# Voices from the crowd: the Colosseum in reviews in English and Italian. A comparative analysis.

Sandro Stancampiano[1]

[1]Istat – sandro.stancampiano@istat.it

## Abstract

According to the official statistics, the tourism industry represents an important source of income for Italy. In fact, Italian state museums registered a steady increased revenue up to date despite huge management problems.
In this research, we study the practical use of Reinert's methodology in cultural heritage management, using as a basis the findings of earlier studies.
We want to highlight differences and similarities between the English and the Italian corpus.
We collected all the reviews published on TripAdvisor in the same period written in both languages: English and Italian. Data collection and Corpus creation have been performed with Diogenes.
We had analyzed the two corpus using IRaMuTeQ and compared the main topics covered by users.
In order to identify the more significant contents we applied multivariate techniques in accordance with ALCESTE methodology.
The framework is built to accommodate new capabilities and grow as research needs evolve.

**Keywords:** cultural heritage, tourism, text mining, big data, social network.

## Abstract

La statistica ufficiale attesta l'importanza del turismo quale risorsa economica fondamentale per l'economia italiana. I dati diffusi dall'Istat confermano la crescita costante degli introiti garantiti dai beni culturali, nonostante i numerosi problemi di gestione che gli organi competenti devono affrontare.
In questo studio abbiamo applicato la metodologia di Reinert alla gestione dei Beni Culturali utilizzando i risultati ottenuti in ricerche condotte in precedenza.
L'obiettivo prinicpale è sottolineare punti di convergenza e differenze che emergono dall'analisi delle recensioni scritte in inglese e in italiano e relative alla visita del Colosseo.
Abbiamo raccolto le recensioni e creato due Corpus; uno in italiano e uno in inglese, utilizzando il software Diogene.
Per applicare il metodo ALCESTE è stato utilizzato il software IRaMuTeQ ed infine si è passati alla fase di analisi dei risultati per comparare gli argomenti principali trattati dagli utenti.
Il framework e i passaggi utilizzati sono validi per condurre analisi su altri punti di interesse aggiungendo ulteriori fonti dati. Utilizzando gli stessi software e la stessa metodologia sarà possible replicare lo studio su larga scala in modo da favorire i processi decisionali.

**Parole chiave:** beni culturali, turismo, text mining, big data, social network.

## 1. Introduction[1]

[1] The views expressed in this paper are those of the author and do not necessarily reflect the policies of ISTAT (Italian National Institute of Statistics).

The tourism industry represents an important source of income for Italy; indeed, Italian state museum registered an increased revenue between 2013 and 2018 as evidenced by the statistics released by Mibact Statistics Office (Table1). The increase was particularly marked between 2016 and 2017 when it went up by 12 percentage points and between 2017 and 2018 when it went up by 18 percentage points.

*Table 1 - Visitors and income by year*

| Year | Visitors | Revenue(€) |
|---|---|---|
| 2014 | 40.744.763 | 135.510.702 |
| 2015 | 43.792.162 | 155.494.415 |
| 2016 | 45.383.873 | 173.440.744 |
| 2017 | 50.169.316 | 193.915.765 |
| 2018 | 55.313.772 | 229.631.099 |

*Source: Mibact – Statistics Office, 2019*

Among the museums, the Colosseum is the most visited and generates the highest proceeds. In 2018, more than 7 million visitors went to the Colosseum guaranteeing over 53 million gross income (Table 2).

*Table 2 - 2018 Top 5 visitors in paid museums*

| Museum | Visitors | Revenue(€) |
|---|---|---|
| Circuito Archeologico "Colosseo, Foro Romano e Palatino" | 7.650.519 | 53.829.956 |
| Area archeologica di Pompei | 3.646.585 | 39.639.574 |
| Galleria degli Uffizi e Corridoio Vasariano | 2.004.358 | 20.133.951 |
| Galleria dell'Accademia e Museo degli Strumenti Musicali | 1.719.645 | 10.689.248 |
| Museo Nazionale di Castel Sant'Angelo | 1.113.373 | 11.645.297 |

*Source: Mibact – Statistics Office, 2019*

Considering the role of Internet is increasingly central to various areas of life and thus in handling our trips as Istat data irrefutably confirm, it is appropriate consider web data to improve the accessibility and usability of museums.

In Table 3, we show data provided by “Trips and Holidays”, a focus included in the Istat Household Budget Survey, certify the wide diffusion of Internet as medium of booking travel (ISTAT, 2019). Reservation made using Internet have increased by 14.2% percentage points

(31.8% in 2014 compared with 46% in 2018)[2].

*Table 3 - Trips by type*

| Year | Holidays | Business | Total |
|---|---|---|---|
| 2014 | 30.2 | 42.8 | 31.8 |
| 2018 | 45.5 | 50.3 | 46.0 |

*Source: Istat – Trips and holidays in Italy and abroad, 2019*

Many visitors assign ratings to places adding considerations on the state of conservation of the monuments, services and disservices they have noticed (Stancampiano, 2018b).
We believe that by analysing these comments, it is possible to deduce valuable information to help civil servants and citizens in decision-making processes (Stancampiano, 2018a). In order to understand the needs of people it is mandatory to listen to their voices and social media are a privileged resource for doing so.

## 2. Theoretical Framework

Text-Mining techniques are an efficient way to look at data provided by people and published on social media. The aim of this research is to verify at how techniques as descending hierarchical classification, cluster analysis and correspondence analysis could summarize the huge amount of textual data on digital platform. We gathered data from TripAdvisor and built two corpus using Diogenes trying to extract useful information on cultural heritage related issues.
We applied the ALCESTE method proposed by Reinert to explore both the Corpus (Reinert M., 1995). A set of reviews composes a Corpus: a review represents an Initial Context Unit (ICU). Descending Hierarchical Classification (DHC) allows switching from ICU to Elementary Context Unit (ECU) that is the unit of analysis. Recurrence of certain words or groups of words in the same discursive context is not a random fact, therefore applying correspondence analysis we can identify the main topics in a homogeneous corpus (Greco 2014; Stancampiano, 2019a).
Applying similarity criterion, the algorithm assign each sentence to a cluster that is about a specific topic. IRaMuTeQ provides the user interface to run DHC and produce a list of the most representative segments for each cluster. The sum of the $\chi 2$ value of the words in a sentence establish the position in the ranking within a cluster. The value of the $\chi 2$ is coefficient calculated with one degree of freedom on the contingency table that crosses the presence/absence of the word in an segment with the fact whether or not this segment belongs to the class considered (Camargo B.V. and Justo A.M., 2013; Reinert M., 1995).
The assumption is that the most significant arguments are expressed using always the same words and text segments are the context of the words, therefore reading these sentences it is possible get a general view about topics discussed in each cluster (Bicquelet, 2017).

## 3. Corpus

[2] Tourism, according to Istat definition, is the activity of travelling made by visitors to a main destination outside their usual environment.

We download and store in the Database reviews about the Colosseum published between 2017 and 2018 on TripAdvisor. In the period have been published reviews available in English more than reviews available in Italian. To analyze corpus of the same size we consider a shorter period for English corpus: this corpus contains reviews from December 2017 to August 2018 while the other contains reviews from February 2017 to August 2018.
English corpus consists of 5047 texts, 190938 occurrences and 5048 forms; Italian corpus consists of 5099 texts, 160632 occurrences and 6250 forms.
Frequency (Table 4) helps to identify keywords to explore the corpus (Bolasco S., 2014).

*Table 4 - First 25 active forms by occurrences*

| | Italian Corpus | | | English Corpus | | |
|---|---|---|---|---|---|---|
| 1 | colosseo | 2500 | nr | tour | 2794 | nom |
| 2 | roma | 2431 | nr | colosseum | 1961 | nr |
| 3 | visitare | 1524 | ver | guide | 1830 | nom |
| 4 | fare | 1181 | ver | visit | 1751 | nom |
| 5 | vedere | 1109 | ver | ticket | 1567 | nom |
| 6 | monumento | 1059 | nom | rome | 1388 | nr |
| 7 | storia | 1057 | nom | line | 1295 | nom |
| 8 | visita | 893 | nom | place | 1110 | nom |
| 9 | mondo | 845 | nom | history | 998 | nom |
| 10 | bello | 819 | adj | time | 988 | nom |
| 11 | romano | 611 | adj | amaze | 920 | ver |
| 12 | unico | 564 | adj | queue | 905 | nom |
| 13 | biglietto | 528 | nom | book | 849 | nom |
| 14 | simbolo | 527 | nom | skip | 797 | ver |
| 15 | entrare | 508 | ver | worth | 720 | nom |
| 16 | solo | 498 | adj | walk | 589 | ver |
| 17 | interno | 470 | adj | person | 589 | nom |
| 18 | consiglio | 464 | nom | great | 582 | adj |
| 19 | anno | 439 | nom | buy | 574 | ver |
| 20 | tempo | 432 | nom | inside | 569 | nom |
| 21 | città | 417 | nom | long | 567 | adj |
| 22 | fila | 405 | nom | recommend | 537 | ver |
| 23 | coda | 385 | nom | day | 523 | nom |
| 24 | foro | 372 | nom | crowd | 438 | nom |
| 25 | guida | 367 | nom | hour | 429 | nom |

Observing the 25 active forms listed in Table 4 we find out words similar but not identical. In both cases, there are theme words as city name and monument name. In both cases we note words related on one side with practical actions and on the other side with the description of the monument visited. Differences appear more obvious if we look at the context performing the hierarchical classification to identify patterns and relationships in the reviews.
ALCESTE relies upon co-occurrence analysis finding within the corpus homogeneous subset of representative sentences (Giuliano, 2013).

## 4. Results

Figure 1 below shows the division in subsets proposed by IRaMuTeQ. DHC divided both corpus into four subsets of sentences. Analysing these sentences we get a clearer picture of the visit experience.

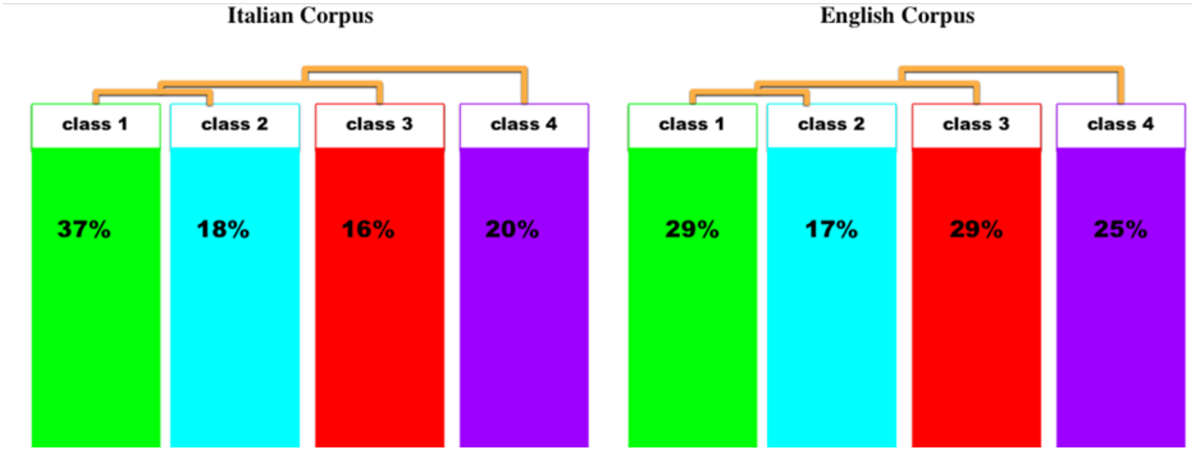


*Figure 1: Corpus division in classes*

The sentences and therefore words strictly correlated with practical aspects as *biglietto*, *fila*, *coda* and *guida* in the Italian corpus and *ticket*, *line*, *queue*, *buy*, *crowd* in the English corpus represent high percentage in both cases. Class 1 account for 37% in the Italian corpus while class 3 account for 28.6% in the English corpus.
We noticed that these words are among the most used but could be more interesting look into the sentences. Reading the sentences, we can better understand what they are talking about, if the visit has been a good experience or maybe they face with problems during the tour.
The main problem is the queue: people tend to find a way to *skip* and *avoid* long *lines*. From the perspective of those who administer cultural heritage it is important understand how people reach the goal. Indeed, we notice that in some cases it may lead to promote behavior undesirable. The need for better management of the booking and purchasing process is clear.
This implies monitoring unauthorised sale of tickets and improving online information channels. No one wants to lost time and money, who offers services, should consider this. Giving more information and more precise both online and on the spot is mandatory to avoid disastrous visiting experiences; in fact, tourists describe in their reviews completely preventable problems.
As stated above practical issues are central in both corpus although reviews available in English also report problems with skip lines solutions, "the line tickets we bought online had

almost 3 hours waiting" and "only delay was the security scanner" without complaining when it comes to security reasons.
We can say that people are willing to spend something more in order not to waste their time in lines "si salta una fila chilometrica prezzo onesto considerato che il biglietto vale 2 giorni".

***Table 5 - Italian Corpus Typical UCE***

| |
|---|
| sicuramente non è luglio il mese ideale per visitare roma ma quello_che offre è talmente tanto e bello che ne vale la pena per saltare una chilometrica coda abbiamo iniziato la visita dal palatino dove abbiamo acquistato il biglietto salta la fila che comprende |
| merita acquistare il biglietto salta fila perché eviti lunghe code e ti rechi direttamente alle casse dalla n 7 alla n 10 per ritirare i biglietti di ingresso e recarti al controllo del metal detector nonostante fosse un giorno di festa il numero dei visitatori |
| consiglio di acquistare on line i biglietti col salta fila e video guida nonostante sia agosto ci sono code lunghissime per entrare i ragazzi poi con le videoguide hanno trovato molto più interessante la visita |
| veramente bello e interessante consiglio di prenotare il biglietto e possibilmente anche visita guidata via internet xché così si salta una fila chilometrica prezzo onesto considerato che il biglietto vale 2 giorni e vale anche x fori romani palatino praticamente ci sono luoghi da visitare |
| per visitare il colosseo occorre fare un biglietto che comprende anche la visita al colle palatino e ai fori imperiali 12 euro il biglietto si può prendere online oppure consiglio di andarlo ad acquistare dove c'è meno fila ovvero all'ingresso del museo palatino per arrivarci |

***Table 6 - English Corpus Typical UCE***

| |
|---|
| skip the line tickets we bought online had almost 3 hours waiting and no less than other normal queue went for group tour instead which was much quicker but had to pay 100 for a family of 4 and 50 we paid for pre booked |
| this place is just out of the world I would definitely recommend for people to buy the ticket in advance online as it_s cheaper and you avoid paying silly money for people who approach you and ask  you to skip the line |
| very interesting and amazing ancient rome largest and most impressive building did the colosseum and forum in one day we skipped the queues with advance tickets bought online the evening before and came there early morning |
| we bought our skip the line tickets from the hop on hop of bus we were able to skip a very long line and walked straight in the only delay was the security scanner the other queue was an hour and a half long be warned we |
| i suggest buying tickets online in advance with skip the line option otherwise you won't be able to visit because of the huge queue the tickets for colosseum includes access to roman forum and palatine hill wear comfortable shoes because you will do a lot |

## 5. Conclusion

The proposed framework can help improve the visit experience. It is a circular path; all began when the user wrote a review for some reason in most cases just to follow a digital ritual and thinking not to be heard (Ho J.Y.C. and Dempsey M., 2008; Ippolita, 2016; Rudder C., 2015).
We can hear to the voices of the crowd by applying the techniques explained. Administrators can respond to expressed problems and translate needs into solutions.
Extraction of information from unstructured data is now possible also in real time using appropriate technologies.
The more you read the more you get; using Diogenes and IRaMuTeQ as tools it is possible replicate this experiment. The next step will be reuse the same procedure with the aim to discover pattern using more data sources about several point of interest.
This project aims, firstly to submit, to the attention of the policy makers, information to guide their actions of government of cultural heritage using social network textual data.

## References

Bicquelet A. (2017). Using online mining techniques to inform formative evaluations: An analysis of YouTube video comments about chronic pain. *SAGE Publications*, Volume: 23 issue: 3, page(s): 323-338.

Bolasco S. (2014). *Analisi Multidimensionale dei dati. Metodi, strategie e criteri d'interpretazione.* Roma: Carocci.

Camargo B.V. and Justo A.M. (2013) *Tutorial para uso do software de análise textual IRAMUTEQ. Universidade Federal de Santa Catrina, Brasil*. [online] Available at: http://www.iramuteq.org/documentation/fichiers/tutoriel-en-portugais [Accessed 22 Jan, 2020].

Giuliano L. (2013) *Il valore delle parole. L'analisi automatica dei testi in Web 2.0*. Roma : Dipartimento di Scienze statistiche [online] Available at: https://www.dss.uniroma1.it/it/node/5868 [Accessed 22 Jan, 2020]

Greco F. (2014). *Integrare la disabilità. Una metodologia interdisciplinare per leggere il cambiamento culturale.* Milano: FrancoAngeli.

Ho J.Y.C. and Dempsey M. (2008). Viral marketing: Motivations to forward online content. *Journal of Business Research,* Volume 63, Issues 9-10, September-October 2010, pp. 1000-1006 [online] Available at: https://doi.org/10.1016/j-jbusres.2008.08.010 [Accessed 27 Jun, 2019].

Ippolita (2016), *Anime Elettriche. Riti e miti social.* Jaca Book.

ISTAT (2019). Trips and holidays in Italy and abroad. Report 11-02-2019, Istat, https://www.istat.it/it/archivio/227018 [Accessed 7 Jan 2020].

Reinert M. (1995). I mondi lessicali di un corpus di 304 racconti di incubi attraverso il metodo "Alceste". In Cipriani R. e Bolasco S. (a cura di), *Ricerca qualitativa e Computer. Teorie, metodi e applicazioni.* Milano: FrancoAngeli.

Stancampiano S. (2018a). Gestire i beni culturali con I Big Data. In Domenica F. Iezzi, Livia Celardo and Michelangelo Misuraca (a cura di), Roma, UniversItalia, *JADT' 18. Proceedings of the 14th International Conference on the Statistical Analysis of Textual Data*, pp. 748-754.

Stancampiano S. (2018b). Misurare, monitorare e governare le città con I Big Data. In Enrico Di Bella, Filomena Maggino and Marco Trapani (a cura di), Fiesole (FI), Genova: UniversityPress, *Libro degli abstract. AIQUAV 2018. Atti del V Convegno dell'Associazione Italiana per gli studi sulla Qualità della Vita. Fiesole (FI), 13-15 Dicembre 2018*, pp. 114-116.

Stancampiano S. (2019a). The monitoring of cultural heritage in real time using Social Media. . In Leonardo Salvatore Alaimo, Alberto Arcagni, Enrico Di Bella, Filomena Maggino and Marco Trapani (a cura di), Fiesole (FI), Genova: UniversityPress, *Libro dei contenuti brevi. AIQUAV 2019. Atti del VI Convegno dell'Associazione Italiana per gli studi sulla Qualità della Vita. Fiesole (FI), 12-14 Dicembre 2019, pp. 241-247*.

Stancampiano S. (2019b). Matera 2019 Text Mining dei Social Network. In Leonardo Salvatore Alaimo, Alberto Arcagni, Enrico Di Bella, Filomena Maggino and Marco Trapani (a cura di), Fiesole (FI), Genova: UniversityPress, *Libro dei contenuti brevi. AIQUAV 2019. Atti del VI Convegno dell'Associazione Italiana per gli studi sulla Qualità della Vita. Fiesole (FI), 12-14 Dicembre 2019, pp. 277-284*.

Rudder C. (2015). *Dataclisma. Chi siamo quando pensiamo che nessuno ci stia guardando.* Milano: Mondadori.