**Tourism attraction recommendations for US major cities**

Bhagyashree Kulkarni

Sunanda Unni

Vidya Patil

**Abstract**

With increasing use of social media, it is noted quite frequently that tourists are researching about the top places to visit in a city. In addition, people love to share their thoughts on the social media about their sentiments over a location. Our goal for this project is to extract tweets from twitter, identify the top locations by doing sentiment analysis, and then predicting a score. Hence, major goal of this project is to study different algorithms for doing sentiment analysis, identifying, and extracting locations to predict the best possible tourist spots within a city.

**INTRODUCTION**

Recommendation engine and branding has been a very common business objective of Twitter sentiment analysis. We have seen a lot of papers where micro blogs (Oku, Hattori, & Kawagoe , 2015), Twitter (Shimada, Inoue, Maeda, & Endo, December 2011), yahoo searches have been extensively mined for analyzing the top tourist spots and tourist behaviors. A lot of travel sites (https://www.stratosjets.com, jatralog- Choudhury, 2016) have also invested in doing an analysis on the same to improve branding and to drive the tourism industry.

The recommendations would help the users to plan his/her trip of the target city effectively.

The characteristics of the target city would help the user in his/her decision making whether to visit the target city or not at the first place based on his inclination or liking. Indirectly this would also help local business to allocate resources in an optimized manner in the major cities for their operations. For example, transportation services like Uber, Lyft can focus on driving their operations primarily near recommended areas within the target cities.

One thing is missing in these researches is mapping the most recommended places within a city with the objective to assist the users in decision making to visit a place, to plan their trip effectively within the target city and to drive local businesses.

**LITERATURE REVIEW**

As a tourist in a city the main question that arises is to find out which places to visit. People used to rely on recommendations of family and friends, but with the popularity of online social media, people are resorting to the online information. The article by Ghani, Alowibdi, Jalal, Mokbel, & Mohamed (2014) discusses about a program called “vacation finder” which is about finding the top locations based on people’s tweets before and after the vacation. The article also discusses about travelling during specific holidays. It’s quite possible that some users might not necessarily tweet before and after leading to lot of missed information. To fill that gap, we will take into consideration user’s current location as well as any tweets for the popular destination irrespective of the location or the time.

Having access to social media has given access to so much information that aids in predicting a pattern for the travelers. The article by Abbasi, Rashidi, Maghrebi, & Waller (2015) not only talks about geolocation tweets, it also discusses about additional attributes such as departure time, traffic, location route, activity duration which helps in predicting certain patterns of the travelers. The article is very informative, but the scope of the project increases requiring more data and further analysis.

Online media is great place to get the information, but it comes with its own challenges. When users tweet messages, they do not necessarily tweet the location. Or they might have never been at the physical location. Hence, analyzing the text data getting the location could pose a challenge. As per the article by Oku, Hattori, & Kawagoe (2015), to overcome this challenge, the article also looked for pictures taken at the tourist spot. This is to ensure that the extracted data is valid, and the system can make a better prediction. The picture at the tourist spot would confirm the location. The article also refers to extracting the latitude and the longitude which provides information for a specific region which helps in getting the exact location of the spot. The article makes some good points about the user tweeting after the vacation, or maybe the tweet before the vacation. It gives some good pointers about excluding such information. The article does not take into consideration the text messages. So, in our case we will consider the text messages.

Pictures are certainly a great way to find out more about a tourist spot and its popularity. We reviewed an article by García-Palomares, Gutiérrez, & Mínguez (2015) which provided great information about extracting images from various websites such as Flickr, Instagram, Panoramio to get the geolocation of the spot with increased accuracy. The article also suggests that people could pretty much create maps referring to these images which is very impressive. The results were based on spatial popularity and the spots were dispersed in a wide region indicating some outliers as well. The pictures taken by local residents were heavily located within the cities whereas the pictures by tourists indicated spots outside the cities. We felt gathering so much images and doing this kind of analysis would be out of scope for us. This article mainly focused on image basis and missed the text messages. In addition to get the tweets with pictures, we will fill the gap of analyzing the tweets as well.

Considering the challenges of extracting information based on images, we plan to focus on text messages that would still provide great information. As per Maghrebi, Abbasi, Rashidi, & Waller (2015), text mining is easier and cheaper to implement utilizing Sentiment analysis based on the context of the text messages. The article discussed the overall advantages of text mining and sentiment analysis which we plan to implement. However, it did not focus specifically on top tourist spots. The article suggested analyzing the data with only the text messages but did not focus on mapping the distances. We will be covering that.

In the research article by Shimada, Inoue, Maeda, & Endo (December 2011), emphasis is on extracting the data from web and analyzing the tweets to find the negative/positive opinion of the tourists to get the feedback on the locations. The article takes the approach of doing sentiment analysis working with unsupervised learning by extracting the seed words from twitter. This article is aligned with our interests of creating a recommendation system for tourists to target the top spots. We found this article to be the most relevant to what we want to achieve. However, the article does not discuss about mapping the most recommended locations and giving some characteristics about the city. We plan to fill that gap.

**DATASET DESCRIPTION**

We collected old tweets from October 2017 to December 2017 for Los Angeles city (3 months data). Tweets were fetched using <https://github.com/Jefferson-Henrique/GetOldTweets-python>

Query: python Exporter.py --near "los angeles" --since 2017-10-01 --until 2017-12-31

**Official Tourism Account Data:**

There might be possibility that users disable or turn off Geolocation service while tweeting. To make sure we did not miss tweets and enrich our data, we augmented that API call to fetch tweets for official tourism user account which publishes vacation ideas, travel tips, & local happenings (@discoverLA).

**To build Characteristics of Data:**

Different twitter user accounts related to places, restaurants, hotels, stadiums, museums etc. found in data for which 500 tweets per user account were separately collected to generate the characteristics of a target city.

**RESEARCH DESIGN AND METHODS**

In the current scope we considered Los Angeles city from USA. The goal of the project was to provide top N attractions as recommendations for a target city based on the analysis of twitter user’s sentiments about the places. The system also highlights the top N characteristics of the target city. For example, “best breakfast”, “best museum”, “bellini landscapes” etc. would be displayed as some of the characteristic of the city.

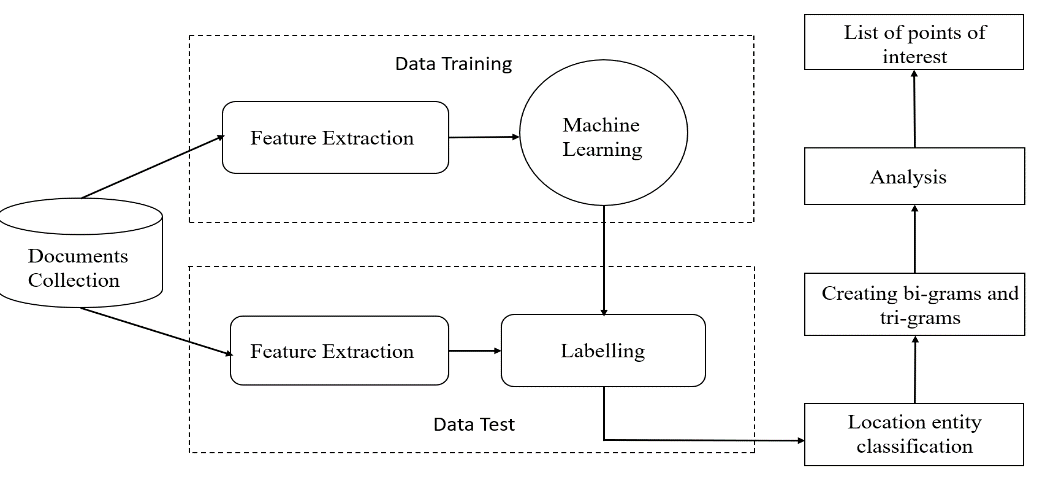
**Preprocessing of Data**

Some manual parsing of the data was required for the csv(delimiter = “;”) since the tweet text would have emoticons which use the same delimiter.

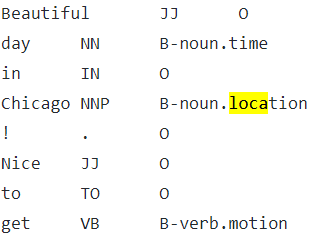
Punctuations, special characters like #,@, digits , http URLs were removed from the tweet data.

3 letter words like com, pic, org etc. were also removed.

**Location Entity Extraction**

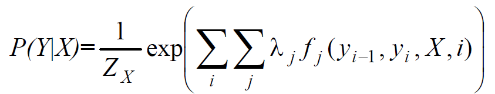


We extract location entities using CRF algorithm. It uses [Ritter dataset](https://github.com/coastalcph/supersense-data-twitter) for training the CRF model which looks like below:



The testing data is our collected twitter data of city which is converted to same format as training data, appending every token to part of speech and predicted label. Feature function is executed on both training as well as testing data.

**Feature Function:**



The function fi takes into account:

* A sentence.
* The position i of a word X in the sentence.
* The label yi of the current word
* The label yi-1 of the previous word
* POS tag of the word
* If the word is in upper case, lower case, alpha-numeric, digit.

Now testing data after feature function is send to the CRF model for predicting the labels. The labelled data is then classified into pre-defined entities. The location entities are extracted for which bi-grams and tri-grams are created. Top 30 locations/point of interest are fetched based on the word count.

## CRF Algorithm

Conditional Random Fields is a statistical modelling method for structured prediction. It not only considers the neighboring samples but also takes into consideration the context of the words. We used Crfsuite which is an implementation of Conditional Random Fields for labelling sequential data. It provides fast training and tagging. Sklearn crfsuite is a python-crfsuite wrapper which is used for cross-validation, hyperparameter optimization and helps dump and load the CRF model using joblib library. So, once we train the model with the training data, we don’t need to run it repeatedly.

**Sentiment Analysis of Locations**

An important part of recommending population locations in a city would require judging the sentiments of the visitors or the people tweeting while visiting the specific location. To achieve this we created a classifier using a supervised sentiment analysis on each tweet collected for the city. Sentiment140 dataset is used for training the tweets, the designers of the classifier used a dataset containing 1,600,000 tweets to train it based on an emoticon tagged dataset, positive tweets were labelled as 4 and negative tweets were labelled as 0. The created classifier model gave an accuracy of 76% with test data.

The trained classifier model was utilized to make sentiment prediction on the collated data for Oct-Dec month for the city using the TFIDF vectorizer. The Dataset was then grouped as per locations identified in each tweet and the mean value of the predicted sentiment was captured. The captured mean values were further normalized using MinMaxScaler for preprocessing the data which varied to extreme differentials.

**Hashtag Extraction**

It would be an appropriate approach to give pointers regarding the city by extracting the most popular hashtags in the city during the particular time period. We achieved this by creating a Directed Graph for network of nodes, with each hashtag marked as a node. Edges are defined between the node if same hashtag have its users matching however it should come from the same tweet.  Two hashtags in same tweet would not be showing or improving upon centrality of the nodes hence we ignored this part. Putting the weights on each node based on total retweet count and follower count is something we could not achieve in the current scope of project and can be considered for long term.

This network was further used for calculating the Degree Centrality, Closeness Centrality and Betweenness Centrality of the hashtags. For this case study we find Betweenness centrality to be of use to unite people new to the city to connect with local populace.

**Characteristics of City**

Different twitter user accounts related to places, restaurants, hotels, stadiums, museums etc. of the target city helped build characteristics of the target city. We used TF-IDF "Term Frequency, Inverse Document Frequency” to score the importance of words (or "terms") in a document, based on how frequently they appear across multiple tweets.

We considered n-gram ranging from 2 to 4 for extraction. Stop words were removed. Tweet data was tokenized, converted to vector space and computed the TF-IDF weight for each tweet in the dataset to return top words having highest scores.

**VISUALIZATIONS**

1. The Fig-1 below is a summary of all the counts that we gathered as part of the twitter data. We looked the favorite counts, retweet counts, hashtag counts, and initial prediction score based on the sentiment analysis of the tweets collected. Overall, we found that the counts were consistently similar. We did see a very slight discrepancy in some locations, but overall it looked similar.



*Fig-1 Favorites, Retweets, Hashtags, Predicted Score*

1. Below visualization was based on final result of all the analysis done on the tweets.



*Fig-2 Top 20 tourist spots in LA*

1. Characteristics of the city “Los Angeles” based on TFIDF scoring done on the Twitter user accounts

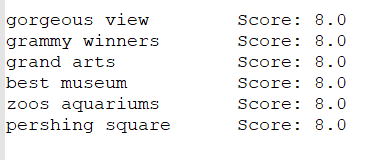
****

Fig-3: TFIDF Characteristics Score for the “LA” city

**RESULTS**

For our final results, we took all the favorite counts, retweet counts, hashtag counts, prediction score then applied the model sklearn.preprocessing.MinMaxScaler. This model estimator scales and translates each feature individually and assigns a scale of 0 to 1. This approach allowed us to get to a final mean score based on all the attributes collected. The top 20 spots that had the highest mean scores were our top 20 recommendations. We decided to plot using a map so that it can visually give a clue to the tourists about the spots that they can plan to visit together. This would help them to narrow down the locations based on the specific areas and allows them for efficient planning of the trip.

|  |  |
| --- | --- |
| **popular\_spot** | **FinalMeanScore** |
| El Pueblo de Los Angeles Historical Monument | 1.0000 |
| santa monica | 0.1571 |
| beverly hills | 0.0467 |
| west Hollywood | 0.0338 |
| Universal Studios Hollywood | 0.0314 |
| venice beach | 0.0179 |
| Hollywood Walk of Fame | 0.0147 |
| north Hollywood | 0.0129 |
| Hollywood Bowl | 0.0128 |
| Hollywood Sign | 0.0086 |
| el monte | 0.0082 |
| universal city | 0.0075 |
| Griffith Observatory | 0.0053 |
| west Covina | 0.0044 |
| Griffith Park | 0.0041 |
| long beach | 0.0041 |
| Urban Light | 0.0011 |
| Petersen Automotive Museum | 0.0008 |
| Venice Canals | 0.0007 |
| Our Lady Queen of Angels Catholic Church | 0.0006 |

**TEAM MEMBERS – CONTRIBUTION**

We identified below areas of work and each section was broken down into 3 parts and handled by each one of them. The best approach was implemented after reviewing the results. So overall each of us contributed to each section in our own way.

Contributed by all:

Research: Study of Twitter API, how to fetch tweets, different parameters.

Reading different papers/previous work related to the topic.

Collection of tweets, preprocessing of the data, deciding the approach, resolving errors , report creation

Coding/ Visualization:

Broad level of contributions:

Vidya : Sentiment Analysis using Affin model, visualization, feature extraction, visualization

Sunanda : Sentiment Analysis using emotion.csv, Hashtag Extraction, feature extraction

Bhagyashree: Location Extraction, Characteristics of a city

**CONCLUSIONS**

The significance of our project is to develop a system that will analyze user’s tweets sentiment over various places and provide accurate ‘top N’ attraction recommendations of the city. Study of such huge enormous twitter data is helpful to understand tweet sentiments around tourist attractions in major US cities. There is still a lot that can be done to improve this model. We gathered 3 months tweets for LA city. We applied the sentiment analysis and generated scores based on the positive or negative emotions. The we collected information such as favorite counts, retweet counts, hashtag counts based on the point of interest. Then a final normalization was done on all the attributes to get to the final mean score. The top 20 with the highest scores were our top recommended spots in LA city.

The big challenge is in analyzing the tweet sentiments. The tweet itself is limited to 140 characters which makes it difficult to analyze the context and sentiment of the tweet. The tweet contained informal language, local language words, misspellings, slangs which needs to be understood and handled properly.The major motivation of this project was to study different algorithms for doing sentiment analysis, identifying and extracting locations (Named entity recognition) to arrive at best possible recommendations.The direct advantage for the end user would be to plan his trip effectively. Besides, the study has great potential to leverage data to build n-day itineraries (1-day itinerary, 2-day or 3-day and so on) for the end user taking into consideration distances between various spots, recommended location ratings, approximate time required to spend at a recommendation. Also, the local businesses would benefit from this system.

**REFERENCES**

Abbasi, A., Rashidi, T. H., Maghrebi, M., & Waller, S. T. (2015). Utilising Location Based Social Media in Travel Survey Methods. *Proceedings of the 8th ACM SIGSPATIAL International Workshop on Location-Based Social Networks - LBSN15*. doi:10.1145/2830657.2830660

Choudhury, N. (2016, December 12). Sentiment Analysis of Twitter Data for a Tourism Recommender System in Bangladesh. Retrieved February 03, 2018, from https://aaltodoc.aalto.fi/handle/123456789/23967

García-Palomares, J. C., Gutiérrez, J., & Mínguez, C. (2015). Identification of tourist hot spots based on social networks: A comparative analysis of European metropolises using photo-sharing services and GIS. *Applied Geography,* *63*, 408-417. doi:10.1016/j.apgeog.2015.08.002

Ghani, Sohaib & Alowibdi, Jalal & F. Mokbel, Mohamed. (2014). VacationFinder: A Tool for Collecting, Analyzing, and Visualizing Geotagged Twitter Data to Find Top Vacation Spots.. . 10.1145/2755492.2755495.

K. Oku, K. Ueno and F. Hattori, "Mapping Geotagged Tweets to Tourist Spots for Recommender Systems," *2014 IIAI 3rd International Conference on Advanced Applied Informatics*, Kitakyushu, 2014, pp. 789-794. doi: 10.1109/IIAI-AAI.2014.159

Kenta Oku, Fumio Hattori, Kyoji Kawagoe, Tweet-mapping Method for Tourist Spots Based on Now-tweets and Spot-photos, Procedia Computer Science, Volume 60, 2015, Pages 1318-1327, ISSN 1877-0509,<https://doi.org/10.1016/j.procs.2015.08.202>.

Maghrebi, M., Abbasi, A., Rashidi, T. H., & Waller, S. T. (2015). Complementing Travel Diary Surveys with Twitter Data: Application of Text Mining Techniques on Activity Location, Type and Time. *2015 IEEE 18th International Conference on Intelligent Transportation Systems*. doi:10.1109/itsc.2015.43

Shimada, K., Inoue, S., Maeda, H., & Endo, T. (2011, December). Analyzing tourism information on twitter for a local city. In *Software and Network Engineering (SSNE), 2011 First ACIS International Symposium on* (pp. 61-66). IEEE.

<https://github.com/coastalcph/supersense-data-twitter>