NYC Multimodal Transportation Recommendation System

Chen Yang
New York University
NYC, US
cy1285@nyu.edu

Zhonghui Hu
New York University
NYC, US
zh1272@nyu.edu

Xuebo Lai
New York University
NYC, US
xl1638@nyu.edu

***Abstract*—Multimodal transportation system has been developed in recent years as the single modal transportation can hardly deal with the multiplex transportation demands. The aim of the paper is to develop a NYC multimodal transportation recommendation system based on exploring the travelling pattern of citizens taking various public transportation tools in NYC. At the stage of exploring user group's preference and travelling patterns, multiple datasets such as taxi data, bike data, subway data and weather data are collected. During the analytics, Spark and Scala are used to fully understand these data to find some useful insights important for building the recommendation model. For the model, the goal of shortest duration and price are set up and a multi-objective optimization model considering the time and weather is built. In this paper, the multi-objective model is transferred into single objective optimization problem using weighted sum method. Floyd-Warshall algorithm is utilized to find the optimal path. Finally, based on the model we built a recommendation system. On the website, the user can choose the origin and destination location and the website could give the "best route" with different transportation modes to the user.**

Keywords—analytics, multimodal route planning, Spark

I. INTRODUCTION

With the growth of economy and urbanization, the total amount of travel has been increasing rapidly. In the meanwhile, more public transportation modes are available nowadays. Subway, buses, taxis, and bicycles make the choice of travel modes no longer a single way.

Different public transportation tools have their own advantages in speed, cost, and comfort. According to the research of [1], when choosing the way of transportation, several factors are taken into considerations, including the time, cost, convenience, safety and comfort. Multimodal transportation recommendations can promote the development of intelligent transportation system, including reducing transfer times, alleviating traffic congestions and distributing traffic flows. As for the online routing applications, the methods predominantly compute the shortest routes, to some extent overlooking users' preferences under different temporal and spatial situations.

To keep the transportation system sustainable, more and more cities' governments begin adding investment on the development of public bicycles, for example, in New York City, the number of Citibike stations increases rapidly in recent years and the bike routes are highly connected, as is shown in Figure 1.1. Sustainable as it can be, many transportation recommendation systems take little consideration on the possible usage of citibikes.

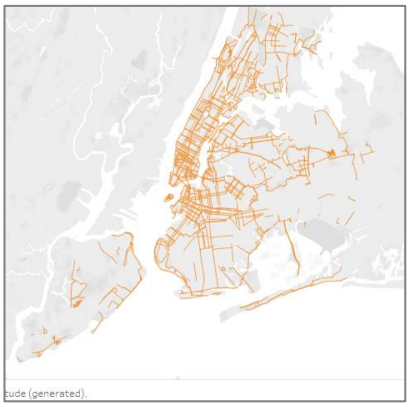

Figure 1.1 Bike routes

Utilizing open datasets in New York City, this paper gives an in-depth analysis to different transportation modes under various temporal and spatial scales, including subway, yellow taxis and Citibikes. Based on the findings from the analysis, a multimodal transportation recommendation algorithm is designed for multiple objectives according to user preference. To validate the result, we will build the user interface for passengers to find the optimal route after typing in the origin and destination. This research is based on Hadoop Spark and use Scala as the programming language. The distributed computing framework allows for a better performance of the computation to large datasets. In the project, we use Spark to profile, clean, analyze the datasets and implement the recommendation algorithm. Our contributions can be summarized as follows:

1. Analyze datasets of different transportation modes in NYC to explore users' preference to transportation and

the system's built-in features under various temporal and spatial conditions.

2. Construct multimodal recommendation method to provide optimal multimodal route considering various factors, such as speed, cost and safety.
3. Design and build website for users to check the recommendation.

The paper is organized as follows: the second section gives the motivation of the research. Related researches have been introduced in the third section. In section IV, each dataset will be elaborated. Description of analytics and design diagram are given in the next two sections to better illustrate the research. In section VIII we will first explore the transportation pattern of NYC in depth to analyze the in-built traffic characteristics and users' preference based on open datasets in NYC. With these findings, the multimodal transportation recommendation algorithm considering different scenarios is constructed. We give the conclusion of the research in the ninth section and more perspectives about the future works will be given in the last section.

II. Motivation

This paper aims at developing a multimodal transportation recommendation model by exploring the pattern of public transportations in New York City, including subway, yellow taxis and Citibikes. By analyzing the transportation pattern and user group preference under different scenarios, a multimodal transportation recommendation algorithm will be constructed, which can fulfill the demand of users in different perspectives. Solution to the recommendation algorithm will be provided then and a website will be built as the user interface, allowing passengers to check the recommendation combination of transportation modes and routes from the origin to the destination.

III. Related Work

Multimodal transportation route planning has been studied over the last years. The goal of multimodal route problem is to seek the best route combined with different transportation modes, given the networks.

Many researchers solve multimodal transportation route planning problem through the basic idea of seeking the shortest path, with different considerations on the cost of the path. When solving the model, Dijkstra's algorithm and A* algorithms are often used to calculate the shortest path where the weights between nodes are correlated with the distances of the edges.

A variety of route planning algorithms have been proposed to fit in different scenarios. Liu[2] etc. take into consideration the time and risk to establish a multi-objective model which aims at minimizing both the risk and the cost. Liu [3] etc. established a multi-objective programming model aiming at minimization of both cost and amount of carbon emissions and integrate the programming problem as a single-objective one. Luo[4] etc. modeled the multimodal transportation problem as a multi-objective planning problem, in which the first objective is minimizing the total cost and the second is minimizing the self-defined time penalty. The researchers designed a heuristic algorithm to solve the planning problem. Floyd algorithm is firstly used to find the relative k shortest paths and then GA[5] algorithm is applied to distribute the optimal mode. While this method takes both the cost and the time into consideration, it is proposed based on three strict hypotheses, requiring that the transfer happens in nodes, there is one transportation between two nodes and the trip is from the origin to the destination, with no division permitted. These hypotheses put tight constraints on the travel modes. What's more, the algorithm proposed overlooks personal preference to the transportation, which under some conditions are vital to the choice of combined transportation tools through traveling.

A bunch of papers put focus on the analysis of users' preference to the transportation and try to personalize the multimodal transportation recommendation. Most of such researches focus on the design and implement of the survey, as well as the analysis to the result of the survey. Liu[6] et al. build a mode choice model by conducting the combined survey of revealed preference survey and stated preference survey. Through the survey, relationship between personal features and trip characteristics, correlation between different distance and the choice of transportation modes are studied. Nested Logit Model is then used to analyze the combined travel utility. Though the findings are of significance for providing insight into the multimodal transportation recommendations, it takes a lot to conduct the survey in order to gather passengers' opinions and may have deviations when it comes to different road and weather conditions. Lucas[7] et al. come to conclusion that built-in characteristics have greater impact on way of traveling, compared with personal attributes, though different user groups have distinct transportation mode choice. By establishing optimal traffic flow model, Si etc. [8] also reveal that the traffic demand is time various. These findings show that it is of importance to explore the transportation pattern before designing the recommendation method, while most of the researches put emphasis on modelling the choice model of users under certain conditions, rather than developing a multimodal way of recommendation.

With the development of machine learning, multimodal transportation recommendation comes to a new stage. Liu[10] et al. proposed a novel learning-based idea for multimodal recommendation. Based on the query data provided by Baidu Map, the researchers first extract a multi-modal transportation graph, consisting of users, Origin-Destination pairs and different transportation modes. The concept of Trans2Vec is proposed, inspired by the Word2Vec method in natural language processing. Researchers then apply joint representation learning method on the graph to study the user preference and OD pairs. Though this method has a good performance, it is highly demanding for large query datasets.

New York City has abundant datasets available in daily public transportation, including yellow taxi, subway, bus and

Citibike. While each kind of data source has been explored in depth, little work has been done to study the correlations between these datasets, leading to insufficient use of the datasets. This paper makes full use of these public transportation datasets to explore the preference of passengers in New York City and the characteristics of traffic conditions under various temporal and spatial conditions. Based on the findings, multimodal transportation recommendation model will be established, aiming at providing the optimal routes with several factors taken into considerations, such as speed, cost, time period and weather condition.

IV. DATASETS

The datasets that have been used are as follows:

- Taxi data
- Subway data
- Bike data
- Weather data

Yellow Taxi data was extracted from New York City TLC trip record data. There are in total 19 fields in the raw data including break-down of the total trip fee, trip distance, start location, ending location, etc. The time range for taxi data that we will be using for this project is from 01/01/2018 to 12/31/2018. After processing the data, eights fields that are relevant to the research are retained. Information about these columns are listed below.

Table 4.1 Yellow taxi dataset

Columns	Type	Max (value/length)	Min (value/length)
start Date	String	2018-12-31	2018-01-01
start Time	String	4	1
end Date	String	2018-12-31	2018-01-01
end Time	String	4	1
trip Dist	Double	99.95	0
pickup	String	265	0
dropoff	String	265	0
amount	Double	999.56	0

Pick up and drop off locations are the number mapped by the TLC taxi zone in New York City. The amount is the aggregated taxi fee for a trip.

In this paper four time ranges are generated: 6:00am to 10:00 am(1), 10:00 am to 16:00 pm(2), 16:00 pm to 21:00 pm(3), 21:00 pm to 06:00 am(4). Trip start and end time have been mapped to number 1/2/3/4 based on which time range the data entry belongs to.

Subway data was downloaded from NYC Open Data. This dataset includes the location of subway stations in NYC. The schema of the data is as follows.

Table 4.2 Subway dataset

Columns	Type	Max Length/Digits	Min Length/Digits
Name	String	34	5
Latitude	Double	14	12
Longitude	Double	11	15
Line	String	15	1

Citibike data is accessed from Citibike official website. Citibike record data in the New York City is collected from 1/1/2018 to 31/12/2018. The schema of Citibike dataset is shown as follows.

Table 4.3 Citibike dataset

Columns	Type	Max (value/length)	Min (value/length)
Duration	Int	19510049	61
Start_time	String	19	19
Stop_time	String	19	19
Latitude	Double	45.506	40.647
Longitude	Double	-73.569	-74.025

Weather data was downloaded from National Centers For Environmental Information. We collected the New York City weather data from 1/1/2018 to 31/12/2018. The size of this dataset is 4.8 MB. It contains lots of useful columns, such as windspeed, temperature and precipitation. The schema of this dataset is as follows.

Table 4.4 Weather dataset

Columns	Type	Max	Min
Date	String	21	21
Temperature	Double	95	5
Precipitation	Double	1.69	0.0
Windspeed	Double	21	0

V. APPLICATION DESIGN

At the first stage of the project, data including taxi, subway, citibike and weather from different sources is collected and stored in Hadoop HDFS. In Spark, the travelling patterns of taxi, subway and citibikes are analyzed under different weather

conditions and time periods. Average velocity and cost of each transportation mode under different weather conditions and time period are generated. With the result of analysis, a multimodal transportation recommendation model minimizing both duration and cost is built and Floyd-Warshall algorithm is utilized to find the optimal path.

After preforming the previous steps, we would gather enough information from the algorithm and analytics to compile the data layers. Data layers has the highly aggregated metadata which are much smaller than original data and running result from the Floyd Washer Algorithm. We created this layer to accelerate the program speed by avoiding running the back-end spark data query code from the beginning and Floyd-Warshall algorithm implementation every time when a user query for the best route information, since the cost for running either spark data query code and Floyd Warshall Algorithm code is very high. Therefore, data layer can be considered as caching the output from data analysis and the algorithm in a sense. However, to keep the data layers up to date, we plan to automate the process of data layers updating itself from the newest data in Spark in the future.

With the data layer constructed as described above, we have the options to build numerous applications on top on it. Because of the limitation of time, we would build a web application as demonstration for the project.

The design diagram is shown in Figure 5.1.

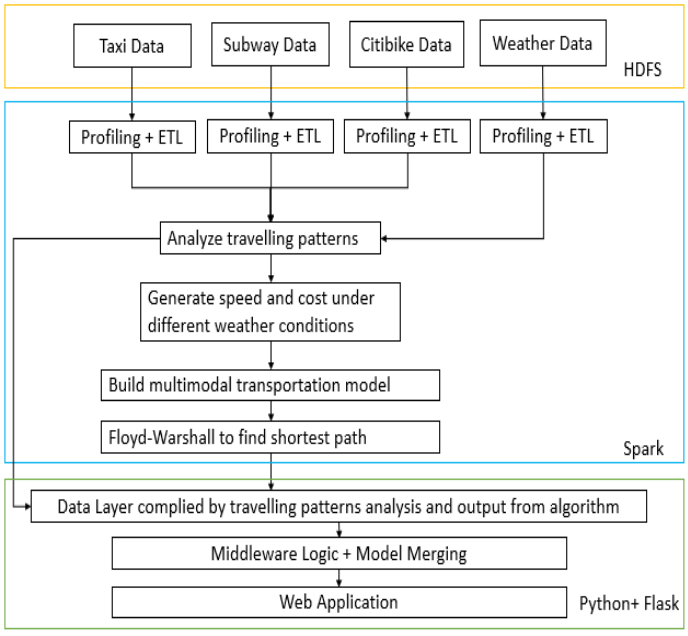

Figure 5.1 Design Diagram

VI. Description of Analytic

1. An overall description of the analytic

There are three major layers in our projects: Data Processing Layer, Implementation of Multi-Layer Floyd Washer Layer for finding shortest path and Aggregated Data Layers which is the metadata complied from the previous steps. The analytics for this project mainly happen in the first layer: Data Processing.

The backend of Data Processing layers is Spark Scala REPL(Read-Evaluation-Print-Loop). We adopted Spark Scala REPL because the interactive shell can help us keep track of data during each step of Data ETL. After performing ETL for each of the dataset, we used Spark DataFrame API and DataFrame API to merge different dataset, aggregate their columns' information by certain standards and query the aggregated data to generate actionable insights discussed in detail on VIII Analytics Section. For example, after the ETL steps for taxi and weather data, we mapped taxi records data to weather condition data based on their date. Then we aggregated the taxi records data using weather condition data and query the aggregated data to understand correlation between taxi's average speed/travel duration/fare amount and weather conditions. Similar data processing and analytics steps were performed on Bike data as well. The findings from this layer are the relations between different data such as correlations between bike/taxi's data and weather condition's data mentioned above. Those actionable findings help us to understand how we should construct the algorithm layers and data layers. Details for all the actionable insights and findings from Data Processing Layers can be found in VIII Analytics Section.

For now, this layer is based on the data we uploaded to HDFS. In the future, we would like to fully automate this process so that the program can automatically download data from their own official database or website. In that case, our program would no longer ask users for inputting data.

2. Taxi Data Analysis

The first analysis for Taxi Data was the pickup and drop off location density analytics.

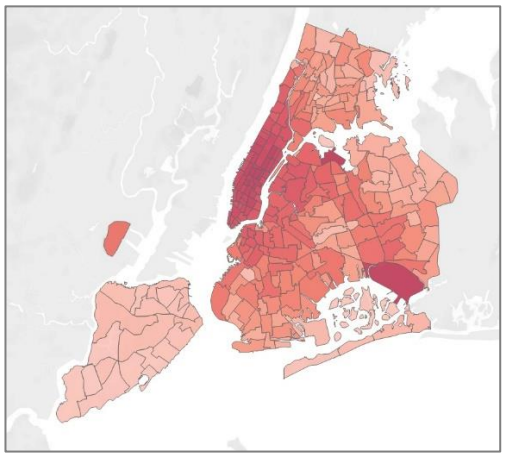

Figure 6.2.1 Taxi pickup by taxi zones

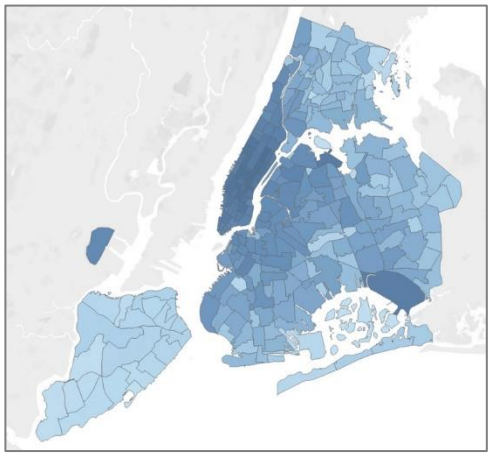

Figure 6.2.2 Taxi drop off by taxi zones

In Figure 6.2.1 pickup heatmap, the more counts an area have, the darker red that area is. As the graph indicates, Manhattan area and the area along Manhattan island have the most counts of pickup. Also, what worth noticing is the area at bottom right of the graph. It has an unusual darker red compared with the areas around it, which is the JFK airport area. Apparently, many people (possibly a good portion of them is tourists) opt to take taxi to airport.

Figure 6.2.2 is the drop off heatmap. The darker blue an area is, the more drop off counts that area has. The densest areas for drop off counts are also in Manhattan and the area along it, as well as JFK and LaGuardia airport. This is a surprising result under the assumption that people would take taxi one-way from one location to another. Here, hypothesis could be raised that a portion of the people tends to take taxi more and use taxi as their regular commuting tools, because the similarity of pickup and drop off heatmap indicates there are many a trip happening between certain areas. This hypothesis will not be the focus point for this paper.

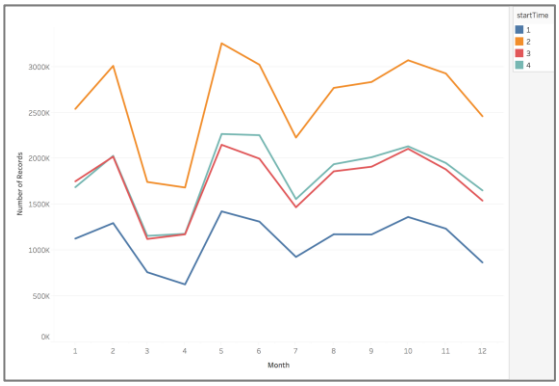

Figure 6.2.3 Usage of taxi

Figure 6.2.3 shows the usage of taxi data across 2018. Taxi data is split based on time during a day and different colors are used for different time during a day. Period 1 is the time from 6 am to 9 am which is considered as the morning traffic peak hours. Period 2 is the time from 10am to 4pm which is considered as the regular hours during daytime. Period 3 is the time from 5 pm to 8 pm, which is considered as the evening traffic peak. Finally, period 4 is from 9 pm to 5 am which is considered as nighttime. The count number for different time period in a day is mapped against each month to get Figure 6.2.3. As can be seen, the total taxi usage reached the peak around February, May and October, and the month of March and April have the least taxi records. It is very intriguing to see sudden surge of taxi records from least records in April to most records in May. Also, as can be seen from the graph, the time period in a day at which people tend to use taxi is fixed across the year. People most likely to use the taxi service at Period 2, which is from 10 am to 4 pm at a day. They are least likely to use taxi for period 1 which is 6 am to 9 am at a day, possibly due to the morning traffic.

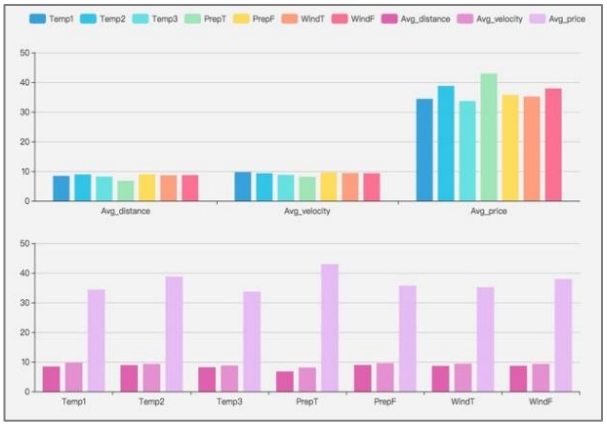

Figure 6.2.4 Information under different weather conditions

Figure 6.2.4 upper describes the average travel distance, average traveling velocity and average price mapping against each of the weather conditions. The weather conditions in the graph are divided to 3 dimensions, including temperature, rainfall or snowfall and wind speed. Temperature can be mapped to cold weather (temp1), regular weather (temp2) and hot weather(temp3). Rainfall or snowfall condition can be mapped to raining or snowing(PrepT), no precipitation(PrepF). Wind condition can be divided to strong wind(WindT) and no strong wind(WindF).

Although not very obvious, it can be seen that taxi's speed is slowest during the time when there is precipitation, and the price reaches the peak during precipitation, which fits into empirical experience since the traffic is usually not good during raining or snowing. From the graph, we can also see that the temperature and whether it is windy independently would barely affect the taxi traveling velocity and prices too much.

Figure 6.2.5 Taxi velocity and duration under different weather conditions

After conducting aggregation of taxi data and weather conditions shown in Figure 6.2.4, we look into the taxi data aggregated by all different weather conditions simultaneously as demonstrated in Figure 6.2.5. The left chart in Figure 6.2.5 denotes the travel velocity and the right chart denotes the average traveling duration. In the Figure 6.2.5, three characters represents the weather conditions: the first character denoting temperatures (1-low, 2-regular, 3-high); the second character means the whether it is raining/snowing; the third character signaling whether it is raining. From the graph above, we can clearly see that when it is cold, raining and windy, the average travel duration reaches the maximum and average travel velocity reaches the minimum. This weather condition causes the most significant difference in taxi traveling data than other weather conditions. What's more, from the chart, we can easily see that raining and low temperature would cause the taxi to slow down traveling speed significantly. From this analysis, we can conclude that multiple weather factors combined have more predicting power and determinacy than independent weather condition by comparing Figure 6.2.5 to Figure 6.2.4. Despite the need to combine multiple weather conditions to determine taxi traveling data, the low-temperature and raining/snowing weather conditions would worsen taxi traveling duration and speed in general.

2. Bike Data Analysis

With the records of Citibikes generated in 2018, the durations of bike riding are counted in 5 minutes intervals. Figure 6.3.1 shows the distribution of different durations.

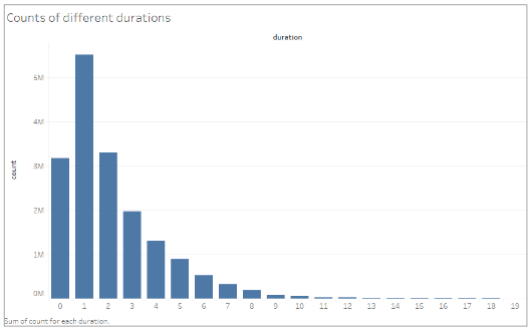

Figure 6.3.1 Distribution of durations

From the histogram it can been seen that most bike riding has a duration within 20 minutes and the number of ridings between 5 to 10 minutes dominates the count of other durations. It can be referred that the purpose of biking riding is mainly short time travelling and transferring.

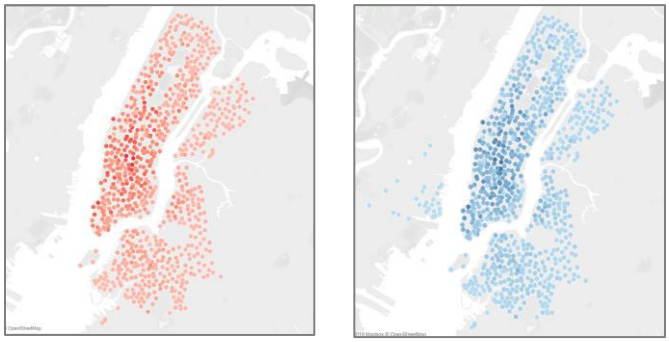

Figure 6.3.2 Citibike pickup and drop off

Figure 6.3.2 shows the pickup and drop off locations for Citibikes in NYC during 2018. The color deepens with the increase of the usage rate of the site. This figure shows that Citibikes are evenly distributed in NYC, especially in Manhattan. Thus, users can basically reach their destination through Citibike in Manhattan. It can be seen that darker spots are mostly concentrated in Midtown Manhattan. Therefore, in these sites with high demand and utilization rate, citibike can increase the number of bicycles appropriately to better meet the needs of customers.

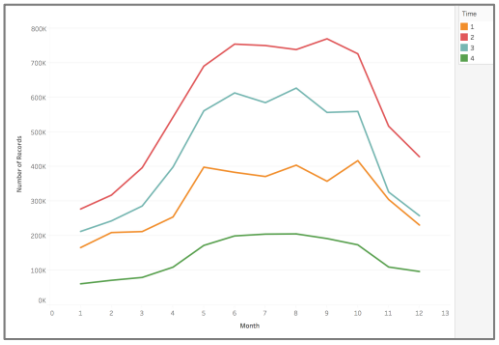

Figure 6.3.3 Usage counts during four time periods

Figure 6.3.3 shows the bike usage counts grouped by four time periods. It is clear that the peak usage of Citibikes appears during May to October. Part of the reason lies in that the weather conditions are mild during this time.

In the same month, the usage of citibike is also very different in different time periods. The time period with the highest usage rate is time period 2(10am-4pm), followed by time period 3(5pm-8pm), then time period 1(6am-9am), and finally time period 4(9pm-5am).

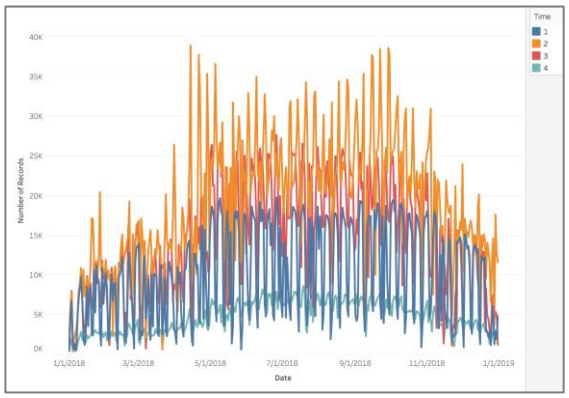

Figure 6.3.4 Usage count in different time period

The above figure shows the usage count grouped by four time period on each day. The lines change periodically. The most obvious is the yellow line chart, and basically every peak is on Saturday. For the comparison of the four time periods, the same as above. Less people ride bicycles at night.

For the purpose of implementing a multimodal transportation recommendation system, which takes the impact of weather condition into consideration, the relationship between the speed of Citibike and weather condition can be further explored.

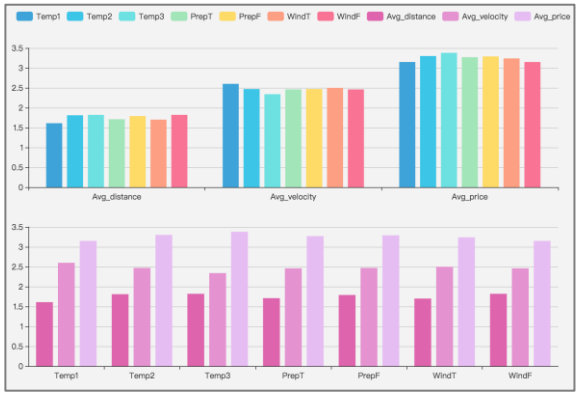

Figure 6.3.5 Information under different weather conditions

The above figure shows the average distance, average velocity and average price for each single weather. First of all, in the three cases of low temperature, rainfall and strong winds, the average distance is low, probably because people will use other modes of transportation under the bad weather conditions. For the average speed, the average speed is slower when the temperature is higher, and this is a reasonable finding. For the average price, it is positively correlated with riding time. The average price is higher at high temperatures, which means the low riding speed makes price higher. Therefore, it is not a good choice for cycling when the temperature is hot.

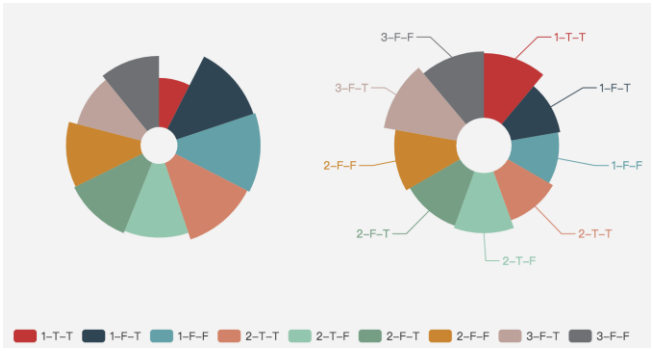

Figure 6.3.6 Impact of single weather factor on Citibike

The above analysis of the impact of a single weather factor on Citibike, the following analysis of the impact of the combination of the three weather factors on the speed and duration of cycling. From the chart on the left, we can clearly see that the speed is the slowest in the case of 1-T-T (the weather is cold and rainy and windy, the three characters had the same meaning as the figure in taxi analysis.) Next, the slower weather combination is 3-F-T and 3-F-F. This shows that in high temperature weather, whether it is windy or not, the speed will be slow. From the chart on the right, we can see that under the combination of weather that makes cycling very slow, the duration time are longer.

From all the analysis of Citibike, we find that both the time period and the weather conditions have significant impact on the usage and speed of Citibike. Thus, we need to know the speed of Citibike under any combination of time and weather. Figure 6.3.7 below shows the result.

startTime	temp	prep	wind	Avg_velocity
4	2	T	F	2.4765044892512096
2	1	F	F	2.4907817457196666
4	2	T	T	2.747594015644973
2	1	F	T	2.436768960334458
3	2	F	F	2.4610921546273117
1	2	T	F	2.6915844025018383
3	2	F	T	2.488845241569071
4	3	F	F	2.44862733954025
1	2	T	T	2.73238817189107
3	1	F	F	2.6158490277218056
3	1	F	T	2.5450484493783494
4	2	F	F	2.5043991962451315
1	3	F	F	2.7223654958248003
2	2	T	F	2.332034445203353
4	2	F	T	2.5752111256650623
2	2	T	T	2.3949505881292654
4	1	F	F	2.7051531130539472
1	2	F	F	2.749400998026502
4	1	F	T	2.647109086128926
1	2	F	T	2.7385223582066534
2	3	F	F	2.2471107348936536
3	2	T	F	2.454509682446299
1	1	F	F	2.748529316852665
2	3	F	T	2.2033824137305347
3	2	T	T	2.5630741812969156
1	1	F	T	2.7391119954440019
2	2	F	F	2.2888743943391945
2	2	F	T	2.331423352185823
3	1	T	T	1.5325911070318845
3	3	F	F	2.3934825981126746

Figure 6.3.7 Average speed under each weather condition

VII. Actuation or Remediation

1. Build Topological Model

Various transportation modes and networks, including walking, citibike, subway, taxi, become the backbones for daily movement in the city. In this paper, the complex road network is split into different layers with each containing only one mode. A network containing subway, taxi and bike can be transformed into 3 layers, as is shown in Figure 7.1.

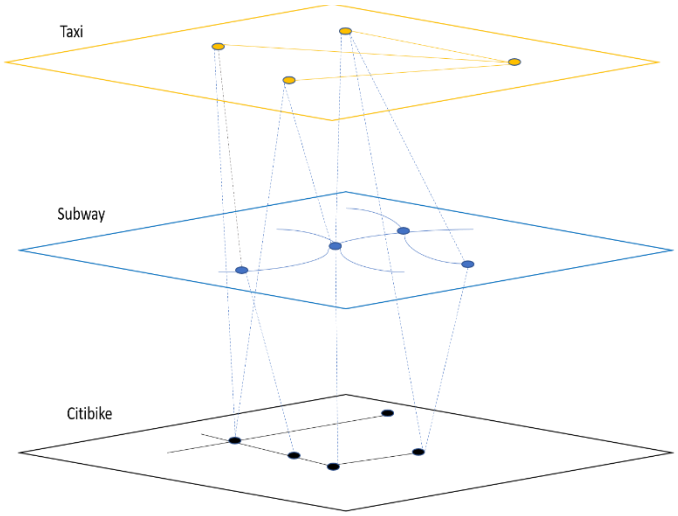

Figure 7.1 3-layer network

Each layer with nodes and edges could be seen as a single network, in which the nodes belongs to one mode.

1.1 Subway layer

In subway layer, each node in the network is a subway station. The edge between the nodes represents whether the two stations are directly connected. If there exists a line that directly connects two stations without passing other stations, the two stations in the graph have an edge. Table n shows part of the adjacency matrix of the subway cost between each OD pair.

Table 7.1 Cost between two stations

O \ D	45	46	47	48	49	50
45	0	inf	inf	inf	inf	inf
46	inf	0	120	inf	inf	inf
47	inf	120	0	inf	inf	inf
48	inf	inf	inf	0	inf	inf
49	inf	inf	inf	inf	0	inf
50	inf	inf	inf	inf	inf	0

1.2 Bike layer

In bike layer, each node in the network is a bike station. The edge between the nodes represents whether the two stations are directly connected. When modelling the edges between bike stations, not only the accessibility between stations should be taken into consideration, but how accessible should be taken into consideration should be considered.

In the analysis of bike duration, the statistics of duration reveal the distribution of user's riding duration. As 0 to 20 minutes are preferred by most users, although each bike station can be accessible from another bike station, the edge between the nodes in bike layer are filtered by duration. Table n shows part of adjacency matrix for building bike layer.

Table 7.2 Cost between bike stations

O \ D	72	79	82	83	119	120
72	0	inf	inf	inf	inf	inf
79	inf	0	678.841	inf	inf	inf
82	inf	678.841	0	inf	1638.78	inf
83	inf	inf	inf	0	895.718	962.652
119	inf	inf	1638.78	895.718	0	1234.92
120	inf	inf	inf	962.652	1234.92	0

1.3 Taxi layer

Taxi layer consists of nodes representing both bike and subway stations and each node in this layer is directly linked to the destination node as taxi is a point to point transportation mode. The cost between nodes in taxi layer can be computed by evaluating the distance between the current location and the destination, with which duration of the trip can be computed and price can be estimated.

1.4 Between layers

To finish a trip with combinations of several transportation modes, the nodes in different layers can be connected. All three layers are connected by edges of different costs, the cost of edges between any two layers is defined as transfer cost. Edges between taxi layer and other two layer have zero cost as the taxi can pick up passenger at almost all locations, and therefore demands little cost for passengers to move exchange from one mode to another.

Edges between subway layer and bike layer is defined as the duration of walking from one bike station to a subway station, and vice versa. Considering the truth and the analysis of duration of bike riding, only those pairs with duration shorter than 30 minutes are considered directly connected. Table n shows part of the transfer cost between different stations.

Table 7.3 Cost of transfer between bike and subway

O \ D	1	2	3	4	5	6
72	inf	inf	525.1617	inf	inf	inf
79	896.3843	273.6463	inf	inf	inf	inf
82	1118.703	424.6209	inf	inf	inf	inf
83	inf	inf	inf	174.1804	inf	inf
119	inf	1572.389	inf	857.1709	inf	inf
120	inf	inf	inf	740.3737	inf	inf

2. Build Multi-objective Model

This paper focus on minimizing the total cost of price and time, considering the impact of travelling time and weather conditions. The goal of optimization is to minimize both the price and duration.

$$\min Z_1 = \sum\sum C_{ij}^k X_{ij}^k + \sum\sum C_i^{kp} Y_i^{ts}$$

$$\min Z_2 = \sum\sum W_{ij}^k X_{ij}^k + \sum\sum W_i^{kp} Y_i^{ts}$$

Subject to

$$\sum\sum X_{ij}^k \geq 1$$
$$\sum\sum Y_i^{ts} \geq 0$$
$$k, p, s = 1,2,3$$
$$i, j = 1,2, \dots, n$$

X_{ij}^k: $equals\ 1\ if\ going\ from\ i\ to\ j\ in\ k\ mode, 0\ otherwise$

Y_k^{ts}: $equals\ 1\ if\ transferring\ t\ to\ s\ \ mode, 0$ otherwise

C_{ij}^k: $price\ between\ moving\ from\ i\ to\ j\ in\ k\ mode$

C_i^{kp}: $price\ of\ transfering\ from\ i\ to\ j\ mode$

W_{ij}^k: $duration\ between\ moving\ from\ i\ to\ j\ in\ k\ mode$

W_i^{kp}: $duration\ of\ transfering\ from\ i\ to\ j\ mode$

In this paper, the multi-objective problem is designed to solved by weighted sum method. Weighted sum method is an easy way to transfer the multi-objective problem into single objective problem. Here duration and price of the travel are considered equally important here, so each is assigned with the weight of 0.5.

3. Algorithm

In this paper, as the multi-objective model is linearized to single-objective model, to minimize the cost, Floyd-Warshall algorithm can be used to calculate the cost of the shortest path. Floyd-Warshall can be used to calculate the shortest path between many origin-destination pair. Given a set of vertices and edges in a graph, Floyd–Warshall algorithm is one of the best algorithms for calculating the shortest path between any two nodes. The framework of the algorithm is shown in table n.

To implement Floyd-Warshall algorithm, two square matrices should be firstly built, with one holding the cost of the edges between origin and destination pairs, and another holding the latest updated intermediate node between the shortest path for a pair of nodes. In the algorithm, the cost matrix and intermediate matrix will be dynamically updated to find the shortest path and node on the path.

Table 7.4 Framework of Floyd-Warshall

Floyd – Warshall algorithm implement

global: $cost[i][j], inter[i][j]$
for i ← 1 to n do
  for j ← 1 to n do
    $if\ cost[i][j]\ !=\ Inf$
      $inter[i][j] = -1$
    $else$
      $inter[i][j] = 0$
    $endif$
  $endfor$
$endfor$
for k ← 1 to n do
  for i ← 1 to n do
    for j ← 1 to n do
      if cost$[i][j] \geq cost[i][k] + cost[k][j]$
      cost$[i][j] = cost[i][k] + cost[k][j]$
      inter$[i][j] = k$
      endif
    endfor
  endfor
endfor

With the cost of the shortest path calculated and the intermediate node recorded, all the nodes in the path can be traced through recursion.

Applying Floyd-Warshall algorithm in our model, the input cost matrix should be built with taxi cost, subway cost, bike cost and transfer cost combined together into one matrix. The matrix has the following form.

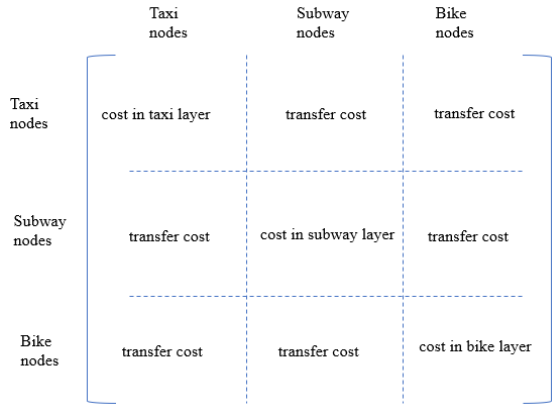

Figure 7.2 Form of input matrix

4. Design

In the analysis to different transportation modes in section V, weather condition and time of the travel impacts the total amount of travelling, average velocity of each transportation mode, as well as the price. In this model, according to the

specific weather conditions, such as whether it rains, whether it snows and which level the temperature is belonged to, there are 12 combinations of various weather conditions. According to the time of travelling, one day is split to 4 period and therefore 48 combinations of various context for travelling is generated. The velocity under each condition is calculated for evaluating the duration of certain OD pairs, with which the cost matrix can be generated. Figure 7.3 shows the design diagram.

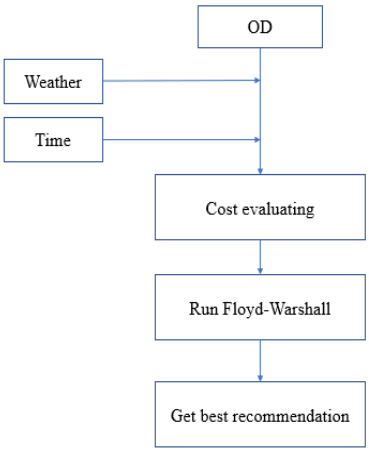

Figure 7.3 Design diagram

As mentioned above, with the data layer caching result from algorithm and data analysis, we can build numerous applications based on the data layer. For now, we planned to build a relatively simply web application named NYC Best Route Recommendation System, as shown in Figure 7.4 and 7.5. The web application is expected to be full-stack with front-end base on Bootstrap and HTML and back-end controlled by Python Flask which queries the data layers.

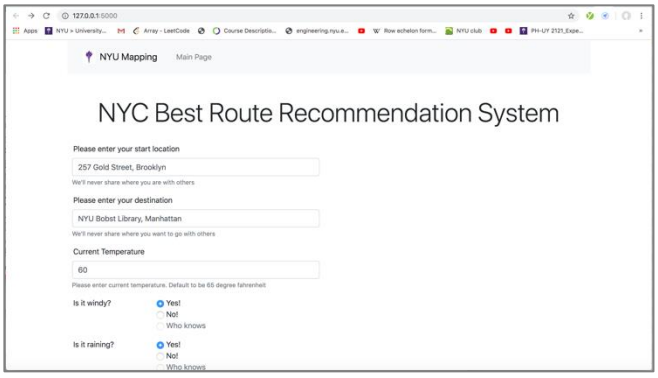

Figure 7.4 Web UI

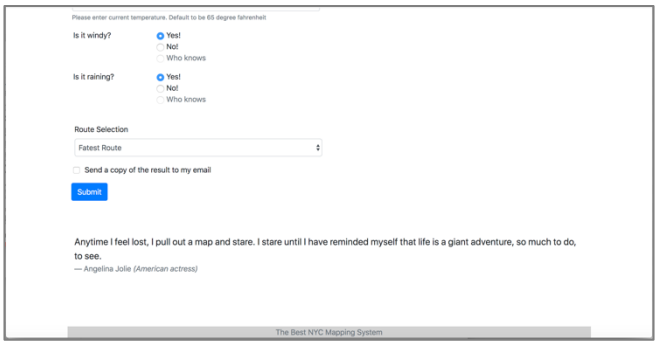

Figure 7.5 Web UI

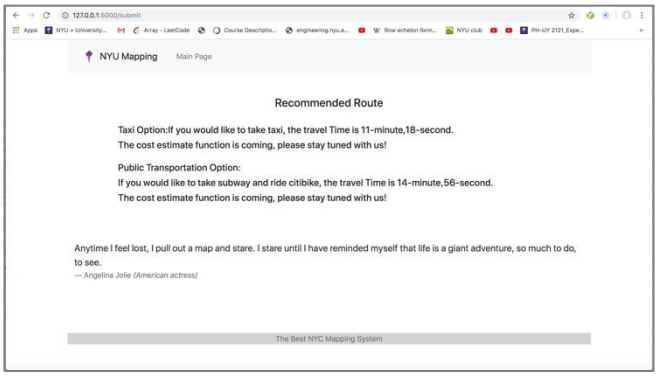

Figure 7.6 Output

The web application would allow user to input their origin location, destination location, current weather condition, etc.(Figure7.4 and Figure 7.5). After user enters all information and click on “Submit”, the web application would query the data layer, generate route recommendation results and present users with the information of the routes, as shown in Figure 7.6.

VIII. Analysis

- Experimental setup:

In this project, the analysis is mainly based on Hadoop Spark. The program is written in Scala. At the first stage, the dataset is stored in HDFS, and then Spark is used to conduct the analytics.

- Problems

Computation load is huge. There are huge number of records in citibike dataset, making the computing fairly slow. The data will cost days for the team member to run, and the computer sometimes collapses.

- Learn

Using Spark to conduct data analysis, we gained our insight into how Spark works and why it speeds up computation.

Gain insight into the challenge of big data.

- Limitations

Lack of available dataset for exploring personal preference on transportation mode choice. Dataset on personal queries on Google Map, and personal travelling chains can hardly be acquired, so personal preference for transportation can hardly be gained, which brings difficulties for the team to explore personalized user's preference and combine it to the recommendation method. Due to the time limitation, real time data are not utilized to get a more precise result.

IX. Conclusion

In this paper, a multimodal transportation recommendation system is built based on the analysis the public transportation datasets in NYC. With Spark, several transportation modes are analyzed combined with weather dataset. With the result of analysis, velocities and price of different transportation modes under different weather conditions are calculated. In the next step, a multi-objective optimization model is built and Floyd-Warshall algorithm is applied to find the optimal path.

X. Future Work

- More factors will be taken into consideration

The recommendation algorithm is designed based on the analysis on NYC open public transportation datasets, so little information about personalized preference can be found. However, personalized recommendations can be of importance in transportation recommendation system due to the personal discrepancy. Besides, factors such crime rate, carbon emission of the transportation mode and road network load balancing could also be considered into the model.

- Algorithm solving multi-objective model could be improved

In this paper, multi-objective model for finding the best recommendation is simply linearized by putting different weights on each dimension. Simply as it can be, the shortcoming lies in that it is difficult to set the weight vector in the weighted sum method. More commonly used method in solving multi-objective model is genetic algorithms.

- Improve the implement of the algorithm

Due to the limited time, the implement of the algorithm use only duration factor whiling neglecting the cost factor. In the future, more improvement on the implement of the algorithm can be finished.

- Use real-time data

Real time data can be utilized to get a more precise computation.

Acknowledgment

Thank to Professor McIntosh for providing previous advice on our project. Thanks for all the team members making effort on the project. Summer vacation is precious and doing something by effort makes it more precious.

References

1. T. White. Hadoop: The Definitive Guide. O'Reilly Media Inc., Sebastopol, CA, May 2012.
2. Liu, Y., & Wei, L. (2018, April). The optimal routes and modes selection in multimodal transportation networks based on improved A∗ algorithm. In 2018 5th International Conference on Industrial Engineering and Applications (ICIEA) (pp. 236-240). IEEE.
3. Lele Liu, Jie Liu. Study on Multimodal Transport Route Under Low Carbon Background AIP Conference Proceedings 1971 050001(2018)
4. Luo, H., Yang, J., & Nan, X. (2018, October). Path and Transport Mode Selection in Multimodal Transportation with Time Window. In 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC) (pp. 162-166). IEEE.
5. Wang Haiying,Huang Qiang,Li Chuantao, et al. Graph theory algorithm and its matlab implementation[M].Bei Jing: Beihang University press,2010:28-35.
6. Liu, Y., Chen, J., Wu, W., & Ye, J. (2019). Typical Combined Travel Mode Choice Utility Model in Multimodal Transportation Network. Sustainability, 11(2), 549.
7. Lucas, K., Phillips, I., Mulley, C., & Ma, L. (2018). Is transport poverty socially or environmentally driven? Comparing the travel behaviours of two low-income populations living in central and peripheral locations in the same city. Transportation Research Part A: Policy and Practice, 116, 622-634.
8. SI B F, YANG X B, GAO L, et al. Urban multimodal traffic assignment model based on travel demand[J]. China Journal of Highway & Transport, 2010, 23(6): 85–91.

   Urban multimodal traffic assignment model based on travel demand. China journal of Highway
9. SI B F, YANG X B, GAO L, et al. Urban multimodal traffic assignment model based on travel demand[J]. China Journal of Highway & Transport, 2010, 23(6): 85–91.
10. Liu, H., Li, T., Hu, R., Fu, Y., Gu, J., & Xiong, H. (2019). Joint Representation Learning for Multi-Modal Transportation Recommendation. AAAI, to appear.