Sentiment Analysis for Tourist Attraction Recommendation

Jin XUE
jinxue@cse.cuhk.edu.hk
Student ID: 1155135535

Yang GUO
yangguo@cse.cuhk.edu.hk
Student ID: 1155136640

Qi LI
1155092087@link.cuhk.edu.hk
Student ID: 1155092087

Zhiqi WANG
zqwang@cse.cuhk.edu.hk
Student ID: 1155092207

ABSTRACT

The popularity of social media provides convenience for people to share their travel experience at any time and any place, which also opens the opportunity to analyze the underlying popularity of tourist attractions and user's preferences. In this project, we generate an objective tourist attraction ranking based on the result of sentiment analysis of a large number of user's comments to help people to choose a tourist attraction. For the people who have specified travel preferences, we also make personalized tourist attraction recommendations by referring to the people who have similar travel preferences.

ACM Reference Format:
Jin XUE, Yang GUO, Qi LI, and Zhiqi WANG. 2021. Sentiment Analysis for Tourist Attraction Recommendation. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

With the development of transportation and improvement of people's material condition, traveling becomes a common choice of relaxation for people in modern society. However, sometimes it is hard for people to decide which city or tourist attraction to visit: People are usually unwilling to go to the city or tourist attraction that they have traveled before and want to experience something new, but choosing a new tourist attraction means that people have to search for a lot of information to help them make a decision. This is a time-consuming process and some wrong information on the Internet may also mislead people to make a disappointing traveling decision. Therefore, a new tourist attraction recommendation system can be a necessary tool to solve this problem.

Past few years, more and more people tend to share their travel experiences like photos or videos on social media such as Flickr, Instagram and Facebook, etc. People can also leave their comments with respect to the specified tourist attraction. These services hold and collect a large number of datasets that contain location information, photograph, time, comments attached to this post, other textual information and etc. For example, there are more than 40 million geotagged photos on Flickr. These huge datasets expose different opinions from different people for different tourist attractions, which opens the opportunity to discover something that can be used to make a tourist attraction recommendation to people by analyzing people's attitudes(comments) towards the tourist attractions. We find that most existing solutions [1, 2] that focus on recommending the tourist attractions are

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

based on user's preferences by understanding the user's historical travel experiences, but these solutions have an assumption that we have enough historical data to analyze the preferences of the user. Therefore, if one user does not have enough historical data, they cannot work. Besides, there is another problem that some people want to try something new and to have some fresh travel experiences, which cannot be satisfied if we only recommend tourist attractions according to their historical preferences. Finally, we can get more accurate ratings for each attractions by sentiment analysis.

To provide high-quality travel recommendations, we also design a solution of preference-independent recommendations for those who do not have enough historical data to obtain their preferences apart from common preference recommendations. For the preference-independent recommendations, we build a popularity ranking to offer recommendations. We use a sentiment analysis model to analyze people's comments to a specified location point to find most people's opinions to this location point. Then, we identify the tourist attraction and aggregate the data points in the region of the corresponding tourist attraction to generate the ranking score for this tourist attraction. We also categorize the tourist attractions by using their tags such as a park.

We design this recommendation system to satisfy the following requirements:

- Build an objective tourist attraction ranking by using sentiment analysis model to analyze people's comments attached to the tourist attraction
- Provide accurate tourist attractions recommendations to the people (including preference-independent and preference-aware)

2 PROBLEM DEFINITION

In this project, the main outcome will be a system that provides scene recommendations to users based on the their geographical location and preferences. The system discovers potential tourist attractions from the data and builds a popularity ranking of these tourist attractions using multiple criteria. The system provides both preference-independent recommendations to users based on geographical locations and also personalized recommendations from the user's travel history.

3 METHOD

3.1 Data Preprocessing

We use Yahoo Flickr Creative Commons (YFCC100M) dataset which is known to be one of the largest assemblages of multimedia check-ins ever created. From this dataset we can find people's opinions and attitudes to the places they have visited, but there are lots of unrelated information such as license name which will not be used in our project or missing attributes. What we care about in this dataset includes photo/video identifier, user information, location coordinates, user comments, user tags. To this end, we first filter these data, which means we only remain the items that contain the needed attributes and the partial attributes of the dataset we will use are showed in table (1). After this operation, the total number of record reduces to 16838559 and takes up 7.9G of hard disk, compared to the original 100 million items and 45G hard disk space. Because there is only longitude and latitude information attached to each record in the original dataset. To analyze the popularity of each tourist attractions, we need to transform these longitude/latitude to the real-world address, but most related geocoding reverser tools are online . Taking the large number of the data size into consideration, the online location-based service API is not acceptable, because it is time-consuming. To solve this problem, we use a python library for offline reverse geocoding called *Reverse Geocoder*. By using this tool, we can get city/town and country code from the location coordinates.

Table 1: Dataset Description

Attribute	Description	Data Type
photo_id	Unique media identifier	String
user_nsid	Unique user identifier	String
user_tags	The features of the tourist attraction given by users	String
description	The content of user’s post	String
longitude	Longitude of the location where the media object was uploaded at	double
latitude	Latitude of the location where the media object was uploaded at	double
date_taken	Date the media object was created	date
comments	Other people’s comments attached to the media media object	String
country	Country code, derived from the location coordinate	String
town	City name, derived from the location coordinate	String
state	State name, derived from the location coordinate	String

3.2 Meta Data Analysis

Before we do the recommendation, we also hope to have some basic knowledge about this dataset, for example which city is liked by most travellers and which kinds of city is more popular and attractes more tourists. These knowledge not only gives us a big picture of the whole dataset, but also can help us build the preference-independent recommendation. In order to handle the large data size, we use MapReduce to accelerate the data processing. At this stage, we compute the popularity of each city by counting the numbers of check-ins in the city and for the popular cities we also explore why they can attract more tourists by analyzing the user tags attached to them.

3.2.1 Top 10 Cities. We use MapReduce to find the most 10 popular cities that attract vistors. As we can see in the chart figure(1), San Francisco is the most popular cities and about half of the most popular cities are in the U.S.(we have to admit that maybe the user distribution also should be considered, but it is hard to get and the result without it seems that still is reasonable). Besides, Tokyo, Paris, London, Brooklyn are also popular cities, which is also consistent with common sense, because they are all world-famous cities and attract the tourists from all over the world.

3.2.2 What attracts tourists? People travel to different places to experience something new or interesting, so what attracts them most? The user tags attached to each media object give us an opportunity to discover the things that attract people. From the result of tag analysis, we find that the tag - nature appears most frequently and it is easy to guess that people who have lived in the city for a long time like to get in touch with nature. The other popular tags include music, art, architecture, beach, park and so on. People likes these things to make relaxation, so the city with these characteristic has high possibility to attract people.

3.2.3 What makes the city popular? For the popular cities, we also analyze the tag distribution for each of them to find out which makes the city popular and distinct. For example, as shown in figure (2), San Francisco is popular and famous for the art, music, party and so on. So San Francisco may be a good choice for those persons who are interested in these things. Another example is Singapore, nature scenery especially marine related (such as seashore, island and marine life) is attractive.

3.3 Fetching Comments

Each row in the YFCC100M dataset contains a photo ID field which is a number that uniquely identifies a photo or a video on Flickr. However,

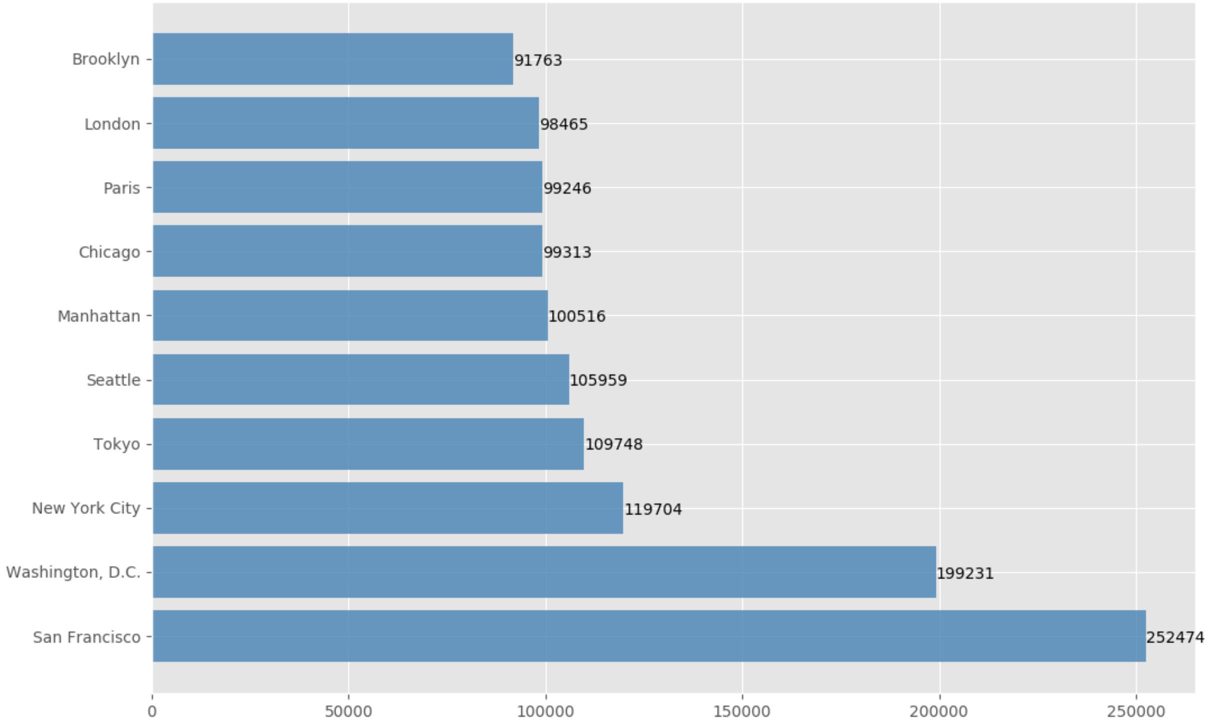

Figure 1: Top 10 Cities in Flickr

the actual comments to a multi-media resource are not included in the dataset.For this reason, the comments must be downloaded through Flickr's flickr.photos.comments.getList API. On success, the API returned the list of comments to the queried photo ID in XML format. Because the total number of samples in the dataset is huge and each sample may contain more than one comments, we utilize MapReduce for the data preprocessing of downloading comments. We use a mapper program without reducers which takes the original dataset as input and send a HTTP request to Flickr API service to download the comment list. The output is a tuple of (photo_id, ns_id, longitude, latitude, comment, tags). If the photo ID has multiple comments, one tuple will be created for each of the comments. We discard all samples that does not have at least one comment and those samples with missing attributes.

3.4 Recommendation

The proposed tourist attraction recommendation method has two phases. In the first phase, we learn a model that predicts the sentiment of short sentences with the Sentiment140 dataset. Essentially tweets are some short pieces of text that express the current thoughts or status or the user that tweets. We believe that tweets and comments collected from Flickr have a similar distribution and thus the model learned from Sentiment140 can generalize well to comments from Flickr and predict the sentiment for those comments. The prediction made by the sentiment analysis model helps us

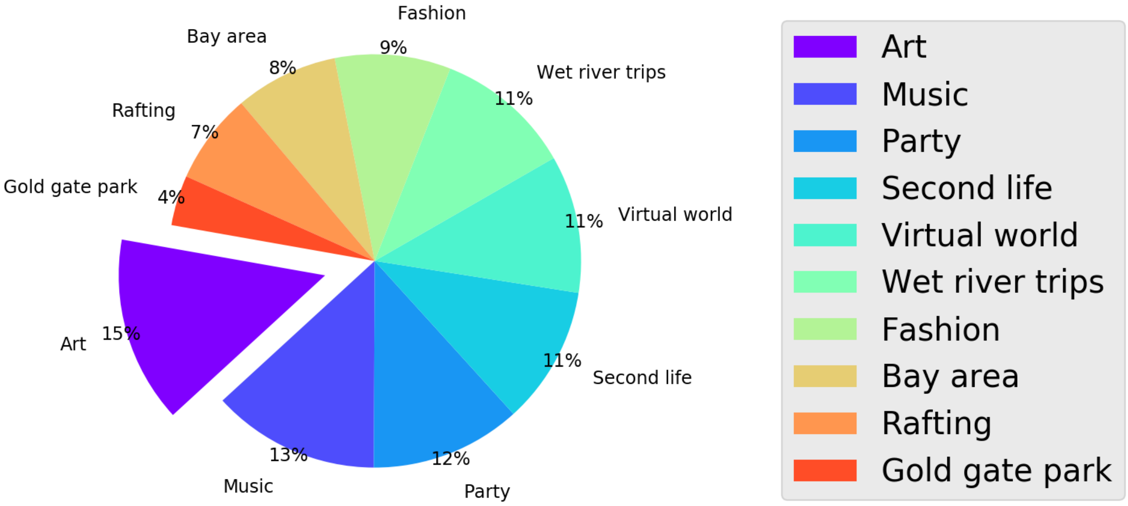

Figure 2: Tag analysis for San Francisco

understand the attitude of other Flickr users towards a specific media item and thus the tourist attraction where the media item is taken or captured. To this end, we plan to learn an embedding from the union of the Sentiment140 corpus and the Flickr comment corpus. After that, a recurrent neural network is trained on the Sentiment140 corpus to predict a single sentiment output from input sequences of variable lengths. Because the Flickr comment corpus has no sentiment information we cannot validate the model directly on the Flickr comment data. Instead, we set aside a part of samples from the Sentiment140 corpus for model validation. For the second phase, we apply the sentiment analysis model on YFCC100M to estimate the popularity of potential tourist attractions and present customized recommendations to the users based on that. The first step is to perform data preprocessing on YFCC100M. YFCC100M only contains metadata for the media contents on Flickr. Among all attributes in the metadata, only some of them are available in all rows in the dataset while the others are optional and may not be present in some items. For example, if the user chooses not to tag the media item with the location where the item is taken then the geographical information(latitude, longitude) will be missing from the row. To handle this, we first perform data cleaning to filter out all rows with incomplete information(e.g. missing geographical location). Because YFCC100M covers a wide range of media contents, we only select the items that are related to tourist attractions. This is done by filtering the rows based on the tags associated with them. We only use rows that have at least one tag related to locations, e.g. park and beach. After data cleaning, we use the Flickr API to pull the user comment list for each of the selected samples. After we obtain all the data required by the recommendation task, the next step is to identify potential tourist attractions from the data and to find the set of samples related to each of them. There are two methods for this identification task. The first method is that we can provide a pre-defined list of coordinates of some popular tourist attractions and collect the samples for each tourist attraction based on some distance metrics. The second method is that we can define some criteria and discover tourist attractions from the data automatically. One approach to do this is to partition the whole geographical region of interest into many small sub-regions and count the total number of samples that are captured in each of them. After that, we can sort the list of sub-regions based on the number of samples to identify some regions with a large number of visitors. Another approach is to first categorize the samples based on their location tag, e.g. park. This will give us a list of samples that are known to be captured in a park. These samples are clustered with some distance criteria to generate some clusters of samples that are captured in the same park. This approach allows us to specify a category when discovering potential tourist attractions. The outcome of tourist attraction identification is a list of potential tourist attractions, each of which is associated with the set of samples related to it. With this information, we can estimate the popularity of each tourist attraction using a set of pre-defined dimensions. These dimensions include statistical information such as the total number of visitors in the area and also some derived features from the dataset. We will also apply the sentiment analysis model that we create in the first phase to analyze the overall attitude towards each potential tourist attraction. Based on such analysis, we can build a popularity ranking for this tourist attraction and present to the user a preference-independent recommendation by finding the nearest tourist attraction to the user with high popularity. Besides the preference-independent recommendations, we can also make personalized tourist attraction recommendations based on the user's historical travel data. For each user in the Flickr corpus, we can match the samples uploaded by the user with the list of tourist attractions discover from the dataset to get the travel history of the user. When predicting recommendations for a new user, we can find some users in the dataset with similar travel history with the new user and recommend the tourist attractions visited by similar users but the new user has not yet visited.

3.5 Sentiment Analysis

The problem of sentiment analysis is to predict the sentiment for a sentence input. The sentiment is represent as a real number $y \in [0, 1]$. A label that is close to 0 means that the sentence has negative sentiment and a label close to 1 means that the sentence has positive sentiment. In this project, we train a deep learning model for the sentiment analysis task. The deep learning model consists of two components. The first component is a Bidirectional Encoder Representations from Transformers (BERT) model for feature extraction. The BERT has a tokenizer model which is used to break the whole sentence into many tokens. After that, the BERT model learns a representation from the dataset which maps each token into a vector. The vector representation extracted by the BERT model is fed into the second component which predicts the sentiment of the input sentence from the vector representations. The second component is a Recurrent Neural Network (RNN) with a Gated Recurrent Units (GRU) cell. The output features produced by the GRU cell go through a Dropout layer and then a fully connected layer with sigmoid activation function is used to transform the GRU output is the final classifier output. To predict the sentiment for a sentence, each word in the sentence is mapped to a vector by the BERT model. After that, the vector representations for all words in the sentence are fed into the GRU cell to update its hidden state. When the sentence ends, a special token is fed into the model and the sentiment of the whole sentence is calculated by the classifier output. For the model training, we use a pretrained model for the BERT model and then fine-tune the RNN output head with the IMDB dataset. The IMDB dataset is divided into a training set of 17500 samples, a validation set of 7500 samples and a testing set of 25000 samples. The RNN head is trained with the training set for 5 epochs.

3.6 Tourist Attraction Discovery

In order to discover all popular tourist attractions in the dataset, we use the k-means clustering algorithm. The k-means clustering is implemented with MapReduce which uses a identity mapper on the comment dataset downloaded in Section. 3.3. The identity mapper simply forwards the input to the reducer. The reducer maintains a list of the current coordinates of the centroids. For each tuple produced by the mapper, the reducer compares the latitude/longitude with the centroids to find which cluster the new tuple is in. After that, the centroid of the new tuple is updated by re-calculating the mean value of all points in the cluster. The MapReduce job is run repeatedly until converge, *i.e.* the delta of centroids between two consecutive iteration is below a threshold. The clustering allows us to discover the most-visited tourist attraction from the dataset. After the cluster centroids are found, we group all samples by the clusters they are in. We apply sentiment analysis to the comments of samples in each cluster to predict the attitude (positive or negative) towards the tourist attraction represented by the cluster centroid.

4 EVALUATION

4.1 Datasets

To provide recommendations for tourist attractions/cities, we have to first create a popularity ranking for a list of selected attractions/cities. To this end, we will use the Yahoo Flickr Creative Commons 100 Million Dataset(YFCC100M) [9] as the data source. The YFCC100M dataset consists of a large collection of online posts gathered from Flickr. The total number of multimedia items in the dataset is 100 million with approximately 99.2 million photos and 0.8 million videos. These collected multimedia items are all uploaded between 2004 and 2014 when the dataset was created. The multimedia items collected in the dataset are published under a Creative Commons commercial or non-commercial license so they are available for research purposes. The dataset is distributed as a compressed archive that can be downloaded from Amazon Web Service. At the time of writing, the dataset contains roughly 14GB of data. The dataset itself only provides the metadata of multimedia items posted on Flickr. Each media object in the dataset is represented as a tuple including some attributes that describe some general information about this item. The most important attributes are the Flickr identifier which is uniquely assigned to each item by Flickr, the user who created this specific multimedia item as well as the date and time when the multimedia item is taken and uploaded to Flickr. Besides such information, each tuple also has the title, the descriptions as well as the direct link to its page on Flickr. Social information like comments, favorites and followers/following information of the author of the multimedia item are not included in the dataset because such information may change day by day. Although the social information is not directly obtainable from the dataset itself, we can use the photo ID provided by the dataset to query the social information about the multimedia item with the Flickr API. The Flickr service exposes a list of APIs for the comments/favorites related to a single multimedia item and also the follower/following list of a specific user. When uploading multimedia items to Flickr, many users prefer to attach a geographical location to the item that describes the place where the item is taken/created. Such location information is also included in the dataset as an optional attribute for each multimedia item. In the dataset, there are 48,366,323 photos and 103,506 videos that have been annotated with a geographical location(more

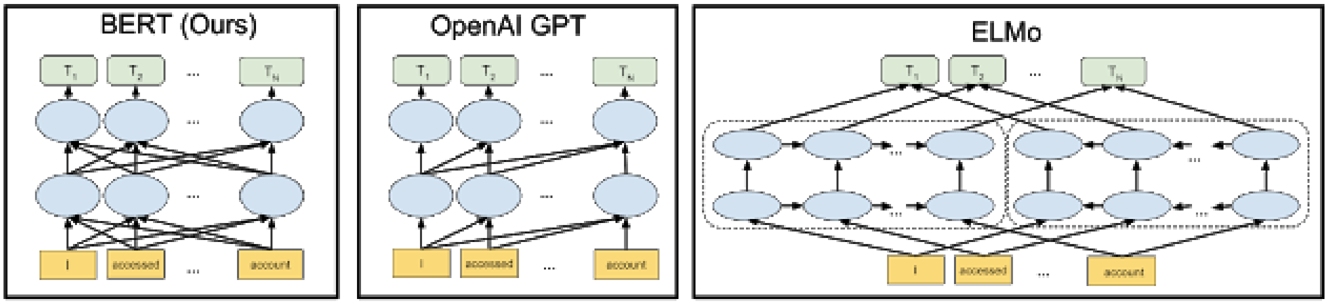

Figure 3: The architecture of BERT

specifically, the latitude and the longitude) either selected by the user manually or provided by GPS automatically. Overall, the dataset contains multimedia items that are captured from 249 different countries/regions and also from international waters and international air space. The geographical location information allows us to obtain features(number of visitors, photos and comments) for a single tourist attraction or a city. Besides geographical locations, a subset of multimedia items in the dataset is also annotated with a set of tags by the user. These tags identify some categories for the entities in the multimedia items. These categories are related to people, animals and locations, e.g. park or beach. Among the whole set of data, a total of 68,552,616 photos and 418,507 videos are annotated by the users. Besides these tags annotated by humans, there are also 3,343,487 photos and 7,281 videos that have tags generated by a machine, e.g. a camera or other automated systems that capture the media item. With the URLs in the metadata, we can retrieve the original media contents from Flickr. These original media contents range from photos of real-world tourist scenes to snapshots of holiday events. The media contents and their comments on Flickr provide a corpus with rich contents that helps us identify potential tourist attractions and estimate the popularity of them. We believe the user comments made to some specific media contents can also help us make more accurate recommendations on tourist attractions. To better utilize such information, we adopt the Sentiment140 dataset [1] for sentiment analysis of user comments. The corpus contains 1.6 million tweets extracted using the Twitter API [3]. Each row in the corpus is a tuple that includes common attributes like tweet ID, the date when the tweet is posted, the user who tweeted and the full text of the tweet. All samples in the corpus are labeled with the user's sentiment. The label can be 0 for negative, 2 for neutral or 4 for positive. The total size of the corpus is 228MB.

4.2 Sentiment Analysis

To train the model for sentiment analysis, we use the IMDB dataset which contains 50000 movie reviews collected from the IMDB website. Each movie review is labeled as positive or negative. We randomly select 25000 samples as the testing set, 17500 samples as the training set and 7500 samples as the validation set. The model is fine-tuned with the training set for 5 epochs. We use a binary accuracy criterion to evaluate the trained model. If the classifier output is below 0.5, the sample is classified as a negative sample and otherwise it is predicted as a positive sample. The validation error after training is 92.3%.

4.3 Sentiment Map

To discover popular tourist attractions in the dataset, we first run k-means on the dataset to cluster the samples into many geographical regions. After downloading the comments for the samples in YFCC100M and removing the ones without any

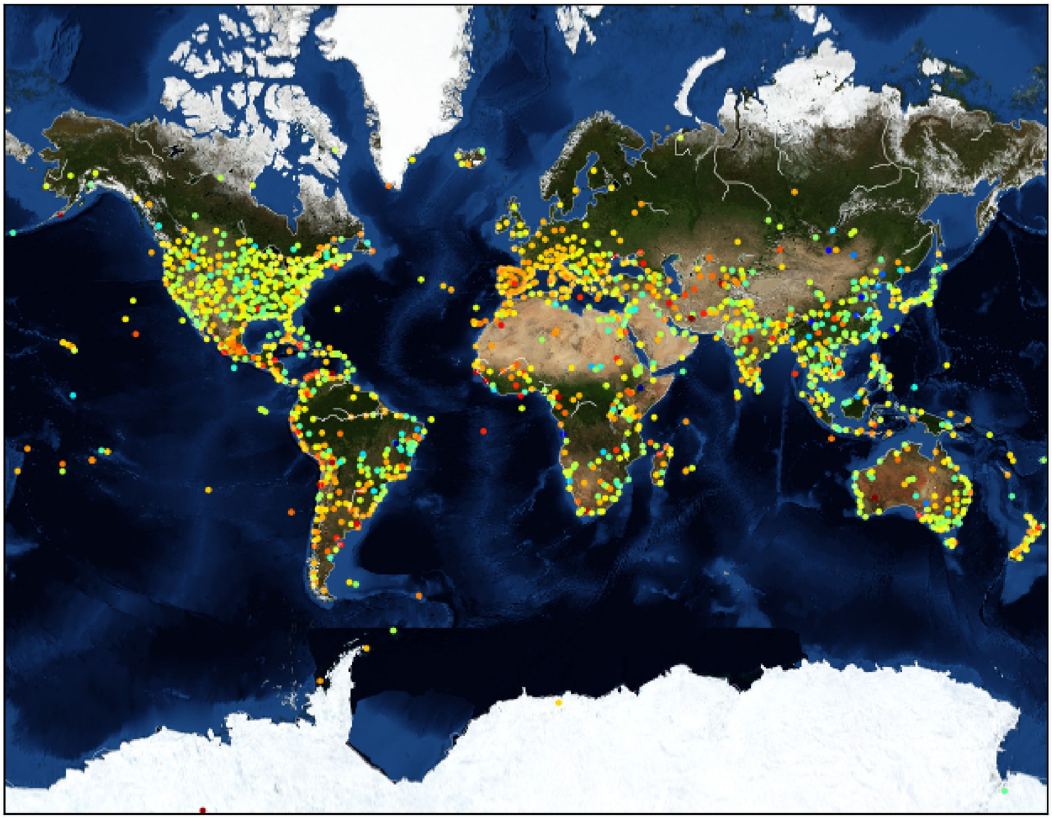

Figure 4: Sentiment Map All Over The World

comments, we have 358127 rows in total. We cluster the sample into $k = 10000$ clusters and remove the clusters without any comments. The number of valid clusters is 1453. After that, we predict the sentiment for the comments in each region and the average sentiment in the region as the score for the region. The sentiment map is shown in Fig. 4. In the figure, each point is a popular tourist attraction. The color of points is the sentiment of the attractions. If an attraction is red in the map then the attitude towards this attraction is positive and the attitude is negative if the point is blue.

5 RELATED WORK

Sentiment analysis. Sentiment analysis has caught a lot of attention and has been applied in many different industries with the rapidly increasing growth of social networks. The tremendous data from social networks have become an important resource for sentiment analysis and the volume of them are keeping increasing. In 2019, twitter has 330 million monthly active users and there are 500 million tweets sent per day [3]. In the third quarter of 2018, Facebook had 2.375 billion monthly active users [8]. These social network applications have produced quantities of data like texts, images and videos, which can be used for sentiment analysis. For text data, [7] has collected 300,000 text posts from Twitter to analyze their opinions on different topics by using n-gram based classifier. As for Image and video data, there are also a lot of work on them by using computer vision and deep neural network to extract the information from them, like [4, 11]

Graph analysis. Generally, social networks can be regarded as a graph with a series of interconnected vertices. Taking the largest social network, Facebook as an example, users, as nodes, are connected by the friendship relationship, as vertices. There are quantities of researches [10] about the properties of Facebook graph, which indicates the global structure and lots of other properties of Facebook graph. And with the increasing scale of social networks, there are also some researches [4] on complex graph processing workflow which can deal with graph with up to trillion edges.

Recommendation algorithms/systems. Recommendation Systems have become one of the most popular topics in recent years. It can be applied in a lot of different industries like video recommendation or product recommendation [5, 6]. In this project, we tried some recommendation algorithms to provide the tourist recommendations.

6 CONCLUSION

In this project, we design a sentiment analysis based system for tourist attraction recommendation. The system provides recommendations by ranking the tourist attractions in the YFCC100M dataset using the sentiment prediction on the comments. Experiment results show that our system can discover some popular tourist attractions in the world and make recommendations accordingly.

REFERENCES

[1] 2009. Sentiment 140 Dataset. http://help.sentiment140.com/home..

[2] 2019. flickr - The App Garden, flickr.photos.comments.getList. https://www.flickr.com/services/api/flickr.photos.comments.getList.html..

[3] 2019. Twitter by the Numbers: Stats, Demographics Fun Facts. https://www.omnicoreagency.com/twitter-statistics/..

[4] Avery Ching, Sergey Edunov, Maja Kabiljo, Dionysios Logothetis, and Sambavi Muthukrishnan. 2015. One trillion edges: Graph processing at facebook-scale. *Proceedings of the VLDB Endowment* 8, 12 (2015), 1804–1815.

[5] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, et al. 2010. The YouTube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems*. ACM, 293–296.

[6] Carlos A Gomez-Uribe and Neil Hunt. 2016. The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)* 6, 4 (2016), 13.

[7] Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining.. In *LREc*, Vol. 10. 1320–1326.

[8] Kit Smith. 2019. 53 Incredible Facebook Statistics and Facts. https://www.brandwatch.com/blog/facebook-statistics/..

[9] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2015. YFCC100M: The new data in multimedia research. *arXiv preprint arXiv:1503.01817* (2015).

[10] Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. 2011. The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503* (2011).

[11] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2015. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *Twenty-ninth AAAI conference on artificial intelligence*.