categories. Separated by category, this allows for each category to be described as a distribution across these topics. Provided this set of learned topics, the LDA model was again applied to each attraction in the C_{Attr} corpus individually. Using these same topics allows us assess the similarity between the topic distribution for a category and the topic distribution for individual attractions. The JSD was calculated between all categories and all attractions producing a dissimilarity value for each pair.

The smallest JSD value for each category is then reported as the most *prototypical* attraction for that category.

Contrary to what one might initially expect, most of the prototypical attractions (those that are most similar to the category overall) are not the #1 ranked attraction in that category. Not only is this the case for the most prototypical attraction of a category across all cities, but it is also true within the home city of the attraction. In fact, most tourists would likely find that the prototypical attraction for a city is not necessarily a famous or notable attraction at all. For example, *Afrata Beach* in Crete, GRC is the most prototypical Beach *and also* the 57th ranked beach in the city. Similarly, the *Saqqara Pyramids* in Cairo, EGY are the most prototypical, but third most popular Ancient Ruins behind the *Keops Pyramid* (first) and *Gizeh Plateau* (second). Finally, *Midhope Castle* is the most prototypical, 7th ranked, and least popular castle in Edinburgh, GBR. Table 2 lists a sample of the most prototypical attractions by category.

Table 2.: A selection of attraction categories along with the top prototypical attractions matching those categories.

Category	**Most Prototypical**
Ancient Ruins	Saqqara Pyramids (Cairo, EGY)
Art Museums	The Museum of Contemporary Art (Los Angeles, USA)
Theme Parks	Six Flags (Mexico City, MEX)
Castles	Midhope Castle (Edinburgh, GBR)
Hiking Trails	Seawall (Vancouver, CND)
Sacred Religious Sites	Civico Tempio di San Sebastiano (Milan, ITA)
Zoos	Auckland Zoo (Auckland, NZL)
Flea & Street Markets	Silom Night Market (Bangkok, THA)
Churches & Cathedrals	Chiesa di Sant'Ignazio di Loyola (Rome, ITA)
Aquariums	Sea Life Melbourne Aquarium (Melbourne, AUS)
Fountains	Fontana dei Quattro fiumi (Rome, ITA)
Beaches	Afrata Beach (Crete, GRC)

As one might expect, not all cities contain instances of all categories, but in our analysis all category topic signatures were compared to all attraction topic signatures. This resulted in some interesting findings such as *Watts Towers* being the most similar attraction to Ancient Ruins in Los Angeles, USA. *Watts Towers* are a set of interconnected sculptural structures within a state park and officially categorized as an Architectural Building. Similarly, the most Beach-like attraction in Beijing (which is not near any substantial natural body of water) is *The Olympic Water Park* which is categorized as a Water Park not a Beach.

In a very few number of cases, the most prototypical attraction for a certain category was not labeled as that category at all. In most of these cases, the categories were very similar (e.g., a Science Museum was chosen for the Children's Museum category). In other cases, the global category was arguably the more correct category for the

specific instance. For example, the *Morro da Urca* is categorized as a Hiking Trail, but is identified by our similarity model as a an example of a Geological Formation. Given that this attraction is a mountain in Rio de Janeiro on which people may hike, it makes sense that the model would identify this as a geological formation as well as a hiking trail. These findings support existing work on assigning place types based on linguistic patterns of place reviews rather than preconceived place type vocabularies (McKenzie et al. 2015b).

4.2. *Prototypical City Attractions*

While prototypical attractions are interesting at a categorical level, they can also be extracted for each city. In this case we find the attraction within a city whose topic distribution is most similar to the overall city topic distribution (as extracted from corpus C_{City}). In all of our sample cities, these were not the top ranked attractions but tended to be attractions that afforded a variety of activities. For example, notable streets or corridors were often identified as highly prototypical for a city. Similarly, significant bodies of water that bordered or cut through a city, such as a river or lake, were found to be prototypical. Given the importance of water in the establishment of many cities, it is makes sense that these could be said to best represent the city. The reviews contributed to these attractions also tended to show a high degree of subject variance. Some visitors discuss prominent sites or attractions along the water or road, while other describe activities that can be done near or on the actual attraction. In many cases, the attraction is so clearly tied to the city (e.g., *Lake Zürich*) that reviewers chose to discuss the city as a whole in the review of the attraction. A sample of cities and their representative attractions are shown in Table 3.

Table 3.: Prototypical attractions for a sample of ten cities.

City	**Prototypical Attraction**
Christchurch, NZL	Willowbank Wildlife Reserve
Edinburgh, GBR	The Royal Mile
Los Angeles, USA	Ripley's Believe It or Not Museum
Madrid, ESP	Calle de las Huertas
Mumbai, IND	Kopar Khairane Holding Pond
Munich, GER	Bier-und Oktoberfestmuseum
Santiago, CHL	Museo Violeta Parra
Shanghai, CHN	Nanxiang Old Street
St. Petersburg, RUS	Winter Palace of Peter I
Vancouver, CND	Yaletown Neighborhood

This analysis also produced novel cases. For example, the second most representative attraction for Christchurch, NZ is *Myuna Farm* located in Melbourne, AUS. Moscow, RUS listed a number of attractions in St. Petersburg, RUS among its most representative attractions. Los Angeles, USA had New York City's *Children's History Museum* amongst it's most prototypical attractions. These demonstrate the strength of national or regional ties amongst attractions demonstrating that reviewers often use the same terms and phrases to describe attractions within the same country or region.

5. Attraction Dominance

In the previous section we explored the *prototypical* attractions associated with different cities along with prototypical attractions of specific categories. In this section we broaden this approach to explore **RQ2**, the level to which attractions or categories of attractions contribute to how travelers depict a city. Specifically, we assess cities on a two dimensional scale. The first dimension shows the degree to which a city is heavily defined by a single attraction versus a broad combination of attractions. Concurrently, the second dimension shows the degree to which a single *category* (e.g. beach, museum, monument, etc.) dominates for each of our sample cities.

5.1. *Number of Reviews*

As a first step we rank attractions in each city by the number of reviews that have been contributed to the TripAdvisor page related to that attraction. We use the number of reviews as a proxy for the popularity of each attraction. Normalizing the number of reviews for each attraction to between 0 and 1 for each city allows us to perform pairwise comparisons of cities. We restricted the number of attractions to the top-40 within a city in order to include smaller cities with a limited number of attractions. We then ranked and plotted the normalized number of reviews and calculated the kurtosis, skewness, and area under the curve (integral) for each city. Summing these measures produced a single value for each city allowing us to differentiate those dominated by a single attraction from those better defined by a range of attractions, and any city in between.

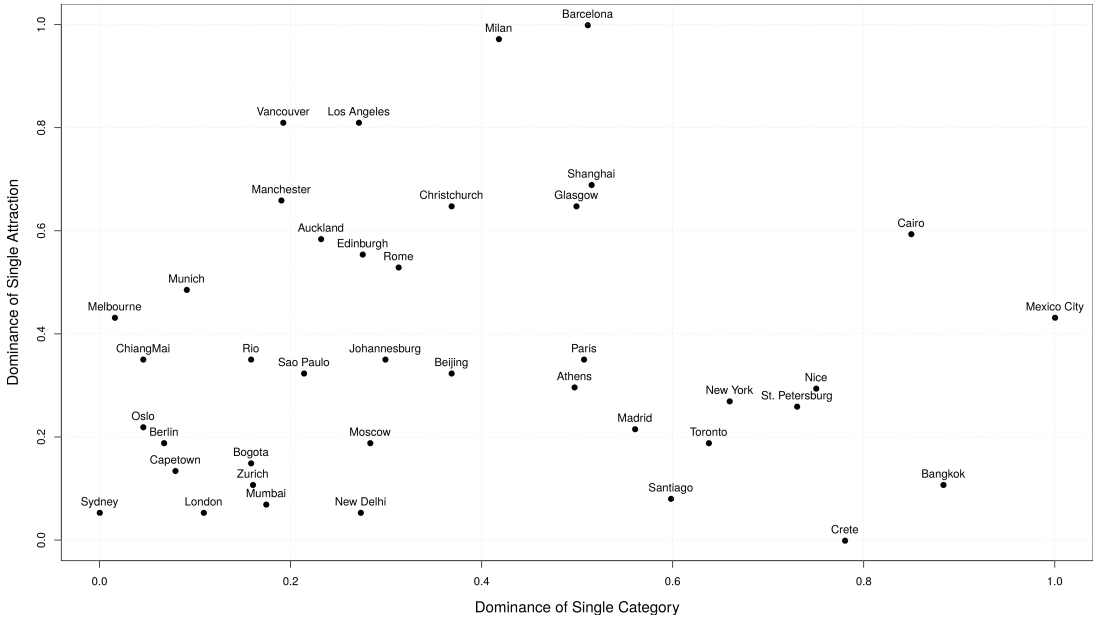


Figure 1.: Cities plotted based on dominance of single attraction instance and dominance of single attraction category. Dominance, in this case, is based on normalized number of unique reviews.

Cities such as Barcelona, ESP and Milan, ITA are dominated by a single attraction (e.g., *Duomo di Milano* in Milan), each receiving close to twice as many reviews as the second most popular attraction. Similarly, cities such as Athens, GRC and Johannes-

burg, ZAF were dominated by two major attractions (e.g., *Gautrain* and *Apartheid Museum* in Johannesburg) with a steep drop in reviews after the top two. Other cities present a smoother distribution of reviews for individual attractions. New Delhi, IND and London, GBR, for example, do not show the same difference in magnitude of reviews, with visits to less prominent attractions also contributing to defining the tourism behavior within the cities.

This review-based approach can also be used to better understand the dominance of attraction *categories* within a city. Some cities are clearly dominated by one or two categories and therefore become focused destinations for a specific group of travelers interested in those types of attractions. For example, reviews in Cairo, EGY and St. Petersburg, RUS are exceptionally high for attractions categorized as *Ancient Ruins* and *Specialty Museums* respectively, and would be well suited for tourists purely interested in those type of attractions. On the opposite end of the spectrum, some cities offer a higher variability in the popularity of their attraction categories. Cities such as Sydney, AUS and Capetown, ZAF are popular in a range of attraction types from *Bodies of Water* to *Architectural Buildings* and *Hiking Trails*. Figure 1 plots each of the cities in our sample set based on the dominance of a single attraction instance (Y-axis) and dominance of a single category (X-axis).

5.2. *Linguistic Analysis*

The number of reviews contributed to each attraction and category within a city provide insight into what draws tourists to certain cities but more refined analysis examines the context of the words and phrases that travelers use to describe each city and its attractions.

Taking the previously generated topic model that describes each *city* as a distribution across the generated topics, we next run an LDA model through the attraction C_{Attr} corpus, using the topic vectors generated from the C_{City} corpus. This again describes each attraction in our set as a distribution across topics, but with the same set of topics used for the city-level topic model. This allows us to compare cities topic signatures to attraction topic signatures and assess the similarity.

In comparing these two sets, we are able to determine not only which attraction is the most similar to the city as a whole, but also measure the variance in similarity across all attractions in a city. In essence, this tells the degree to which the online description of a city is driven by one or many attractions. Figure 2 shows similarity patterns for five cities in our dataset. Sao Paulo, BRA, for instance, has fewer TripAdvisor attractions than London, GBR but shows a significant drop off in similarity between the city topic signature and the attractions within the city (0.81 - 0.30). This implies that the overall description of Sao Paulo is predominantly driven by a few key attractions. London, on the other hand, shows much less variation in the similarity of attractions to London overall (0.89 - 0.46) indicating that many different attractions contribute to the overall description of London. Shanghai, CHN; Barcelona, ESP; and New York City, USA all demonstrate different *fall-offs* in similarity, with Shanghai showing a relatively high similarity to a large number of attractions and few very low in similarity.

Using this same approach to compare *categories* instead of individual attractions, we again assess the variance in JSD between each city and reviews aggregated at the category level. Again, we find a range of variance values with cities such as Capetown, ZAF and Glasgow, GBR showing low variance values and high means while Athens,

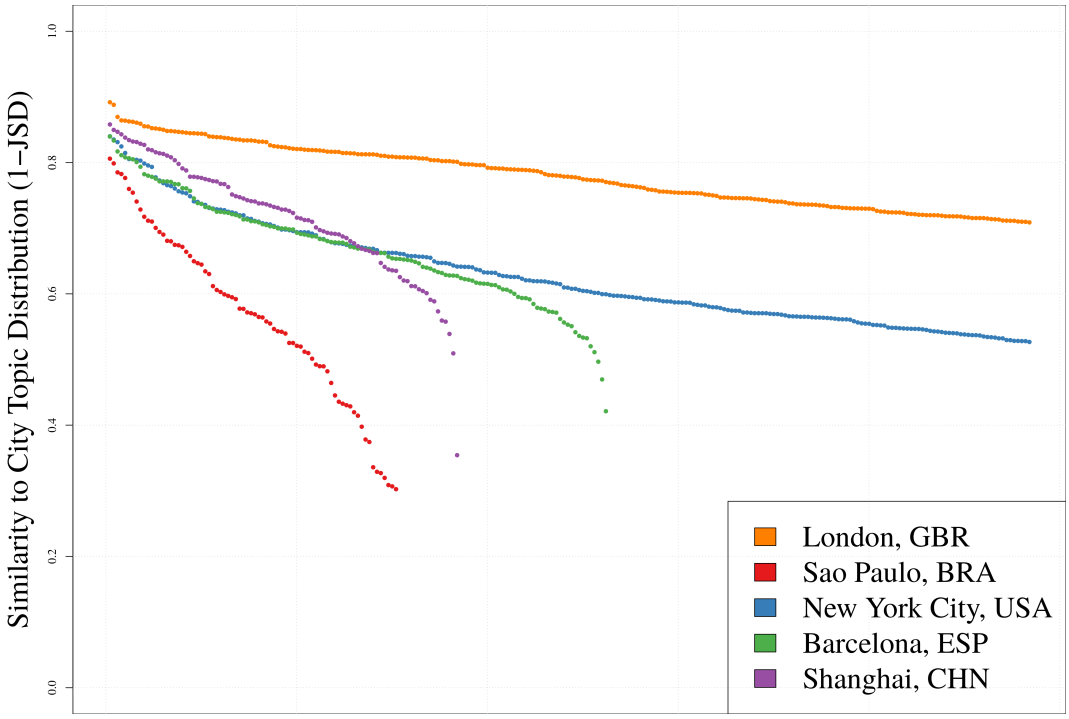


Figure 2.: Similarity of attraction topic signatures to city topic signatures for five sample cities. The x-axis is each individual attraction within the city ordered by similarity (1-JSD).

GRC and Bogotá, COL report a larger variance. The measure of the variance across categories again indicates the degree to which a city can be understood by one or two dominant categories or is more broadly defined through a contribution of attractions from a wide spectrum of categories. This linguistic analysis is more nuanced and influenced by the term choices made by reviewers of attractions instead of review counts, hence there is some disagreement between the attraction and category dominance in this analysis and the review-count-based analysis presented in the previous section.

6. Comparing Cities

In this section we decrease our spatial resolution and look at how attractions affect the similarities of cities as a whole. In addressing **RQ3**, we present a number of approaches for assessing the similarity of cities based on the categories, rank, and text reviews of the attractions within each city.

6.1. *Category Occurrence*

An initial method for quantifying a city by its attractions is to examine the distribution of attraction categories within the city. To do this, we first created a list of all possible unique categories across cities. For each city, the occurrence of an attraction from each category in the list is counted. This array of category counts is normalized to account for different numbers of total attractions across cities and JSD is used to measure the dissimilarity between cities based on these attraction category occurrence arrays.

For instance, using a selection of three cities, we find that based on the count of each category, Vancouver, CND and Toronto, CND have a dissimilarity measure of 0.088 while Vancouver and Chiang Mai, THA (the most dissimilar cities) have a JSD value of 0.357. On further examination of the attraction category distribution within each of these cities we see that Chiang Mai has many more *Beaches* than either Vancouver or Toronto while both of the latter cities have more *Art Galleries*. This initial approach to comparing cities is well suited to travelers that are specifically interested in a select few categories of attractions as it places emphasis on the count in each category.

6.2. *Overall City Topics*

While the first approach offers one method for assessing the similarity of cities based on TripAdvisor-defined categories, it is overly strict in comparing these attraction categories. For example, it assumes that *Art Galleries* and *Art Museums* are as similar to one another as *Art Galleries* and *Biking Trails*, which humans inherently understand is not the case. In this second method, each city is approached as an aggregate of the terms and phrases used to define the attractions within a city. This removes the strict categorical constraint and allows for flexibility through exploitation of the nuanced differences between cities based on how individuals describe attractions.

To this end, an LDA topic model is constructed to extract topics from the C_{City} corpus allowing for each city to be described as a distribution across these topics. Each city is compared to every other city by measuring the JSD value between each pair of topic distributions, which produces a dissimilarity matrix for all combinations of cities. For our sample case of Vancouver, the results of this analysis show that Toronto is still the most similar city based on words and terms, yet Chiang Mai is somewhere in the middle with Sao Paulo, BRA now being the least similar city to Vancouver.

6.3. *Category Rank*

This next approach considers the rank of an attraction within the city. TripAdvisor attractions are ranked by contributors to the platform with the more popular and higher ranked attractions rising to the top. We argue that cities with the same categories at higher rankings are more similar than those with the same attractions at lower rankings since more prominent attractions contribute more to the tourist-viewed make-up of the city.

We consider a number of variations for this ranked approach which are outlined in the following sections. All of these employ Equation 1 but vary the dissimilarity weight (w_i) parameter in some way. N represents the total number of attractions in each city and i is the rank from 1 (highest) to 100 (lowest) in each city. We set $N = 100$ in these cases in order to include all cities in our analysis.

$$City_{sim} = \sum_{i}^{N} log(N/i) \cdot (1 - w_i) \tag{1}$$

6.3.1. Boolean Category Rank.

The boolean categorical rank approach assumes $w_i = 0$ if there is a one-to-one category match at the same rank and $w_i = 1$ if the categories do not match at that rank. The

influence of the log function on the rank means that categorical matches that occur higher in rank produce a greater similarity value than the same number of matches lower in rank do. This also means that a high number of matches at lower ranks can also lead to a high similarity value. While this approach does produce a measure for comparing cities, it is categorically strict and does not consider nuanced similarities between categories, ignoring any category mismatch.

6.3.2. Similarity Weighted Category Rank.

Acknowledging the issues with the boolean approach, the similarity weighted category rank method considers the nuanced similarities between categories of attractions as weights. Building an LDA model based on the C_{Cat} corpus, we compare all categories to all other categories producing a JSD dissimilarity matrix for all combinations of categories. The category-to-category JSD values are include here as weights w_i in our model. When two cities are compared, the similarity between $City_A Attraction_i$ and $City_B Attraction_i$ is applied to the model as a weight. This means that if two categories are the same at a certain rank, w_i is set to 0 (no dissimilarity) while two very different categories produce a w_i close to 1.

This approach is more sensitive to differences in ranked attraction categories than the boolean approach but still does not allow for any slight categorical rank difference between two cities. For example, $City_A$ may have a *Beach* at rank 2 while $City_B$ lists a *Beach* at rank 3. To account for this, we introduce a moving window comparison between city attraction ranks. w_i is the minimum value of a moving window which includes the category of the attraction one rank higher and one rank lower for each city. In other words, this is the minimum weight value of a 3×3 category comparison.

6.4. *A Combined Model*

Each of the previously listed approaches for quantifying a city based on attractions uses its own unique dimension of the data and results in a slightly different similarity value between cities. To account for these differences we combine the different methods into one single similarity measure. First, all values within each dissimilarity matrix are normalized to between 0 and 1. The values are then averaged across each dataset to produce the final combined dissimilarity matrix. Ward's method of hierarchical clustering (Ward Jr 1963) is applied to the data showing the similarities between the different cities based on the combined approach to assessing similarity. Ward's method is a general agglomerative hierarchical clustering approach which merges pairs of clusters in a step-formation based on a distance (similarity) matrix. Figure 3 shows the step-based clusters from no clustering at step 0 to two clusters at step 4.

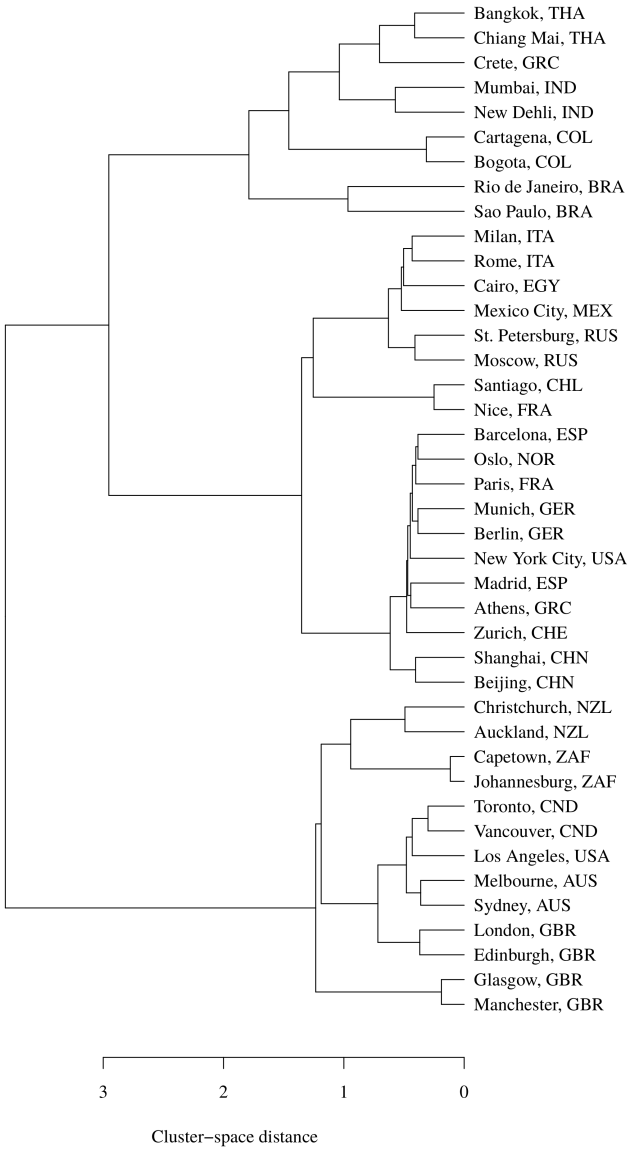


Figure 3.: Hierarchical dendrogram of cities using Ward's clustering.

This hierarchical graph demonstrates a number of interesting city similarities. In general, we find a high degree of similarity between cities within the same country (e.g., China, South Africa, Columbia, Germany) and non-English speaking cities (e.g., Nice and Santiago, Moscow and Milan). In other cases, city similarity appears to match with major tourism capitals that have similar distributions of certain categories (e.g., Rome, Athens, Cairo are high in *Ancient Ruins*) and in some cases Commonwealth of Nations countries (e.g., South Africa, New Zealand, Canada, Australia). Canadian and Australian cities are grouped together which supports previous research on the high degree of similarity between these two countries (Mackay 2014; Wattenberg 1982). We also find Los Angeles in this group whereas the other U.S. city, New York City, is found to be more similar to other non-English speaking cities such as Paris and Mexico City, supporting the notion that NYC is a very much a *melting pot*. There are also some

curious findings as well such as Crete, GRC being very dissimilar to most other cities. This is possibly due to it's unique feature of being the only island city in the dataset.

7. Reviewer Regional Characteristics

A final area of interest for travel behavior research is gaining a better understanding of the tourists who visit these sites and contribute reviews. While the previous sections focus on reviewers as a single contributing entity, in this section, we split reviewers based on their home country and examine the differences between how tourists from different countries approach attractions and cities.

7.1. *Reviewer Location*

Our sample set of roughly 1.7 million reviews was contributed by 548,573 users, with a mean number of 2.8 contributions per user and a median of 1. This follows the typical long-tail distribution often found with user-contributed content where a small number of people contribute a lot while a large number contribute little. Aside from the data cleaning mentioned in previous sections, all reviews contributed from the generic user named *A TripAdvisor Member*[7] were removed as they do not include a georeferencable location.

The GeoNames geocoding webservice[8] was used to identify the home country of each remaining contributor. The *location* field offered to reviewers is a free-text form, allowing for any information to be entered. While most users entered georeferencable location information, 17.1% of reviewers did not enter a location or what they entered could not be reliably geocoded, e.g., `Location=Earth`. Of the remaining reviewers, 35.4% claimed to be from the United States, 15.7% from the United Kingdom of Great Britain, 5.5% from Australia, and 4.7% from Canada. Given that TripAdvisor is an American-based travel website primarily written in English, the dominance of reviewers from these locations is to be expected. Figure 4 shows a sample of six cities split by the ratio of countries contributing reviews. The dominant contributing visitors in the four countries with the highest overall contributions are from the city's own country. Interestingly, the third highest contributing country to Sydney, AUS is New Zealand, showing that there is a proximity effect. Examining cities outside of these four top reviewing countries, we find that Rome, ITA and Beijing, CHN reviewers are made up of roughly 10% locals from their respective countries, the third highest contributing regions.

7.2. *Thematic Differences*

We next focus on **RQ4**, namely investigating the differences between reviewers from different regions. In this section, we specifically center on the nuanced differences in the words, terms, and topics used to describe our set of cities and their attractions. All review text were first aggregated by a combination of city-of-interest and home country of the reviewer. Analysis was restricted to the top four contributing countries, namely the United States, United Kingdom, Canada, and Australia in order to ensure there was enough text to generate robust and accurate topic models. The initial LDA topic model included *all* of the text in the review corpus regardless of contributor location. As with the topic models in previous sections, this produced a range of topics, from

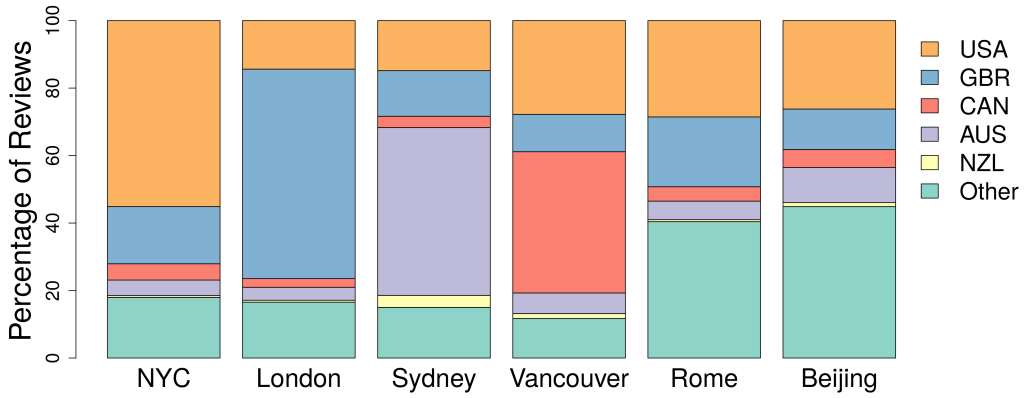
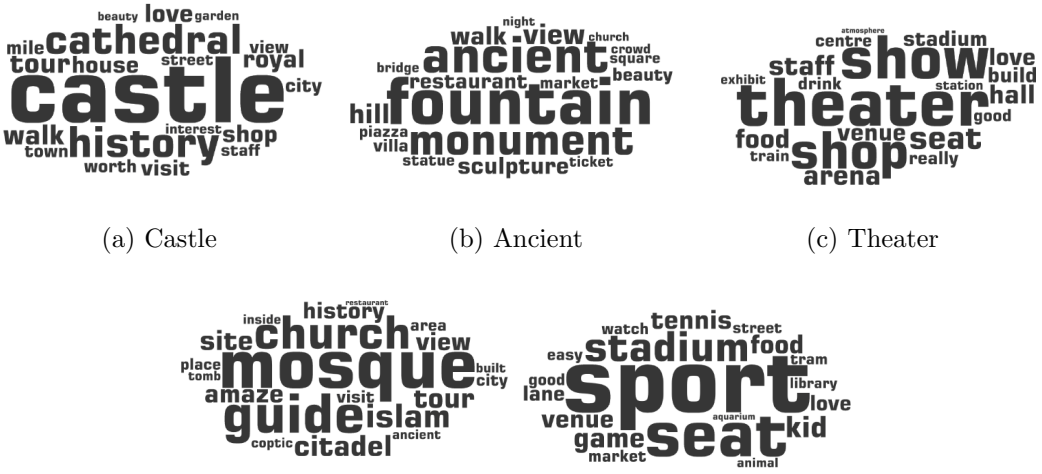


Figure 4.: A sample of six cities shown with the percentage of contributors from the most popular contributing countries. Note that contributors from Italy and China make up roughly 10% of reviews contributed to Rome and Beijing respectively.

those comprised of a high number of adjectives and adverbs to topics high in activity-related terms. Five of these topics are depicted as word clouds in Figure 5.

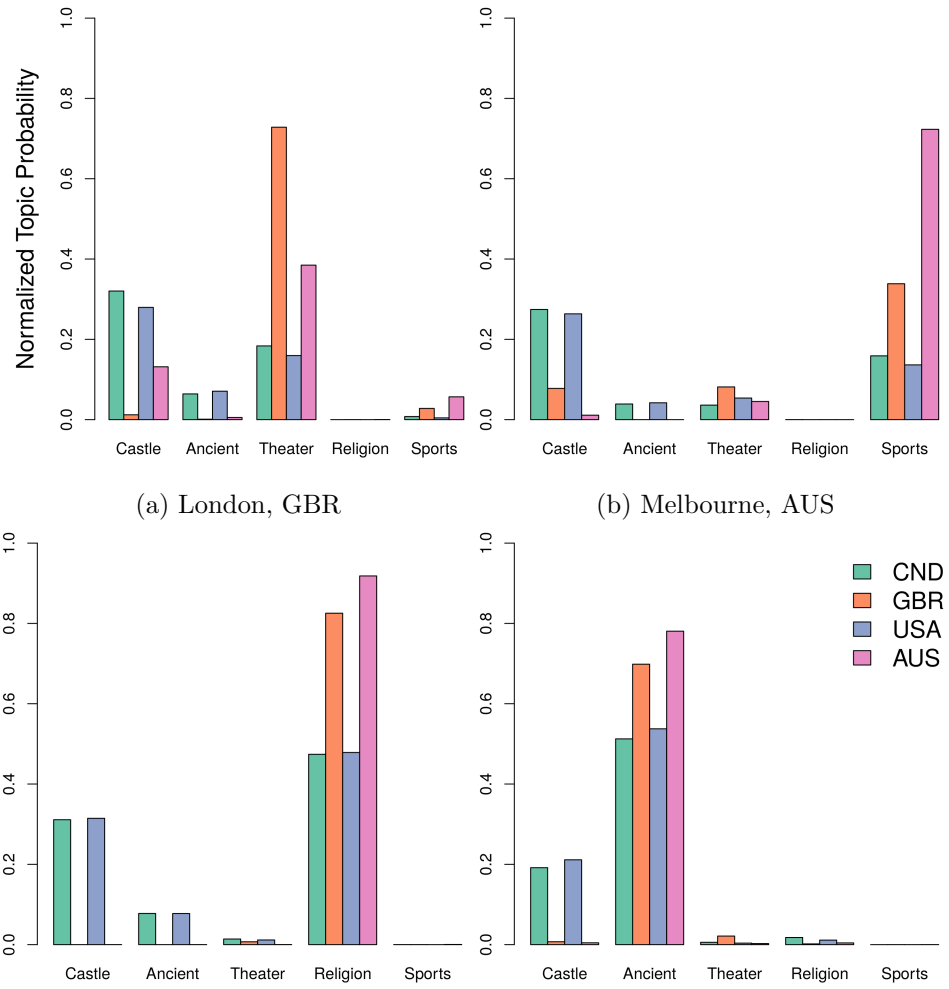


(a) Castle (b) Ancient (c) Theater

(d) Religion (e) Sports

Figure 5.: Five topics visualized as word clouds showing the top twenty terms associated with each topic (ranked by size). Summarizing labels have been assigned for easy reference.

The individual country-specific corpora were then trained on the set of topics produced from the *all-reviews* LDA model. Each model constructed a set of country-specific distributions for each city, across all topics. For example, reviews from Canadians contributed to London, GBR generated a unique distribution over 40 topics whereas reviews from American generated a different distribution across that *same* set of topics. This allows us to compare contributors across the same topic space. Figure 6 shows these distributions, reduced to the five topics shown in Figure 5. These demonstrate some subtle (and not so subtle) differences in how visitors from different



(a) London, GBR

(b) Melbourne, AUS

(c) Cairo, EGY

(d) Rome, ITA

Figure 6.: Four cities described by their distribution across five topics, split by country of reviewers.

countries think about cities. Take London, GBR (Figure 6a), for instance. Reviewers from the United Kingdom tend to use words that are most often associated with the *Theater* topic, more so than Australians and considerably more than Americans or Canadians. In contrast, reviewers from the United Kingdom use fewer terms related to *Castles* than the other three countries. This supports the argument that points of interest such as castles are less interesting to locals–those that see them more often–whereas an activity such as going to the theater may be more prevalent in the mind of a GBR reviewer.

A similar *local* effect is found in Melbourne, AUS (Figure 6b). For this city, Australian reviewers are far more focused on *sporting* terms than historical castle or cathedral terms whereas the opposite is true for contributors from the other three countries. Notably, in both of these cases, Australians and reviewers from GBR follow similar patterns in their topic probabilities as do Americans and Canadians. Cairo, EGY and Rome, ITA (Figures 6c and 6d) show increased topic probabilities in *religious terms* and *ancient monument*-related terms respectively, with negligible interest on *sports*

or *theater*. In all cases, Americans and Canadians use terms related to the topic of *historical castles* or *cathedrals* much more than Australians or reviewers from GBR which, given that all topic probabilities sum to 1, means the probabilities are slightly decreased in some of the other topics.

These four cities depicted in Figure 6 are those that show the largest difference between reviewers from different regions, as measured via JSD. Cities such as Edinburgh, GBR and Oslo, NOR demonstrated the least amount of measurable difference between contributing reviews. The former being high in topics related ancient monuments and castles and the latter being dominated by topics related to water, art, and museums.

7.3. *Linguistic Similarities*

Understanding the nuanced thematic differences between reviewers from different regions exposes interesting insight into how certain reviewers think about tourist attractions and potentially some of the reasons why they choose to travel to certain locations. Having focused on *differences* in the previous section, we now investigate *similarities*, namely contextual linguistic similarities between attractions and cities. For example, given a famous tourist attraction, what other attractions do Australians find similar and do they differ from those found to be similar by Canadian contributors?

Table 4.: The most similar attractions to ten notable attractions as described by reviewers from four different regions.

Attraction	**USA**	**GBR**
Eiffel Tower (FRA)	Hollywood Sign (USA)	The Shard (GBR)
Big Ben (GBR)	London Eye (GBR)	Tower Bridge (GBR)
Mutianyu (CHN)	Badaling (CHN)	Badaling (CHN)
Opera House (AUS)	Griffith Observatory (USA)	Harbour Bridge (AUS)
Colosseum (ITA)	911 Memorial (USA)	Palatine Hill (ITA)
Sugarloaf Mt. (BRA)	Corcovado Mt. (BRA)	Treptower (Germany)
Table Mountain (ZAF)	Mt. Evans (USA)	Arthurs Seat (GBR)
Brandenburg Gate (GER)	Piazza Venezia (ITA)	Reichstag Building (GER)
Hollywood Sign (USA)	Griffith Observatory (USA)	Parliament Hill (GBR)
Stanley Park (CAN)	Lake Zürich (CHE)	Tiergarten (DEU)
Attraction	**CAN**	**AUS**
Eiffel Tower (FRA)	CN Tower (CAN)	Harbour Bridge (AUS)
Big Ben (GBR)	Fitzroy Gardens (AUS)	Royal Palace (GBR)
Mutianyu (CHN)	Connaught Place (IND)	Juyong Pass (CHN)
Opera House (AUS)	Burrard Inlet (CAN)	Harbour Bridge (AUS)
Colosseum (ITA)	Forbidden City (CHN)	Louvre (FRA)
Sugarloaf Mt. (BRA)	Botafogo Beach (Brazil)	Mitropolis Church (Greece)
Table Mountain (ZAF)	Hollywood Sign (USA)	North Head Park (AUS)
Brandenburg Gate (GER)	Pont Neuf (FRA)	Syntagma Square (GRC)
Hollywood Sign (USA)	Manhattan Skyline (USA)	Harbour Bridge (AUS)
Stanley Park (CAN)	False Creek (CAN)	Darling Harbour (AUS)

Word embeddings are the modeling foundation on which we ask these questions and are based on the assumption that terms, e.g., tourist attractions, are more sim-

ilar if they exist in similar linguistic contexts. Based on this assumption, we employ *Word2Vec*, a robust method for producing word embeddings based on linguistic context (see Section 3.2.2 for further details). This approach allows us to measure the similarity between attractions based on the descriptive terms through which attraction names are referred. The same is possible for cities.

This analysis identifies some interesting country-specific findings. Table 4 shows ten world-famous attractions from ten different countries. Word2Vec was used to find the most similar attractions within each country-specific review corpus. On average, reviewers from each of the four countries find the top most similar attractions to be those from their home country. For example, American reviewers of the *Eiffel Tower* describe it in a manner most similar to the *Hollywood Sign* in Los Angeles, USA. Australians and Canadians find the *Harbour Bridge* in Sydney, AUS and the *CN Tower* in Toronto, CND most similar to the Eiffel Tower, respectively. *The Shard* in London, GBR is deemed most similar based on word embeddings for reviewers from the United Kingdom. There are clearly exceptions to this and in many cases, other attractions within the same city and country as the focal attraction are also found to be quite similar. For example, *Mutianya*, a portion of the Great Wall of China is found to be most similar to other parts of the Great Wall of China by reviewers from all countries except Canada. Overall, these attraction similarities suggests that tourists outside their home country tend to find parallels to attractions within their home country, using similar contextual words and phrases to describe these attractions. This supports existing research on ethnocentric and nationalistic tendencies in online tourism (Reid 2014).

At a city level, we again find similarities and differences between reviewers from different regions. In most cases, reviewers from one country describe the cities from their country in a similar fashion. For example, Canadians describe Toronto, CND as most similar to Vancouver, CND and vice versa. Those from the United Kingdom describe similarities between Edinburgh, GBR; Glasgow, GBR; and London, GBR whereas an Australian's language related to attractions in Sydney, AUS and Melbourne, AUS show high degrees of similarity. The word embedding approach also shows that contributors tend to relate cities outside of their home country back to cities within their home region. London, GBR for example, is most similar to Sydney, AUS for Australians whereas Vancouver, CND is contextually similar to Auckland, NZL.

8. Discussion

The previous sections each focus on a novel approach to measure the similarity or differences between cities, attractions, and travel reviewers around the world. These methods aim to help us answer how tourist attractions influence a traveler's perspective on a city. As there is no absolute quantitative measure for this, we instead chose to report measures relative to other attractions and cities. In many ways, this is akin to the approach taken by advertising agencies and marketing companies: identify similar and competing products. By understanding the differences and similarities between products, an agency can develop strategies to promote their product. From a local tourism perspective this could mean targeting visitors to a city similar to your own or promoting a local tourist attraction in a foreign city high in that category of attraction. Travel forums are littered with messages from travelers requesting alternatives to existing destinations (e.g., Hawaii but cheaper, Seattle with less rain, etc.), and the data-driven approaches presented in this research can address these types of questions.

Methodology aside, this work also exposes some interesting findings. The results of the city similarity analysis show that there is still a high degree of similarity between territories of the Commonwealth of Nations.[9] This is supported, to some degree, in the agreement found between reviewers from those countries. For example, Australia, the United Kingdom, and Canada displayed a higher degree of similarity than the United States to any one of those countries, other than Canada. The similarity between Canada and the United States is presumably due to the influence of spatial proximity, though further research is necessary to confirm this.

Another important finding is that analysis of reviewers from the four countries identified nationally-local attractions as more similar to popular international attractions overall. The results shown in Table 4 indicate that when faced with the task of describing an attraction, reviewers write in a manner that is linguistically similar to an attraction in their home country. From a tourism perspective one could use these findings for nationally-specific targeted advertising (e.g., *"Older than the CN Tower"* - Parisian Tourism Board advertising the Eiffel Tower to Canadian travelers). At a minimum, these results confirm the notion that there is a degree of either conscious or unconscious nationalism or ethnocentrism present in travel review platforms.

Overall, this work demonstrated some of the ways that large-scale linguistic analysis and more broadly, data science research, can be used to supplement existing methods used to differentiate cities, attractions, and travelers.

9. Conclusions & Future Work

Data science has an important role to play in understanding the choices that travelers make when selecting a destination. The rise of user-generated content in the form of attraction reviews offers a remarkable opportunity to investigate the similarities and differences between cities and attractions as identified by visitors to these locales. In this work we presented a range of methods for assessing the similarities between cities, the importance of attractions and categories in defining a city, and the notable differences between reviewers from different parts of the world. Finally, we outlined the importance of these methods to the travel and tourism industry and discussed how the findings of this work can be used to augment existing data collection methods.

The bias of user-contributed content is a limitation of this work. As is the case when working with any user-contributed content, only a certain subset of the travel community contribute to online platforms via ratings and reviews. While many reviewers simply wish to share their experiences via the platform, other reviewers are motivated by negative experiences or financial incentives. Little is known about the actual socio-economic status of the reviewers, since much of the information (including their home location) is self-reported. This makes it difficult to quantify the biases or provide detail on specific demographics (e.g., age or gender). Fake reviews are also a concern in analysis of user-contributed content and while effort was made to reduce the number of fake reviews, future work should continue to explore methods to reduce the impact of these data on similarity models.

Future Work

Future work in this area will focus on including additional sources of data in the analysis. While TripAdvisor is one of the leading online review platforms, other web and mobile applications offer different lenses through which to explore travel experi-

ences. Since TripAdvisor.com is an overwhelmingly English language focused platform, comparing these results with multi-language platforms, and less popular cities or attractions is of considerable interest. Furthermore, this work aims to highlight the value of user-contributed content in assessing city, attraction, and reviewer similarity. Future work in this area should focus on the accuracy of the methods as well as compare these similarity approaches to other similarity measures. Last, future work will compare the findings from the presented methods with traditional survey-based results, with the goal of combining approaches for a more robust knowledge base on which to make travel related decisions.

Acknowledgments

Dr. Adams was partially supported by the Building Research Association of New Zealand through the National Science Challenge *Building Better Homes, Towns and Cities: Ko ngā wā kāinga hei papakainga (BBHTC).*

References

Adams, Benjamin, and Grant McKenzie. 2013. "Inferring thematic places from spatially referenced natural language descriptions." In *Crowdsourcing geographic knowledge*, 201–221. Springer.

Adams, Benjamin, Grant McKenzie, and Mark Gahegan. 2015. "Frankenplace: interactive thematic mapping for ad hoc exploratory search." In *Proceedings of the 24th International Conference on World Wide Web*, 12–22. ACM.

Adams, Benjamin, and Martin Raubal. 2014. "Identifying salient topics for personalized place similarity." In *Research@Locate14*, edited by S. Winter and C. Rizos, 1–12. CEUR-WS. `http://ceur-ws.org/Vol-1142/`.

Arsal, Irem, Sheila Backman, and Elizabeth Baldwin. 2008. "Influence of an online travel community on travel decisions." *Information and communication technologies in tourism 2008* 82–93.

Ayeh, Julian K, Norman Au, and Rob Law. 2013. ""Do we believe in TripAdvisor?" Examining credibility perceptions and online travelers attitude toward using user-generated content." *Journal of Travel Research* 52 (4): 437–452.

Ballatore, Andrea, and Benjamin Adams. 2015. "Extracting Place Emotions from Travel Blogs." In *Proceedings of AGILE*, Vol. 2015, 1–5.

Banerjee, Snehasish, and Alton YK Chua. 2016. "In search of patterns among travellers' hotel ratings in TripAdvisor." *Tourism Management* 53: 125–131.

Blei, David M. 2012. "Probabilistic topic models." *Communications of the ACM* 55 (4): 77–84.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent dirichlet allocation." *Journal of machine Learning research* 3 (Jan): 993–1022.

Borràs, Joan, Antonio Moreno, and Aida Valls. 2014. "Intelligent tourism recommender systems: A survey." *Expert Systems with Applications* 41 (16): 7370–7389.

Calantone, Roger J, C Anthony Di Benedetto, Ali Hakam, and David C Bojanic. 1989. "Multiple multinational tourism positioning using correspondence analysis." *Journal of travel research* 28 (2): 25–32.

Dolnicar, Sara, Christian Laesser, and Katrina Matus. 2009. "Online versus paper: Format effects in tourism surveys." *Journal of Travel Research* 47 (3): 295–316.

Fang, Bin, Qiang Ye, Deniz Kucukusta, and Rob Law. 2016. "Analysis of the perceived value of online tourism reviews: Influence of readability and reviewer characteristics." *Tourism Management* 52: 498–506.

Filieri, Raffaele, Salma Alguezaui, and Fraser McLeay. 2015. "Why do travelers trust TripAdvisor? Antecedents of trust towards consumer-generated media and its influence on recommendation adoption and word of mouth." *Tourism Management* 51: 174–185.

Firth, John R. 1957. "A synopsis of linguistic theory, 1930-1955." *Studies in linguistic analysis* .

Gao, Song, Krzysztof Janowicz, Daniel R Montello, Yingjie Hu, Jiue-An Yang, Grant McKenzie, Yiting Ju, Li Gong, Benjamin Adams, and Bo Yan. 2017. "A data-synthesis-driven method for detecting and extracting vague cognitive regions." *International Journal of Geographical Information Science* 1–27.

Hall, David, Daniel Jurafsky, and Christopher D Manning. 2008. "Studying the history of ideas using topic models." In *Proceedings of the conference on empirical methods in natural language processing*, 363–371. Association for Computational Linguistics.

Hays, Stephanie, Stephen John Page, and Dimitrios Buhalis. 2013. "Social media as a destination marketing tool: its use by national tourism organisations." *Current issues in Tourism* 16 (3): 211–239.

Hung, Kam, and Rob Law. 2011. "An overview of Internet-based surveys in hospitality and tourism journals." *Tourism Management* 32 (4): 717–724.

Kim, Junchul, Maria Vasardani, and Stephan Winter. 2017. "Similarity matching for integrating spatial information extracted from place descriptions." *International Journal of Geographical Information Science* 31 (1): 56–80.

Lee, Hee Andy, Rob Law, and Jamie Murphy. 2011. "Helpful reviewers in TripAdvisor, an online travel community." *Journal of Travel & Tourism Marketing* 28 (7): 675–688.

Lin, Jianhua. 1991. "Divergence measures based on the Shannon entropy." *IEEE Transactions on Information theory* 37 (1): 145–151.

Lu, Eric Hsueh-Chan, Ching-Yu Chen, and Vincent S Tseng. 2012. "Personalized trip recommendation with multiple constraints by mining user check-in behaviors." In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, 209–218. ACM.

Mackay, Tim. 2014. "The remarkable similarities between Australia & Canada." `https://www.linkedin.com/pulse/20141001234402-10016207-the-remarkable-/similarities-between-australia-canada`.

Marine-Roig, Estela, and Salvador Anton Clavé. 2016. "Destination Image Gaps Between Official Tourism Websites and User-Generated Content." In *Information and Communication Technologies in Tourism 2016*, 253–265. Springer.

McCallum, Andrew Kachites. 2002. "Mallet: A machine learning for language toolkit." .

McKenzie, Grant, and Krzysztof Janowicz. 2017. "The Effect of Regional Variation and Resolution on Geosocial Thematic Signatures for Points of Interest." In *AGILE 2017: Societal Geo-innovation*, 237–256. Springer.

McKenzie, Grant, Krzysztof Janowicz, Song Gao, and Li Gong. 2015a. "How where is when? On the regional variability and resolution of geosocial temporal signatures for points of interest." *Computers, Environment and Urban Systems* 54: 336–346.

McKenzie, Grant, Krzysztof Janowicz, Song Gao, Jiue-An Yang, and Yingjie Hu. 2015b. "POI pulse: A multi-granular, semantic signature–based information observatory for the interactive visualization of big geosocial data." *Cartographica: The International Journal for Geographic Information and Geovisualization* 50 (2): 71–85.

Menner, Thomas, Wolfram Höpken, Matthias Fuchs, and Maria Lexhagen. 2016. "Topic detection: identifying relevant topics in tourism reviews." In *Information and Communication Technologies in Tourism 2016*, 411–423. Springer.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* .

Nielsen Company LLC. 2016. "Global Connected Commerce Survey." Insights Report. `http://www.nielsen.com/us/en/insights/reports/2016/global-connected-commerce.html`.

O'Connor, Peter. 2008. "User-generated content and travel: A case study on Tripadvisor. com." *Information and communication technologies in tourism 2008* 47–58.

Pan, Bing, Tanya MacLaurin, and John C Crotts. 2007. "Travel blogs and the implications for destination marketing." *Journal of Travel Research* 46 (1): 35–45.

Porter, Martin F. 1980. "An algorithm for suffix stripping." *Program* 14 (3): 130–137.

Preoţiuc-Pietro, Daniel, Justin Cranshaw, and Tae Yano. 2013. "Exploring venue-based city-to-city similarity measures." In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, 16. ACM.

Rahmani, Kamal, Juergen Gnoth, and Damien Mather. 2017. "Tourists? Participation on Web 2.0: A Corpus Linguistic Analysis of Experiences." *Journal of Travel Research* 0 (0): 1–13. https://doi.org/10.1177/0047287517732425.

Reid, Stuart RM. 2014. "Lost in translation: ethnocentric tendency in website communication." *International Journal of Cultural and Digital Tourism* 1 (1).

Rosch, Eleanor. 1978. "Principles of Categorization." In *Cognition and Categorization*, edited by Eleanor Rosch and B. B. Lloyd, 27–48. Hillsdale, NJ: Erlbaum.

Schuckert, Markus, Xianwei Liu, and Rob Law. 2016. "Insights into suspicious online ratings: direct evidence from TripAdvisor." *Asia Pacific Journal of Tourism Research* 21 (3): 259–272.

Shin, Seung-Hun, Sung-Byung Yang, Kichan Nam, and Chulmo Koo. 2017. "Conceptual foundations of a landmark personality scale based on a destination personality scale: Text mining of online reviews." *Information Systems Frontiers* 19 (4): 743–752.

Shin, Seunghun, Namho Chung, Doyong Kang, and Chulmo Koo. 2016. "How Far, How Near Psychological Distance Matters in Online Travel Reviews: A Test of Construal-Level Theory." In *Information and Communication Technologies in Tourism 2016*, 355–368. Springer.

Sparks, Beverley A, Helen E Perkins, and Ralf Buckley. 2013. "Online travel reviews as persuasive communication: The effects of content type, source, and certification logos on consumer behavior." *Tourism Management* 39: 1–9.

Toral, SL, MR Martínez-Torres, and MR Gonzalez-Rodriguez. 2017. "Identification of the Unique Attributes of Tourist Destinations from Online Reviews." *Journal of Travel Research* 0047287517724918.

Tourism New Zealand. 2016. "About the Industry." Online. http://www.tourismnewzealand.com/about/about-the-industry/.

TripAdvisor, Inc. 2017. "About TripAdvisor: Log Statistics." Accessed 2017-11-24. https://tripadvisor.mediaroom.com/US-about-us.

Urry, John. 1992. "The tourist gaze ?revisited?" *American Behavioral Scientist* 36 (2): 172–186.

Vermeulen, Ivar E, and Daphne Seegers. 2009. "Tried and tested: The impact of online hotel reviews on consumer consideration." *Tourism management* 30 (1): 123–127.

Ward Jr, Joe H. 1963. "Hierarchical grouping to optimize an objective function." *Journal of the American statistical association* 58 (301): 236–244.

Wattenberg, Martin P. 1982. "Party identification and party images: a comparison of Britain, Canada, Australia, and the United States." *Comparative Politics* 15 (1): 23–40.

Xiang, Zheng, and Ulrike Gretzel. 2010. "Role of social media in online travel information search." *Tourism management* 31 (2): 179–188.

Yan, Bo, Krzysztof Janowicz, Gengchen Mai, and Song Gao. 2017. "From ITDL to Place2Vec–Reasoning About Place Type Similarity and Relatedness by Learning Embeddings From Augmented Spatial Contexts." In *ACM SIGSPATIAL 2017*, .

Zhang, Binru, Xiankai Huang, Nao Li, and Rob Law. 2017. "A novel hybrid model for tourist volume forecasting incorporating search engine data." *Asia Pacific Journal of Tourism Research* 22 (3): 245–254.

Zhou, Sha, Stephan Winter, Maria Vasardani, and Shunping Zhou. 2017. "Place descriptions by landmarks." *Journal of Spatial Science* 62 (1): 47–67.

Notes

[1]For consistency, we use the TripAdvisor designation *Attraction* here to indicate points of interest that are popular amongst travelers.

[2]See http://www.mbie.govt.nz/info-services/sectors-industries/tourism/tourism-research-data/ivs

[3]e.g., https://www.tripadvisor.com/Attractions-g28970-Activities-_Washington_DC_District_of_Columbia.html

[4]Custom scripts were written to access the content through the public TripAdvisor web platform.

[5]A current list of TripAdvisor categories is accessible here: https://developer-tripadvisor.com/content-api/business-content/categories-subcategories-and-types

[6]Note that due to the uniqueness of cities, there was a range in overlap of attraction categories between cities.

[7]The name given to those that contributed reviews anonymously

[8]http://www.geonames.org/export/free-geocoding.html

[9]The Commonwealth of Nations is an organization of 52 member states, mostly former territories of the British Empire.