PREDICTING INTERNATIONAL TRAVEL DESTINATIONS FROM GEOTAGGED FLICKR PHOTOS

by

Ugochukwu Francis Umeokafor, B.Sc.

A directed research report submitted to the Geography Department of Texas State University in partial fulfillment of the requirements for the degree of Master of Applied Geography with a specialization in Geographic Information Science

December 2017

Committee Members:

Dr. Yihong Yuan, Chair

Dr. T. Edwin Chow

FAIR USE AND AUTHOR'S PERMISSION STATEMENT

Fair Use

Duplication Permission

ACKNOWLEDGEMENTS

First I would like to thank my advisor, Dr. Yihong Yuan, who spent uncountable hours reviewing my work in order to give explanations, comments, and revisions. I thank her for all her guidance, support, encouragement and motivation that she provided throughout my directed research process. I would also like to thank my committee member, Dr. T. Edwin Chow for his tremendous support, comments and questions regarding my research. Great thanks also go to Guixing Wei for all his help with data processing. Finally, I would like to thank my parents for all their love, support and encouragement. I will not be where I am today without their support. Thank you.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER 1

Introduction

The emerging use of social media and rapid development of mobile technologies has created a large volume of geotagged photos to be shared online from all over the world. The boom in mobile phone usage and social media in the past few years has increased the number of photos shared on social media sites like Facebook, Twitter, Instagram, Flickr, Weibo, etc. [1]. A large number of these shared photos posted by users are geotagged, which makes it possible to extract the location, date and time of when the photos were taken [2]. This boom has prompted a recent increase in human mobility research and location-based social media studies. The scale of information and data provided by social media far exceeds the traditional method where travel data was collected by surveys and global positioning system (GPS) [3]. Social media data has helped minimize the problems of privacy concerns in human mobility studies[18]. It is also more scalable compared to the traditional methods. Social media data has been used by travel companies like Expedia, Trivago, Venere, Travelocity, Orbitz, and HomeAway to recommend a list of destinations for a traveler who is looking for vacation destinations or places of interest. Social media data has been used for individual purposes like travel destinations planning [4], and it is also essential in tourism planning by city and state governments and studying of flow of travel between different countries.

Modelling and understanding human mobility patterns at different spatial and temporal scales is an important research topic in many application fields, ranging from urban and transportation planning to tourism planning, resource allocation and prediction of migration flows [16]. The challenging task is how this enormous amount of information can be utilized on a worldwide scale, providing an understandable and useful resource to manage travel destination and tourism [13].

The purpose of this research is to explore the historical travel records of frequent travelers from the United States based on the photos they uploaded on Flickr. We used the travel records to study user behaviors, and travel flows between different countries. Secondly, using the same dataset, we used machine learning algorithm Multinomial Naïve Bayes model to predict users' next travel destinations (visits) based on their historical visits.

Our results provide useful travel recommendation for future travelers to explore possible destinations, as travelers with similar interests are likely to visit similar locations [2,3]. Our result shows the visiting pattern between different countries and will benefit policymakers regarding urban and transportation planning, tourism planning, and resource allocation. It will also provide a useful resource for individuals and tourist companies in recommending new destinations to prospective travelers based on where travelers from the same country have visited in the past.

CHAPTER 2

Literature review

2.1 Travel Destinations Prediction Overview

Location has played a pivotal role in recent location-based social media research as information about the location of users has enabled numerous compelling location-based services [5]. The availability of GPS-equipped mobile devices today has made it possible to determine the locations of users from the photos they share on social networking websites like Flickr, Instagram, Twitter, Foursquare and Facebook [6] or from recorded GPS data on taxis and trucks [5]. Travel destinations have been predicted by previous studies using different machine learning algorithms, for example, Multinomial Naïve Bayes Model, Markov Model, Support Vector Machine, Topic model, etc. Some of the previous research that used prediction algorithms for destination prediction includes a study by Krumm and Horvitz [5]. The authors used GPS data collected from 169 different drivers who participated in Microsoft Multiperson Location Survey (MSMLS) and Bayesian inference to predict the destination of a driver as a trip progresses. In another work by John Krumm [7] on real-time destination prediction based on efficient routes, Krumm used GPS data gathered from drivers and Bayes rule to predict the destination of a driver halfway through the drive. Similarly, Karbassi & Barth [8] developed a model to process historical GPS data from shared-use vehicle systems to extract the most common routes between five pre-designated locations. Their model uses the highest probability from pre-computed route probabilities and map-matching algorithm to estimate routes as the vehicle is moving. Given the destination, their goal is to predict the route in order to estimate arrival times. Another study by Schmandt & Marmasse [9] on user-centered location model developed an algorithm called comMotion.

ComMotion is a location-aware computing environment which links personal information to location in its user's life and predicts where the user is going and estimate time to destination [9]. The authors used several pattern recognition models including the Bayes classifier, Histogram modeling and Hidden Markov Model for route learning. Similarly, Ashbrook & Thad [10] in their research aims to learn significant locations and predict movement across multiple users, used movement history from GPS dataset and Markov model to predict a user's future movements. They tested their model on GPS data collected from two different locations - one in Atlanta and another in Zurich. Their model generated consistent results across multiple users.

2.2 Modelling User Behavior from Geotagged Photos

Photo-sharing sites such as Flickr, Instagram, Twitter, Weibo and Facebook contain vast amounts of potential information about our world and human behavior [14]. Travel behaviors vary among different groups of tourists [3]. Travelers from different countries will have different preferences in choosing travel destinations.

Analysis of movement is traditionally performed by GPS devices which lacks semantic meaning and background information [12]. For instance, previous studies on location-based social media, travel destinations prediction and user behaviors were mostly conducted based on manually collected GPS data and survey-based travel diary data [18]. Traditional data collection method provided useful information for city planners to understand the origin, destination, and flow of tourists [18] but it lacks information on the behavior of a tourist. E-tourism service providers present travelers with estimated visit duration to different cities. The estimates are provided by field experts and users, but it varies in different cities [19].

To understand the behavior of a tourist, we relied on the boom in social media sites; these sites provide information about the behavior of a tourist by helping us to understand the amount of time spent at different locations in a city based on the timestamp of posting pictures on the social sites. With Location based social media research, researchers have been able to estimate the amount of time a tourist spends at different landmark locations and predict their favorite landmark at the end of their trip [19].

For example, Girardin et al., [18] studied tourist dynamics using platform explicitly disclosed location information from Flickr. They designed geo-visualization models to reveal the tourist activity and flows in space and time. The authors retrieved 81,017 photos taken by 4,280 photographers over a period of two years from the popular photo-sharing web platform Flickr. Based on the time and the disclosed location of the photos, they extracted records of their presence and performed statistical analysis that separates visitors from inhabitants of the city being used in the study (Florence). These outputs allow the evaluation of the potential of using people-generated geographically referenced information to contribute to understanding how people travel and experience the city [18]. The authors identified four constraints that affect capturing of mobility data for travel surveys. The four constraints are scalability, longitudinal studies, individual consents and privacy issues affect the following methods of performing travel surveys; GPS, GSM (device –based), GSM (aggregated network-based) and Bluetooth. The authors argue that while social media data does not overcome all the constraints, it has improved human mobility and human behavior studies. Their result shows that social media data from explicitly disclosed spatio-temporal data coming from public web platforms can overcome these constraints and provide additional insights in understanding the dynamics of travelers.

Movement of mobile phone users can be captured as trajectories using the time-stamped and geotagged photos or check-in records shared by the users on social networking sites. The study by Kurashima et al. [11] used geotagged photos from social media photo-sharing websites for route recommendation. Their recommendation was characterized by the introduction of traveler's interest and the desired time to spend at a location instead of relying on the construction of routes based on a user rating system [11]. The authors used four probabilistic models (Multinomial model, Markov model, Topic model and Markov-Topic model) to predict the next landmark to be visited by a traveler. The studies by Kisilevich et al. [12] and Kurashima et al. [11] used Flickr geotagged photos to draw a route recommendation. Procedures in their studies involved: the extraction of photos through the Application Program Interface (API) of the social networking site and "running analysis to cluster, infer relationships and extract features from the collected data" including the information stored in tags which are provided by user input [13, 11,14, 4, 1]. However, clustering techniques can show the attractiveness of places without the involvement of temporal dimension [12] while user inputs (tags) can be utilized for individual route recommendation.

The studies by Memon, et al. [2], Sun, et al., [15] and Serdyukov, et al. [16] all used geotagged photos in travel recommendation, but none of the research predicted the next destination of a frequent traveler. Memon et al. [2] used the user's travel history to predict the user's preference in Chinese cities. The authors used density-based clustering algorithms to cluster user locations and associated geo-tags and then modeled user preferences and user similarities. The authors were able to predict tourist preferences at the city level more precisely.

In the study of Sun et al. [15] "Road-based travel recommendation using geotagged images”, they presented a new travel recommendation approach integrating landmarks and travel routings. The authors proposed a novel approach in which the primary unit of routing searching is separate road segments instead of the traditional GPS enabled devices. Using spatial clustering method, the authors generated clusters of images, identified essential landmarks from the clusters, and ranked the landmarks in terms of tourism popularity. The authors also calculated the tourism popularity of roads and identified geotagged images posted in those locations. They generated appropriate routing between two landmarks using recommendation index (a variable that quantifies the overall popularity of the road). Their model demonstrates that the approach can recommend the user suitable routings considering the images, points of interests (POI) and road length [15].

A lot of studies have developed different models which have been used in the study of geotagged photos, landmark identification, user behaviors and destination predictions. For example, in the study of Beira et al. [16] to predict human mobility through the assimilation of social media traces into mobility models, they proposed a hybrid model of human mobility that integrates a large-scale publicly available dataset from Flickr with classical gravity model under a stacked regression procedure. The result shows that the hybrid models afford enhanced mobility prediction at different spatial scales.

Popescu & Grefenstette [19] used geotagged Flickr photos to study the duration of time it takes a tourist to visit a tourist attraction site, which varies between different tourist attraction sites. Their study provided more insight into the duration of time it takes a tourist to visit a particular site based on the timestamp on the photos posted to social networking sites. In a similar study, Hung et al. [6] also presented a new trajectory pattern mining framework called Clustering Clues of Trajectories (CCT), which is used to discover trajectory routes that represent frequent movement behaviors of a Flickr user. The authors proposed two methods, one clue-aware trajectory similarity which is used to measure the clues between two trajectories and clue-aware trajectory clustering algorithm to cluster similar trajectories into groups to capture the movement behaviors of the user.

Other studies have used geotagged photos posted on the social networking sites by tourists to study and identify popular landmarks in different cities. In a study by Zheng et al. [20], they developed a landmark recognition engine which automatically mines frequently photographed landmarks from a large collection of geotagged photos. The authors performed clustering on mined hashtags, GPS coordinates and images of popular landmarks in photo sharing websites like Picasa and Panoramio in other to extract landmark names. They also retrieved landmark names from the travel guide articles from websites, such as Wikitravel. The landmark recognition engine developed by Zheng et al. is used to recognize the presence of a landmark in an image and also contributes to a worldwide landmark database that organizes and indexes landmarks, regarding geographical locations, popularities, cultural values and social functions. Zheng et al. [20].

Similarly, Girardin et al., [18] leveraged on explicitly disclosed location information to understand tourist dynamics. In their research, they analyzed tourist flows in the Province of Florence, Italy based on a corpus of geo-annotated Flickr photos. The authors used a geo-visualization model to reveal the tourist concentration, points of interests and spatiotemporal flow of tourists in the city of Florence. Also, Popescu & Grefenstette [19] deduced trip-related information from Flickr photos by using an automatically constituted gazetteer [21] and sets of geotagged and time-stamped photos to extract information about trips described by Flickr photos. The author's used geographical gazetteer from Popescu et al., [21] which has a broad coverage of tourist site names, site types and their GPS coordinates to validate their results. The result shows that from many individual behaviors, they can thus estimate user behavior by averaging all the times found for each visitor attraction, and automatically add these estimations of visit times to the tourist sites descriptions in the gazetteer from Popescu et al., [21]. Their result deduced visiting durations for attraction sites in four major cities in London, New York, Paris and San Francisco.

Other studies have used geotagged photos to predict travel behaviors of tourists. A study by Clements et al. [23] used Flickr geotagged photos to predict user travel behavior. The author's used mean shift algorithm and Gaussian kernel to predict a user’s favorite locations in a city based on the users Flickr geotagged photos uploaded in other cities. Another study by Rattenbury et al., [22] extracted geographic and events database from Flickr metadata using a burst analysis technique by which 85% of the automatically mined place names were correct and were also approximately situated. Similarly, a study by Kisilevich et al. [12] on spatiotemporal clustering derived user’s trajectories based on geotagged photos and time span given by the photo session (time spent from the first taken photo to the last). The authors used trajectory clustering algorithm to study clustering across domains. In the same way, Crandall & Snavely [14] used online photo collections to reconstruct information about the world and its inhabitants by developing automatic algorithms that analyzed large collections of imagery to understand and model people and places at a global scale and local scale. The result of their model for analyzing the world at a global scale can automatically create annotated world maps by finding the most photographed cities and landmarks, inferring place names from text tags, and analyzing the images themselves to identify “canonical” images to summarize each place [14].

2.3 Multinomial Naïve Bayes Model

Multinomial Naive Bayes is a first-order probabilistic classifier and a specialized version of Naïve Bayes. According to McCallum & Nigam, [24], Bayesian probabilistic approaches make strong assumptions about how the data is generated, and posit a probabilistic model that embodies these assumptions; then they use a collection of labeled training examples to estimate the parameters of the generative model. Multinomial Naïve Bayes Model captures word frequency information in documents (in our case country codes). In the Multinomial Model, a document is an ordered sequence of word events, drawn from the same vocabulary V. We assume that the lengths of documents are independent of class. The authors made a similar Naive Bayes assumption that the probability of each word event in a document is independent of the word's context and position in the document. Thus, each document d_i is drawn from a Multinomial distribution of words with as many independent trials as the length of d_i. This yields the familiar “bag of words" representation for documents. N_{it} is defined as the number of times word w_t occurred in document di. Then, the probability of a document is simply the Multinomial distribution:

$$\mathrm{P}\left(\mathrm{d_i}\middle|c_j;\theta\right) = P(|\mathrm{d_i}|)|\mathrm{d_i}|!\prod_{t=1}^{|V|}\frac{P\left(\mathrm{w_t}\middle|\mathrm{c_j};\theta\right)^{\mathrm{N_{it}}}}{\mathrm{N_{it}}!} \tag{1}$$

In the study by Domingos & Pazzani [25] on the optimality of the simple Bayesian classifier under zero-one loss, the authors argued that the question about the optimality of the simple Bayesian classifier has not been answered because it performs well only in domains containing explicit attribute dependencies. The authors show that, although the Bayesian classifier’s probability estimates are only optimal under quadratic loss if the independence assumption holds, the classifier itself can be optimal under zero-one loss (misclassification rate) even when this assumption is violated by a wide margin [25]. The authors addressed that Bayesian classifier has a much greater range of applicability than previously thought. They show in their study that Bayesian classifier was shown to be optimal for learning conjunctions and disjunctions. Bayesian classifier even when it is not optimal may still perform better than classifiers with greater representational power, such as C4.5, PEBLS, and CN2 [25]. According to Domingos & Pazzani, [25] Bayesian classifier often out-performs more powerful classifiers for common training set size and numbers of attributes.

Another study by Friedman et al. [26] argued that supervised learning has shown that a surprisingly simple Bayesian classifier with strong assumptions of independence among features, called Naive Bayes, is competitive with state-of-the-art classifiers such as C4.5. This fact raises the question of whether a classifier with less restrictive assumptions can perform even better [26]. According to Friedman et al. [26], One of the most effective classifiers, in the sense that its predictive performance is competitive with state-of-the-art classifiers, is Naive Bayes classifier. This classifier learns from training data the conditional probability of each attribute. Classification is then done by applying Bayes rule to compute the probability of C given the particular instance of A1,..., An, and then predicting the class with the highest posterior probability. This computation is rendered feasible by making a strong independence assumption: all the attributes A1…., An are conditionally independent given the value of the class C. By independence we mean probabilistic independence, that is, A is independent of B given C whenever Pr(A|B,C) = Pr(A|C)for all possible values of A, B, and C, whenever Pr(C) > 0 [26].

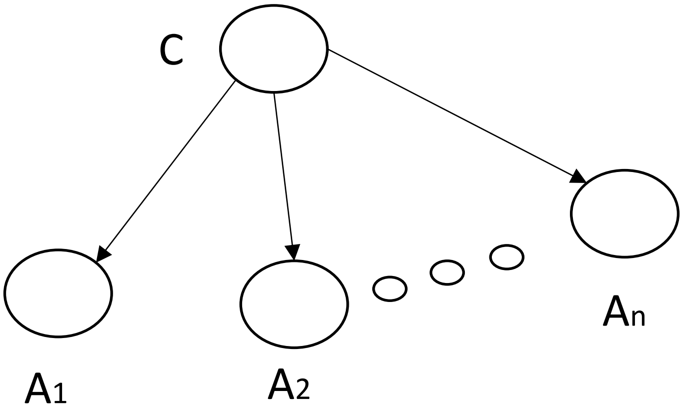

Figure 1. The structure of the Naive Bayes network (Friedman et al. 1997).

Multinomial Naïve Bayes has been used in a lot of location-based studies. Hobel et al. [27] adapted Multinomial Naïve Bayes in their study on deriving the geographic footprint of cognitive regions. The authors described Multinomial Naïve Bayes as a powerful machine learning model. The authors [27] used a semantic representation of geographic regions extracted from a GIS and passed over to machine learning algorithm (Naïve Bayes) which learns from the pre-classified samples and locates other areas according to semantic similarity.

In a study on monitoring public health concerns health information visualization by Ji et al. [28], the authors used tweets to keep track of spreading epidemics, the locations of disease spread and also understand the concerns of the population on the disease outbreak. In their study, Ji et al. [28] employed three machine learning-based algorithms; Naïve Bayes, Multinomial Naïve Bayes and Support Vector Machine. Multinomial Naïve Bayes achieved overall best results and took significantly less time in building the classifier. Multinomial Naïve Bayes has been shown to perform well in a variety of domain in machine learning, Bo et al. [29] adapted Multinomial Naïve Bayes in Geolocation prediction for two reasons (1) it incorporates a class prior, allowing it to classify an instance in the absence of any features shared with the training data; and (2) generative models outperform discriminative models when training data is relatively scarce [29, 30], other studies for example (Ashbrook & Starner [10], Marmasse & Schmandt [9], Kurashima et al. [11], Kurumm & Horvitz [5, 7] all adapted multinomial Naïve Bayes in their respective studies.

We choose multinomial Naïve Bayes as our model for this research because it has been proven by many studies (e.g.; Hobel et al. [27], Ji et al. [28], Bo et al. [29], Kurashima et al. [11]) to out-perform other powerful machine learning algorithms. Multinomial Naïve Bayes performs well with a large training dataset. Our Flickr dataset contains over 17 million features. The number of distinct countries visited by all users in our training dataset will be the vocabulary size, while the travel history of the users and the countries visited by the users were used as the training features. Multinomial Naïve Bayes is a very good machine learning model because it learns from pre-classified samples (training vectors) and classifies other unclassified feature vectors according to their similarity with the given training vectors [27]. Multinomial Naive Bayes model assumes that the features are conditionally independent of one another, it incorporates a class prior, allowing it to classify an instance in the absence of any feature shared with the training data and generates a posterior probability from the test data [29]. Multinomial Naïve Bayes model is very efficient because it takes less time to build as compared to other machine learning or classification algorithms [28].

Chapter 3

Methodology

3.1 Dataset and Pre-processing

In this research, we used publicly available Creative Commons dataset published by Flickr in 2014. The dataset contains over 48 million photos posted by over 216,000 users randomly sampled globally between 2004 – 2014[1]. The geotagged photos contain attributes supplied by the user when they uploaded the photo. These attributes include date and time, longitude and latitude, country code, etc.

We cleaned the noise in the data by selecting only the users from the United States who uploaded more than three geotagged photos outside the United States. This is to improve the accuracy of the model, and we want to make sure that the result will not be skewed by users who visited only one or two countries or users who uploaded photos with no country code.

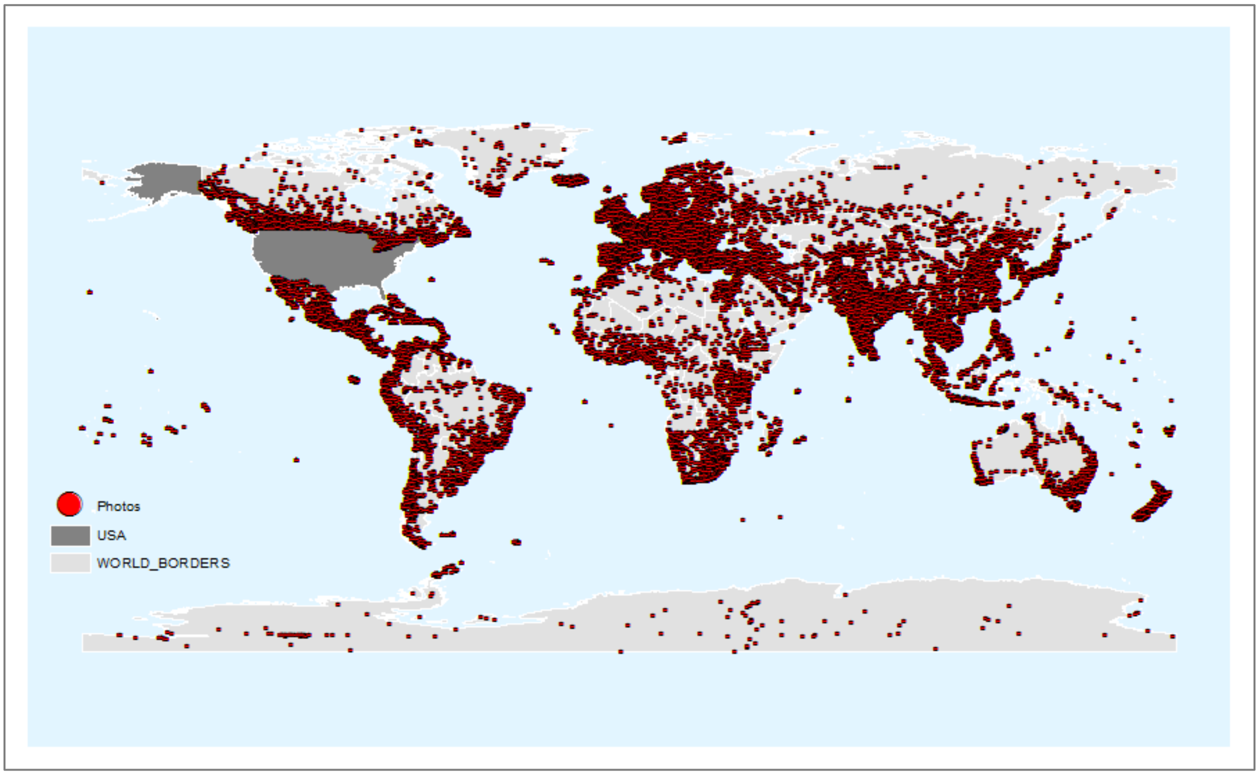

Figure 2. Location of photos posted by United States users

From the users that met all the above requirements, we selected ten countries with the highest number of photos uploaded as shown in Table 1. We also calculated number of photos posted in each country.

Table 1. Top 10 countries according to photos uploaded

Country Name	Country Code	Number of Photos Uploaded
United Kingdom	206	4,740,726
Spain	187	2,219,985
France	64	1,921,398
Germany	71	1,766,505
Canada	23	1,606,389
Italy	85	1,375,926
Japan	88	1,233,758
Australia	8	1,017,326
Netherlands	153	823,044
China	29	565,325

As part of summarizing the descriptive statistics, we used probability theory to calculate the direction of travel and different probabilities of visits among the ten countries selected by U.S travelers based on their past travel history. For example, we calculated the probability of a user from the United States visiting other countries in Europe after their first visit to the United Kingdom or the probability of visiting other countries in Asia after visiting Japan. For each of the countries, we calculated the total number of United States visitors who visited the country with the study period as shown in Figure 3.

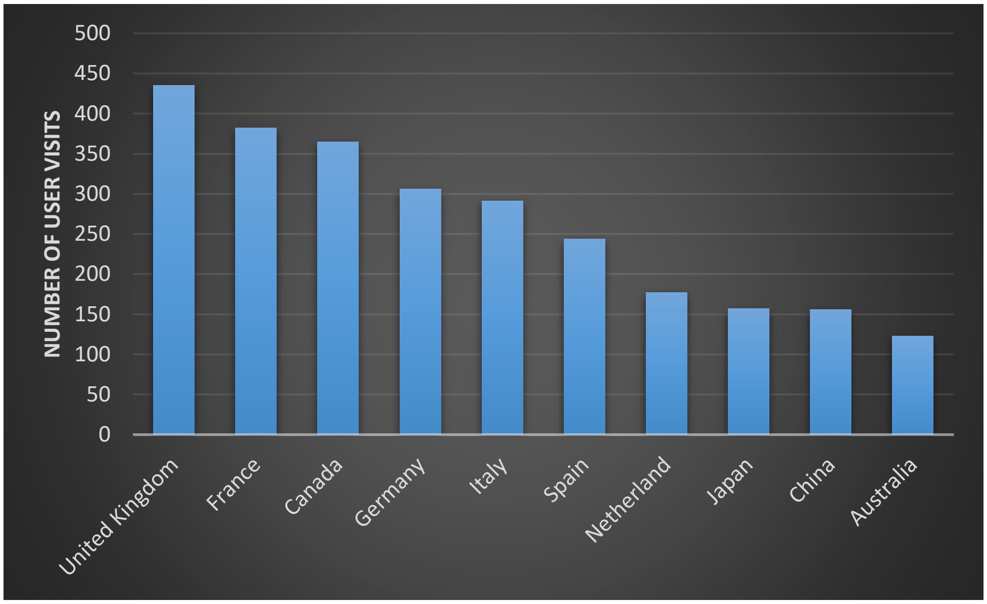

Figure 3. Visiting patterns for United States travelers

To calculate the probability of visits, for each country pair among the ten countries selected, we calculated the total number of users that visited the first country before visiting the second country based on the timestamp of when the photo was uploaded; we call that number N1. The total number of users with geotagged photos that met the requirements of our analysis is N2.

For probability calculation, $\mathrm{P} = \frac{N1}{N2}$, we repeated this for all the countries selected. Table 4 shows travel directions between different countries.

3.2 Analysis Framework

Bayes theorem uses the combination of prior and likelihood probabilities from training data to calculate posterior probabilities for the test data.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \qquad (2)$$

Bayes rule describes the likelihood of an event happening based on prior events. P(A) is the prior probability. P(B/A) is the likelihood or conditional probability, which is the probability of observing event B given that A is true. P(A) and P(B) is the probability of event A or B happening, and it is independent of other events. P(B) is called model evidence or marginal likelihood, that is, the factor is the same for all possible events being considered.

The Multinomial Naïve Bayes classifier is a probabilistic classifier based on Bayes theorem with strong and naïve assumptions that the features used in classification are independent. This is called conditional independence [31]. Multinomial Naïve Bayes has been used widely in text classification, spam email detection, personal email sorting, document categorization, language detection and prediction. Multinomial Naïve Bayes takes into account multiple occurrences (count) of features used in classification; this is the main difference between Multinomial Naïve Bayes, Binarized Multinomial Naïve Bayes, and Bernoulli Naïve Bayes. We applied this rule to our training dataset. The document is the timestamped ordered sequence of countries where photos were uploaded (hereafter referred to as features) by the users. The document size is the number of users in the training dataset. The class is the ten countries selected. The features are the country codes. Again, we make a Naïve Bayes assumption that the probability of each feature in a class is independent of the features context and position in the document. Thus our training set is drawn from a multinomial distribution of features with as many independent features as the length of the training set.

3.2.1 Data Training

The training dataset contains 80% records randomly selected from the dataset while remaining 20% were used for testing. We chose an 80/20 division for the training/testing set because this is a commonly used ratio used in machine learning research, asMultinomial Naïve Bayes classifier performs better with more training data. Although other ratios have also been adopted. For example, McCallum & Nigam [24] in their research used 70/30 ratio and in another study 50/50 ratio, while other studies have used 67/33 and 60/40. The training dataset contains attributes like usernames, date and time, longitude, and latitude and country codes of countries where the photos were uploaded. For example, if we want to train the classifier for users who uploaded photos in the United Kingdom (country code 206), we count all the users who have feature 206 in their trajectory and mark them as “visited_206”, we use the features before 206 as the training features. In the same way, all the users who have no feature 206 will be marked as “no visit to_206” and all the features in their trajectory used as training features for not visiting 206. The trained model produced prior probabilities and likelihood probabilities of visiting and not visiting 206 respectively. The two probabilities will be used to calculate the posterior probability in the test dataset. An example is shown below.

To compute prior probabilities for a feature belonging to class A, we use Equation 3,

$$P(A) = \frac{\sum V}{U} \tag{3}$$

P(A) is the prior probability of features belonging to class A. ∑V is the sum of the count of features that belong to class A. U is the size of the document.

We also calculated the prior probabilities for a feature not belonging to class A using Equation 4,

$$P(notA) = \frac{\sum \mathrm{V}}{\mathrm{U}} \tag{4}$$

P(notA) is the prior probability of a feature not belonging to class A. ∑V is the sum of the count of features not belonging to a class. U is the size of the document.

To compute the likelihood probabilities of features (in this case B) belonging to a class A and likelihood probabilities of features not belonging to class A, we use Equations 5 and 6 below.

$$P(B/A) \; = \frac{\sum \mathrm{C}}{\sum \mathrm{D} + \sum \mathrm{E}} \tag{5}$$

Equation 5 is used to calculate the likelihood probability of feature B in class A, i.e., P(B/A). $\sum \mathrm{C}$ is the sum of the count of occurrence of feature B in class A. Then $\sum D$ is the sum of count of features B before feature A while $\sum \mathrm{E}$ is the sum of number of unique features in the dataset.

$$\mathrm{P(B/notA)} \; = \frac{\sum C_{notA}}{\sum D_{notA} \; + \sum \mathrm{E}} \tag{6}$$

Equation 6 is used to calculate the likelihood of feature B not belonging to class A, i.e., P(B/notA). $\sum C_{notA}$ is the count of occurrence where features B does not belong to class A, $\sum D_{notA}$ is the count of occurrence where feature A is not in the trajectory while $\sum \mathrm{E}$ is the sum of number of unique features in the dataset.

We demonstrate a simple worked example from Table 2, using country code 206 as an example.

Table 2. Sample training data

User_id	Countries visited
10006374@N03	{119,23}
10014738@N05	{238,85,206}
10016029@N04	{81,206}
10017016@N03	{64,85}
10017201@N02	{85,23}
10017367@N03	{50,206,71,85}
100460312@N04	{71,206}
10019779@N00	{82,185,187}

Prior probabilities of features belonging to class 206 and not belonging to class 206, using

Equation 3,

P(206) = $\frac{4}{8}$ = 0.5 ≈ 50%.

For the probability of not belonging to 206, using Equation 4,

P(N206) = $\frac{4}{8}$ = 0.5 ≈ 50%.

For all the features in the training dataset, to calculate the likelihood probabilities of features belonging to class 206, using Equation 5,

$P(119|206) = \frac{0+1}{5+12} = \frac{1}{17}$, P(23|206) = $\frac{0+1}{5+12} = \frac{1}{17}$ etc., when a particular feature does not appear in a document, its conditional probability is equal to 0, to avoid this problem, we use add-one or Laplace smoothing by adding 1.

To calculate the likelihood probabilities of features not belonging to class 206, using Equation 6,

P(119|N206) = $\frac{1+1}{9+12} = \frac{2}{21}$, P(23|N206) = $\frac{2+1}{9+12} = \frac{3}{21}$ etc, again we added Laplace smoothing by adding 1.

As shown, the trained model produced four probabilities, two prior probabilities and two likelihood probabilities for user 1 with two features (119 & 23).

3.2.2 Testing

We used the remaining 20% of our dataset as test data. We used the results from the training set to calculate the posterior probability and then used the selected posterior probability to test the accuracy of the model. We did this using Equations 7, 8, 9 and 10.

$$P(A/B) \propto P(A) \text{ x } P(B/A) \tag{7}$$

$$P(A_{not}/B) \propto P(A_{not}) \text{ x } P(B/A_{not}) \tag{8}$$

Equations 6 and 7 are varying the proportionality over A for a given B.

$$P\left(A/test\right) = \frac{P\left(A/B\right)P\left(A\right)\ x\ P\left(B/A\right)}{\{P\left(A/B\right)P\left(A\right)\ x\ P\left(B/A\right)+P\left(A_{not}/B\right)P\left(A_{not}\right)\ x\ P\left(B/A_{not}\right)\}} \tag{9}$$

$$P\left(A_{not}/test\right) = \frac{P\left(A_{not}/B\right)P\left(A_{not}\right)\ x\ P\left(B/A_{not}\right)}{\{P\left(A_{not}/B\right)P\left(A_{not}\right)\ x\ P\left(B/A_{not}\right)+P\left(A/B\right)P\left(A\right)\ x\ P\left(B/A\right)\}} \tag{10}$$

Using Equations 9 and 10, for each user in the test dataset, we calculated the posterior probabilities of belonging to class 206 and posterior probability of not belonging to class 206 respectively and compare it to the actual travel history of each user to check the accuracy of our model. Equations 9 and 10 is a combination of priors and likelihood probabilities from Equations 7 and 8 respectively.

For our prediction, we compare the two posterior probabilities; the highest probability is used to classify a user to either belonging to class 206 or not belonging to class 206. The same steps will be repeated for all the classes and users in the dataset.

Again, we demonstrate a simple worked example from Table 3, using country code 206 as an example.

Table 3 Sample testing data

User_id	Countries visited
10020416@N06	{153,23,206}
100212960@N05	{82,185,50}
10021381@N05	{50,153,206}
10022497@N03	{81,71,85}

For user 1 in the test dataset, the posterior probabilities of belonging and not belonging to class 206 is given as;

P(206|test) α $\frac{4}{8} * \frac{1}{17} = \frac{4}{136}$

P(N206|test) $\alpha \frac{4}{8} * \frac{3}{21} = \frac{12}{168}$

P(206|test) $= \frac{\frac{4}{136}}{(\frac{4}{136} + \frac{12}{168})} = 0.292 \approx 29\%$

P(N206|test) $= \frac{\frac{12}{168}}{(\frac{12}{168} + \frac{4}{136})} = 0.71 \approx 71\%$

From the example above, given the trajectory of user1, it shows that the user belongs to class 206 which is the truth, but our model predicts that there is only a 29% probability of the user being classified to class 206 and there is a 71% probability of the user not being classified to class 206. The poor accuracy of the prediction is because we trained the model on only two features. We are only using this example as an illustration of how the Multinomial Naïve Bayes classifier works in document classification.

CHAPTER 4

Results

4.1 Travel Flow

The results of our probability analysis calculating travel flow between different countries shows that majority of U.S traveler’s destinations are Canada and United Kingdom. Also, the result shows that European countries have the highest probabilities for a U.S tourists to visit different countries when they visit Europe. The reason for this is because European countries are close to each other and the train system makes it easy to ride a train from one country to another. Table 4 shows nine probabilities of a U.S visitor visiting other countries after their first visited country (the first country visited is the first column in Table 4). From Table 4, France has the highest probability of visit (23%) after a visit to Netherland, Germany has the 21% probability of visit, the United Kingdom has 20% probability of visit and China has the least probability of visit (7%) after a visit to Netherland. Australia has the lowest probability of being visited (4%) after a visit to Canada, followed by Japan 5% and China 5%.

Table 4. Probabilities of a US Flickr user visiting different countries

Top 10 countries visited	United Kingdom	Spain	Germany	Italy	France	Canada	Australia	Japan	Netherland	China
United Kingdom	0	11.45%	15.25%	12.73%	21.19%	15.84%	5.51%	6.44%	10.14%	6.41%
Spain	16.77%	0	15.28%	15.48%	18.93%	14.94%	5.75%	7.23%	10.62%	6.63%
Germany	17.18%	12.47%	0	15.43%	21.06%	15.09%	4.56%	6.79%	13.00%	7.47%
Italy	16.19%	13.96%	15.76%	0	18.76%	14.61%	4.42%	6.05%	9.92%	6.11%
France	17.90%	13.86%	16.09%	16.92%	0	15.26%	4.38%	6.50%	10.99%	7.10%
Canada	11.85%	6.20%	8.92%	7.73%	10.50%	0	3.63%	4.88%	5.65%	5.08%
Australia	14.25%	10.56%	12.21%	11.07%	16.28%	16.28%	0	9.92%	7.76%	9.67%
Japan	13.67%	8.73%	12.52%	8.48%	13.67%	14.00%	5.52%	0	6.84%	12.44%
Netherland	19.57%	14.29%	20.69%	14.84%	23.01%	14.84%	5.19%	7.70%	0	7.42%
China	11.83%	9.39%	10.39%	8.58%	11.47%	14.27%	6.05%	8.85%	7.05%	0

4.2 Prediction

To predict if a user has been to a certain country using Multinomial Naïve Bayes model, the total number of users from the dataset that met the requirements for analysis was 5,132. For training, 4,136 users were used to train the model while 996 were used to test the model. The prior probabilities for features belonging to a class as shown in Table 5 for the ten classes selected show that fewer users from the training set visited the ten classes. The prior probabilities of no visit were high for all the ten classes.

Table 5 prior probabilities of visit and no visit to the ten classes for the training set

Class	Visit to Class	No Visit to Class
29	14%	86%
88	15%	85%
206	42%	58%
8	11%	89%
64	39%	61%
71	31%	69%
85	27%	73%
187	22%	78%
23	37%	63%
153	18%	82%

The likelihood probabilities of features belonging to the classes also show a big gap between the likelihood of a feature belonging to a class and a feature not belonging to a class, which is in agreement to the prior probabilities. The number of users who visited the ten classes are less than the number of users who did not visit the ten classes. As shown in figure 4 and 5, classes 29 and 88 are China and Japan respectively.

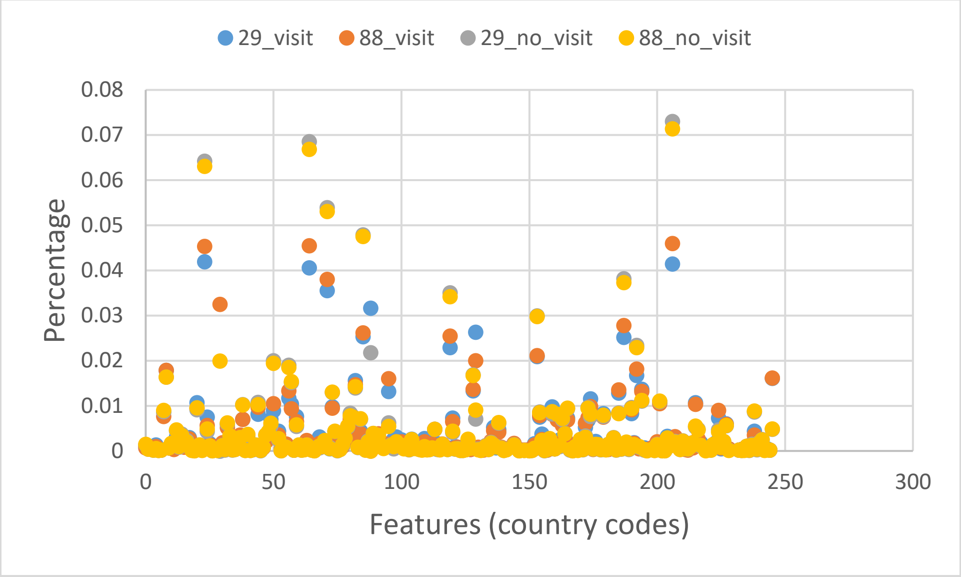

Figure 4 likelihood probabilities of features belonging to class 29 & class 88 (Table 5).

For testing, we used the priors and likelihoods results from the training data to calculate the posterior probabilities of the users belonging to a class. Figure 5 shows the posterior probabilities for all the users for two classes. The posterior probabilities show that the higher probabilities are for the no visit which corresponds to the results from the training set.

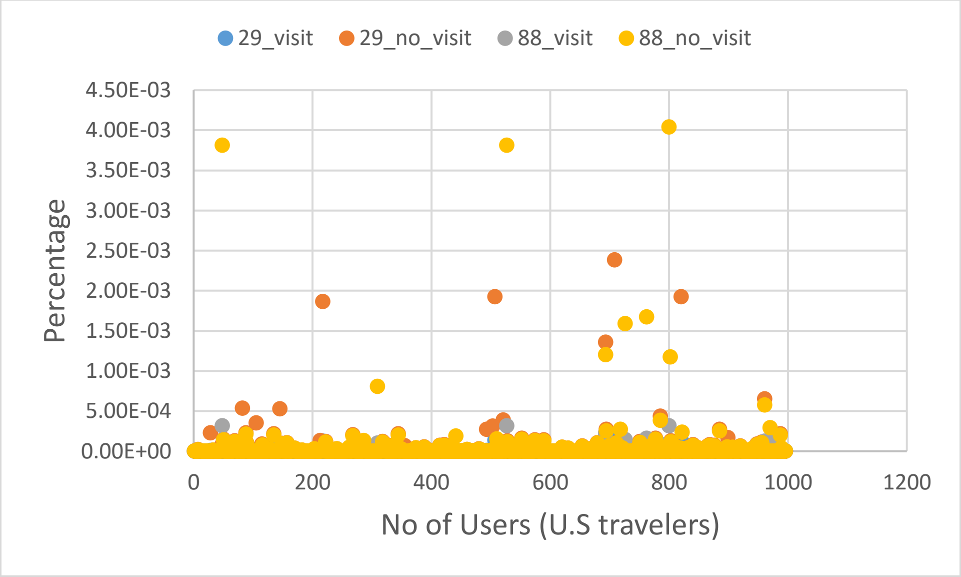

Figure 5 posterior probabilities of features belonging to class 29 & class 88 (table 5).

We classified the users using the posterior probabilities. To classify a user as visiting a class, we compared the posterior probability of visit and the posterior probability of no visit, if the posterior probability of visit is higher than the posterior probability of no visit, that user is classified as visit to that class, if the posterior probability of no visit is higher than the probability of visit, that user is classified as no visit to that class. Those users classified as visit to a class is the prediction. As shown in Table 6, the prediction table is sorted from highest to lowest predicted number of user visits for each class.

Table 6 predicted users for each class

Class	Predicted users to class	Accuracy
206	119	58%
64	103	62%
71	90	66%
85	64	71%
29	48	76%
88	45	76%
23	26	83%
187	24	83%
8	22	85%
153	12	87%

We tested the accuracy of the model by dividing the difference between the truth and predicted results by the number of users in the test dataset.

It is interesting to notice that the classes that had the most number of photos uploaded from Table 1(206) also had the most number of users predicted to that class but with the lowest accuracy. The reason for that is because most of the photos uploaded in 206 were uploaded by a few users, and during training and testing, we only used one photo per user and not all the photos from each user. We did this to make sure that our result is not skewed as some users uploaded thousands of photos in one country. Netherland (153) had the most accurate prediction because, from the test training data, most of the users actually visited Netherland.

The results from the prior and likelihood show that the probability of users that did not belong to a class was higher than the probability of users that belong to the classes, this again is because of the number of users that meet the requirements for our analysis. If the number of users in the training and test dataset is increased significantly, the prior and likelihood probabilities of visit as well as the posterior probabilities will significantly increase with a better accuracy result.

CHAPTER 5

Conclusion

This research generated valuable information in identifying user behavior patterns which will be beneficial to the tourism industry in different countries, travel agencies and future U.S travelers. The results also lead to a better understanding of travel patterns of a frequent traveler and the order of their visits to different countries. Our results will be a useful resource for travel agencies and U.S travelers in planning their future travel destinations. Our results from the travel flow analysis between countries show that European countries are very popular destinations for tourists from the United States. In future research, we will studyglobal travel flow pattern to better understand the travel preferences of travelers from other countries around the world. For example, it will be interesting to model the travel flows and countries visited by travelers from China, the Caribbean, South America and Africa.

For our prediction, we used Multinomial Naïve Bayes a machine learning algorithm to analyze location-based social media data from Flickr with advanced data processing techniques and predict if a user is likely to visit a particular country (class) based on their past travel history. Our results show that the accuracy of the model is satisfactory.

In future research, we will improve the sample size and conduct a sensitivity analysis for the training and test dataset ratios to improve the model accuracy. We will also incorporate more attribute into the training and test dataset to improve the accuracy of the model. Some of the attributes will include adding more features and increasing the number of classes in our training and test dataset.

The accuracy of the model will improve if we add more dataset as Multinomial Naive Bayes is known to perform better with a larger dataset. Another thing we will do in our future research is to train the same data with another classifier like Support Vector Machine or Markov Model and compare the predicted results and the accuracy of the models to see which of the machine learning algorithms that will perform better in classifying Flickr data.

References

1. Yuan, Yihong, and Monica Medel. "Characterizing International Travel Behavior from Geotagged Photos: A Case Study of Flickr." *Plos One* 11.5 (2016).
2. Memon, Imran, Ling Chen, Abdul Majid, Mingqi Lv, Ibrar Hussain, and Gencai Chen. "Travel Recommendation Using Geotagged Photos in Social Media for Tourist." *Wireless Personal Communications* 80.4 (2014): 1347-362. Web.
3. Vu, Huy Quan, Gang Li, Rob Law, and Ben Haobin Ye. "Exploring the Travel Behaviors of Inbound Tourists to Hong Kong Using Geotagged Photos." *Tourism Management* 46 (2015): 222-32. Web.
4. Yanai, Keiji, Keita Yaegashi, and Bingyu Qiu. "Detecting cultural differences using consumer-generated geotagged photos." Proceedings of the 2nd International Workshop on Location and the Web - LOCWEB 09 (2009).
5. Krumm, John, and Eric Horvitz. "Predestination: Inferring Destinations from Partial Trajectories." Lecture Notes in Computer Science UbiComp 2006: Ubiquitous Computing (2006): 243-60. Web.
6. Hung, Chih-Chieh, Ling-Yin Wei, and Wen-Chih Peng. "Clustering Clues of Trajectories for Discovering Frequent Movement Behaviors." *Behavior Computing* (2012): 179-96. Web.
7. Krumm, John. "Real Time Destination Prediction Based On Efficient Routes." SAE Technical Paper Series (2006).
8. Karbassi, Abdolreza, and Mathew, Barth. "Vehicle route prediction and time of arrival estimation techniques for improved transportation system management." IEEE IV2003 Intelligent Vehicles Symposium. Proceedings (Cat. No.03TH8683) (2003).

9. Schmandt, Chris, and Natalia, Marmasse. "User-centered location awareness." Computer 37.10 (2004): 110-11. Web.

10. Ashbrook, Daniel, and Thad Starner. "Using GPS to learn significant locations and predict movement across multiple users." Personal and Ubiquitous Computing 7.5 (2003): 275-86. Web.

11. Kurashima, Takeshi, Tomoharu Iwata, Go Irie, and Ko Fujimura. "Travel Route Recommendation Using Geotags in Photo Sharing Sites." *Proceedings of the 19th ACM International Conference on Information and Knowledge Management - CIKM '10*(2010).

12. Kisilevich, Slava, Florian Mansmann, Mirco Nanni, and Salvatore Rinzivillo. "Spatio-temporal clustering." Data Mining and Knowledge Discovery Handbook (2009): 855-74. Web.

13. Mamei, Marco, Alberto Rosi, and Franco Zambonelli. "Automatic Analysis of Geotagged Photos for Intelligent Tourist Services." *2010 Sixth International Conference on Intelligent Environments* (2010).

14. Crandall, David, and Noah Snavely. "Modeling People and Places with Internet Photo Collections." *Communications of the ACM* 55.6 (2012): 52. Web.

15. Sun, Yeran, Hongchao Fan, Mohamed Bakillah, and Alexander Zipf. "Road-based Travel Recommendation Using Geotagged Images." *Computers, Environment and Urban Systems* 53 (2015): 110-22. Web.

16. Beiro, Mariano, Andre Panisson, Michele Tizzoni, and Ciro Cattuto. "Predicting Human Mobility through the Assimilation of Social Media Traces into Mobility Models." *EPJ Data Science* 5.1 (2016).

17. Wolf, Jean, Randall Guensler, and William Bachman. "Elimination of the Travel Diary: Experiment to Derive Trip Purpose from Global Positioning System Travel Data."*Transportation Research Record: Journal of the Transportation Research Board* 1768 (2001): 125-34. Web.

18. Girardin, Fabien, Filippo Dal Fiore, Carlo Ratti, and Josep Blat. "Leveraging Explicitly Disclosed Location Information to Understand Tourist Dynamics: A Case Study." *Journal of Location Based Services* 2.1 (2008): 41-56. Web

19. Popescu, Adrian, and Gregory Grefenstette. "Deducing Trip Related Information from Flickr." *Proceedings of the 18th International Conference on World Wide Web - WWW '09* (2009).

20. Zheng, Yan-Tao, Ming Zhao, Yang Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, Tat-Seng Chua, and H. Neven. "Tour the World: Building a Web-scale Landmark Recognition Engine." *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009).

21. Popescu, Adrian, Grefenstette, Gregory, Moëllic, Pierre-Alain. Gazetiki: “automatic construction of a geographical gazetteer.” In Proc. of JCDL (2008) (Pittsburgh, PA, June 2008).

22. Rattenbury, Tye, Nathaniel Good, and Mor Naaman. "Towards Automatic Extraction of Event and Place Semantics from Flickr Tags." *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '07* (2007).

23. Clements, Maarten, Pavel Serdyukov, Arjen P. De Vries, and Marcel J.t. Reinders. "Using flickr geotags to predict user travel behavior." Proceeding of the 33rd

international ACM SIGIR conference on Research and development in information retrieval - SIGIR 10 (2010).

24. Andrew McCallum, Kamal Nigam. “A Comparison of Event Models for Naive Bayes Text Classification” *Source: CiteSeer* 05 (2001): n.pag. Web.
25. Pedro Domingos, Michael Pazzani “On the Optimality of the Simple Bayesian Classifier under Zero-One Loss” Machine Learning, 29, 103–130 (1997).
26. Nir Friedman, Dan Geiger, Moises Goldszmidt. “Bayesian Network Classifiers” Machine Learning, 29, 131–163 (1997).
27. Hobel, Heidelinde, Paolo Fogliaroni, and Andrew U. Frank. "Deriving the Geographic Footprint Of Â Cognitive Regions." Geospatial Data in a Changing World Lecture Notes in Geoinformation and Cartography (2016): 67-84. Web.
28. Ji, Xiang, Soon Chun, and James Geller. "Monitoring Public Health Concerns Using Twitter Sentiment Classifications." ACM Digital Library. IEEE Computer Society (2013): Web. 24 July 2017.
29. Bo, Han, Cook, Paul, and Baldwin, Timothy. “Geolocation prediction in social media data by finding location indicative words”. (2012).
30. Ng, Andrew Y., and Michael I. Jordan. "On Discriminative vs. Generative Classifiers: A comparison of logistic regression and Naive Bayes." Advances in Neural Information Processing Systems 14 (NIPS) (2001).
31. Zhang, Harry. “Exploring Conditions for The Optimality of Naive Bayes.” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 19, no. 02, 2005, pp. 183–198., doi:10.1142/s0218001405003983.