Big Data Trip Classification on the New York City Taxi and Uber Sensor Network

Huiyu Sun[1], Siyuan Hu[1], Suzanne McIntosh[1], Yi Cao[2]

[1] Department of Computer Science, New York University, USA
[2] Jiangsu Engineering Centre of Network Monitoring & School of Computer and Software, Nanjing University of Information Science and Technology, China
hs2879@nyu.edu, sh3291@nyu.edu, sm4971@nyu.edu, caoyinuist@163.com

Abstract

Millions of trips are made every day by taxis and Uber in New York City. We first employ big data technologies to analyze this vast dataset: Apache Spark is used for data processing and classification, Apache Hive is used for data storage, and MapReduce is used for data profiling. Since taxis and Uber are equipped with GPS sensors, we then visualize a mobile sensor network over New York City separated into fine-sized regions each acting as a mobile sensing node. Each location on the network falls into a region and is classified into one of three categories based on which service dominates the particular region: Yellow taxi, Green taxi, or Uber. We utilize logistic regression to classify a region into one of the three categories. Our classification algorithm is then used to analyze the interaction between taxi and Uber, for example to quantify the expansion of Uber. Experiments run on the Spark cluster show our classifier achieves an accuracy of over 85% scored on the 2014 taxi and Uber dataset. Finally, we propose a trip recommendation system for users using classification results together with a web service application.

Keywords: Big data, Classification, Mobile sensor network, NYC taxi, Uber

1 Introduction

With the invasion of Uber into New York City (NYC) in 2011, we investigate its expansion in the past few years and the effect it has on yellow and green taxi trips. Taxis and Uber in NYC are all equipped with GPS sensors and fare collection systems that collect data and upload them on to a server. All trip data are made publicly available on the NYC Taxi & Limousine Commission website. Since the summer of 2013, TLC introduced a new Borough Taxi program, adding thousands of Green taxis. The taxi dataset includes trip records from all trips completed in yellow and green taxis in NYC between 2009 and 2015. Many works have been proposed analyzing the taxi dataset [1-6]. The Uber dataset includes trips completed in 2014 and 2015. The Uber dataset was analyzed in [7-9]. We utilize big data technologies (MapReduce, Hive and Spark) to classify trips and make prediction on the taxi and Uber network.

Each taxi and Uber acts as a sensor node in a mobile sensor network that covers New York City. Deri and Moura [10] analyze NYC taxi data by considering a network approach. Ganti et al. [11] consider taxis as a part of a large participatory sensor network. Aslam et al. [12] considers a city into a roving sensor network and utilizes the information to predict traffic. Mobile sensing [13-14] has been utilized in many works to monitor environment, determine activities [15-16] and offer services. In this paper, NYC is split into small regions each acting as a sensor node in addition to individual taxis and Uber. We aim to classify each region into one of three categories (yellow taxi, green taxi or Uber) based on the most active service in terms of pickup numbers inside the region. We use logistic regression as our classifier. In terms of classification, a number of related works [17-21] have also been proposed.

2 Pipeline

We analyze the taxi and Uber dataset in several stages. The pipeline describing each stage is shown in Figure 1. We first ingest the taxi and Uber dataset onto our High Performance Computing cluster and store the data on HDFS and Hive databases. Big data analytics are then carried out on the Hadoop cluster where we obtain a set of initial results such as taxi and Uber pickup shares by districts. Other analysis results are investigated in more detail in [22]. We utilize these initial big data results as a baseline for obtaining more complex results.

*Corresponding Author: Huiyu Sun; E-mail: hs2879@nyu.edu
DOI: 10.3966/160792642018031902027

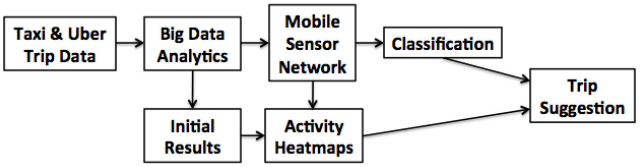


Figure 1. Pipeline of our system

Next we use the GPS sensors [23-24] on taxis and Uber to generate heatmaps over the city providing insight into the taxi and Uber network. We split the network into small regions and use classification algorithms on each region to learn the most dominate service (Yellow taxi, Green taxi or Uber) using historical data. We can associate each historical and future taxi and Uber trip with a region and classify the region into one of three categories: Yellow, Green or Uber. By doing so, each possible trip location can be associated with one of the three categories. This allows us to achieve trip suggestion: suggesting to passengers which service to use given a location.

3 Big Data Processing

In order to analyze the vast taxi and Uber dataset, the data is first ingested into our Hadoop [25-26] cluster and stored in the Hadoop Distributed File System (HDFS) where we can efficiently process it using Hadoop Spark [27], MapReduce [28], and Hive [29]. The Cloud can also be used as an efficient storage agent [30]. Security is a major concern and a lot of works [31-35] have been proposed to address the issue. Spark's MLlib [36] also provide efficient classification tools such as logistic regression. The high-level design diagram is shown in Figure 2. Apache Spark is a powerful tool in executing batch-processing jobs and stream processing. We also use Apache Hive, which is a data warehouse infrastructure built on top of Hadoop, to store large-scale datasets.

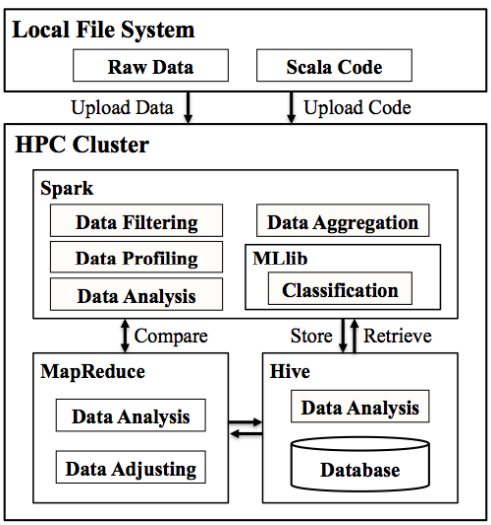


Figure 2. High level design diagram for analyzing taxi and Uber data using big data tools

A number of big data analyses were conducted in our previous paper [22] where we exclusively focused on the taxi (yellow and green) dataset. In this paper, we mainly focus our analysis on the Uber dataset and obtaining a comparison between yellow taxi, green taxi and Uber. Detailed analysis of boroughs and zones are described in the next section.

4 Taxi and Uber Sensor Network

Taxis and Uber play an important role as mobile sensors. All taxi and Uber have the GPS sensor installed on them which records their precise location at all times. Not only that, the time of the day is also recorded. So retracing the coordinates and their respectively time shows a map with the route traversed by each taxi and Uber. This can offer a number of useful information regarding traffic [37-38] and the surrounding environment [39]. Although the data is not uploaded in a real-time fashion, taxi sensors can also offer many insights. We break down the Uber pickup numbers by boroughs and zones, discuss the formation of sensor networks over these boroughs and zones and then analyze the expansion of Uber trips on the network.

4.1 Boroughs and Zones Network

The NYC network consists of five boroughs. We plotted the daily and monthly Uber pickups over a 6-month period in 2015 by each borough. This is shown in Figure 3 where the top graph shows the daily numbers and the bottom graph shows the monthly numbers. From the daily graph, we see that pickup numbers fluctuate with a large margin in the Manhattan borough, follows a periodic pattern in Brooklyn and Queens, and mostly remains flat in Bronx. To get rid of the daily noise, we analyze the monthly data. For the monthly pickups, we only show results for Manhattan, Brooklyn and Queens as more than 95% of Uber pickups occur in these three boroughs. We observe that the pickup numbers is gradually increasing for all boroughs, but the rate of increase of Brooklyn and Queens are higher than Manhattan.

The taxi and Uber sensor network over NYC can be split into five parts, each part covering one of the five boroughs. The smallest borough being Manhattan with a size of 59.1 million square meters and the largest borough Queens with 280 million square meters. On the most abstract level, these boroughs each form a sensor network and are connected to each other to form a parent network. Each borough consists of various districts, called zones. The zones vary in size, with most zones between 1 and 5 million square meters in size. These zones act as sensor nodes and form its own network over each borough. Zones in term can be split into smaller regions and finally into individual taxis

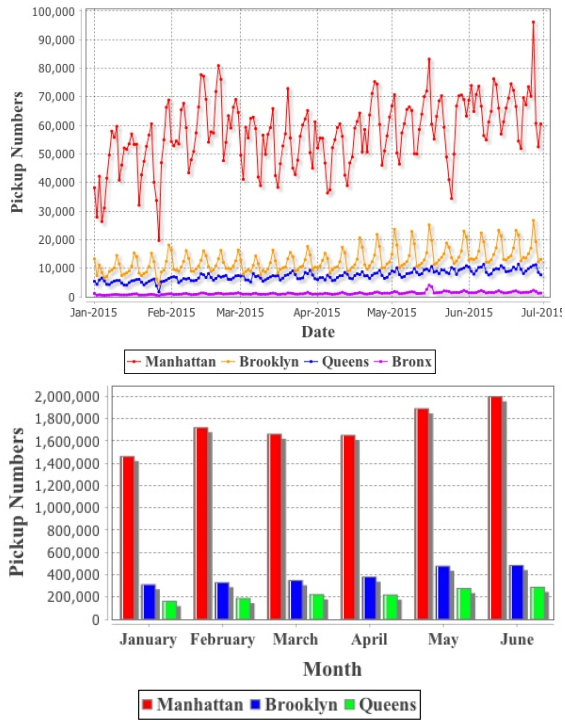


Figure 3. Daily (top) and monthly (bottom) Uber pick-up numbers between January and July 2015

and Uber operating all over the city. We take this hierarchical approach to form mobile sensor networks over the city.

To gain insight into the distribution over each zones, we plotted the number of Uber pickups by different zones in 2015. Figure 4 shows the pickups by zones break down in Manhattan, Brooklyn and Queens. For Manhattan, there are many zones with a high number of pickups indicating an even spread of service over the area. For Brooklyn and Queens, on the other hand, there are only a few zones with a high number of pickups indicated by the few spikes in the graph. This is especially true for Queens, as only two zones (JFK airport and LaGuardia airport) have a high pickup number, indicating most Uber activity in Brooklyn and Queens are concentrated in small areas. In terms of mobile sensing, this means that when we connect individual zones over the network there will be problems of data sparseness in some of these zones, making Queens and Brooklyn less effective in terms of network coverage. For Manhattan on the other hand, since activities are spread out all over the borough, it makes the perfect candidate to form an interconnected mobile sensor network within it.

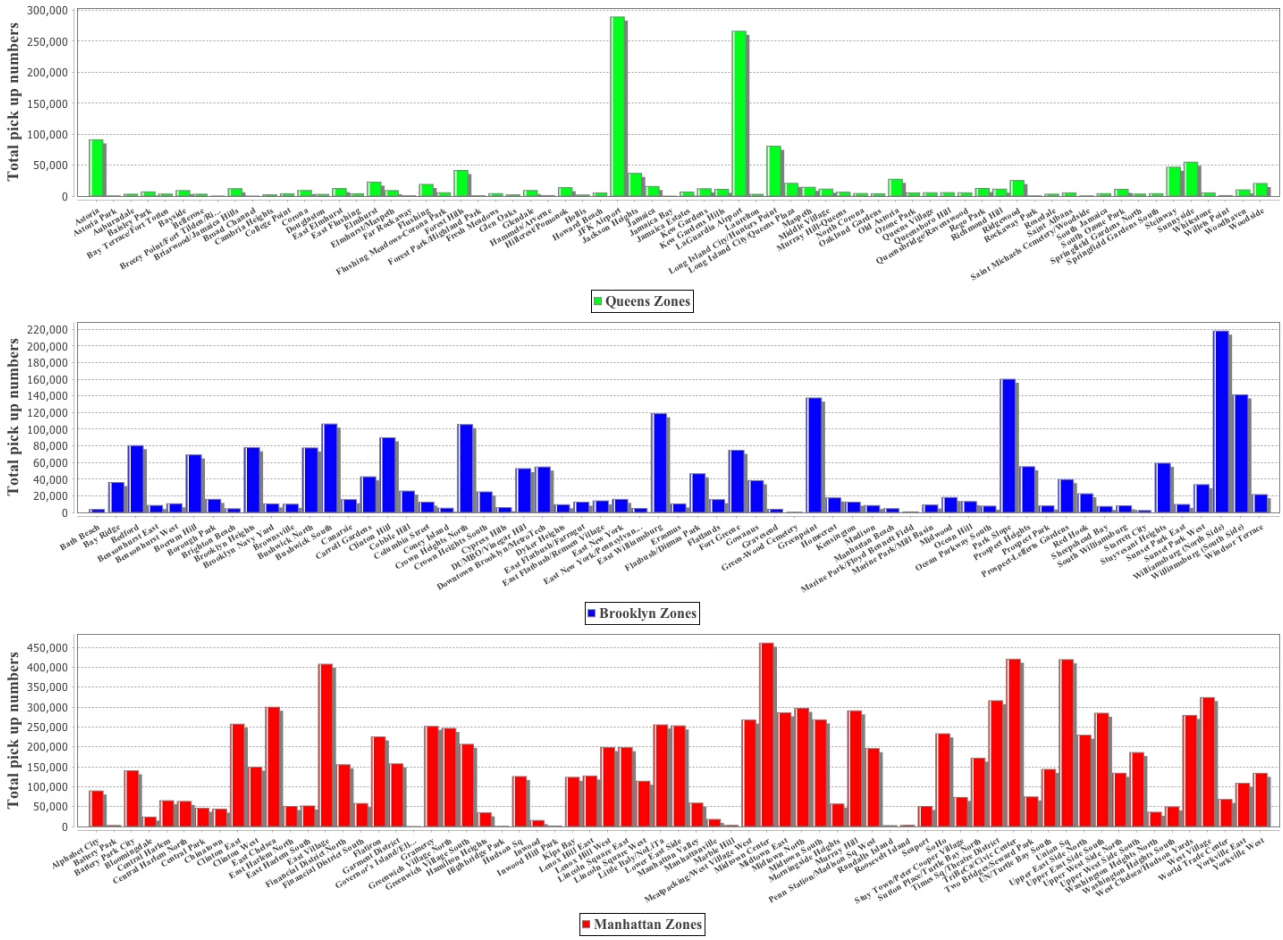


Figure 4. Uber pickup numbers by zones inside Manhattan (top), Brooklyn (middle) and Queens (bottom) in 2015

4.2 Analyzing the Expansion of Uber

The average monthly pickups made by yellow taxi, green taxi and Uber in 2014 are respectively 14 million, 1.2 million, and 0.7 million. In 2015, the average monthly pickup numbers are respectively 13 million, 1.5 million, and 2.5 million. The general trend sees yellow taxi pickup numbers declining, green taxi numbers increasing, and Uber increasing. Yellow taxi numbers fell by 1 million and Uber numbers rose by almost 2 million, more than tripling its pickup numbers from 2014. Comparing the percentages of shares between the three services reveals more information. In 2014, yellow taxis account for roughly 89% of all trips made in NYC, green taxi account for 8% of trips and Uber account for 3% of trips. Between 2014 and 2015 for the three services, yellow taxi shares fell by 15%, Uber shares rose by 14%, and green taxi shares rose by 1% over the course of the year. Now Uber has overtaken green taxi in pickup numbers and is beginning to catch up to yellow taxi numbers. This implies that the rise in Uber pickup numbers is resulting in the decline of yellow taxi numbers. To test this hypothesis, we conduct a series of detailed analyses next.

To understand how Uber has affected yellow taxi pickup numbers, we generated heatmaps for the two services over the course of 5 months. We used the Google Maps API to generate heatmaps of New York City according to pickup locations. We focused on the Manhattan borough of New York City since most yellow taxi pickups occur there. Yellow taxi heatmaps on a day in April 2014 and then 5 months later in September are shown in Figure 5. Red areas indicate the most activity, dark blue areas indicate moderate to high activity, light blue areas indicate moderate activity, and other areas indicate minor or no activity. We noticed that the heatmap shows a strikingly similar trend between the two, even though they are 5 months apart. Locations with the most pick-ups, indicated by red, occur on the same spots on the two maps; locations with high and moderate pick-ups are very similar too. This shows that yellow taxi has not changed much in terms of expanding their pickup locations.

We generated heatmaps for Uber over the same 5 months period in 2014 as shown in Figure 6. Here we can clearly see a rapid expansion of many pickup hotspots for September compared to five months earlier in April. In April 2014, there are only a few pickup hotspots in Manhattan. Now in September the map is filled with heavy Uber activities all over the city. This reveals that Uber is rapidly expanding and taking over many taxi hotspots, showing Uber's business strategy has directly influenced taxi pickup numbers. We next use classification algorithms to quantify the expansion of Uber and the decline of yellow taxi.

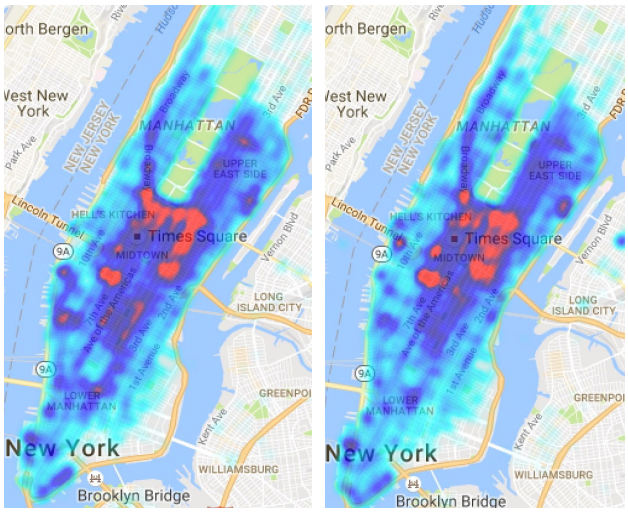


Figure 5. Heatmaps of pickup locations for yellow taxis on a day in April 2014 (left), and five months later in September (right)

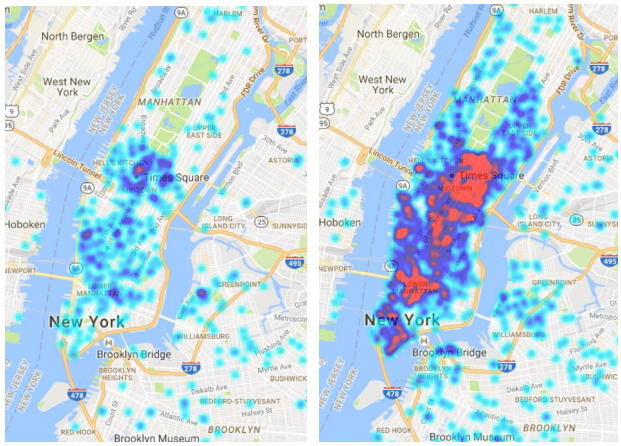


Figure 6. Heatmaps of pickup locations for Uber on a day in April 2014 (left), and five months later in September (right)

5 Classification

Given a location point in New York City, our goal is to classify this point to be one of three categories: yellow taxi, green taxi, or Uber. First, we split the map into small regions. Within each region, we utilize the pickup data to learn which service is the most prevalent in the region. We want the region sizes to be small enough so that each region can act as a mobile sensor and each region is connected to others to form a large grid. This also forms a mobile sensor network over New York City. We first discuss the choice of regions then introduce the classification algorithm used.

5.1 Region Selection and Mobile Sensing

As discussed in the previous sections, we use a hierarchical approach to breakdown the sensor network over the city to regions capable of sensing a small area. A borough covers an area of more than 60 million square meters and a zone usually cover more than 1

million square meters. Both are way too large to sense any information on the scale of individual cars. We further divide zones into regions and each region should be small enough so it can act as a mobile sensor. Manhattan is mostly made up of straight intersecting roads forming city blocks, and each block is roughly an 80m by 250m area. One city block covers an area of roughly 20,000 square meters. Having regions be a similar size to blocks provides a number of advantages such that it's small enough to offer useful services but not too small as the margin for error would be too large.

Since we are dealing with coordinates, we considered the latitude and longitude coordinates of each pickup location rounded to decimal places and used that to split the map up into regions. A latitude change in 3 decimal places from, for example, 40.730 degrees to 40.731 degrees translate to a distance change of roughly 115 meters. A longitude change in 3 decimal places from, for example, -74.000 to -74.001 equates to roughly 85 meters. Splitting the map up with coordinates rounded to 3 decimal places make each region roughly 10,000 square meters in size. Going even smaller to 4 decimal places make each region 100 square meters in size. This is way to small and would not offer any useful information because location coordinates are not this precise. Going to 2 decimal places, on the other hand, would make each region too big, almost as big as the size of the zones. Therefore, we decided to go with coordinates rounded to 3 decimal places and use that to split the map into regions. This way, each region covers 10,000 square meters and the Manhattan borough would be split into roughly 60,000 regions. Next we train classifiers to find the dominated category (yellow, green or Uber) for each region.

5.2 Algorithm

We use logistic regression as our classifier. Learning in logistic regression involves choosing the parameters w which makes the probability of the observed y values in the training dataset to be the highest, given the observations x as expressed by Equation (1). Other machine learning algorithms such as Maximum Entropy has been utilized in a number of works [40-41].

$$\hat{\omega} = \arg\max_{\omega} P(y^{(i)} \mid x^{(i)}) \quad \textbf{(1)}$$

This in term can be generalized to finding the weights which result in the maximum log-likelihood as expressed by Equation (2). We use the quasi-Newton method L-BFGS for learning the weights. Spark's MLlib provide the necessary logistic regression functions we require to perform the classification. Other works [42-43] have been proposed to tackle related problems.

$$\hat{\omega} = \arg\max_{\omega} \sum_{i} y^{(i)} \log \frac{1}{1+e^{-w \cdot f}} + (1-y^{(i)}) \log \frac{e^{-w \cdot f}}{1+e^{-w \cdot f}} \quad \textbf{(2)}$$

5.3 Performance

Each dataset consists of all pickup coordinates for yellow taxi, green taxi, and Uber over a month in 2014. We formed 6 monthly datasets in total: April to September 2014. Then for each month, we grouped the pickup coordinates by regions and each region forms its own dataset. In each month, we randomly selected 80% of the dataset as training data, the rest 20% serves as testing data. We trained classifier on the training dataset for each region using logistic regression. Then we used the test data to score the accuracy of the classification. Next, classification accuracy for each region is averaged up to arrive at an overall classification accuracy for each month. Our classifier gets an average accuracy of 85% for the April to September 2014 datasets with the highest accuracy of 90% on the April dataset.

For example, the input points (40.7600, -74.0100), (40.7700, -73.9700) and (40.0500, -69.0500) are classified respectively as Uber, yellow taxi and green taxi. This means that the first coordinate falls in a region where the most active service inside it is Uber, and the other two coordinates fall respectively into yellow and green taxi heavy regions.

We picked 3 million random location points and trained model on the April 2014 dataset using our classifier. Of those 3 million, around 600,000 points were classified into the Uber category, meaning around 600,000 points were located in regions dominated by Uber pickups. We then trained model on the September 2014 dataset using the same 3 million points. Now, around 800,000 points were classified into the Uber category, meaning around 800,000 of those points are in regions dominated by Uber pickups. An increase of 200,000 points was observed here. This shows that at 200,000 locations, the regions surrounding them was previously dominated taxi but are now dominated by Uber.

This result directly supports the observations from the pickup numbers and heatmaps, and shows that the expansion of Uber into previously taxi-dominated region is very successful. On the other hand, taxis have not yet responded with a change of strategy, reflected in the similar distributions in the yellow taxi numbers. Classifications done on the following months reveal a similar pattern, more and more regions previously classified as Yellow or Green taxi are now being classified as Uber. This gives direct insight into areas of the city where all three services provide a large number of pickups (e.g. Midtown areas in Manhattan). Now we are able to learn exactly which service dominates the areas.

6 Applications on the Mobile Sensor Network

6.1 Trip Suggestion

For riders, given any location in NYC, we can suggest to them whether yellow, green or Uber is the optimal choice in terms of wait time, distance, and chance of pickup. For taxi drivers, we can tell them to spatially avoid areas dominated by Uber and temporally focus on regions with higher historical pickups. For Uber drivers, they can spatially focus on zones with higher number of historical pickups. Given any location point in New York City, we use our classifier to output a category (yellow, green or Uber) reflecting which service is dominating the given point's surrounding region (approximately 10,000 square meters in size), allowing suggestions of service based on locations which is further exploited in the next sections.

Figure 7 shows our proposed user interface for the web application. The input of the model is an address of any location given by user in NYC, and the output is a suggestion to user on which transportation service would be the optimal choice at that specific time and location in terms of historical pickup frequency. More specifically, the front-end enable users to type in an address in the search bar, and we used Google Geocoding API to translate the address to a longitude-latitude coordinate that would be sent to the backend classifier. After the backend receives the given coordinate, it will take an adequate amount of points around the given point and the model will take those random points as input to generate prediction for each point. After running the classification model on random points, each point gets a predicted label and by summarizing which label has the highest count, the model can then claim which company historically had the highest pickups around that location.

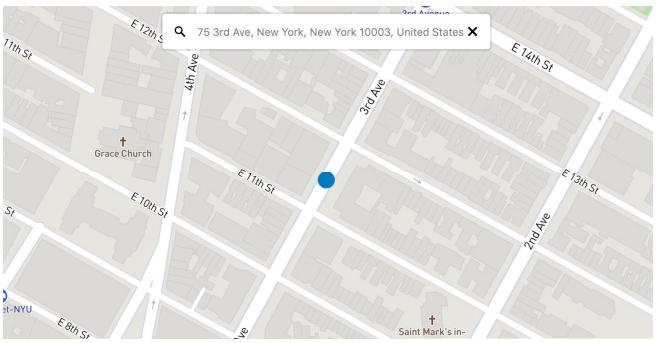


Figure 7. User interface of the map page of our proposed web service application.

7 Conclusion

We first conducted data analytics and visualizations on the NYC taxi and Uber trip dataset using big data technologies including Spark, Hive and MapReduce in order to understand traffic and travel patterns of taxi and Uber pickups. Each taxi and Uber acts as a sensor node on the interconnected mobile sensor network over NYC. We divided the city into fine-sized regions and determined the optimal region size to cover around 10,000 square meters. We then used logistic regression to train classifiers on the 2014-2015 taxi and Uber dataset with a training/testing data split of 80%/20%. Each trip is associated with a region and is classified into one of three categories based on the most active service inside the region: Yellow taxi, Green taxi or Uber. For example, a trip classified as Yellow taxi means that in the region surrounding the trip, Yellow taxi is the most active service. Our classifier achieved an average accuracy of 85% tested on the 2014-2015 monthly taxis and Uber dataset with the highest accuracy of 90% achieved on the April 2014 dataset. We next used our classifier to verify Uber's rise directly resulted in the decline of taxi pickups. Finally, we proposed applications describing how our analysis results can provide actionable insights to users of our services. Our trip suggestion application integrates the backend system producing analytics and classification results with a web front-end to provide users with meaningful visualizations and make recommendations on future rides.

References

[1] J. A. Deri, F. Franchetti, J. M. F. Moura, Big Data Computation of Taxi Movement in New York City, *2016 IEEE International Conference on Big Data*, Washington, DC, 2016, pp. 2616-2625.

[2] J. Freire, C. Silva, H. Vo, H. Doraiswamy, N. Ferreira, J. Poco, Riding from Urban Data to Insight Using New York City Taxis, *IEEE Technical Committee on Data Engineering*, Vol. 37, No. 4, pp. 43-55, December, 2014.

[3] S. Osswald, N. Brueckel, C. Brickwedde, M. Lienkamp, M. Schoell, Taxi Checker: A Mobile Application for Real-Time Taxi Fare Analysis, *Adjunct Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, Seattle, WA, 2014, pp. 1-6.

[4] V. Salnikov, R. Lambiotte, A. Noulas, C. Mascolo, *OpenStreetCab: Exploiting Taxi Mobility Patterns in New York City to Reduce Commuter Costs*, https://arxiv.org/abs/1503.03021.

[5] J. W. Y. Chan, V. L. N. Chang, W. K. Lau, K. T. Law, C. J. Lei, Taxi App Market Analysis in Hong Kong, Journal of Economics, *Business and Management*, Vol. 4, No. 3, pp. 239-242, March, 2016.

[6] S. Ma, Y. Zheng, O. Wolfson, Real-Time City-Scale Taxi Ridesharing, *IEEE Transactions on Knowledge And Data Engineering*, Vol. 27, No. 7, pp. 1782-1795, July, 2015.

[7] L. Poulsen, D. Dekkers, N. Wagennar, W. Snijders, B. Lewinsky, R. Mukkamala, R. Vatrapu, Green Cabs vs. Uber

in New York City, *Proceedings of the 2016 IEEE International Congress on Big Data (BigData Congress)*, San Francisco, CA, 2016, pp. 222-229.

[8] L. Chen, A. Mislove, C. Wilson, Peeking Beneath the Hood of Uber, *Proceedings of the 2015 ACM Conference on Internet Measurement Conference*, Tokyo, Japan, 2015, pp. 495-508.

[9] J. Cramer, A. B. Krueger, Disruptive Change in the Taxi Business: The Case of Uber, *The American Economic Review*, Vol. 106, No. 5, pp. 177-182, May, 2016.

[10] J. A. Deri, J. M. F. Moura, Taxi Data in New York City: A Network Perspective, *49th Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, 2015, pp. 1829-1833.

[11] R. Ganti, I. Mohomed, R. Raghavendra, A. Ranganathan, Analysis of Data from a Taxi Cab Participatory Sensor Network, *MobiQuitous 2011*, Copenhagen, Denmark, 2012, pp. 197-208.

[12] J. Aslam, S. Lim, X. Pan, D. Rus, City-Scale Traffic Estimation from a Roving Sensor Network, *SenSys'12*, Toronto, Canada, 2012, pp. 141-154.

[13] R. Mizouni, M. E. Barachi, Mobile Phone Sensing as a Service: Business Model and Use Cases, *Seventh International Conference on Next Generation Mobile Apps, Services and Technologies*, Prague, Czech Republic, 2013, pp. 116-121.

[14] L. Zhang, J. Liu, H. Jiang, Y. Guan, SensTrack: Energy-Efficient Location Tracking With Smartphone Sensors, *IEEE Sensors Journal*, Vol. 13, No. 10, pp. 3775-3784, October, 2013.

[15] H. Sun, S. McIntosh, Phone Call Detection Based on Smartphone Sensor Data, *2nd International Conference on Cloud Computing and Security*, Nanjing, China, 2016, pp. 284-295.

[16] H. Sun, S. McIntosh, B. Li, Detection of In-Progress Phone Calls Using Smartphone Proximity and Orientation Sensors, *International Journal of Sensor Networks*, Vol. 25, No. 2, pp. 104-114, 2017.

[17] B. Gu, V. S. Sheng, A Robust Regularization Path Algorithm for v-Support Vector Classification, *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 28, No. 5, pp. 1241-1248, May, 2017.

[18] B. Gu, X. Sun, V. S. Sheng, Structural Minimax Probability Machine, *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 28, No. 7, pp. 1646-1656, July, 2017.

[19] J. Li, X. Li, B. Yang, X. Sun, Segmentation-based Image Copy-move Forgery Detection Scheme, *IEEE Transactions on Information Forensics and Security*, Vol. 10, No. 3, pp. 507-518, March, 2015.

[20] C. Yuan, X. Sun, R. Lv, Fingerprint Liveness Detection Based on Multi-scale LPQ and PCA, *China Communications*, Vol. 13, No. 7, pp. 60-65, July, 2016.

[21] Y. Zhang, X. Sun, B. Wang, Efficient Algorithm for K-barrier Coverage Based on Integer Linear Programming, *China Communications*, Vol. 13, No. 7, pp. 16-23, July, 2016.

[22] H. Sun, S. McIntosh, Big Data Mobile Services for New York City Taxi Riders and Drivers, *Proceedings of the 5th IEEE International Conference on Mobile Services*, San Francisco, CA, 2016, pp. 57-64.

[23] G. Pan, G. Qi, Z. Wu, D. Zhang, S. Li, Land-use Classification Using Taxi GPS Traces, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 14, No. 1, pp. 113-123, March, 2013.

[24] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, Y. Huang, T-drive: Driving Directions Based on Taxi Trajectories, *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, San Jose, CA, 2010, pp. 99-108.

[25] Apache Hadoop, http://hadoop.apache.org.

[26] V. K. Vavilapalli, A. C. Murthy, C. Douglas, S. Agarwal, M. Konar, R. Evans, T. Graves, J. Lowe, H. Shah, S. Seth, B. Saha, C. Curino, O. O'Malley, S. Radia, B. Reed, E. Baldeschwieler, Apache Hadoop Yarn: Yet Another Resource Negotiator, *Proceedings of the 4th Annual Symposium on Cloud Computing*, Santa Clara, CA, 2013, Article No. 5.

[27] Apache Spark, http://spark.apache.org.

[28] J. Dean, S. Ghemawat, MapReduce: Simplified Data Processing on Large Clusters, *Communications of the ACM*, Vol. 51, No. 1, pp. 107-113, January, 2008.

[29] Apache Hive, http://hive.apache.org.

[30] Q. Liu, W. Cai, J. Shen, Z. Fu, X. Liu, N. Linge, A Speculative Approach to Spatial-temporal Efficiency with Multi-objective Optimization in a Heterogeneous Cloud Environment, *Security and Communication Networks*, Vol. 9, No. 17, pp. 4002-4012, November, 2016.

[31] Z. Fu, K. Ren, J. Shu, X. Sun, F. Huang, Enabling Personalized Search over Encrypted Outsourced Data with Efficiency Improvement, *IEEE Transactions on Parallel and Distributed Systems*, Vol. 27, No. 9, pp. 2546-2559, September, 2016.

[32] Z. Fu, X. Sun, Q. Liu, L. Zhou, J. Shu, Achieving Efficient Cloud Search Services: Multi-keyword Ranked Search over Encrypted Cloud Data Supporting Parallel Computing, *IEICE Transactions on Communications*, Vol. E98-B, No. 1, pp. 190-200, January, 2015.

[33] Z. Fu, X. Wu, C. Guan, X. Sun, K. Ren, Towards Efficient Multi-keyword Fuzzy Search over Encrypted Outsourced Data with Accuracy Improvement, *IEEE Transactions on Information Forensics and Security*, Vol. 11, No. 12, pp. 2706-2716, December, 2016.

[34] Z. Xia, X. Wang, X. Sun, B. Wang, Steganalysis of Least Significant Bit Matching Using Multi-order Differences, *Security and Communication Networks*, Vol. 7, No. 8, pp. 1283-1291, August, 2014.

[35] Z. Xia, X. Wang, L. Zhang, Z. Qin, X. Sun, K. Ren, A Privacy-preserving and Copy-deterrence Content-based Image Retrieval Scheme in Cloud Computing, *IEEE Transactions on Information Forensics and Security*, Vol. 11, No. 11, pp. 2594-2608, November, 2016.

[36] X. Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkatarama, D. Liu, J. Freeman, D. B. Tsai, M. Amde, S. Owen, D. Xin, R. Xin, M. J. Franklin, R. Zadeh, M. Zaharia, A. Talwalkar,

Mllib: Machine Learning in Apache Spark, *The Journal of Machine Learning Research*, Vol. 17, No.1, pp. 1235-1241, April, 2016.

[37] R. Sen, A. Maurya, B. Raman, R. Mahta, R. Kalyanaraman, N. Vankadhara, S. Roy, P. Sharma, Kyun Queue: A Sensor Network System to Monitor Road Traffic Queues, *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems*, Toronto, Canada, 2012, pp. 127-140.

[38] C. Wenjie, C. Lifeng, C. Zhanglong, T. Shiliang, A Realtime Dynamic Traffic Control System Based on Wireless Sensor Network, *ICPP 2005 Workshops*, Oslo, Norway, 2005, pp. 258-264.

[39] J. Eriksson, L. Girod, B. Hull, R. Newton, S. Madden, H. Balakrishnan, The Pothole Patrol: Using a Mobile Sensor Network for Road Surface Monitoring, *Proceedings of the 6th International Conference on Mobile Systems, Applications, and Services*, Breckenridge, CO, 2008, pp. 29-39.

[40] H. Sun, R. Grishman, Y. Wang, Active Learning Based Named Entity Recognition and Its Application in Natural Language Coverless Information Hiding, *Journal of Internet Technology*, Vol. 18, No. 2, pp. 443-451, March, 2017.

[41] H. Sun, R. Grishman, Y. Wang, Domain Adaptation with Active Learning for Named Entity Recognition, *2nd International Conference on Cloud Computing and Security*, Nanjing, China, 2016, pp. 611-622.

[42] B. Chen, H. Shu, G. Coatrieux, G. Chen, X. Sun, J. L. Coatrieux, Color Image Analysis by Quaternion-type Moments, *Journal of Mathematical Imaging and Vision*, Vol. 51, No. 1, pp. 124-144, January, 2015.

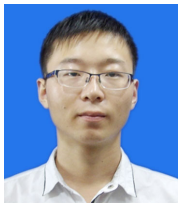
[43] Z. Zhou, Y. Wang, Q. J. Wu, C. N. Yang, X. Sun, Effective and Efficient Global Context Verification for Image Copy Detection, *IEEE Transactions on Information Forensics and Security*, Vol. 12, No. 1, pp. 48-63, January, 2017.

Biographies

Huiyu Sun is a Masters student in the Department of Computer Science at New York University. His research interests are in natural language processing, deep learning, wireless sensor networks and big data technologies.

Siyuan Hu is an undergraduate student majoring in Computer Science at New York University. His research interests include distributed systems and wireless networks.

Suzanne McIntosh is an Adjunct Professor at New York University, a technology consultant with Cloudera, and previously worked at IBM T. J. Watson Research Center where she led cross-disciplinary research teams in virtualization, data center energy optimization, and security research. She is Guest Editor of Transactions on Services Computing.



Yi Cao received the BE degree in Nanjing University of Information Science & Technology in 2016, China. He is currently working towards the MS degree in Nanjing University of Information Science & Technology, China. His research interest includes network and information security.