Automated Analysis of Online Reviews to Improve Visitor Experience in New York State Parks

Hari Prasad Udayapuram
Srinagesh Gavirneni
Samuel Curtis Johnson Graduate School of Management
Cornell University
Ithaca, NY 14853
{hpu2,nagesh}@cornell.edu

Abstract
As more and more patrons are registering their facility visit experience in the form of online reviews and blogs, there is an opportunity for facility management to improve the service operations by addressing the concerns. This data however, lacks structure, is voluminous and is not easily amenable to manual analysis. In this project, we design, develop, and implement software systems that download, organize, and analyze the unstructured text from the reviews and help the facility managers identify strategies to improve the visitor experience at their facilities.

- A live system for review collection, processing and storage.
- A software model for identifying positive and negative sentiments in reviews.
- A methodology to extract themes of interest from positive and negative sentiments.

This report was prepared for the New York State Water Resources Institute (WRI) and the Hudson River Estuary program of the New York State Department of Environmental Conservation, with support from the NYS Environmental Protection Fund.
Introduction
The New York State Park system consisting of 214 state parks and historic sites is one of the best and most diverse in the nation. Encompassing over 335,000 acres, 2,000 miles of trails, 67 beaches, and 8,355 campsites, it attracts approximately 60 million visitors every year. The Office of Parks, Recreation, and Historic Preservation (OPRHP) is responsible for strategic planning, operation, and maintenance of the state park system and it has identified “Increase, Deepen, and Improve the Visitor Experience” as one of its priorities. Receiving, analyzing, and responding to visitor feedback is integral to achieving this objective and the existing approaches (public meetings, web-based surveys, and comment cards) have proven to be inadequate. There is a large volume of visitor feedback on online platforms such as Yelp, TripAdvisor, and Google. The State Park System could reap huge benefits if it can harness this source of information that has been ignored so far. This data however, lacks structure, is voluminous and is not easily amenable to manual analysis. We design, develop, and implement software systems that download, organize, and analyze the unstructured text from reviews and help the park managers identify strategies to improve the visitor experience at their facilities.

Methods
Sentiment Analysis
For any given review, we classify each phrase in the text as a positive, negative or neutral sentiment on a scale of 0-1 in increments of 0.1. The software uses an established dataset that has pre-classified words according to their sentimental value. For example, phrases such as dazzling, excellent, mesmerizing, fantastic, beautiful are categorized as positive ingredients whereas phrases such as sucks, terrible, awful, boring are categorized as negative ingredients. The software is made robust via enhancements that adjust for various nuances in writing style, most importantly negation. In our implementation, we use a well-tested and a popular software tool to perform the sentiment analysis.

Theme Extraction
Once the reviews have been segregated, they are analyzed to tabulate the themes they are emphasizing and to do this, we use the technique known as probabilistic topic modeling. This statistical model assumes that each review emphasizes topics with different weights; each word in the review is drawn from one of the topics selected according to the aforementioned weights. The software is then designed to identify the most likely topics, the distribution of words in each topic, and the weight for each topic in a review. For example, a topic could contain the words “mountain, hike, views, trails, walk, tower” and refer to the hiking trails in the park. Another topic could contain the words “pool, zoo, picnic, tables, crowded, pretty” and refer to the picnic facilities. The advantage of probabilistic topic modeling is that these topics can be automatically distilled and then used to determine managerial actions targeted towards improving the visitor experience.

Results & Discussion
For the prototype, we collected 600 reviews for 5 facilities viz. Hudson River Park, Bear Mountain State Park, Harriman State Park, Pierson Pier and Bash Bash Falls from Yelp, TripAdvisor and Google reviews which publish crowd-sourced reviews about local businesses.
To demonstrate the usefulness of ideas described above, we developed the software system consisting of the aforementioned features and performed an analysis of the reviews collected for Bear Mountain State Park located on the west bank of the Hudson River. It attracts more than two million visitors per year and is visited for a wide variety of activities like fishing, swimming, hiking, and boating. For this park, we found (as of June 2014) 70 reviews on Yelp, 191 on TripAdvisor, and 34 on Google, and downloaded them into a database that could be searched and sorted in order to facilitate cursory manual review. However, given the large volumes (more than 25,000 words) of text found in the data, we needed to perform automated analysis in the areas of sentiment rating and topic modeling.

We performed the sentiment analysis on each of the available reviews. Figure 1 below illustrates the distribution of sentiment among the reviews. Observe that about 45% of the reviews have a sentiment rating of 0.7 or higher. As an example, a review that received a sentiment rating of 0.9 contained phrases such as “beautiful attraction”, “absolutely gorgeous”, and “fun to watch.” A different review that received a sentiment rating of 0.4 contained phrases such as “incredibly crowded and dirty”, “trash all over”, and “bathrooms were disgusting.” Having this ability to quantify the sentimental tendency of the reviews enables the facility manager to sort them according to these values and analyze the text content within. However, given the voluminous nature of the data, we propose that this be done using appropriate software tools.

We separated the negative and the positive reviews and performed the probabilistic topic modeling on both the sets. From the negative reviews, we were able to distill the following major topics:

**Crowded**

....The park grounds were incredibly crowded and dirty...

....Arriving at the pool...we had to make a line because it was crowded going in.....

**Parking**

....Worn out, run down, poor access/parking, filled with drunk young punks........parking was $8! I have been to a bunch of parks and this is the first one I've had to pay

**Trash**

....There was trash all over the place....

....There were tons of plastic cups & trash (p.s. trash bin literally 3ft away) on the floor & even in the pond....
From the positive reviews, we were able to distill the following topics:

**Trails**

….Great outdoors for all to enjoy all seasons. Trails, zoo and lake walk to be enjoyed….

….There are lots more trails and good bird watching at Iona Island….

**Animals**

….Took my granddaughter and was thrilled to see all the natural beauty .... with it's natural rock, forna and animals….

….we were able to enjoy the zoo (which is small, but cute with plenty of animals….

**Hike**

….It was a nice hike (LOTS of steps, our butts felt it later!) with a beautiful view at the top....

....Lake, carousel, hikes, play fields, zoo, WOW!

This analysis helps the facility manager to recognize that in order to improve the visitor experience, he/she must take steps to alleviate crowding, improve parking, and clean up the trash. On the other hand, he/she must realize that the trails, the zoo, and the hikes are the reviews that have been labeled as positive or negative. But no review is completely positive or negative. Each review contains segments that are positive alongside segments that are negative. Thus in order to get a more accurate collection of positive and negative themes, it is necessary to perform topic modeling on collections of positive/negative review segments as opposed to whole reviews.

**Topic Modeling Incorporating Expert Feedback:** A topic is a collection of words and when the topics are chosen by software, some of the words in the topic may not belong in it. In such cases, an expert can be used to identify words that should be dropped from a topic and after doing that, the topic modeling can be re-run. Such a recursive approach can lead to a more accurate extraction of themes.

**Direct Solicitation of Reviews:** Currently we are using reviews from third party platforms and they are unsolicited and often cannot be verified. Having established the usefulness of these reviews, we are now developing a system that collects reviews solicited directly from visitors that have recently visited one of the recreation sites.

---

**Conclusion**

We proposed, designed, developed, and implemented a software system for automated analysis of text contained in online reviews for New York state parks. Using Bear Mountain State Park as an example, we demonstrated the usefulness of the tool. We believe that this tool can and should be enhanced in the following directions:

**Automatic Downloading of Reviews:** The social platforms continue to increase in popularity and as a result, news reviews are being submitted at a fast pace. Manual downloading is not feasible and it is necessary to develop a tool that automatically downloads and organizes these reviews into the database.

**Topic Modeling on Negative/Positive Review Segments:** Current system extracts themes from whole reviews that have been labeled as positive or negative. But no review is completely positive or negative. Each review contains segments that are positive alongside segments that are negative. Thus in order to get a more accurate collection of positive and negative themes, it is necessary to perform topic modeling on collections of positive/negative review segments as opposed to whole reviews.

---

**Outreach**

Connected with New York State Department of Environmental Conservation (DEC), closely working with Water Research Institute (WRI) Cornell, proposing to work with with New York State Parks (NYS).

Attended and presented at CaRDI conference in Ellenville, NY and Research and Policy Briefs – CaRDI, Ithaca, NY (forthcoming); planning to attend Hudson River Summit 2015. Tried connecting with OPRHP and I Love NY.

Developed and deployed the tool prototype online – www.rtmso.com