Geovisualization of Sentiment Extracted from Place Descriptions

Rosa Isela Soria Monzón

Masters of Science in Geographic Information Science (MSGISci)

Department of Geography, California State University, Long Beach

Introduction

This project explored positive and negative sentiments associated with the description of specific places in New York City. New York City was selected as the study area because the New York Times has a well-documented set of developer's tools. Furthermore, sites in New York City were more likely to be mentioned in their publication than sites in other places. For this reason, ten landmarks in New York City were deemed relevant based on popularity or significance to the City and were selected for analysis (Figure 1). A Python script collected article content using the New York Times Article Search API and used natural language processing as well basic sentiment analysis techniques to assess the positivity or negativity of adjectives used in the articles. The results of this project may be useful in assessing the opinion of certain locations and might assist in decision-making processes that affect those locations.

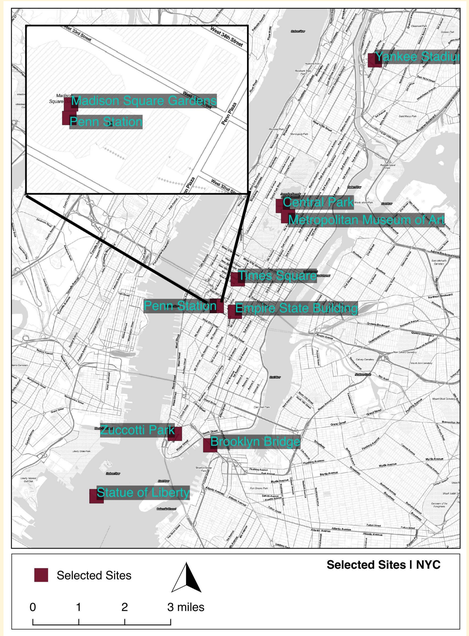

Figure 1. *Ten significant locations in New York City were selected for analysis.*

Data and Data Sources

The data for this project came in the form of plain text. The geosemantics for the ten landmarks selected were geocoded and stored as a shapefile.

The attributes for the shapefile were derived from words in the corpora. Words were tagged by part of speech and those tagged as adjectives were used to derive the positivity ratio.

Table 1. *Data sources.*

Dataset	Source
Corpora	The New York Times Article Search API
10 New York City Landmarks	Geocoding script
Positive/Negative Adjective List	University of Chicago

Methodology

The first task in the methodology was to geocode the ten landmarks by creating a KML with the landmark names. The plain text that eventually became the dataset was acquired from the New York Times. A *Python 2.7* script that made requests to the New York Times Article Search API gathered articles based on query terms, in this case the ten landmarks. The script parsed the JSON data returned by the API and articles with the query term in their headline were returned as a list of URLs. Later the script iterated though this list and used the library *BeautifulSoup* for parsing through HTML mark-up and collecting only the articles' contents. The contents were stored in a plain text file which became the corpora used for analysis.

A second script used basic natural language processing (NLP) and sentiment analysis to analyze the positivity or negativity of adjectives used in the corpora based on the positive and negative adjective list obtained at the outset of the project. This script was also written in *Python 2.7* and used the library Natural Language Toolkit (NLTK) for natural language processing. A ratio was calculated in order to normalize the results for comparison. These results were plotted on a web map.

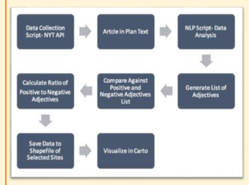

Figure 2. *Process workflow*

Figure 3. NLP script

Timeline

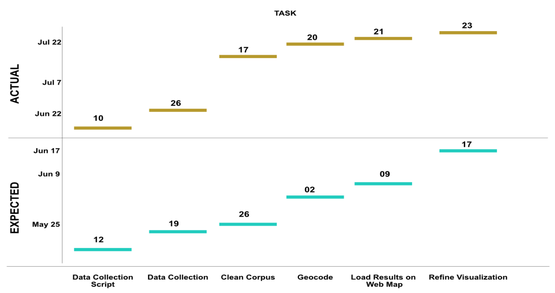

Figure 4. *Project timeline*

Results

Because of the exploratory nature of this project, the results varied. The first results of the Python script were the corpora. The corpora are the collection of articles that were stored as a plain text file. This result can be conceptualized as the "raw data." **Figure 5** shows the corpus of the Brooklyn Bridge search.

The second set of results came in the form of a Python list. The NLP script assigned structure to the plain text and analyzed it in order to extract adjectives found in the corpus. An example of this can be seen in **figure 6.**

The final result was the positivity ratio visualized on a web map, as seen in **figure 7.** One interesting finding was that places known to have negative connotations were not strikingly marked as such by the script. This was the case with Penn Station – articles displayed a clear negative sentiment when read in context but the ratio determined by the script was average. This can likely be attributed to negation in language. In other words, negation occurs when words neutralize or counteract the meaning of other words in the same sentence. For instance, "the renovations at Penn Station come at a less than ideal time for commuters, which should make for a great deal of setbacks." In that sentence, the meaning or connotation of "renovations" and "great" are cancelled out by the surrounding words in the context.

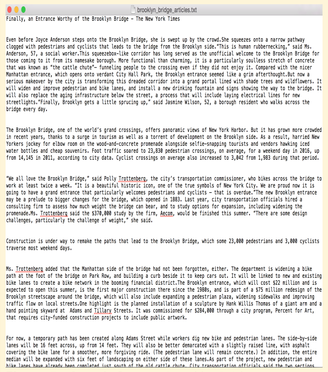

Figure 5. *Sample Corpus- Brooklyn Bridge*

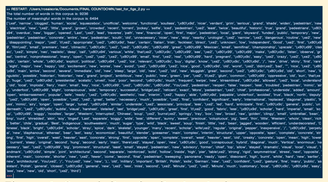

Figure 6. *Sample list of adjectives in the corpus.*

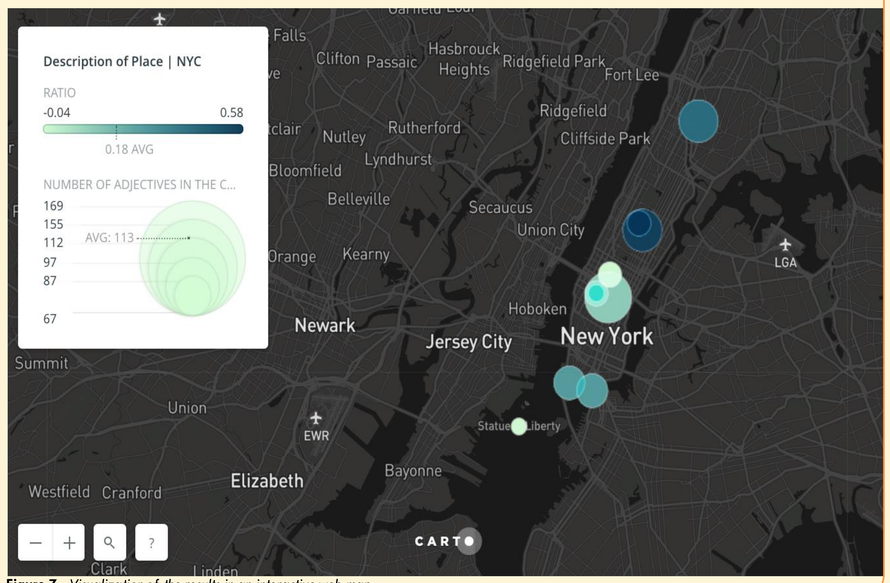

Figure 7. *Visualization of the results in an interactive web map.*

Discussion

The results of the project serve primarily as the foundation for a larger study. While the project accomplished the task of providing a basic sentiment analysis of places, there are many limitations in the natural language processing methods. NLP is extremely intricate and the results of a ratio of positive to negative adjectives cannot fully capture the breadth of its complexity.

One significant limitation is the NLP script does not account for negation or sarcasm in the text. Nor can it determine if a location is being used to refer to the site, or if it is being used figuratively. An example of the latter is the term Wall Street used to refer to a location in New York vs. used to represent the U.S. financial services industry.

Regardless of the NLP script limitation, the relevance of the project lies in its ability to be modified and scaled. With a refined methodology, the study has the potential to be used in a wide range of disciplines. One significant use can be tracking the reception of a new product (via reviews or social media posts) using location intelligence.

Conclusion

The findings of this project have set the foundation for a more complex study that can explore geosemantics more deeply. While the results do show limitations, the issues identified can certainly move the project in the right direction.

The current methodology has plenty of room for improvement. One main avenue for improvement is the development of a more interactive user interface that allows users to pass query terms. This will ensure the map remains relevant and useful.

Submitted in partial fulfillment of the requirements of the Masters of Science in Geographic Information Science(MSGISci), August 12, 2017.

For additional information please contact: Rosa I. Soria.
rosaisoria@gmail.com | http://www.meetrosasoria.com/geosemantics