Listening to the Crowd: Data Science to understand the British Museum visitors

Taha Yasseri
Oxford Internet Institute
Dec 2017

There is more to The British Museum than Egyptian mummies and the Rosetta Stone - more than 6 million people walk through the doors each year, travelling from every corner of the globe to see the Museum's collection and get a better understanding of their shared histories. Those visitors offer us a unique test bed for data science and social science experimentation at scale. In order to address some of the challenges of welcoming such a large number of visitors, the British Museum is constantly gathering feedback and information about the visiting experience. Research about our visitors informs decisions made by teams around the Museum and help the Museum evolve along with its audience. The tools at our disposal include direct feedback channels (such as email or comment cards), ticketing data, wifi data, audio guide data, social media conversations, satisfaction surveys, on-site observation and conversations on online review sites such as Trip Advisor.

Trip Advisor reviews are one of the largest and richest qualitative datasets the Museum has access to. On average, over 1,000 visitors review their visit on the platform every month. These reviews are written in over 10 languages by visitors from all parts of the world, and historical data stretches back over two years. In these comments, visitors discuss the positive and negative aspects of their visits, make recommendations to others, and rate their satisfaction. The data set is an opportunity for the Museum to learn more about its visitors, to understand what the most talked about topics are, and which factors have the biggest impact on satisfaction. This research project aims to dig into a rich set of qualitative data, uncovering actionable insights which will have a real impact on the Museum. The research will have an immediate and tangible effect and will help the organization improve the visiting experience currently on offer at the Museum. The Museum is currently undergoing pivotal strategic change, and the insights will also feed into future iterations of the display and audience strategies. As far as we know, the British Museum is the first institution of its kind to take a programmatic approach to this kind of qualitative data. This pioneering research could potentially impact the rest of the cultural sector and show the way to a new method of evaluation and visitor research.

Some of the questions we hope to answer with this data are:
- Understanding satisfaction – what it means, how it affects propensity to recommend, and which aspects of a visit have the biggest impact on overall satisfaction.
- Analyzing the different topics talked about in different languages. Do positive and negative experiences vary according to language?
- Analyzing which parts of the collection or objects visitors talk about the most, and how feedback differs from one area of the Museum to another - Tracking comments regarding a variety of key topics, and understanding how they relate to one another (tours and talks, audio guides, access, facilities, queues, overcrowding…).
- Understanding and anticipating external factors which might impact decisions made to visit (economy, weather, security concerns, strikes, politics…).

The Museum has recently set up a partnership with Trip Advisor, which gives us access to the reviews in an XML format. This file includes the date and URL of the reviews, as well as their title, score, language and full review text. Scope of the research The Museum could take a manual approach to tagging and analysing reviews, but we believe that more insight can be generated through computational approaches. The proposed research project will therefore involve heavy use of modern Natural Language Processing (NLP) techniques. The complete corpus of review text consists of approximately 7,500,000 words in 50,000 distinct reviews. Recent advances in machine learning and NLP provide a wide range of potential approaches to the subject, but suggested methods include:

- Topic modelling
- Clustering/classifying reviews by topic or sentiment
- word2vec style approaches to training/using word embeddings
- Automating the tagging of new reviews
-Time series analysis and principal component analysis