Tweet Geolocation Error Estimation

Erik Holbrook[°], Gupreet Kaur[°], Jared Bond[°], Josh Imbriani[°], Christan Grant[°], Elaine Nsoesie[ω]

[°]*School of Computer Science, University of Oklahoma*

[ω]*Institute for Health Metrics and Evaluation, University of Washington*

Introduction

Accurately assigning a geolocation to a tweet is useful to a broad range of applications, such as, disease surveillance through social media [5]. However, only about 1% of tweets contain accurate location information [2].

Previous methods use machine learning models trained on the textual content of the tweet and the user's past tweets to identify city-level location information. These methods are heavily dependent on heuristics and the granularity of the training data; any extension would require intense restructuring of the classifiers. They also require access to the users' timeline.

In this paper, we examine the reliability and accuracy of predicting tweet origin locations based on meta-content of the of the tweet itself. Specifically, the user-supplied *location* and *description* fields, as well as propose a lightweight system for identifying the location of a user.

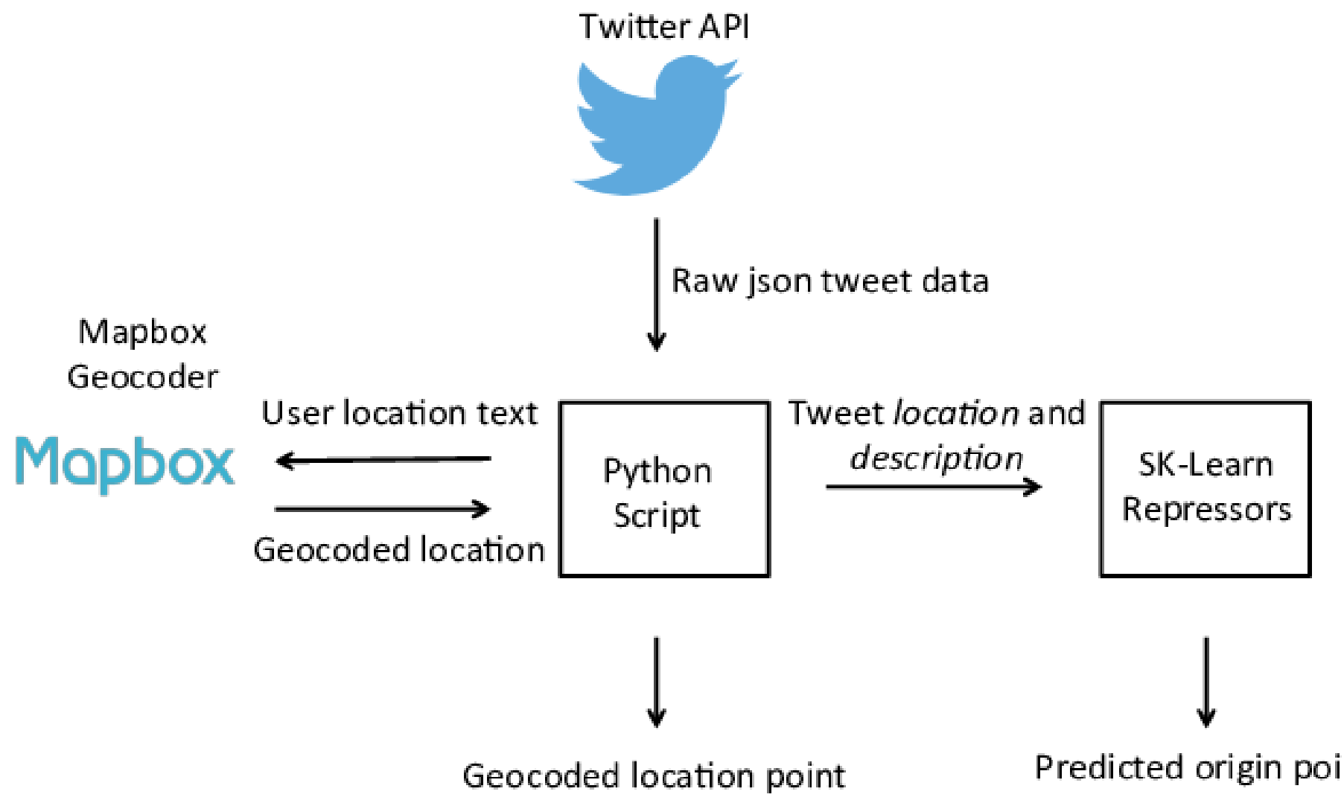
Related Work

Identifying the locations of these trends has been examined from several different angles [1, 8, 7]. The common focus is predicting the home location of a Twitter user based on their past tweets.

These methods require a large amount of operational overhead, i.e. a very large number of tweets from each user and other data obtained from the Twitter API, or a complicated prediction mechanism. Also each study adopts a different approach to error measurement, thereby making comparison of method performance difficult.

Paper	Full Timeline	Graph	Distance Error	Tweets
Compton, et al.[1]	Y	Y	N	25B
Mahmud, et al.[4]	Y	N	N	1.5M
Priedhorsky, et al.[7]	N	N	Y	30k
Zhang, et al.[8]	N	N	N	1k
This work	N	N	Y	326k

Table 1: Summary of techniques used in literature for tweet distance estimation.

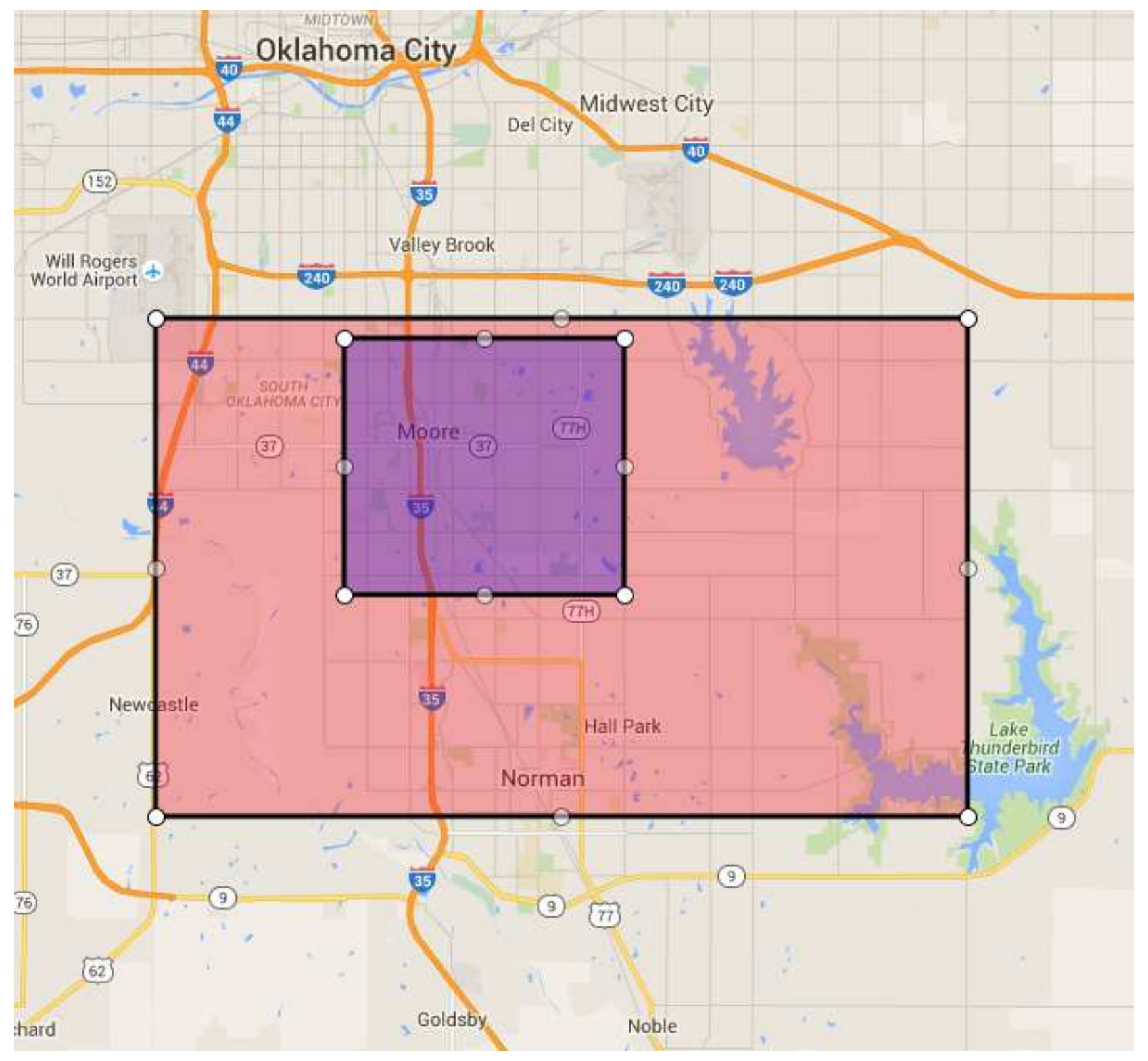


Methods

Our data consists of tweets collected from April 11 to April 19, 2016. The Twitter API was used to identify users in Oklahoma who supplied location information with their tweets. 325788 tweets were collected during this period. A python program sent the unaltered text in the user 'location' fields to the Mapbox geocoder[1]. In parallel, two Scikit-learn [6] SVR regressors were trained on the 'location' and 'user description' fields of the tweets.

Experiments

First, we determined what percentage of users in our data set reported a location which could be resolved (not necessarily correctly) with the Mapbox API. We found that approximately 56% of users' locations could be resolved. Next, we examine the accuracy of these locations as compared to the actual origin. The location field is geocoded with the Mapbox geocoder API. If a location is identified, its coordinates are determined via the API. Finally, we built two estimators with the Scikit-learn library to estimate the origin of the tweet based on the meta data to predict the latitude and longitude, respectively.



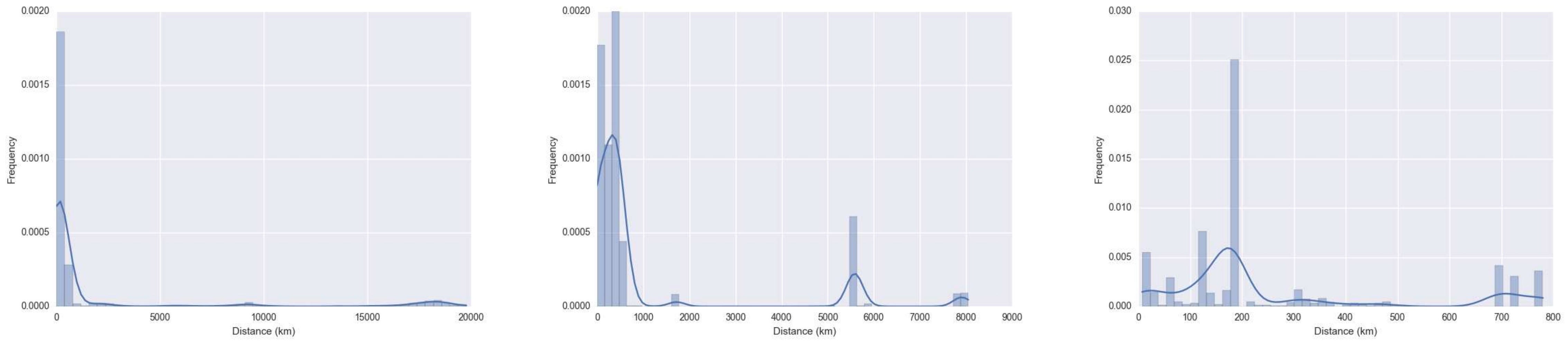
Since both the geocoder and the Twitter API describe users' locations as bounding boxes, we require an error calculation method that can describe an average distance, and also the potential range of that distance to deal with the problem of overlap.

We model the distance between the actual and predicted regions as the average distance between any two random points selected respectively from the two region. We assume that any point within the regions could be the origin point with an equal probability.

The user's 'location' and 'user description' fields were tokenized into uni- and bi-grams, and each regressor was trained independently on the first 90% of the tweet dataset, leaving the remaining 10% for testing. Therefore, the two bounding boxes are simply two uniform probability distributions, whose mean distance is simply the distance between the two centroids: $\mu_{AB} = \mu_{center_A} - \mu_{center_B}$.

To address the problem of overlap and containment, we propose examining the standard deviation of the mean of $\mu_{AB} : \sigma_{AB} = \sqrt{\sigma_A^2 + \sigma_B^2}$, where σ^2 is in the latitudinal or longitudinal direction. If two regions are centered exactly on the same point, the mean distance will be zero, but this calculation reveals how much the two regions differ in size.

Left: A bounding box diagram indicating the difference between predicted origin (pink) and ground-truth origin (purple). **Below, left to right:** Overall regression error frequencies from ML algorithm. Standard Deviation Frequencies. Error when tweets have been slightly filtered to eliminate noise.



Discussion

This method of error calculation relies on two assumptions. First, the tweet could have originated from any point with uniform probability. Second, both regions are rectangles, whose respective sides are parallel.

We found almost 30% of the tweets with *location* field data can be resolved to city-level accuracy by simply geocoding the text with no alterations. Our mean accuracy was approximately 1941 km.

Hecht et al. [3] reported that approximately 34% of users had either non geographic information or simply nothing entered in the 'location' field. Our results suggest that automated techniques could eliminate these from the dataset and thus improve overall prediction accuracy.

Summary & Future Work

In this work, we made several steps toward location prediction of social media users. We have examined the accuracy and reliability of geocoding users' location information as a way of predicting tweet origin locations. We have also demonstrated the potential utility of this information through machine learning on the textual content. Finally, we have established a robust and straightforward method for accuracy measurement in terms of the distance between the predicted location and the ground-truth location of tweets. In the future, we will engineer several techniques to improve accuracy and evaluate the results on a larger data set.

Acknowledgments: Thanks to Le Gruenwald for her helpful comments and the Robert Wood Johnson Foundation for their generous support.

References

[1] Ryan Compton, David Jurgens, and David Allen. Geotagging one hundred million twitter accounts with total variation minimization. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 393–401. IEEE, 2014.

[2] Aron Culotta. Estimating county health statistics with twitter. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 1335–1344, New York, NY, USA, 2014. ACM.

[3] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H Chi. Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 237–246. ACM, 2011.

[4] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. Home location identification of twitter users. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):47, 2014.

[5] Michael J Paul and Mark Dredze. You are what you tweet: Analyzing twitter for public health. *ICWSM*, 20:265–272, 2011.

[6] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830, November 2011.

[7] Reid Priedhorsky, Aron Culotta, and Sara Y Del Valle. Inferring the origin locations of tweets with quantitative confidence. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 1523–1536. ACM, 2014.

[8] Wei Zhang and Judith Gelernter. Geocoding location expressions in twitter messages: A preference learning method. *Journal of Spatial Information Science*, 2014(9):37–70, 2014.

Contact information: Christan Grant, DEH 234, 110 W. Boyd, Norman, Oklahoma, 73071, Computer Science, University of Oklahoma

Phone: 405–325–5408; Email: *cgrant@ou.edu*; Web: *https://www.cs.ou.edu/ cgrant*