Journal of Scientific Temper
Vol 9(1&2), Jan-June 2021, pp. 21-32

ARTICLE

Visualizing museums through the visitors' eye: An n-gram model-based text analysis approach

SHATADAL GHOSH, MANASH BAGCHI, JAYANT GANGOPADHYAY, NATARAJ DASGUPTA, ANURAG KUMAR

Central Research & Training Laboratory, National Council of Science Museums, Ministry of Culture, Government of India
Email: jayantgang@gmail.com

ABSTRACT

All over the world Museums are developing a Customer Service mindset characteristic of large businesses. Part of this shift can be seen in the rapid adoption of museum visitor surveys that gather feedback about a visitor's experience in relation to their demographic characteristics — such as likes, dislikes, etc. as per their age, gender and income. The concept of listening to the visitor through such census data collection is not new, it was there from the very beginning, but in certain aspects, it appears to be deficient. For example, it becomes arduous and difficult to identify if the visitors find the place "memorable". Did they feel "welcome" from the beginning? Do they think "science" is given a due share in that science museum? Can we build an at-a-glance graphical representation of the museum's image in the visitor's mental map?

Census type data collection is very manpower intensive and is often biased as the visitors feel righteous to answer the questionnaire more "correctly" than "frankly". They are not generally considered "easy in-person survey tools". In this paper, we will discuss a very new way of listening to the visitors and reconstructing their free and fair post-visit overall mental map of the museum. In this new effort, we have tried to construct an automated AI-based word-cloud image from visitors' feedback offered voluntarily in social networks like, say, in Google review. This image is helpful in summarizing and analysing a large amount of feedback data on a single page by focusing on the keywords and phrases adopted by the visitor in describing her experience. This approach is especially useful for centralized analysis of feedback data on museums having a large number of branches.

Hence, we have applied this method to process feedback on the National Council of Science Museums, the largest body of science museums in India managing twenty-five centres all over the country. For the very first time applying this method, we have been able to visualize our museums through the visitor's eye rendering the analytic result in the form of a word-cloud that reflects what the visitors fancy most when they recall the museum experience after the visit.

KEYWORDS: Science museum, NCSM, Communication, Word cloud, Google review, n-Gram, MATLAB

Introduction

Standing in the era of the twenty-first century we cannot just rely on the manual feedback collection and analysis system that helped museum professionals for many years[1]. With the revolution of communication network and introduction of social media people now have more flexibility to express their views[3]. As museum professionals of the world's largest science museum network[4] we also wanted to utilize this golden opportunity to listen to the unbiased voice of our visitors. The process of collecting and analysing visitors' feedback is a statistical process[5] so it's very important to be unbiased for getting the real picture.

The most common method of collection of feedback by asking questions has many drawbacks[6]. Hence, for this study we have selected the Google review data which is given by visitors on their own volition. But this review data is not structured unlike the data collected by the question-based system. This makes us explore a text analysis based approach like n-gram text modelling[7]. The idea was to focus on the mostly used keywords in the discussion. So text analysis based word cloud[8] was identified to solve our problem. Through the word cloud we are able to see our centres through the visitors' eyes.

In the paper first we discuss about the proposed method for dealing with this huge unstructured data. Then a mathematical validation is done through simulation and after that the method is applied to real life data to draw some conclusions.

Proposed Methodology

We are working with a data which is highly unstructured and the quantity of the feedback may go up to forty-seven thousand for bigger centres. Table 1 below shows the quantity of available

Google review data for four national level centres and one Science City, Kolkata under the aegis of NCSM.

Table 1: Quantity of available Google review data for four national level centres and one Science City, Kolkata*

S.No.	Name of the Centre	Number of available review data
1.	National Science Centre, Delhi	11756
2.	Birla Industrial & Technological Museum	4513
3.	Nehru science Centre, Mumbai	10,408
4.	VITM, Bangalore	10,392
5.	Science City, Kolkata	47,856
*data as on 10-12-2019		

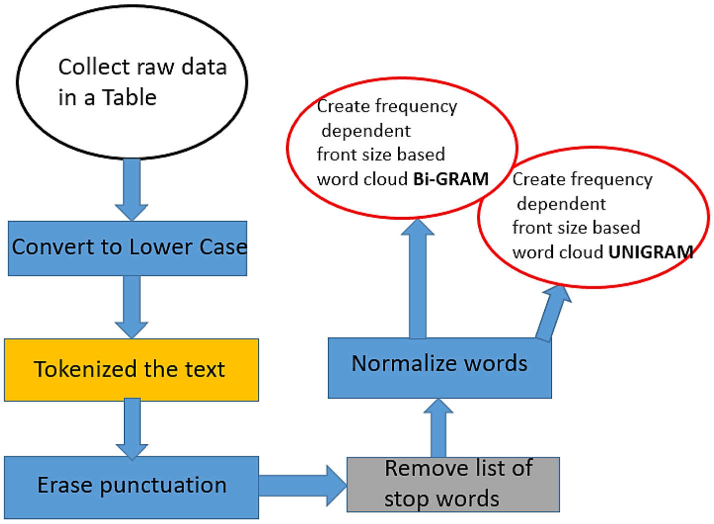

Collect raw data in a Table
Convert to Lower Case
Tokenized the text
Erase punctuation
Remove list of stop words
Normalize words
Create frequency dependent front size based word cloud **Bi-GRAM**
Create frequency dependent front size based word cloud **UNIGRAM**

Figure 1: Proposed methodology

In order to make a compressive image from this huge unstructured data, we have applied some pre-processing operation as shown in Figure 1 below. First the raw data is converted to all lower case, then the text is tokenized and punctuation is erased. After that all stop words (e.g. *and, for, or* etc.) are removed and text is normalized. This normalized data is then used for creating word cloud using n-gram model of text analysis. In this study we have used n = 1, 2, 3 which is unigram, bigram and trigram. Unigram counts the frequency of single word appearances which gives an overall view of keywords mainly used by visitors while giving feedback. Bigram counts the frequency of double word combinations giving a better comprehension of the feedback. But from trigram the information is again diluted with noise as combination of three words in same pattern within a random paragraph are unlikely. Hence, though we have tried generating up to three combinations of words, for analysis purpose we will be using up to bigram only.

Simulation Result

The very first step of validating any new method is best done by simulation. Our proposed methodology is also validated using computer simulation by MATLAB. For the simulation we have taken the review data of Goa science centre and then with actual data noise is mixed following the Gaussian probability curve (Eq. 2) as the overall paragraphs are nothing but additive sum of keywords (Eq. 1). The overall process of simulation is described in figure.

$$paragraph\ for\ word\ cloud = key\_word_1 + Key\_word_2 + \cdots + Key\_word_n \quad (1)$$

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad (2)$$

With both the data, after pre-processing, we got more or less similar word cloud (Figure 2) which validated the applicability of the proposed methodology in the presence of noise.

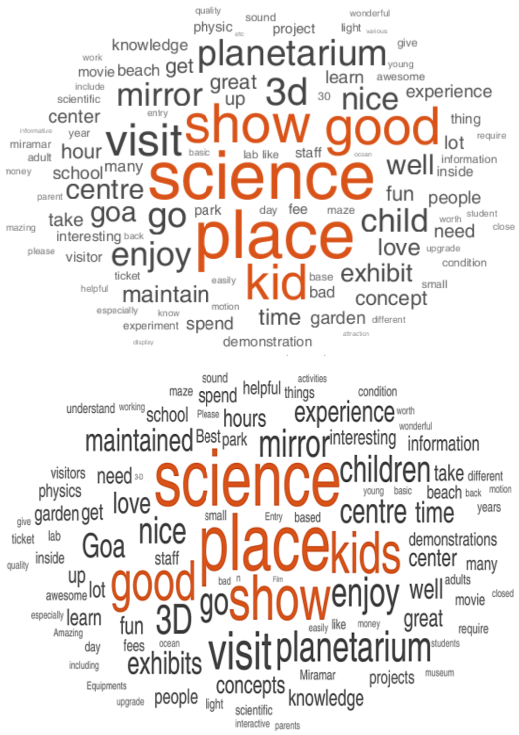


Figure 2: Word cloud with noisy data (above) and without noisy data (below)

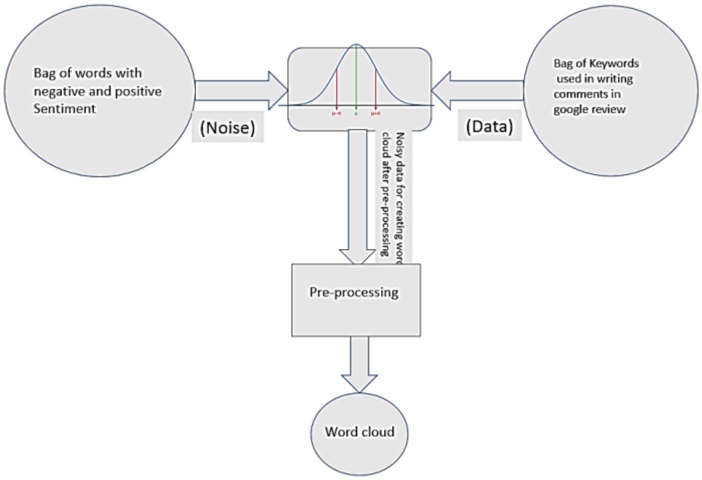

Bag of words with negative and positive Sentiment
(Noise)
Bag of Keywords used in writing comments in google review
(Data)
Noisy data for creating word cloud after pre-processing
Pre-processing
Word cloud

Figure 3: Simulation process of proposed methodology

Experimental Result and Analysis

Simulation shows that the important issues can be reconstructed from the word cloud created by noisy data. Hence experimentation is done using the review comment given by various visitors. A CSV format Table is created with the comments from Google review; then the data is fed into our algorithm realised using MATLAB. For this proposed study, we have analysed the data from two national level science centres – National Science Centre Delhi and Nehru Science Centre Mumbai, one district level science centre of Digha, one Regional level science centre and planetarium of Calicut and one large and unique Science Experience Park – Science City, Kolkata. Figures 4–8 below show the unigram and bigram models of the above mentioned centres.

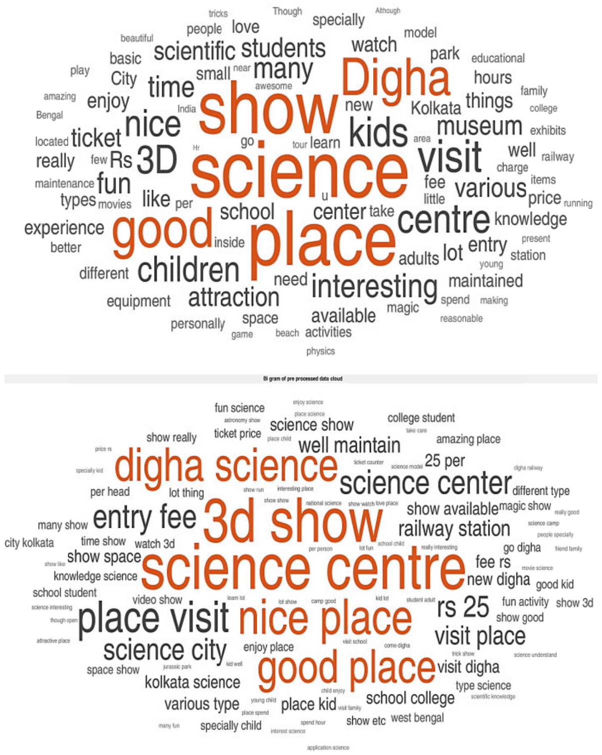


Figure 4: Unigram and bigram model of Digha Science Centre

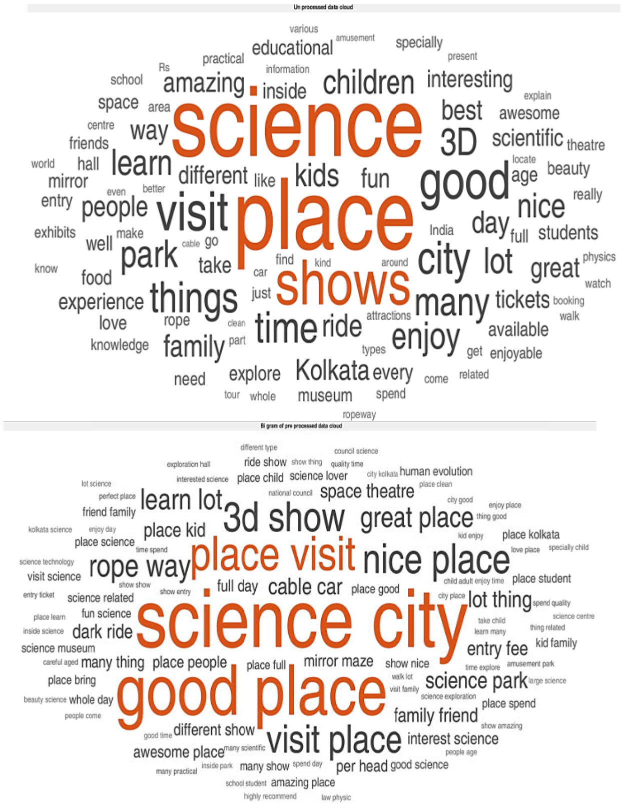
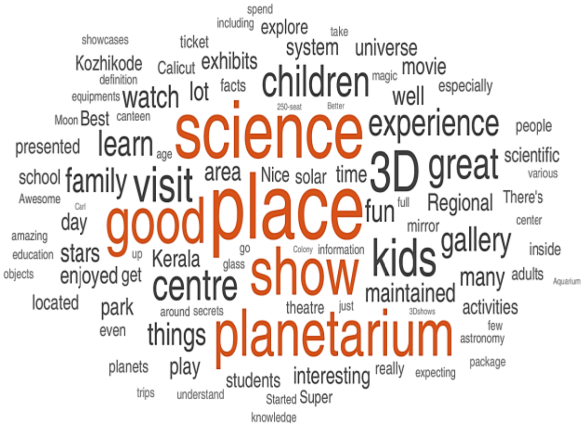


Figure 5: Unigram and bigram model of Science City



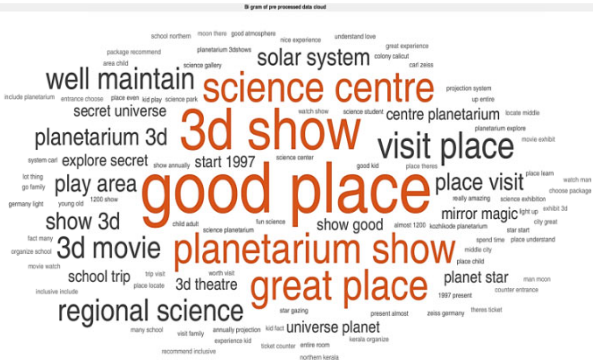


Figure 6: Unigram and bigram model of Regional Science Centre Calicut

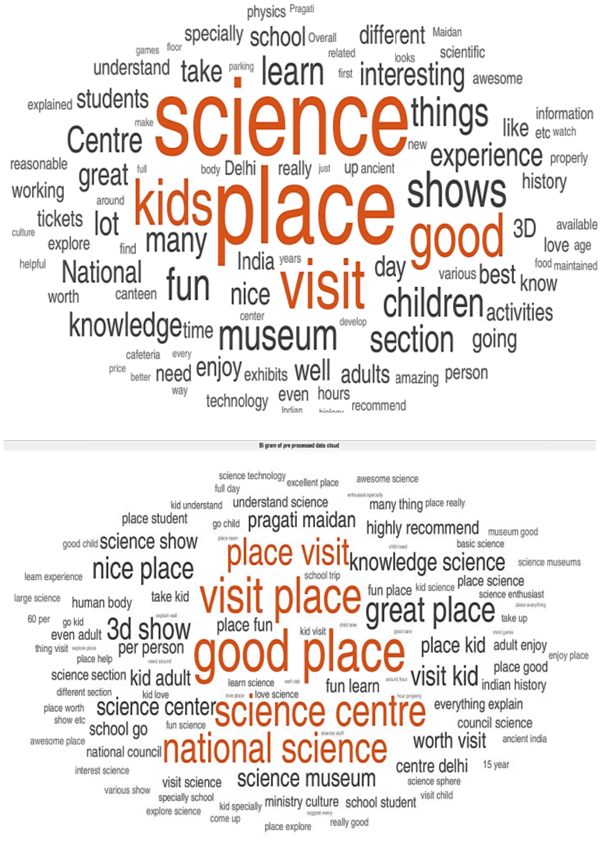


Figure 7: Unigram and bigram model of National Science Centre Delhi

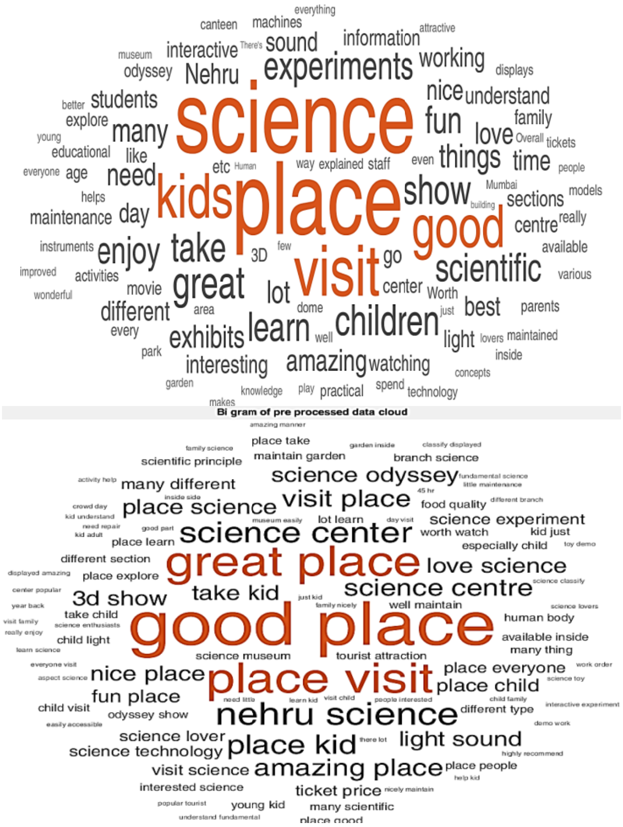


Figure 8: Unigram and bigram model of Nehru Science Centre Mumbai

After studying all the word clouds, we discovered many aspects of visitor perception for our science centres. Below are some of the important observations:

1. Every word cloud has an overall similarity despite differences in dimensions and themes.
2. All our science centre “Shows” are very popular. Hence the word Shows comes in word cloud with loud bold font. Bigram analysis gives a deeper insight and we realize that the most popular show in all our science centres is the 3D show; after that the Science Shows follow in popularity.
3. In visitors’ minds all the centres are mostly very useful for Kids. Very few people consider this as a place for family

visit, which is a concern we need to work at. We also need to highlight and popularize the fact that science centres are meant for grown-ups as well because here the visitors are also made aware about the issues about which they need to take informed decisions. However, Science City, as expected, is considered better suited for family visit in comparison to other units. The overall bold face appearance of the words Kids and Children represent the fact that we are primarily considered as Children's institution. We propose our policymakers to review this point so as to see if the perception is matching the objective or we need to reorient the efforts further.

4. The most important and generic observation is our failure to make our visitors understand the genesis and existence of different levels of science centres run by the National Council of Science Museums. If we compare the two unigrams of Digha Science Centre and Science City, there is no significant difference. But as soon as we concentrate on bigram word cloud of Digha Science Centre, we find that a large number of visitors compared it with the Science City, Kolkata without knowing the fact that both are established and run by the same organization and both have very different purposes of existence. This comparison has two different aspects – if we see it from the point of view of the Digha centre, it may be thought as negative feedback. Whereas for Science City, Kolkata it is a very positive feedback. This is happening as the locations of both the centres are not very far off and Science City Kolkata has become a brand on its own.

It is felt that even though individual centres earn their popularity, their connection with NCSM is not clear in popular perception. Hence the Brand NCSM has a poor reflection in popular perception. Thus, based on this analysis, we propose that we use NCSM identifier before the name of every science centre (e.g. NCSM-Digha Science Centre). With this common thread, if we keep a centralized TV-based introduction in front of all centres along with brief purpose of existence of different levels of science centres and highlight what new is happening all over our network, the brand NCSM may find its place more readily in the minds of our visitors. This may help our visitors to realize the importance of the existence of NCSM which in turn will help to convey the importance of scientific temperament. This prior knowledge will also help the visitor to access the centre with more realistic expectation. For example, if Big Bazar of future group existed with a local name everywhere like Kolkata Bazar, Digha Bazar, Goa Bazar etc., they had to work hard towards establishing their identity for every single local market. But with the promotion of the brand Big Bazar, they have established the common thread of goodwill and hence local level market establishments became redundant.

Conclusion

We can conclude that this unique method of visitor feedback analysis will be very much helpful for visualizing and analysing the museums through the visitor's eye. This method is helpful for comparing and understanding the individual and group responsibility of a network of museums like NCSM.

Many more aspects of science centres like difference in image of national level centres versus district level centres, district to RSC and National to Science City can be done using the same word cloud. Also this Google review feedback is a continuous process. Hence in order to further extend the work we may create an LDA model for automatic sentiment analysis of the given comments. Data can also be compared for all the centres with respect to time and space.

References

Introduction to Visitor Surveys. Te Papa National Services, JUNE 2001, ISSN 1175-6462, Issue no. 4.

Museum of New Zealand Te Papa Tongarewa (2001). 'Know Your Visitors'. Te Papa National Services Resource Guides 3.

Natalia Dudareva (May 2014). Museums in Social Media. Master Thesis, Cand.soc. in Management of Creative Business Processes, Copenhagen Business School.

Jayanta Sthanapati (August 2017). Project Report: History of Science Museums and Planetariums in India. *Indian Journal of History of Science,* 52(3).

https://www.nla.gov.au/content/improving-organisational-performance-through-visitor-feedback https://surveyanyplace.com/questionnaire-pros-and-cons/

Rashmi Agrawal & Mridula Batra (January 2013). A Detailed Study on Text Mining Techniques, *International Journal of Soft Computing and Engineering (IJSCE),* ISSN: 2231-2307, 2(6).

Florian Heimerl, Steffen Lohmann, Simon Lange & Thomas Ertl (2014). Word Cloud Explorer: Text Analytics based on Word Clouds. Proceedings of the 47th Hawaii International Conference on System Science (HICSS 2014), January 6-9, 2014 IEEE Computer Society.