MANAGEMENT OF CULTURAL HERITAGE USING BIG DATA

Sandro Stancampiano

1. Introduction[1]

The focus of this paper is about management of cultural heritage. We want to show the huge potential of data on the web to produce statistics in order to optimize decision-making processes.
We have chosen three of the most important Italian cities of art, characterised by stunning works of art but also by many administrative difficulties.
We build a corpus - in Latin language - composed of recent reviews about St Mark's Basilica (Venice), Colosseum (Rome) and Norman Palace (Palermo).
We want to discover regularity in the text examined by cluster analysis (BOLASCO, 2014).
The tourism industry represents an important source of income for the Italy, as a matter of fact italian state museums registered an increased revenue between 2013 and 2018, according to the statistics released by Mibact Statistic Office (Table 1).

Table 1 – *Visitors and income by year.*

Year	visitors	revenue (€)
2014	40.744.763	135.510.702
2015	43.288.366	155.494.415
2016	45.383.873	173.383.941
2017	50.169.310	193.915.765
2018	55.504.372	229.360.234

Source: Mibact - Statistic Office, 2019.

[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).

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

Table 2 – *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 (including those visiting museums in their city of residence) assign ratings to places, adding considerations on the state of conservation of the monuments, services and disservices they have noticed.
We believe that by analysing these comments, it is possible to deduce valuable information (STANCAMPIANO, 2018).
This paper is structured as follows: in section 2 we present the theoretical framework; in section 3 we describe the corpus and the methods; in section 4 we illustrate the main results; in section 5 we delineate the conclusions and the future steps.

2. Theoretical framework

The texts we are analysing in this study have been collected using the Diogenes software[3].
The Diogenes software is developed by the author in the Java language: allows to store on a relational Data Base texts extracted from the web using web scraping techniques [4].
Within the activities related to this study were automatically collected and stored more than 6000 documents (each document is a review).

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

[3] http://diogenes.statsapp.it/ [Accessed 26 Oct, 2019]

[4] https://en.wikipedia.org/wiki/Web_scraping [Accessed 26 Oct, 2019]

Tripadvisor was chosen among the many websites used by users to produce content. Registered user use the site to write their reviews on the places they went to sharing their experiences and exchanging advice.
The reasons why people publish content on the web are partly unknown; although some potential reasons have been investigated (HO and DEMPSEY, 2008).
We believe that, regardless of the reasons why this happens, studying and analysing this content is an opportunity to be adequately valued (CERON *et al*., 2014).
In this study we applied the methodology proposed by Reinert: the speaker exposes a series of mental places in sequence that impose their vocabulary. The statistical examination of the distribution of this vocabulary makes it possible to identify the "mental rooms" subsequently inhabited by the speaker (REINERT, 1995).
We can identify the main topics of the corpus highlighting the “lexical worlds”: this happens because the recurrence of certain words or groups of words (in the same discursive contexts) is not a random fact.
To understand the content of a text, without reading it, Reinert proposes the study of co-occurrences.
We use the software IRaMuTeQ, created by Pierre Ratinaud, in order to apply the ALCESTE (Analyse Lexicale para Context d’un Ensemble de Segments de Texte) method, that allows us to classify the contents of the corpus analysing the words full (nouns, verbs, adjectives and some adverbs) reduced to their lexeme (GRECO, 2016).
IRaMuTeQ is a graphical interface of the statistical software R developed in the Python language (SOUZA *et al*., 2018).
Applying the method implemented in IRaMuTeQ we obtain the dendrogram in which each class represents a theme of the corpus.
The analyzed corpus is composed of a set of reviews dealing with similar topic, in this case the account of a cultural experience.
Descending Hierarchical Classification (DHC) allows you to switch from Initial Context Unit (each review is an ICU) to Elementary Context Unit (ECU) or text segments. Text segments (TS) are the words context; the corpus division in TS is automatically done.

3. Corpus and Methods

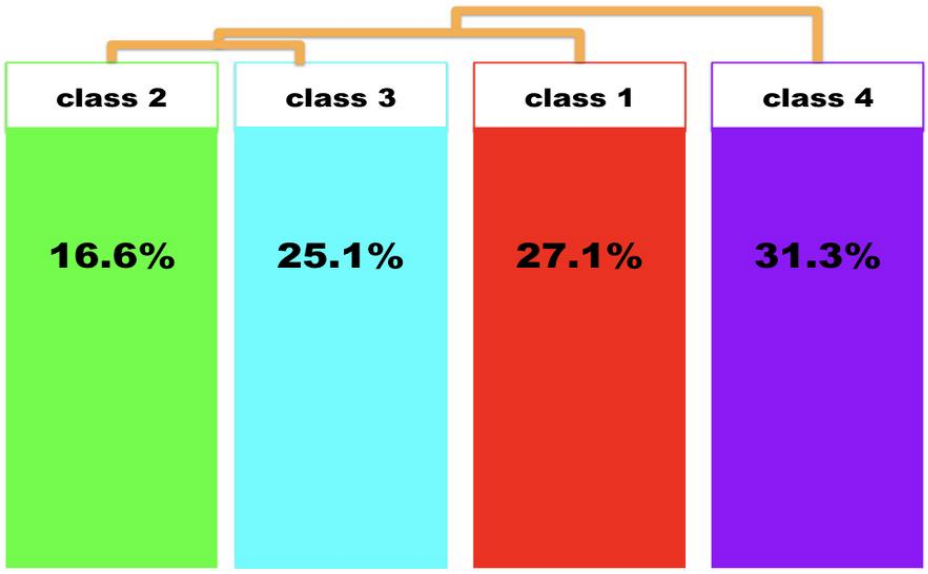
The 6569 reviews collected, published between 2015 and 2019, are divided as follows: Colosseum 2421 (36.8%), Norman Palace 2049 (31.1%), and St Mark’s Basilica 2099 (31.9%).
The corpus analyzed with the purpose identifying different arguments about the

same topic consists of 7125 texts segments, 231229 occurrences and 13122 forms. The number of active forms with a frequency greater than 3 is 2868.

DHC was performed taking into account only full words (adjectives, adverbs, nouns and verbs) and produced a division in four clusters of text segments. Each cluster represents a recurring topic within the texts. The algorithm maximizes similarity between statements in the same class.IRaMuTeQ divided the corpus into two subcorpus classifying ranking 6416 out of 7125, more than 90% a very efficient result (CAMARGO and JUSTO, 2018).

The first subcorpus consists of 2006 ECU that correspond to 31.3% of the total while the second subcorpus had a further division in three classes: class 1 with 1739 ECU (27.1%), class 2 with 1062 ECU (16.6%) and class 3 with 1609 ECUs (25.1%) (Figure 1).

Figure 1 – *Dendrogram of the classes.*



Source: Iramuteq data.

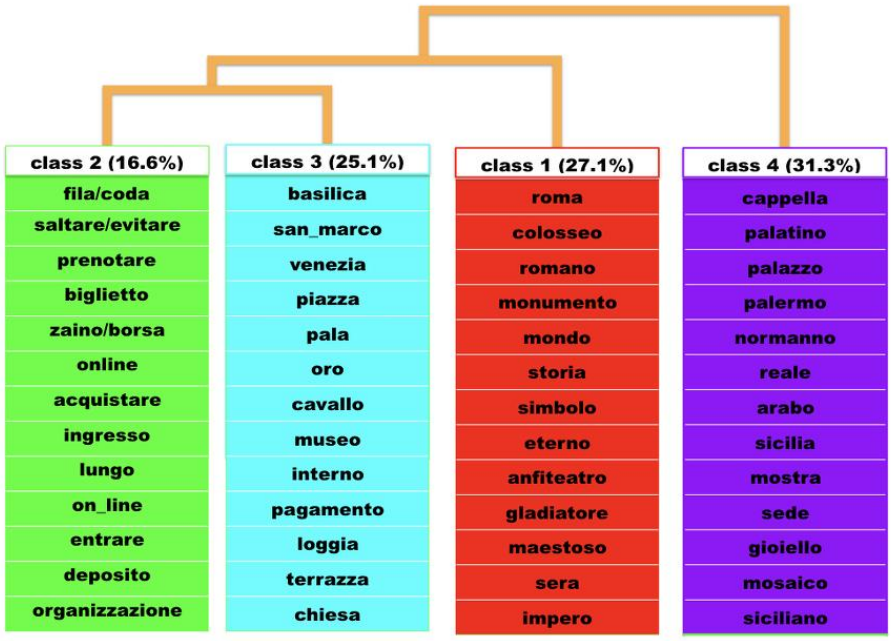
The analysis made it possible to identify groups of segments that are homogeneous within them and heterogeneous among themselves regrading the "concepts" expressed in the whole corpus.

The classes, identified by hierarchical clustering, divide the vocabulary and allow the definition of the lexical worlds. Alceste textual analysis methodology allows to extract the most important contents because repeated by many users in their stories. Users by observing what happens and telling their experience on social network can become sentinels and facilitate the task of those who must manage cultural heritage.

4. Main results and Discussion

Cluster analysis allows us to group statistical units by maximizing the cohesion and homogeneity of the words included in each group and at the same time minimizing the logical link between those assigned to different groups / classes.

Figure 2: *Dendrogram of most representative words*



Source: Own processing on Iramuteq data

The dendrogram (Figure 2) shows the division of the corpus into 4 classes. The words contained in each class make it possible to identify the types of topics covered in the corpus, applying the Alceste methodology proposed by Reinert and

implemented in the IRaMuTeQ software (REINERT, 1995). In Figure 2 we observe the words belonging to the four groups and how they are related to the three point of interest. These word clusters integrate several segments according to the vocabulary distribution.

Clusters 1, 3 and 4 concern respectively the visit of the Colosseum, St Mark's Basilica and Norman Palace as we can imagine from the characteristic words.The words are ordered by the $\square^2$ of association between word and cluster: coefficient calculated with one degree of freedom on the contingency table that crosses the presence/absence of the word in an ECU with the fact whether or not this ECU belongs to the class considered (REINERT, 1995). Obviously we find words that are closely related to the place visited in the first positions. There are theme words like *roma, colosseo, simbolo, anfiteatro* and *gladiatore* concerning cluster 1, *basilica, san_marco, venezia, pala* and *oro* regarding cluster 3 and words as *cappella, palatino, palazzo, palermo and normanno* in cluster 4; moreover in these three clusters there are words that represents TS that express positivity and actions related to the visit.

Table 3 − *Typical ECU of cluster 2.*

prima domenica del mese ingresso gratis la biglietteria apre alle 8 30 ma già alle 8 c'è fila. Tutt'intorno è pieno di venditori più o meno abusivi di visite guidate con opportunità di saltare la fila quella all'ingresso non quella alla biglietteria
colosseo fantastico, da rivedere l'organizzazione, monumento di bellezza più unica che rara merita assolutamente la visita, difetta di organizzazione nella vendita dei biglietti visto che polizia municipale ed un addetto mi hanno fatto fare la fila al botteghino salvo poi scoprire che il biglietto si comprava direttamente all'ingresso.
caotico per accedere a questa basilica è obbligatorio fare code chilometriche e bisogna informarsi molto bene degli orari, noi l'abbiamo fatta per poi scoprire che era chiusa, inoltre occhio a non avere zaini o borse grandi altrimenti verrete respinti sinceramente mi aspettavo migliore organizzazione
la basilica merita sicuramente qualche minuto di coda ed è gratis non ci si può accedere con gli zaini ma viene offerto un deposito gratuito presso il quale ti danno un pass che dura un'ora e che ti permette anche di saltare la fila
nonostante gli sforzi per permettere l'accesso ai diversamente abili motori, resta molto da fare. La "pala" è visibile solo con accompagnatori così come il "tesoro"
Belle le stanze e davvero da non perdere la cappella palatini. Accessibile anche a chi ha difficoltà motorie, è davvero un monumento da vedere.

Source: Own processing on IRaMuTeQ data.

From our point of view the most interesting group is the cluster 2. We observe that in this cluster we find words like *fila, saltare, prenotare, biglietto, deposito* and *organizzazione*. The text segments related to these words represents economic and practical aspects that in some cases may cause discomfort during the visit.

The underlying topics are related to the cost of the ticket, waiting list and the manner of the visit with both positive and negative connotations depending on the particular situation described by the user.

In Table 3 we present some of the typical text segments of the cluster. The segments cover all three attractions examined.

As we can read these reviews are full of useful and precise indications regarding times and visit arrangements as well as regarding the usability of the monuments.

5. Conclusion

The issues highlighted are of interest to public administrators, who can hear directly from the voice of citizens what are the main problems from the point of view of users.
Based on this type of analysis, the decision maker can evaluate if and how to intervene to improve the management of cultural sites and heritage.
The flow of information starts from the citizen who at the end of the process can obtain tangible benefits thanks to the data that he has put on the web.
The process described in this paper shows a classic use of Big Data: data produced with a specific purpose are subsequently used to achieve other objectives, bringing an undeniable added value (RUDDER, 2015).
The applied text mining techniques allowed to enhance information that otherwise would have remained unused.
Further and more detailed analysis can be carried out using the same methodology and the same software used in this work.
You can continue monitoring, increasing the corpus to conduct further analysis on these same monuments or study other cities and other cultural assets in order to improve management policies and optimize decision-making processes.

References

BOLASCO, S. 2014. *Analisi Multidimensionale dei dati. Metodi, strategie e criteri d'interpretazione*. Roma: Carocci editore.

CAMARGO B.V., JUSTO A.M. 2018. Tutorial para uso do software de análise textual IRAMUTEQ. Universidade Federal de Sata Catarina, Available at: http://www.iramuteq.org/documentation/fichiers/tutoriel-portugais-22-11-2018 [Accessed 27 Jun, 2019].

CERON A., CURINI L., IACUS S. M. 2014. *Social Media e Sentiment Analysis. L'evoluzione dei fenomeni sociali attraverso la Rete.* Springer Italia.

GRECO, F. 2014. *Integrare la disabilità. Una metodologia interdisciplinare per leggere il cambiamento culturale*. Milano: FrancoAngeli.

HO J.Y.C., DEMPSEY M. 2008. Viral marketing: Motivations to forward online content, Journal of Business Research, Vol. 63, pp. 1000-1006, https://doi.org/10.1016/j-jbusres.2008.08.010 [Accessed 27 Jun, 2019].

ISTAT 2019. Trips and holidays in Italy and abroad. https://www.istat.it/en/archivio/22702.

REINERT, M. 1995. I mondi lessicali di un corpus di 304 racconti di incubi attraverso il metodo "Alceste". In CIPRIANI R. e BOLASCO S. (Eds), *Ricerca Qualitativa e Computer. Teorie, metodi e applicazioni*. Milano: FrancoAngeli, pp. 203 - 223.

RUDDER, C. 2015. *Dataclisma. Chi siamo quando pensiamo che nessuno ci stia guardando.* Milano: Mondadori.

SOUZA M.A.R., WALL M.L., THULER A.C.M.C., LOWEN I.M.V., PERES A.M. 2018. O uso do software IRAMUTEQ na análise de dados em pesquisas qualitativas, *Revista da Escola de Enfermagem da USP*, Vol. 52, http://dx.doi.org/10.1590/S1980-220X2017015003353 [Accessed 27 Jun, 2019].

STANCAMPIANO, S. 2018. Misurare, monitorare e governare le città con i Big Data. In IEZZI DOMENICA F., CELARDO L., MISURACA M. (Eds), *JADT' 18.Proceedings of the 14thInternational Conference on the Statistical Analysis of Textual Data*, Roma: UniversItalia, pp. 748 - 754.

SUMMARY

Management of cultural heritage using Big Data

The results of this research can support the administrators in the management of services dedicated to users of cultural heritage in the area. The experiment uses text analysis to extract information from reviews downloaded from the web using web scraping techniques. The flow of information starts from the citizen who at the end of the process can obtain tangible benefits thanks to the data put on the web.
The process described in this paper shows a classic use of Big Data: data produced with a specific purpose are subsequently used to achieve other objectives, bringing an undeniable added value.
The applied text mining techniques allowed to enhance information that otherwise would have remained unused.

Sandro STANCAMPIANO, Istat, stancamp@istat.it