Extracting web data from TripAdvisor as a support for tourism indicators development in Minas Gerais

Rafael Almeida de Oliveira
Fundação João Pinheiro

Renata Maria Abrantes Baracho
Federal University of Minas Gerais

EXTRACTING WEB DATA FROM TRIPADVISOR AS A SUPPORT FOR TOURISM INDICATORS DEVELOPMENT IN MINAS GERAIS

Rafael Almeida de Oliveira[1]
Renata Maria Arantes Baracho Porto[2]

ABSTRACT

The research aims to study the phenomenon called "Big Data" and the possibility of using free web data extraction tools (web scrapers) to help the development of indicators about tourist attractions in Minas Gerais State (Brazil) registered in the world's most famous travel-related website known as "TripAdvisor". After the literature review about the theme Big Data, we used a web scraper tool to collect data from TripAdvisor, searching for key information of Minas Gerais' tourist attractions and turning them into a structured database. Thus, it was possible to extract information such as the division of tourist attractions by categories from the state and municipalities, the number of evaluations, visitors' profiles, satisfaction levels, and the period of most visits at each of the attractions. After that it was made a brief study using the web scraper methodology about the case of Pampulha Modern Ensemble. The attraction was chosen by UNESCO as a World Heritage Site during the research and it was possible to check the impact of its nomination into the TripAdvisor's user's opinions. We expect this methodology to assist the state authorities and municipalities to extract strategic information that is already available on the web at low cost, improving actions and ensuring an improvement in the use of public resources in tourism policies.

Keywords: web scraping, tourism, big data, minas gerais, tripadvisor, information retrieval, information management.

1 INTRODUCTION

Much of the information produced today is in digital format, sustained by multiple users, collected and stored by computers, being organized and extracted only by computational tools (PUSCHMANN; BURGESS, 2014). They end up giving meaning to databases that could not be understood by an individual or a group of individuals. Therefore, tools for "Big Data Analysis" are created. According to Andrejevic (2014, p.1675) "(…) Big Data denotes the moment when automated forms of pattern recognition known as data analytics can catch up with automated forms of data collection and storage".

The Big Data phenomenon enabled managers and decision makers to know their business environments better. Besides that, it transformed the acquired knowledge into an increase in performance and more precise strategies (MCAFEE; BRYNJOLFSSON, 2012). As a complement to this thought, "Big Data brings extremely important insights into the company. However, the manager´s decision and

[1] Master's degree student in Information Science at School of Information Science. Universidade Federal de Minas Gerais (UFMG). E-mail: rafalolbh@hotmail.com
[2] Professor in the department of Theory and Information Management. School of Information Science. Universidade Federal de Minas Gerais, Brasil. E-mail: renatabaracho@eci.ufmg.br

vision, which are much more grounded in data nowadays, will always be essential for the company." (NOVO; NEVES, 2013, p.34, translated by the author).

From this context, it is important to search for new ways of extracting and interpreting digital data, as well as in the tourism sector, to optimize processes and lower public resources in research. As an example, Minas Gerais state (Brazil) has the research into tourist demand – carried out by the State Department of Tourism (SETUR-MG) – as a way to trace the visitors profile and their degree of satisfaction with the tourism products and services in the region. Nonetheless, for being an applied research made from the hiring of individuals who should visit several municipalities on site, the quality and number of questionnaires each year depend directly on budgetary resources which, if not available, end up in the disrupt of the historical series and the difficulty of measuring tourism performance indicators in the state (MORAIS, OLIVEIRA, PAIM. 2015). As a consequence, it becomes essential to look for alternatives to data collection that enable a lower-cost assistance to monitor the impact of the activity and thus, to policy development.

In the age of Big Data, data extraction becomes a major tool for gathering information, mainly from the accumulation of material posted on social network websites and content over the Internet (CHEN; CHIANG; STOREY, 2012) that can help managers in decision making. According to Devika and Surendran (2013, p.278), “most business applications depend on web to collect information that is crucial for decision making process”. Thus, the data extraction in digital format, especially from websites, became one of the most used techniques in Big Data (MARRES; WELTEVREDE, 2012).

Considered the world´s largest travel information sharing site, TripAdvisor becomes a rich source of data for satisfaction analysis with tourist destinations and products. It also enables methodologies that can extract this data and transform them into information of the utmost importance to help public and private managers in decision making.

2 METHODOLOGY

In this study, the extraction of information and data processing for the development of indicators were treated quantitatively, working with the attractions information and Minas Gerais destinations registered on TripAdvisor website, by using a data extraction tool known as "import.io”.

The choice of using this application in this study was made because it is a data extraction tool available for free on the web and mainly for not requiring previous knowledge in programming language. Therefore, it has been considered the most handy tool for this research purpose. For the extraction of information, we used the web version of import.io, which does not require software installation on a computer for its operation.

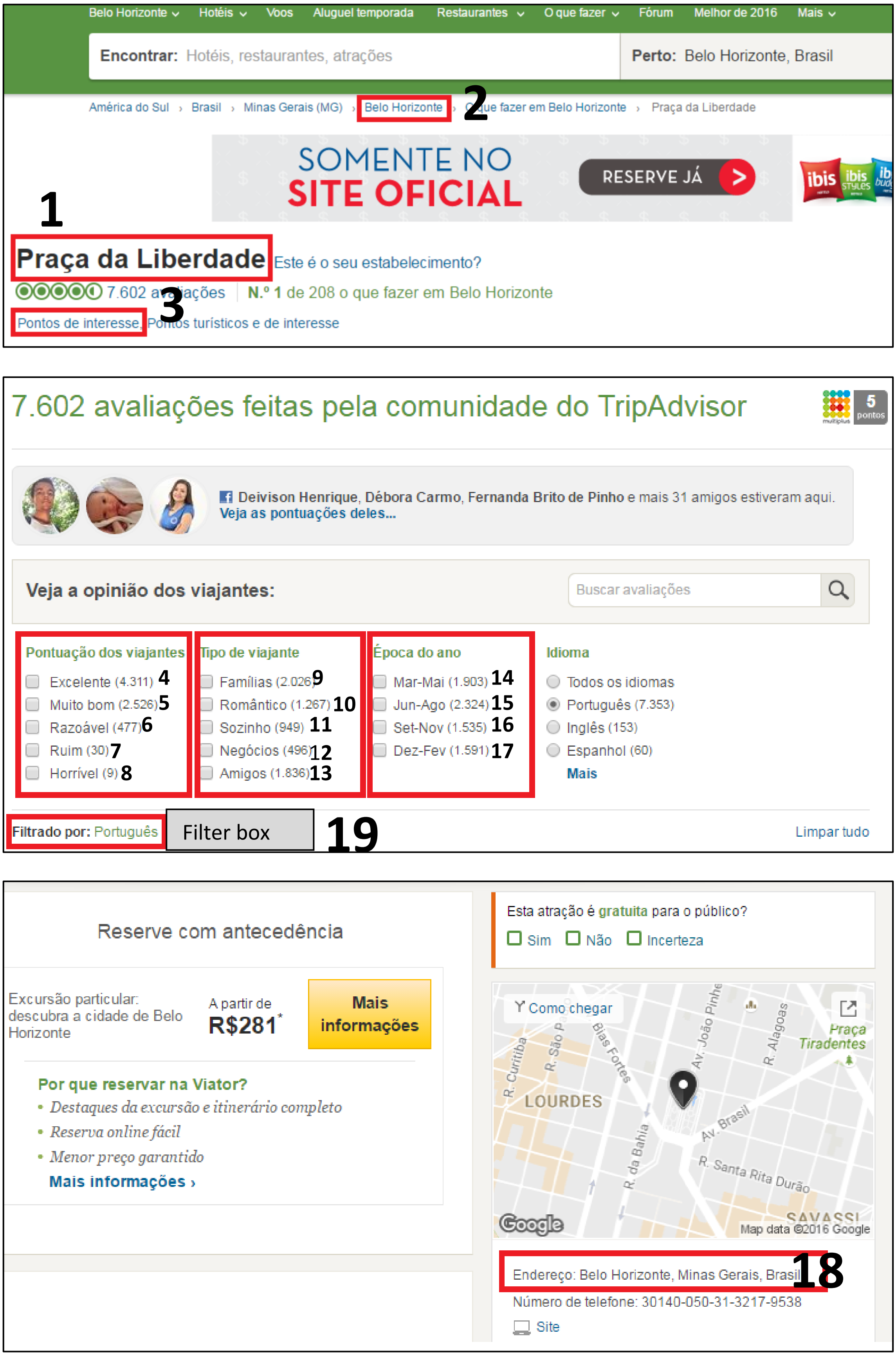
To collect all the selected information of Minas Gerais tourist attractions, a two-step work was carried out. The first one aimed to select information that would be collected from a page template of a Minas Gerais attraction. At the second stage, all URLs from other attractions in the state been registered on TripAdvisor were considered in order to replicate the extraction of the model page information to other pages, enabling the database creation.

The selected information for the capture were highlighted as numbered in Figure 1 and listed as, attraction’s name (1) , municipality (2) , attraction’s category (3), number of ratings by ranges of satisfaction (excellent (4), very good (5), average (6), poor (7), terrible (8)) , motivation for visiting the attraction (families (9), couples (10), solo (11), business (12) or friends (13)) in addition to the travel season (March to May (14), June to August (15), September to November (16) and December to February (17)) and address (18).

It is important to note that in Figure 1, under "3", it shows that TripAdvisor allows viewing more than one option of the attraction’s categorization (in the example, Praça da Liberdade is categorized as "points of interest & landmarks " and "sights & landmarks"). However, it was noticed in some tests before this work’s final extraction that the first classification shown on the page (from which was extracted the information of "attraction’s category" for this work) was considered the most significant to classify the attraction, as the second category is seen as more general than the first, and the third one is even more generalized than the second and so on. Therefore, it was preferred to extract only the first category as it was seen as the most segmented rating for the attractions (in the example it was considered only the rating of "points of interest & landmarks").

Another point highlighted in Figure 1 refers to the numbers ranged between 4 and 17. You can tell that the number of ratings displayed on the site is related to the comments filtered by the language used by the traveler, which in this case was Portuguese (as mentioned in item 19). If the search is carried out on TripAdvisor site in English (www.tripadvisor.com or www.tripadvisor.co.uk), the filter automatically changes it into English, and the numbers shown for each of the reviews would be limited to the reviews in this language, greatly reducing the results related to Minas Gerais attractions.

Figure 1 : Selected information for extraction on TripAdvisor



Belo Horizonte ⌄ Hotéis ⌄ Voos Aluguel temporada Restaurantes ⌄ O que fazer ⌄ Fórum Melhor de 2016 Mais ⌄

Encontrar: Hotéis, restaurantes, atrações **Perto:** Belo Horizonte, Brasil

América do Sul › Brasil › Minas Gerais (MG) › Belo Horizonte › O que fazer em Belo Horizonte › Praça da Liberdade

SOMENTE NO SITE OFICIAL RESERVE JÁ >

Praça da Liberdade Este é o seu estabelecimento?

●●●●◐ 7.602 avaliações | **N.º 1** de 208 o que fazer em Belo Horizonte

Pontos de interesse, Pontos turísticos e de interesse

7.602 avaliações feitas pela comunidade do TripAdvisor

5 pontos

Deivison Henrique, **Débora Carmo**, **Fernanda Brito de Pinho** e mais 31 amigos estiveram aqui.
Veja as pontuações deles...

Veja a opinião dos viajantes: Buscar avaliações

Pontuação dos viajantes
- □ Excelente (4.311)
- □ Muito bom (2.526)
- □ Razoável (477)
- □ Ruim (30)
- □ Horrível (9)

Tipo de viajante
- □ Famílias (2.026)
- □ Romântico (1.267)
- □ Sozinho (949)
- □ Negócios (496)
- □ Amigos (1.836)

Época do ano
- □ Mar-Mai (1.903)
- □ Jun-Ago (2.324)
- □ Set-Nov (1.535)
- □ Dez-Fev (1.591)

Idioma
- ○ Todos os idiomas
- ● Português (7.353)
- ○ Inglês (153)
- ○ Espanhol (60)
- **Mais**

Filtrado por: Português Filter box Limpar tudo

Reserve com antecedência

Excursão particular: descubra a cidade de Belo Horizonte A partir de R$281* **Mais informações**

Por que reservar na Viator?
- *Destaques da excursão e itinerário completo*
- *Reserva online fácil*
- *Menor preço garantido*

Mais informações ›

Esta atração é **gratuita** para o público?
□ Sim □ Não □ Incerteza

Endereço: Belo Horizonte, Minas Gerais, Brasil
Número de telefone: 30140-050-31-3217-9538
Site

3 THE DATABASE ORGANIZATION

Once the information was extracted, it was downloaded from the database to Excel, where the data was organized for further analysis. In total, the data from 1,482 URLs was collected, and 98 (6.6%) was discarded because these were URLs from groups of attractions or services (and not from a single attraction or service), which are posted on some travel destinations pages, redirecting to other attractions.

Thus, it was considered a total of 1,384 attractions of 253 registered municipalities for the database creation. However, it was realized that some of the extracted attractions had a very small number of assessments, which would hinder the analysis of quantitative information, for example, the percentage of people who evaluated the attraction as excellent, very good, average, poor or terrible. Therefore, it was decided to make a cut in the number of attractions from the total number of evaluations of each (a sum of ratings from excellent, very good, average, poor or terrible).

In order to realize this cut, it was considered the average of the number of the attractions ratings. The results showed that on average, each attraction had 106 reviews and therefore it was considered for the analysis of the results only the attractions that had numbers equal to or above that average.

After this cut, 235 attractions of 46 municipalities were considered for the analysis of the results, that is, 17% of the number of attractions and 18% of the municipalities initially extracted.

4 RESULTS

The methodology used enabled to collect information from each of the selected attractions and work with them efficiently in an *Excel* spreadsheet. Overall, it was possible to identify:

- The offer of attractions by municipality and by the attraction’s category (museums, churches, shopping centers, etc);
- The attractions, municipalities and categories with the greatest number of assessments, average scores, besides the percentage of ratings per satisfaction range (excellent, very good, etc);
- The profile of visitors to each city, attraction and category (family, solo, business, etc)
- The time when each municipality, attraction or category received more visitors.

For instance, Table 1 shows which are the main tourist attractions in the city of Diamantina, divided into categories. From the data presented, it can be noted that the "historical sites" category represents 31.33% of the total assessments of the attractions submitted. The category "architectonic works" represents 27.09% of all ratings. Within this category, it can be seen that the Municipal Market (Mercado Central) has more evaluations.

It can also be noticed that from all the city's attractions, the highest score of satisfaction was achieved by Parque Estadual do Biribi and the lowest by Casa da Chica da Silva (3.6).

Table 1 – Sample of information taken from Diamantina city

Attractions	Assessments	Rating	% Families	% Couples	% Solo	% Business	% Friends
Diamantina	**100,00%**	**4,1**	**31,3**	**34,4**	**5,3**	**3,4**	**25,6**
Walking tour areas	13,64%	4,4	30,6	35,0	2,9	3,4	28,2
Vila de Bibiri	13,64%	4,4	30,6	35,0	2,9	3,4	28,2
Historic sites	31,33%	3,9	33,1	34,3	6,5	3,5	22,5
Casa da Chica da Silva	14,00%	3,6	34,0	34,0	6,1	2,8	23,1
Casa de Juscelino Kubitschek	17,33%	4,2	32,3	34,6	6,9	4,2	21,9
Specialized museums	8,61%	3,7	34,3	38,0	5,8	2,9	19,0
Museu do Diamante	8,61%	3,7	34,3	38,0	5,8	2,9	19,0
Architectonic works	27,09%	4,1	29,0	33,9	5,6	3,9	27,6
Casa Gloria	11,27%	4,2	25,6	35,7	6,5	4,8	27,4
Mercado Municipal (dos Tropeiros)	15,82%	4,0	32,5	32,1	4,6	3,0	27,8
Parks	19,33%	4,6	30,2	31,2	3,9	2,8	31,9
Parque Estadual do Biribiri	19,33%	4,6	30,2	31,2	3,9	2,8	31,9
Overal score	**100,00%**	**4,1**	**31,3**	**34,4**	**5,3**	**3,4**	**25,6**

Source: TripAdvisor
Created by the author

Regarding the travelers' profiles, it was observed that Diamantina has a similar distribution in the profiles of people traveling as a couple (34.4%), with family (31.3%) and friends (25.6%).

Another example is the analysis of the main specialized museums in Minas Gerais, with Inhotim representing 35.11% of all assessments (Table 2).

As far as users' satisfaction is considered, it was noticed that Museu Dona Beja in Araxá made lowest score. This result helps directly the public sector and museums managers identify where the points of greater attention are in order to propose improvements. Once satisfaction problems were identified at Museu Dona Beja, a qualitative analysis of the comments on the site was made and it was realized that the museum was closed for renovations for more than a year, causing many users' dissatisfaction. Another example is of Museu do Diamante in Diamantina, about which the biggest complaints were the lack of infrastructure and poor archives. Thus, policies can be created to help improve the museums conditions in a timely manner in order to boost the users' satisfaction over time. If a renovation or improvement in visitation conditions is carried out, the indicator may reflect a rise in the ratings, making it possible to analyze quantitatively the impact of a particular action for such enhancements.

Table 2 - Sample of information taken for the "specialized museums" category in Minas Gerais

Attractions	Assessments	Rating	% Mar-Mai	% Jun-Aug	% Sep-Nov	% Dec-Feb
Specialized museums	**100,00%**	**4,3**	**26,0**	**27,6**	**22,6**	**23,8**
Casa da Memoria de Chico Xavier	1,17%	4,6	31,5	20,6	18,8	29,1
Casa Kubitschek	0,80%	4,4	33,6	27,4	16,8	22,1
Estacao Ferroviaria de Ouro Preto	1,10%	4,2	29,0	23,2	27,1	20,6
Inhotim	35,11%	4,8	26,3	26,9	22,5	24,2
Memorial Minas Gerais Vale	11,83%	4,7	25,9	29,4	24,2	20,6
Museu das Minas e do Metal	7,58%	4,5	24,9	27,6	23,5	24,1
Museu Casa dos Inconfidentes	0,99%	4,6	25,7	25,0	19,3	30,0
Museu Casa Guimaraes Rosa	0,98%	4,4	23,9	23,9	23,9	28,3
Museu da Liturgia	3,02%	4,5	23,2	32,6	20,4	23,7
Museu da Mineralogia	3,56%	4,5	24,9	32,0	23,7	19,5
Museu de Arte da Pampulha	1,54%	3,9	28,0	26,6	20,6	24,8
Museu De Artes & Oficios	3,82%	4,6	23,7	30,6	24,7	21,0
Museu da Escola de Minas/UFOP	1,73%	4,5	23,0	30,7	23,0	23,4
Museu de Sant'Ana	4,24%	4,5	25,4	27,1	23,1	24,4
Museu do Aleijadinho	1,38%	4,2	27,7	25,6	23,1	23,6
Museu do Automovel	3,47%	4,1	23,3	31,5	21,5	23,7
Museu do Diamante	1,01%	3,7	26,1	28,2	21,8	23,9
Museu do Oratorio	3,03%	4,5	25,2	29,2	22,9	22,7
Museu Ferroviario	1,81%	4,3	27,0	24,6	24,2	24,2
Museu Historico Dona Beja	1,72%	3,6	22,6	28,4	22,6	26,3
Museu Historico e Geografico	1,08%	4,2	17,0	27,5	26,1	29,4
Museu Mariano Procopio	2,64%	4,1	31,6	27,8	21,3	19,2
Palacio da Liberdade	6,40%	4,5	25,0	25,0	27,4	22,6
Overal score	**100,00%**	**4,3**	**26,0**	**27,6**	**22,6**	**23,8**

Source: TripAdvisor
Created by the author

This data extraction enabled to see the periods with the highest number of visitors to the museums. Thus, it is possible to check possibilities of creating integrated visitation itineraries to museums that have low ratings during a certain period of time (December to February, for example), creating initiatives such as free events, better dissemination of the spaces, among others, as the case of Memorial Minas Gerais Vale (20.6%), Palácio da Liberdade (22.6%) and Museu de Artes e Ofícios (21%), all located in Belo Horizonte.

5 UNESCO’S WORLD HERITAGE SITE: THE PAMPULHA MODERN ENSEMBLE

The Pampulha Modern Ensemble was created in the 1940s, in Belo Horizonte, in order to foster the development of the capital of Minas Gerais state through the creation of several buildings and gardens angled towards leisure and culture. In addition to the artificial Pampulha Lake, the complex has the old Casino (transformed into the current Pampulha Museum of Art), the Golf Yacht Club, a ballroom, the Kubitschek House and São Francisco de Assis church, all designed by the architect Oscar Niemeyer (in collaboration with several other artists such as Portinari and Burle Marx). It is considered a landmark of modern architecture in the country and served as inspiration for several works around the world, from the potential of reinforced concrete. Its importance as a world heritage site was recognized by the United Nations Educational, Scientific and Cultural Organization (UNESCO) on 17 July 2016.

This date favored analyze a possible impact on the performance of the tourist attractions within the Pampulha area by the methodology presented in this work as the first extraction was taken 05 days before UNESCO’s announcement and the second one, about 20 days later. The comparative data between these two periods allowed us to evaluate the variation in the level of satisfaction and the number of reviews on TripAdvisor.

When comparing the number of evaluations of all 50 attractions registered in Belo Horizonte, it was noted that on average, each attraction has increased their rating between July and August at 5.95%. The attraction "Lagoa da Pampulha" was exactly the one which got the biggest change of all, having an increase of 14.36% in the number of evaluations. It is also noticed that three attractions in the area were among the 11 with the highest rate of change for the period (Parque Guanabara, Igreja São Francisco de Assis e Conjunto Arquitetônico da Pampulha), as shown in Table 4.

By analyzing the average score of the assessments, other results were also important for the Pampulha area. As shown in Table 5, on average, 50 tourist attractions of Belo Horizonte had an increase in performance of 0.04% in August compared to July 2016. If we analyze the 10 attractions with the highest growth rates, 06 of them are directly linked with the Pampulha area (highlighted in bold). Interestingly, within one month after the announcement by UNESCO, there was no infrastructure investment or significant improvements in the region that caused the variation become positive above average. It can be suggested that the announcement meant that the visitors’ appreciation of the attractions within the area has increased. When analyzing the comments of the attractions on TripAdvisor website individually, it was clearly noted references to the recognition of the Pampulha Modern Ensemble.

Table 4 - Number of absolute assessments and the percentage change of the main attractions in Belo Horizonte between July and August 2016 on *TripAdvisor*

Attractions	Assessments Jul/16	Assessments Aug/16	Variation (%)
Belo Horizonte	**48155**	**51018**	**5,95**
Lagoa da Pampulha	**926**	**1059**	**14,36**
Teatro Topazio	151	167	10,60
Edificio Maletta	747	821	9,91
Parque Guanabara	**534**	**585**	**9,55**
Edificio Niemeyer	364	398	9,34
Praca Israel Pinheiro	357	390	9,24
Museu de Ciencias Naturais	138	150	8,70
Mercado Central de Belo Horizonte	6632	7163	8,01
Lagoa Santa	208	223	7,21
Igreja Sao Francisco De Assis	**2057**	**2198**	**6,85**
Conjunto Arquitetonico da Pampulha	**2168**	**2316**	**6,83**

Source: TripAdvisor
Created by the author

Table 5 - Average of ratings and the percentage change of the main attractions in Belo Horizonte between July and August 2016 on *TripAdvisor*

Attractions	Ratings Jul/16	Ratings Ago/16	Variation (%)
Belo Horizonte	**4,260**	**4,261**	**0,04**
Casa do Baile	**4,009**	**4,033**	**0,60**
Parque Guanabara	**3,968**	**3,986**	**0,46**
Lagoa Santa	4,183	4,202	0,46
Praca Raul Soares	3,225	3,239	0,44
Conjunto Arquitetonico da Pampulha	**4,387**	**4,403**	**0,37**
Lagoa da Pampulha	**4,266**	**4,281**	**0,37**
Teatro Topazio	4,464	4,479	0,35
Museu de Arte da Pampulha	**3,872**	**3,885**	**0,35**
Minascentro	4,250	4,261	0,26
Igreja Sao Francisco De Assis	**4,445**	**4,455**	**0,23**

Source: TripAdvisor
Created by the author

This methodology will enable a broader analysis of the results, month by month, to deepen the discussion on future work and see if in fact there is a direct impact of the UNESCO recognition in the heritage concerned, especially when compared to other attractions of the city. Preliminarily it is interesting to note that there is a positive signal of the visitors towards a greater appreciation of the heritage, even without significant changes in the tourist site structure. It is expected that in the long term, there will be greater intervention by public authorities in the area, aiming to maintain the recognition and attract more visitors to the city. Moreover, monitoring the data evolution may serve as a tool to assess the effectiveness of promotional activities or structural improvements.

REFERENCES

ANDREJEVIC, Mark. The Big Data divide. **International Journal of Communication**, nº 8, 2014.

BARACHO, R. M. **A. Sistema de recuperação de informação visual em desenhos técnicos de engenharia e arquitetura:** modelo conceitual, esquema de classificação e protótipo. Tese de Doutorado. Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2007.

CHEN, Hsinchun; CHIANG, Roger HL; STOREY, Veda C. **Business intelligence and analytics:** from Big Data to Big Impact. MIS quarterly, v. 36, n. 4, p. 1165-1188, 2012.

DEVIKA, K.; SURENDRAN, Subu. An overview of web data extraction techniques. **International Journal of Scientific Engineering and Technology**, v. 2, n. 4, 2013.

MARRES, Noortje; WELTEVREDE, Esther. Scraping the social? Issues in real-time social research. **Journal of Culture Economy** (subm), p. 1-52. Goldsmiths Research online, 2012. Available in http://eprints.gold.ac.uk/6768/. Access: 3 mai. 2016.

MCAFEE, Andrew; BRYNJOLFSSON, Erik. Big data. The management revolution. **Harvard Bus Rev**, v. 90, n. 10, p. 61-67, 2012.

MORAIS, Raul S.; OLIVEIRA, Rafael A.; PAIM, Alessandra H. C. A. As Pesquisas de Demanda Turística em Minas Gerais: evolução histórica, principais desafios e perspectivas de futuro. **Revista Turismo e Análise.** V. 26, n.1, p. 21-37. Universidade de São Paulo (USP), 2015.

NOVO, Rafael; NEVES, J. M. Souza. Inovação na inteligência analítica por meio do Big Data: características de diferenciação da abordagem tradicional. **VIII Workshop de pós-graduação e pesquisa do Centro Paula Souza**: Sistemas produtivos: da inovação à sustentabilidade. São Paulo, October, 9th to 10th, 2013.

PUSCHMANN, Cornelius; BURGESS, Jean. Metaphors of Big Data. **International Journal of Communication,** nº 8, 2014.