AoCMM 2017 - #725

The 2017 Annual Association of

Computational and Mathematical Modeling Essay Sheet

Contents

AoCMM 2017 - # 725

Problem one

Integrated Cosine Similarity and Euclidean Distance Metric Learning for Identification of Typing Patterns

1. Summary

It has been shown typing patterns can be used to identify a person. If you want to determine whether the two paragraphs are the same person to type, the key is to calculate the similarity between the new sample and the original sample. To confirm this idea, we collected typing patterns from eleven of our officers using two different typing methods: fourteen short English quotes and six paragraphs composed of random characters.

Before establishing the mathematical model, we set the three hypotheses. We assume that typing habits of all persons are different and have not typing the given paragraphs before. We assume that the circumstance that all person type is the same. There are no differences about all keyboard sensitivity and the status of all persons. We assume that the typing paragraphs are independent with each other. In order to match the second and third categories to the officers ,we analyze some factors include the frequency of the character errors, the typing word per minute and the accuracy that may indentify the typing pattern. The similarity matching model is established. Matching algorithm based on cosine similarity and matching algorithm based on threshold of the euclidean distance were used to verify the model. According to the strengths and weaknesses of two algorithms, we proposed a new matching algorithm based on feature fusion. The experiment proofs that the new algorithm can make up the shortcomings of above two algorithms and get a good result. Matching Algorithm Based on Feature Fusion integrates the cosine similarity with euclidean distance. Feature vector represents the frequency of the character errors and feature vector represents the WPM and the accuracy were considered. Different feature matching methods are used for different feature vectors. This matching method is robust and effective.

For our model，the results can be affected by the following factors: The number of samples is small. It may affect the accuracy of the identification. The distance between the vectors can be calculated in a number of ways. Different ways may get different results. The weights of frequency of the character and the accuracy are different. The change of weight may affect the accuracy of matching.

2. Introduction and Problem Restatement

Question:

It has been shown typing patterns can be used to identify a person. To confirm this idea, we collected typing patterns from eleven of our officers using two different typing methods: fourteen short English quotes and six paragraphs composed of random characters. The quotes and paragraphs are grouped into three categories: eight quotes and three paragraphs where the officer's identity is known, six quotes where the officer's identity is unknown, and another three paragraphs where the officer's identity is also unknown. Based on the first category, how would one match the second and third categories to the officers .

Introduction

With the development of the Internet, e-commerce has become a new model for people to carry out business activities. Biometrics authentication method is more secure than traditional authentication methods. The application of biological behavioral characteristics is low and efficient. It can be completed through the existing equipment, such as keystroke behavior data without adding new hardware. It can conducive to the implementation and promotion of the system.

Joyce and Gupta have described an user authentication system by using keystroke. They are identifying user by comparing the keystroke latencies of a fixed string, i.e. the password, with the previously stored samples. For confirming that typing patterns can be used to identify a person, we collected typing patterns from eleven of our officers using two different typing methods. The quotes and paragraphs are grouped into three categories: eight quotes and three paragraphs where the officer's identity is known. The six quotes and three paragraphs are test set.

Based on the first category, we would establish mathematical modelling to match the second and third categories to the officers. The frequency of the character errors , the different mistakes about the character such as bad cases，bad ordering，doublet, the typing word per minute and the accuracy are the important factors to reflect the typing patterns. If you want to determine whether the two paragraphs are the same person to type, the key is to calculate the similarity between the new sample and the original sample. Each influencing factor can be quantified as a vector. The simplest, direct and efficient method of calculating the similarity between different vectors is the distance method and the value of cosine. The cosine similarity is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. Distance similarity is measured by the length of the differential. The most common distance methods are Manhattan distance, mahalanobis distance, Euclidean distance and so on. In this paper，we will use the value of cosine and the Euclidean distance as the measure of the metric. Through identifying the typing patterns, we can match the second and third categories to the officers.

3. Assumptions and Justifications

1.We assume that typing habits of all person are different and have not typing the given paragraphs before.

Justification: Our goal is to demonstrate that typing patterns can be used to identify a person. Therefore, the typing habits of all person should be different so that it could be to distinguish. Also, the persons didn't type the given paragraphs before so it reveals the real typing patterns of different person.

2.We assume that the circumstance that all person type is the same. There are no differences about all keyboard sensitivity and the status of all persons.

Justification: This paper is more concerned with the typing pattern .The keyboard sensitivity and the status of person should be ignored so that the conclusion can be meaningful.

3.We assume that the typing paragraphs are independent with each other.

Justification: Through scanning the paragraph, the test data are independent with each other so that the cross test is valid.

4. Definitions

G: a feature vector that represents the frequency of the character errors.

Cos<A,B>: the measure of similarity between the vector A and the vector B of an inner product space that measures the cosine of the angle between them.

D:The euclidean distance between the two vectors

the word per minute(W): The number of words per minute that person types

the accuracy rate (A): The wrong characters account for the total number of characters.

T=<W,A> : the number of words per minute that person types and the wrong characters account for the total number of characters form a two-dimensional vector.

T1: the matching rate according to the ranking in the algorithm 1

T2: the matching rate according to the ranking in the algorithm 2

T: the matching rate according to the ranking in the algorithm 3

5. Clear explanation of solution

we count the wrong type of characters and the number of characters in the paragraphs . The wrong type of characters include a, A, b, c, d, e, E, f, F, g, h, H, i, I, j, k, l, m, M, n, N, o, O, p, q, r, s, t, u, v, w, x, y, Y, z.

For every person, we set a feature vector that represents the frequency of the character errors.

G= < a, A, b, c, d, e, E, f, F, g, h, H，i, I, j, k, l, m, M, n, N, o, O, p, q, r, s, t, u, v, w, x, y, Y, z>

For example ,in the Given person 1,the feature vector is (2, 0, 0, 3, 2, 2, 0, 2, 0, 0, 4, 0, 5, 0, 0, 0, 3, 1, 0, 4, 0, 2, 0, 1, 0, 3, 2, 4, 0, 0, 0, 0, 1, 0, 0).

In the test quotes A person ,the feature vector is (3, 0, 0, 0, 3, 3, 0, 1, 0, 0, 3, 0, 5, 0, 0, 0, 2, 2, 0, 3, 0, 4, 0, 0, 0, 1, 6, 2, 3, 2, 1, 0, 0, 0, 0)

Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0° is 1, and it is less than 1 for any other angle. It is thus a judgment of orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1, two vectors at 90° have a similarity of 0, and two vectors diametrically opposed have a similarity of -1.

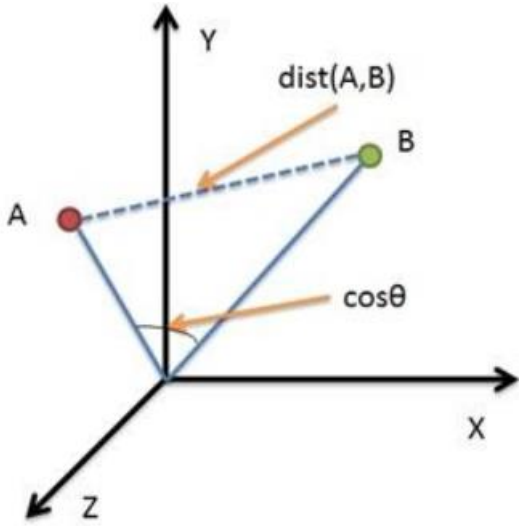

Figure 1 Cosine similarity

5.1 Algorithm 1: Matching Algorithm Based on Cosine Similarity

Cosine similarity (CS) between two vectors x and y is defined as:

$$\text{Similarity}=\cos(\theta)=\frac{A\square B}{\|A\|_2\|B\|_2}=\frac{\sum_{i=1}A_iB_i}{\sqrt{\sum_{i=1}^{n}A_i^2}\sqrt{\sum_{i=1}^{n}B_i^2}} \quad (1)$$

Where A_i and B_i are components of vector A and B respectively.

Cosine similarity has a special property that makes it suitable for metric learning: the resulting similarity measure is always within the range of -1 and +1

.

Strengths

Feature vector represents the frequency of the character errors has 35 dimensions. The more uniform of direction of two vectors, the more similiarity of two persons. The cosine distance uses the cosine of the angle between the two vectors as a measure of the size of the differences between the two individuals. Compared to Euclidean distance, cosine distance pays more attention to the difference between the two vectors in the direction.

Weaknesses

The number of words per minute and the accuracy rate were not considered but they also reflect the typing patterns. They have absolute numerical values. Cosine distance reflects the distinction between the directions but it is not sensitive about the numerical values.

If you want to determine whether the two paragraphs are the same person to type, the key is to calculate the similarity between the new sample and the original sample. If the similarity is higher than a certain threshold, we consider that two sample are typed by the same person. The word per minute and the accuracy rate can make up a feature vector. The simplest, direct and efficient method of calculating the similarity between different vectors is the distance method, and the similarity is measured by the length of the distance. At present the Manhattan distance, Mahalanobu distance and Euclidean distance are used in the industry and academia. In this study, we use the Euclidean distance to calculate the similarity.

We assume that there are two vectors A (x1, y1), B (x2, y2).Manhattan distance, Euclidean distance, Mahalanobis distance calculation method can be expressed as follows:

Manhattan Distance

$$d=|x_1-x_2|+|y_1-y_2| \quad (2)$$

Euclidean Distance:

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \qquad (3)$$

Mahalanobis Distance

$$d = \sqrt{(A-B)^T S^{-1}(A-B)} \qquad (4)$$

S is covariance matrix

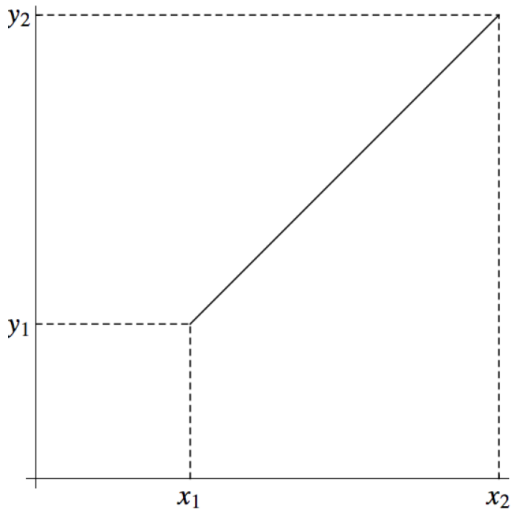

Figure 2 Euclidean distance in R2

5.2 Algorithm 2: Matching Algorithm Based on Threshold of the Euclidean Distance

Input: A person's WPM and Accuracy about short English quotes and random characters (T).For every person ,Given that contains k paragraphs was Training Data, Test that contains 11 persons about quotes and 10 persons about letters was test data.

Step1:For T1,T2,…,T11, (K=11,for person 1,2,3,5,6,7,8,9,10,11;K=8,for person4)Calculate the distance between two vectors and then get the distance array d=< d1,d2,…,dm>, m=C_k^2

Step2: Sort the distance array d from large to small and get a descending sequence B.(B=<b1,b2,…,bm> ； b1< b2<…bm) and select 85% of the quantile as a threshold L.

Step3：For test data, we calculate the average of each paragraph's typing speed and accuracy as a vector that represents the person.

Step4: For Test data Tk+1, Tk+2, … Tk+n (n=11 or 10), calculate the distance with the historical data respectively and get the distance array D=< D1,D2,…,D11> or D=< D1,D2,…,D10>

Step5: If the calculated distance D1 or D2 or…Dn is less than the threshold, maybe that person is considered to be the same person.

Step6: Sort all possible persons according to the value of distance D. The smaller

value of D, the highest possible matching of the person.

Output: For every data to be matched, we rank person according to the matching similarity.

Results:

Letter
Q - 9
R - 3
S - 3
T - 8
U - 7
V - 4
W - 9
X - 5
Y – 9
Z - 10

Quote：
A - 4
B - 6
C- 1
D - 6
E - 4
F - 4
G - 1
H - 1
I - 1
J - 7
K – 1

Strengths

The word per minute and the accuracy are the two factors that can reflect the typing patterns. Different person has the differences in the numerical size of two dimensions that contain word per minute and the accuracy. Euclidean distance can reflect the absolute difference of individual numerical features.

Weaknesses

The Euclidean metric needs to ensure that the dimensions are at the same scale level because it bases on the absolute value of each dimension feature. The scale level of word per minute and the accuracy are different. So the use of Euclidean metric for distinguishing typing patterns may invalidate the results.

Fusing the feature vector that represents the frequency of the character errors and the feature vector that represents person's WPM and accuracy, it can provide more comprehensive information about typing pattern.

5.3 Algorithm 3 Matching Algorithm Based on Integrated Cosine Similarity and Euclidean Distance

According to the ranking in the algorithm 1, every person has a matching rate (R1). According to the ranking in the algorithm 2, every person has a matching rate (R2). There are 11 persons. So the value of matching rate was set to be 90% or 81%,

72% or 63% or 54% or 45% or 36% or 27% or 18% or 9% or 0% in order.

The final matching rate R= a* R1+b* R2. （5）

In this algorithm, we set a=1, b=1.

The person who gets the highest matching rate was considered as the right matched person.

Results:

Letter
Q - 7
R - 6
S - 11
T - 2
U - 9
V - 3
W - 5
X - 1
Y - 8
Z - 10

Quote
A - 4
B - 8
C- 1
D - 10
E - 4
F - 8
G - 4
H - 8
I - 1
J - 2
K - 1

Strengths

Matching Algorithm Based on Feature Fusion integrates the cosine similarity with euclidean distance. Feature vector represents the frequency of the character errors and feature vector represents the WPM and the accuracy were considered. Different feature matching methods are used for different feature vectors. This matching method are robust and effective.

Weaknesses

Training set and test set are too small. It may affects the effectiveness of the matching method.

6. Sensitivity Analysis

For our model，the results can be affected by the following factors:

The number of samples is small. It may affect the accuracy of the identification.

In this experiment, the similarity between the typing patterns is measured by the distance between the vectors and the distance between the vectors can be calculated in a number of ways. In this study, we use the Euclidean distance. If we use the Manhattan distance, and Mahalanobis distance, the results may have some differences.

The use of matching algorithm based on cosine similarity or matching algorithm based on euclidean distance alone cannot identify a person effectively. The distinction is small. The matching algorithm based on feature fusion takes into account all possible factors that reflect the typing patterns. The matching of the second and third categories to the officers is good. It is a one-to-one correspondence.

We set the weight of frequency of the character errors to be a and set the weight of accuracy and word per minute to be b.In the algorithm 3, we set a and b to be 1. Actually, the weights of frequency of the character and the accuracy are different. The change of weight may affect the accuracy of matching.

7. Conclusion

In order to match the second and third categories to the officers ,we analyze some factors include the frequency of the character errors, the typing word per minute and the accuracy that may indentify the typing pattern. The similarity matching model is established. Matching algorithm based on cosine similarity and matching algorithm based on threshold of the euclidean distance were used to verify the model. According to the strengths and weaknesses of two algorithms, we proposed a new matching algorithm based on feature fusion. The experiment proofs that the new algorithm can make up the shortcomings of above two algorithms and get a good result.

8. Citations

Song D, Venable P, Perrig A. User Recognition by Keystroke Latency Pattern Analysis[J]. Retrieved on, 1997.

http://www.cnblogs.com/chaosimple/archive/2013/06/28/3160839.html

https://en.wikipedia.org/wiki/Euclidean_distance

https://en.wikipedia.org/wiki/Cosine_similarity

http://www.cnblogs.com/heaad/archive/2011/03/08/1977733.html

http://www.ruanyifeng.com/blog/2013/03/cosine_similarity.html

https://en.wikipedia.org/wiki/Euclidean_distance

AoCMM 2017 - # 725

Problem two

Best strategy for taxi driver and the head of the taxi company

1. Summary

It is concluded that taxi drivers and taxi company managers think differently. The former needs to consider how much revenue the day will be, while the latter will need to consider how to carry out vehicle scheduling to get more passengers. For the former, we use mathematical expectations to make decisions, and decide which direction to go by comparing which direction of income is higher. And for the latter, what we're going to do is figure out how much of a region's demand for taxis to decide whether need to dispatch taxis to the area.

We build a mathematical model whose core is the expected value of total_amount for taxi drivers, which is used as a standard to make decisions. In the solving process, we will simplify the total_amount expected formula, until reduced to the most basic form, and then continuously optimize the parameters in the formula on the basis of the known data samples. Finally we get the result.

We set up another set of mathematical model for taxi company managers. We obtain the relative quantity of the demand for taxis in the region by calculating the difference between the number of times of the taxi in a certain area and the number of times of the car. The smaller difference indicates that the demand for taxis is greater, and as a criterion, we can come up with a scheduling strategy.

2. Introduction and Problem Restatement

There are a lot of taxis in NYC, but drivers can easily blindly wander aimlessly, resulting in wasted hours. According to common sense, we will choose the area where are more people to wait for passengers because it is easier to get passengers. But there are a few questions: first, where there are many people, must there be many passengers? Second, where there are so many people, is the number of mileages for the passenger bigger? We all know the longer the ride, the more the taxi driver earns. Third, where are the potential passengers? We have the idea that when our taxi is empty, we should go downtown to get more passengers. But that doesn't necessarily make sense, at least the data doesn't tell us directly that we need to compute the results. Only the results tell us which is reasonable and true, and all speculation is groundless.

We need to decide for two identities both taxi drivers and the taxi companies, and the decisions must be different. According to the New York taxi management mode, the taxi driver only need to pay a fixed to the taxi company rent every day, to be sure that they have a taxi driving license, so taxi companies get income from taxi drivers are fixed. All daily income of the taxi driver is the fare passengers pay. As the result the purpose of the taxi driver should be effectively drive more mileage, namely passengers mileage. The distance directly influences the fare.

Taxi companies aim to get more passengers on their cabs, even if each passenger has a short range of miles to drive, but to pick up as many passengers as possible.

So when we are taxi drivers, we should think about how we can make more money by going to places where we can make more money. And when we work as a taxi company header, we should be able to dispatch more taxis to those areas where there is a shortage of taxis and sufficient demand.

3. Assumptions and Justifications

Assumption one : There is no need to consider the oil cost

Justifications: The current taxis in New York are commonly used in Ford, Vitoria crown and Nissan NV200, with an exhaust capacity of 1.6L and fuel consumption of 8L/100km, while the average oil price is 0.75USD/L, and the cost of oil per kilometer is

$$0.75 * 8 / 100 = 0.06 \text{ USD}$$

According to the known data, we can calculate the average income per kilometer of $3.98, so we found that, 0.06<<3.98, we can ignore the cost of oil. It can be not taken into consideration in the decision making.

Assumption two : There is no need to consider the impact of payment methods

Justifications: According to known data, the use of the third, fourth, fifth and sixth payment methods only accounted for less than 0.1%, so it can be considered that cash and credit cards are the default means of payment for passengers.

Assumption there : The manager of the taxi company need to consider the number of businesses instead of the single business income

Justification: Based on the taxi operating mechanism of New York, taxi drivers need to pay a fixed amount of company license rental fee without part of the income. When there is a fixed number of taxi drivers, the taxi company gets a fixed income from the taxi drivers. So the manager of the taxi company needs to consider how to compete with other taxi companies, especially competition with internet car-hailing, such as Uber. Make sure that the company's taxi appear in everywhere there are taxi passengers to maximize the company's taxi market.

Assumption four : There is no need to consider the effect of rate_code on the taxi fares

Justifications： According to the data, it only accounts for about 0.03% when the rate_code is not standard, which is not enough to affect the overall amount of revenue. So it can be not taken into consideration in the decision making.

4. Clear explanation of solution

4.1 Model Introduction

According to the explanation of the problem, we can consider this problem as a decision problem, that is, to find the most favorable solution from several alternatives. This problem has two questions, that is to say, we have two roles: taxi drivers and the manager of the taxi company. For the different roles, we have different judges on the advantages and disadvantages.

For the role of taxi drivers in Question 1, we should pass judgment on the personal income. The taxi drivers should go to the direction which can earn more money. And we can use the mathematical expectation of an area's income as a judge of the area's income level. According to the known data, we can get the mathematical expectation of each area. Then the taxi drivers can choose the area with the highest expectations.

And when we turn to the role of the manager of the taxi company, we should pay more attention to ensure that more customers can get the service of our taxi company and obtain greater market share to win the competition with other taxi companies. Therefore we should pass judgment on taxi scheduling. So we should put more taxi into the area with more taxi demand.

4.2 Solution of Q1

4.2.1 Regional Classification

So the first step, we're going to divide New York into a couple of regions, so that we can use the k-means algorithm to plot the region as the direction of the decision. We use the k-means algorithm to divide the region.

K-means algorithm is one of the top ten data mining algorithms that is the most classical clustering method based on division. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. Through the iterative method, the values of each cluster center are updated, until the best clustering results are obtained.

We take the latitude and longitude coordinates of all the data as the sample set, and divide the sample set into c categories. The algorithm process is as follows:

Select the initial center of class c appropriately;

In the kth iteration, compute the distance to each center for each sample and classify it into the center of the shortest distance;

Updated the central value of each class by means of means;

For all c cluster centers, stop the iteration if the value of center remain the same after step 2 and step 3,otherwise continue step 2.

After several iterations, we can divide all of the boarding points into seven areas as seen in the graph below.

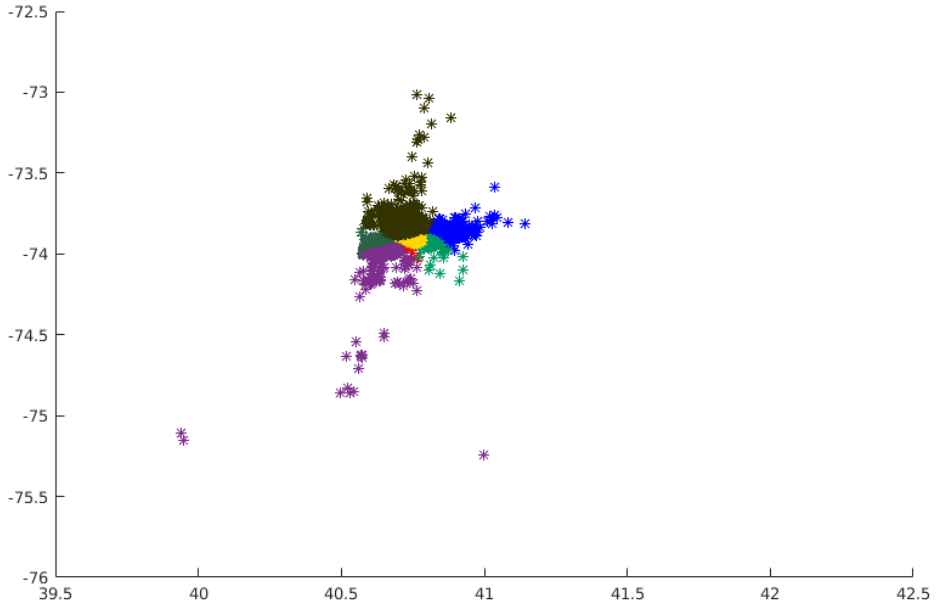

Figure 1 result of K-means algorithm

We use these seven areas as the driving directions for our taxi and the corresponding central coordinate points are shown below.

Table 1 Regional Classification

ID	colour	Longitude	Latitude
C1	black	-73.95250569	40.71975764
C2	blue	-73.89979545	40.85066373
C3	dark green	-73.94525171	40.66754154
C4	light green	-73.94644765	40.80735845
C5	red	-73.90477420	40.75483281
C6	yellow	-73.98502071	40.68485142
C7	purple	-73.84140209	40.72484642

4.2.2 Time division

According to common sense and data, we all know that there is a difference between the number of cabs at different times of the day and the general direction of traffic flow due to the intuitive impact of the moment. And all decisions at different times are not necessarily the same or even the opposite.

We also found that working days and rest days have a big impact on the taxi situation. Therefore, we divide all data into several categories according to working days and rest days and different time periods.

The average situation of the taxi ride at different times in weekdays and weekends is shown in table below.

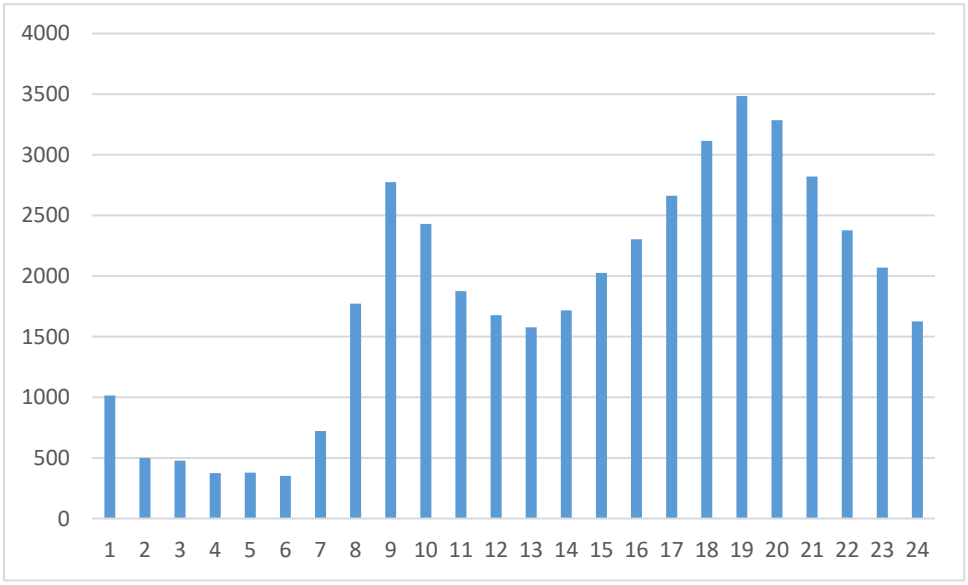

Figure 2 average situation of the taxi ride at different times in weekdays

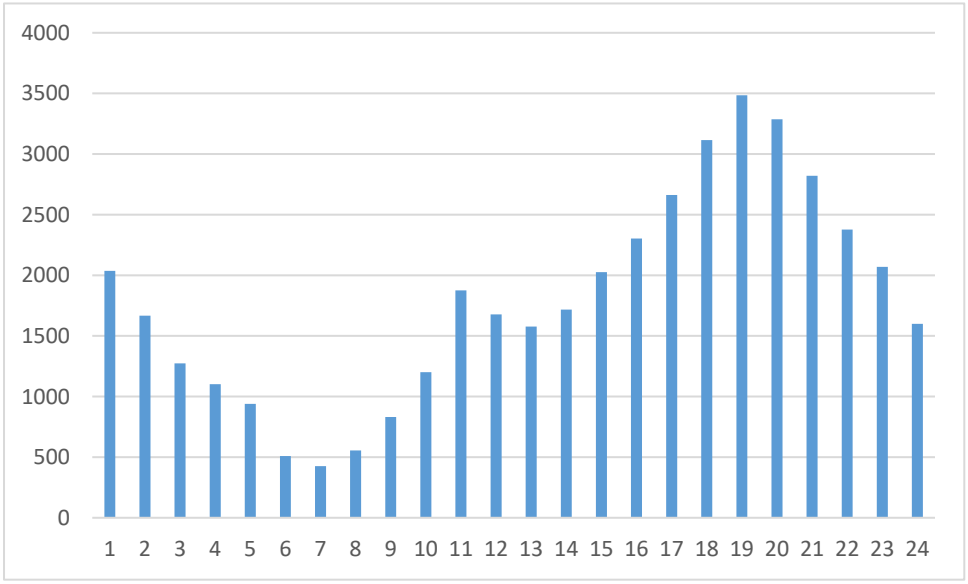

Figure 3 average situation of the taxi ride at different times in weekends

As a result we divide the time periods as shown in the table and make decisions at each time.

We divide the time into four parts like these: 0:00-7:59, 8:00-12:59, 13:00-19:59, 20:00-23:59.

4.2.3 Introduction of Model

According to the first two sections, we divided New York City into seven areas according to the location of the car, and prepared to make corresponding decisions and judgments in accordance with the time period. In this section we will set up a judging standard to decide as a taxi driver, which direction should I go?

As a taxi driver, the cost per day that I should cover includes cost of time, licence fee and oil. The first two are basically fixed, and the oil fee is not a factor affecting the decision making. Therefore, I only need to consider how to get a bigger income. The corresponding data item for income is total_amount, and we want to get the expected value of total_amount.

The total_amount value in our model is discrete, so the formula for it is as fellow:

$$E（\text{total\_amount}) = \sum_{x \in X} xP(x)$$

x represents the total_amount for each trip,X is the total set of samples,P(x) is the probability of the x happening. According to the above formula, we can find the expectation of total_amount in each region and make decisions by comparing expectations.

Next, we analyze that how to calculate E(total_amount). From the data given, we can know that total_amount consists of three parts: the fare- amount, fixed- fee and tip-amount. The fixed-fee includes three parts: extra, MTA_tax and improvement_surcharge, but their addition are fixed to $1.3. So we can consider it as a constant——fixed-fee. Accordingly, we obtain the following formula to calculate total- amount.

$$total\_amount = fare\_amount + fixed\_fee + tip\_amount$$

According to the expected calculation property, we can derive the following formula to calculate the expectation of total- amount.

$$E(total\_amount) = E(fare\_amount) + E(fixed\_fee) + E(tip\_amount)$$

E(fixed_fee) equals 1.3. So we turn to compute E(fare_amount)+E(tip_amount) ——the sum of the expectations of the fare_amount and tip_amount.

We observe the value of fare_amount, which is influenced by the path length, payment method, and rate_code. However, because of assumption 3, we think that rate_code is not an influence factor, so we don't consider this factor. Since we assume 2, we only consider the cash and credit card in the payment method as two payment methods. Therefore, the value of the fare_amount is related to the length of the ride and the means of payment.So we set up the following relationship.

$$fare\_amount = trip\_distence * price(payment\_type)$$

Because trip_distance and price are unrelated, the expectation of fare_amount can be expressed as follows:

$$E(fare\_amount) = E(trip\_distence) * E(price(payment\_type))$$

payment_type represents the impact coefficient of each payment method on unit price. We analyze the impact factors of the two payment methods by calculating the average price of different payment methods by means of all trips, and we can obtain the influencing factors. The result is shown as tables below.

Table 2 fare_amount in each payment type

payment method	Average trip_distance	Average fare_amount	Average price
Credit card	1.36	13.10	9.61
Cash	1.37	10.47	7.65

As you can see from the table above, it is slightly more expensive to take cash

than credit card. The approximate price is 9.61/7.65=1.25. It can be concluded that the amount of cash paid is 1.25 times the amount of credit card.

The expectation of price can be expressed as:

$$E(price(payment)) = price' * E(payment)$$

The price represents the base fee which is the amount of cash that is charged. Payment is medium of payment. We consider two types ——cash and credit card, so the payment conforms to a model similar to the 0-1 distribution. According to the previous ratio, we have the expectation of payment:

$$E(payment) = 1 * p(cash) + 1.25 * p(credit\_card)$$

p(cash) represents the probability of using cash. p(credit_card) is the probability of using credit card. We can simplify this to:

$$E(price(payment)) = price' * (1 + 0.25 * p(credit\_card))$$

Then simplified expected formula of fare_amount is:

$$E(fare\_amount) = E(trip\_distence * price') * (1 + 0.25 * p(credit\_card))$$

The trip_distance*price can be simplified as fare_cash meaning the cost of a cash payment. So the further simplification is as follows:

$$E(fare\_amount) = E(\text{fare\_cash}) * (1 + 0.25 * p(credit\_card))$$

Fare_cash is a discrete random variable. So the expected value of it is the average,the expected formula is:

$$E = \sum fare\_cash / n$$

n represents the number of data. In conclusion, the expected formula for the final fare_amount is:

$$E(fare\_amount) = \sum fare\_cash / n * (1 + 0.25 * p(credit\_card))$$

We continue to look at tip_amount, which is a discrete random variable, so the formula for its expected value is:

$$E(tip\_amount) = \sum (tip * p(tip)) / n$$

Then we get the final result of E(total_amount):

$$E(total\_amount) = \sum fare\_cash / n * (1 + 0.25 * p(credit\_card)) + \sum tip / n * p(tip) + 1.3$$

4.2.4 Conclusion

When it is weekdays.

Table 3 E(fare_amount) in weekdays

region	0:00-7:59	8:00-12:59	13:00-19:59	20:00-23:59
1	12.33	12.11	11.66	10.41
2	13.08	11.93	12	16.7
3	13.79	14.82	14.68	12.54
4	10.88	10.93	10.75	10.17
5	9.78	10.63	9.93	9.14
6	14.39	13.25	12	11.45
7	12.69	15.12	12.84	11.02

Table 4 E(tip_amount) in weekdays

region	0:00-7:59	8:00-12:59	13:00-19:59	20:00-23:59
1	1.58	1.64	1.46	1.5
2	0.59	0.57	0.54	0.76
3	1.72	1.42	1.78	1.57
4	1.03	1.06	0.99	0.96
5	0.69	0.84	0.73	0.74
6	2.01	1.83	1.47	1.54
7	0.78	1.15	1.05	0.88

Table 5 p(credit_card) and P(tip) in weekdays

region	p(credit_card)	1+0.25*p(credit_card)	P(tip)
1	0.58	1.14	0.52
2	0.30	1.08	0.18
3	0.50	1.13	0.40
4	0.44	1.11	0.36
5	0.32	1.08	0.28
6	0.57	1.14	0.52
7	0.37	1.09	0.30

Table 6 E(total_amount) in weekdays

region	0:00-7:59	8:00-12:59	13:00-19:59	20:00-23:59
1	16.18	15.96	15.35	13.95
2	15.53	14.29	14.36	19.47
3	17.57	18.61	18.60	16.10
4	13.75	13.81	13.59	12.93
5	12.06	13.02	12.23	11.38
6	18.75	17.36	15.74	15.15
7	15.37	18.13	15.61	13.58

Then we know from the list , when our car is vacant in weekdays, we should

drive to the region 3/6 in 0:00-7:59; we should drive to the region 3/6/7 in 8:00-12:59; we should drive to the region 3 in 13:00-19:59; we should drive to the region 2 in 20:00-23:59.

When it is weekends

Table 7 E(fare_amount) in weekends

region	0:00-7:59	8:00-12:59	13:00-19:59	20:00-23:59
1	12.65	14.07	12.86	11.66
2	12.7	14.41	13.53	12.21
3	13.76	13.81	14.46	13.48
4	11.43	10.31	10.74	9.87
5	10.89	11.72	10.64	9.57
6	13.73	13.78	12.15	12.02
7	11.96	16.53	13.92	13.21

Table 8 E(tip_amount) in weekends

region	0:00-7:59	8:00-12:59	13:00-19:59	20:00-23:59
1	1.58	1.64	1.46	1.5
2	0.59	0.57	0.54	0.76
3	1.72	1.42	1.78	1.57
4	1.03	1.06	0.99	0.96
5	0.69	0.84	0.73	0.74
6	2.01	1.83	1.47	1.54
7	0.78	1.15	1.05	0.88

Table 9 p(credit_card) and P(tip) in weekends

region	p(credit_card)	1+0.25*p(credit_card)	P(tip)
1	0.61	1.15	0.53
2	0.34	1.09	0.19
3	0.50	1.13	0.38
4	0.43	1.11	0.34
5	0.30	1.08	0.27
6	0.55	1.14	0.47
7	0.34	1.08	0.27

Table 10 E(total_amount) in weekends

region	0:00-7:59	8:00-12:59	13:00-19:59	20:00-23:59
1	16.71	18.51	16.95	15.48
2	15.27	17.18	16.21	14.70
3	17.45	17.35	18.32	17.18
4	14.27	13.06	13.55	12.56
5	13.26	14.20	13.00	11.82
6	17.69	17.69	15.81	15.67
7	14.48	19.40	16.58	15.86

Then we know from the list , when our car is vacant in weekends, we should

drive to the region 3/6 in 0:00-7:59; we should drive to the region 1/2/3/6/7 in 8:00-12:59; we should drive to the region 3 in 13:00-19:59; we should drive to the region 3 in 20:00-23:59.

4.3 Solution and conclusion of Q2

When we are the header of the taxi company, we decide the strategy by the difference value of the pick-up number and drop-off number in every region. Then we can easily get which region needs more taxi. We compute the average number of pick-up and drop-off everyday, and then compute the difference value of them. The minimize is what we want. We then send more taxis to those regions whose difference values are smaller.

When it is weekdays

Table 11 average value of pick-up in weekdays

	0:00-7:59	8:00-12:59	13:00-19:59	20:00-23:59
1	795	778	1360	1241
2	365	688	1121	337
3	557	739	1347	804
4	1490	3727	5984	1898
5	1206	1742	3167	1701
6	746	1985	3717	1911
7	432	676	1898	1002

Table 12 average value of drop-off in weekdays

	0:00-7:59	8:00-12:59	13:00-19:59	20:00-23:59
1	660	587	1041	1025
2	363	564	904	289
3	509	566	1028	651
4	1596	3032	4921	1638
5	1217	1399	2586	1500
6	737	1518	2893	1591
7	406	516	1472	868

Table 13 difference value of pick-up and drop-off in weekdays

	0:00-7:59	8:00-12:59	13:00-19:59	20:00-23:59
1	-135	-191	-319	-216
2	-2	-124	-217	-48
3	-48	-173	-319	-153
4	106	-695	-1063	-260
5	11	-343	-581	-201
6	-9	-467	-824	-320
7	-26	-160	-426	-134

Then we know from the list , when our car is vacant in weekdays, we should send more taxis to the region 1 in 0:00-7:59; we should send more taxis to the region

4/6 in 8:00-12:59; we should send more taxis to the region 4/5/6/7 in 13:00-19:59; we should send more taxis to the region 4/6 in 20:00-23:59.

When it is weekends

Table 14 average value of pick-up in weekends

	0:00-7:59	8:00-12:59	13:00-19:59	20:00-23:59
1	2202	482	1340	1246
2	447	465	1111	334
3	804	560	1234	810
4	1523	2634	6003	1879
5	2040	1140	3522	1820
6	848	1313	3601	2032
7	645	568	1912	1300

Table 15 average value of pick-up in weekends

	0:00-7:59	8:00-12:59	13:00-19:59	20:00-23:59
1	1737	383	102	1320
2	377	366	920	423
3	653	434	1024	830
4	1294	2140	4933	1930
5	1738	937	2582	1736
6	699	1018	2585	1920
7	554	437	1520	1054

Table 16 difference value of pick-up and drop-off in weekends

	0:00-7:59	8:00-12:59	13:00-19:59	20:00-23:59
1	-465	-99	-438	-126
2	-70	-99	-191	-111
3	-151	-126	-210	-180
4	-229	-494	-1070	-149
5	-302	-203	-940	-284
6	-149	-295	-1016	-312
7	-91	-131	-392	-446

Then we know from the list , when our car is vacant in weekdays, we should send more taxis to the region 1 in 0:00-7:59; we should send more taxis to the region 4 in 8:00-12:59; we should send more taxis to the region 4/5/6 in 13:00-19:59; we should send more taxis to the region 5/6/7 in 20:00-23:59.

5. Sensitivity Analysis

Our model does not take some special cases into consideration, such as the third, fourth, fifth and sixth payment methods. And we only consider the situation when the rate code is standard. Only in very few cases the income of taxi drivers will be affected by these conditions. We divided the day into four periods according to the taxi flow, and the edge of each period may be affected by the previous period. In January 1, 2016, there are so many numbers of taking a taxi, much different from the other days. We think it is due to New Year's Day, so we don't take it as a sample data and we think the special holidays will also affect our results. At last, our model does not take the impact of weather or major events into consideration.

6. Citations

Julia Stoyanovich,Matthew Gilbride, Vera Zaychik Mott. Zooming in on NYC taxi data with Portal[J]. Retrieved on, 2017.

https://www.linkedin.com/pulse/nyc-green-cab-ridership-consumer-data-analysis-insights-chen-liu/

http://blog.knowi.com/2016/09/analyzing-12-billion-nyc-taxi-rides_26.html

https://techxplore.com/news/2017-01-mathematical-taxi-manhattan.html

https://www.ocf.berkeley.edu/~dlevitt/2015/12/13/final-project-nyc-taxi-and-uber-data/

Appendix of problem one

G: a feature vector that represents the frequency of the character errors.

T=<W,A> : the number of words per minute that person types and the wrong characters account for the total number of characters form a two-dimensional vector.

The vector G and the vector T of all persons were shown below：

Given:

G_{QPk} represents a feature vector the frequency of the character errors about Quotes

(k=0,1,…,11)

G_{LPk} represents a feature vector the frequency of the character errors about Letters (k=0,1,…,11)

G_{QP1}=(2， 0， 0, 3, 2, 2, 0, 2, 0, 0, 4, 0, 5, 0, 0, 0, 3, 1, 0, 4，0,，2, 0, 1, 0, 3, 2, 4, 0, 0, 0, 0, 1, 0, 0)

G_{LP1}=(0， 0，0, 0, 0, 2, 0, 1, 0, 0, 0, 0, 2, 0, 0, 2, 0, 0, 0, 0，0, 0, 0, 0, 0, 2, 0, 3, 0, 1, 0, 0, 0, 1, 0, 0)

p2

G_{QP2}=(14, 0, 1, 0, 5, 17, 0, 4, 0, 0, 7, 0, 6, 0, 1, 1, 7, 4, 0, 7, 0, 10, 0, 1, 0, 5, 9, 24, 6, 1, 3, 0, 5, 0, 0,)

G_{LP2}=(1, 0, 0, 1, 3, 0, 3, 0, 1, 1, 0, 4, 0, 0, 1, 1, 0, 0, 0, 0, 3, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0)

p3

G_{QP3}=(5, 0, 0, 1, 1, 3, 0 2, 0, 0, 7, 0, 2, 0, 2, 0, 4, 2, 0, 5, 0, 4, 0, 1, 2, 4, 4, 1, 7, 4, 5, 0, 2, 0, 0)

G_{LP3}=(0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)

p4

G_{QP4}=(3, 0 0, 1, 1, 3, 0, 1, 0, 1, 4, 0, 6, 0, 0, 1, 5, 2, 0, 5, 0, 1, 0, 1, 1, 1, 3, 5, 1, 1, 0, 0, 1, 0, 0)

G_{LP4}=(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)

p5

G_{QP5}=(2, 0, 0, 1, 1, 6, 0, 1, 0, 15, 1, 0, 4 0, 0, 0, 2,

2, 1, 3, 0, 3, 0, 0, 0, 2, 4, 5 1, 0, 0, 0, 2, 0, 0)
G_{LP5}=(10, 0 0, 0, 2, 0, 2, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,
0, 0, 0, 2, 0, 0, 2, 1, 2, 0, 0, 0, 0, 0, 0, 0, 0,)

p6
G_{QP6}=(6 0, 2, 1, 0, 7, 0, 4, 0, 1 4, 0, 3, 0, 0, 0, 3,
4, 0, 3, 0, 6, 0, 0, 0, 3, 5, 5, 3, 1, 3, 0, 1, 0, 0)
G_{LP6}=(10 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0,
0, 0, 0, 0, 0, 0, 2, 0, 2, 0 0, 0 2, 0, 0, 0, 0)

p7
G_{QP7}=(14,0, 0, 2, 2, 10, 0, 0, 0, 0, 4, 0, 3, 0, 0, 2, 2, 3,
0, 4, 0, 7, 0, 1, 0, 3, 4, 8, 3, 0, 3 0, 2, 0, 0)
G_{LP7}=(1, 0 0 0 0 3 0 2 0 0, 0, 0, 0, 0, 0, 0, 1, 2,
0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 1, 0, 2, 0, 4, 0, 0, 0, 0)

p8
G_{QP8}=(4, 0, 0, 0, 2, 7, 0, 1, 0, 1, 1, 0, 4, 0, 0, 0, 2,
6, 0, 3, 0, 2, 0, 1, 0, 4, 7, 8, 2, 0, 4, 0, 0, 1, 0)
G_{LP8}=(2, 0, 0, 0, 1, 4, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 2,
0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 3, 0, 2, 0, 0, 0, 0)

p9
G_{QP9}=(9, 0, 1, 5, 1, 11, 0, 5, 1, 4, 12, 1, 12, 1, 0, 0, 9,
2, 0, 8, 1, 8, 1, 1, 0, 9, 11, 21, 4, 2, 3, 2, 4, 0, 1)
G_{LP9}=(0, 0, 0, 0, 3, 2, 0, 0, 0, 0, 0, 0, 1, 0, 1, 2, 0,
0 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 2, 0 0, 0, 0)
p10
G_{QP10}=(6, 1, 3, 3, 6, 10, 0, 1, 0, 1, 6, 0, 9, 0, 0, 1, 4,
0, 0, 5, 1, 5, 0, 1, 5, 7, 6, 5, 5, 1, 3, 0, 0, 0, 0)
G_{LP10}=(0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1,
0 0, 0, 0, 2 0, 2, 2 0, 2, 0, 0, 0, 2, 0, 0, 0, 0)
,
p11
G_{QP11}=(6, 0, 3, 2, 6, 8, 1, 2, 0, 3, 3, 0, 4, 0, 0, 1, 4,
2, 0, 7, 0, 5, 0, 1, 1, 7, 3, 5, 3, 2, 1, 0, 0, 0, 0)
G_{LP11}=(1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1,
0, 0, 0, 0, 1, 0, 1, 3, 2, 0, 0, 1, 0, 0, 0, 0, 0, 0)

Test: Quotes

G_{Qk} represents a feature vector the frequency of the character errors about Quotes

(k=0,1,…,11)

A

G_{QA}=(3, 0, 0, 0, 3, 3, 0, 1, 0, 0, 3, 0, 5, 0, 0, 0, 2, 2, 0, 3, 0, 4, 0, 0, 0, 1, 6, 2, 3, 2, 1, 0, 0, 0, 0)

B

G_{QB}=(4, 0, 1, 0, 3, 8, 0, 2, 0, 1, 3, 0, 4, 0, 1, 1, 6, 0, 0, 1, 0, 5, 0, 1, 0, 1, 1, 6, 2, 1, 4, 0, 1, 0, 0)

C

G_{QC}=(3, 1, 2, 1, 0, 3, 2, 0, 0, 1, 4, 0, 4, 0, 1, 1, 1, 1, 0, 2, 0, 2, 0，1, 0, 5, 2, 1, 1, 0, 1, 0, 2, 0, 0)

D

G_{QD}=(1, 1, 0, 0, 3, 11, 0, 5, 0, 1, 7, 0, 10, 0, 0, 3, 4, 9, 1, 5, 0, 12, 2, 1, 0, 6, 3, 8, 9, 1, 1, 0, 2, 1, 0)

E

G_{QE}=(8, 0, 2, 1, 0, 5, 0, 0, 0, 1, 1, 0, 2, 0, 0, 0, 4, 4, 0, 3, 0, 2, 0, 0, 0, 0, 3, 5, 0, 1, 1, 0, 1, 0, 0)

F

G_{QF}=(5,0, 0, 2, 1, 2, 0, 1, 0, 0, 3, 0, 4, 0, 0, 2, 6, 1, 0, 3, 1, 5, 0, 0, 0, 0, 3, 6 1, 1, 3, 0, 1, 0, 0)

G

G_{QG}=(3, 1, 2, 1, 1, 6, 0, 1, 0, 2, 1, 0, 6, 0, 0, 0, 1, 0, 0, 4, 0, 4, 0, 0, 0, 2, 3, 1, 1, 1, 2, 0, 1, 0, 0)

H

G_{QH}=(4, 0, 0, 0, 2, 6, 0, 0, 0, 2, 2, 0, 5, 0, 0, 0, 1, 0, 0, 3, 0, 3, 0, 1, 0, 2, 4, 0, 0, 2, 2, 0, 0, 0, 0)

I

G_{QI}=(2, 0, 1, 1, 0, 2, 0, 0, 0, 0, 4, 0, 3 0, 0, 0, 1, 1, 0, 4, 0, 1, 1, 0, 0, 0, 3, 1, 1, 2, 3, 0 1, 0, 0)

J

G_{QJ}=(8, 2, 0, 0, 5, 15, 0, 0, 0, 5, 8, 0, 6, 0, 0, 1, 9, 2, 0, 4, 0, 14, 0, 3, 0, 4, 3, 12, 5, 5, 4, 0, 1, 0, 0)

K

G_{QK}=(3, 0, 1, 2, 1, 2 3, 0, 0, 1, 3, 0, 2, 0, 2, 0, 0,

0, 0, 0, 0, 0, 1, 2, 0 2, 0, 1, 0, 0, 1, 0, 0, 0, 0)

Test:L

G_{Lk} represents a feature vector the frequency of the character errors about Letters (k=0,1,…,10)

Q

G_{LQ}=(0, 0, 0, 0, 3, 2, 0, 1, 0, 0, 0, 0, 2, 0, 2, 3, 1
0, 0, 0, 0 2, 0, 0, 1, 0, 0, 0, 3, 0, 1, 0, 0, 0, 0)

R

G_{LR}=(0,0, 0, 0, 1, 2, 0, 1, 0, 0, 0, 0, 1, 0, 3, 0, 1, 0,
0, 0, 0, 2, 0, 2, 1, 2, 1, 0, 0, 0, 1, 0, 0, 0, 0)

S

G_{LS}=(0,0, 0, 0, 3, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0,
0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0)

T

G_{LT}=(0,0, 0, 0, 1, 5, 0, 2, 0, 0, 0, 0, 1, 0, 1, 0, 4, 0,
0, 0, 0, 0, 0, 3, 0, 2, 0, 0, 5, 0, 0, 0, 0, 0, 0)

U

G_{LU}=(0, 0, 0, 0, 0, 2, 0, 2, 0, 0, 0, 0, 1, 0, 0, 1, 3,
0, 0, 0, 0, 2, 0, 0, 2, 0, 0, 0, 2, 0, 1, 0, 0, 0, 0)

V

G_{LV}=(0,0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0,
0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0)

W

G_{LW}=(0, 0, 0, 0, 2, 2, 0, 0, 0, 0, 0, 0, 2, 0, 0, 1, 0,
0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0)

X

G_{LX}=(0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,
0, 0, 0, 0, 2, 0, 1, 1, 0, 2, 0, 1, 0, 1, 0, 0, 0, 0)

Y

G_{LY}=(0,0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 4, 0, 0,
0, 0, 0, 1, 0, 0, 2, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0)

Z

G_{LZ}=(0,0, 0, 0, 3, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 3, 0, 0, 0, 0, 0, 0, 0, 0, 2, 1, 2, 0, 2, 0, 1, 0, 0, 0, 0)

T_{Qk}=<W,A> represents the number of words per minute that person types and the wrong characters account for the total number of characters about quotes(k=0,1…11)

T_{Lk}=<W,A> represents the number of words per minute that person types and the wrong characters account for the total number of characters about Letters(k=0,1…11)

P1
T_{Q1} =(83.325, 97.06625)
T_{L1} =(41.02666667, 92.32333333)

P2
T_{Q2} =(59.52875, 91.26)
TL2 =(41.60666667, 92.29333333)

P3
T_{Q3} =(60.28625, 96.15125)
T_{L3} =(44.30333333, 98.12)

P4
T_{Q4} =(104.72625, 96.23)

P5
T_{Q5} =(102.81625, 94.6825)
T_{L5} =(44.45, 94.80666667)

P6
T_{Q6} =(101.445, 94.88125)
T_{L6} =(58.01666667, 95.89666667)

P7
T_{Q7} =(83.26875, 95.205)
T_{L7} =(48.39333333, 91.68)

P8
T_{Q8} =(60.28625, 96.04875)
T_{L8} =(36.17333333, 93.63666667)

P9
T_{Q9} =(60.0675, 89.00875)

T_{L9} =(36.02, 93.51333333)

P10
T_{Q10} =(63.78875, 92.8975)
T_{L10} =(22.85666667, 94.99333333)

P11
T_{Q11} =(75.0775, 94.86875)
T_{L11} =(40.80333333, 92.69666667)

T_{Qk}=<W,A> represents the number of words per minute that person types and the wrong characters account for the total number of characters about quotes(k=A,B…K)

T_{Lk}=<W,A> represents the number of words per minute that person types and the wrong characters account for the total number of characters about Letters(k=Q,R…Z)

Quotes to be matched:
A -T_{QA} =(102.75, 95.605)
B -T_{QB} =(66.14333333, 94.725)
C -T_{QC} =(82.87333333, 96.17333333)
D -T_{QD} =(55.43166667, 89.72666667)
E -T_{QE} =(110.13, 96.36833333)
F -T_{QF} =(56.45666667, 95.37333333)
G -T_{QG} =(100.9016667, 96.70833333)
H -T_{QH} =(59.44666667, 96.60833333)
I -T_{QI} =(77.99833333, 97.02666667)
J -T_{QJ} =(56.37666667, 90.045)
K -T_{Qk} =(81.1, 98.16)

Letters to be matched:
Q-T_{LQ} =(47.57, 92.23)
R-T_{LR} =(54.21333333, 92.92)
S-T_{LS} =(40.14666667, 96.37)
T-T_{LT} =(41.65666667, 91.01666667)
U-T_{LU} =(35.5, 94.05)
V-T_{Lv} =(45.07666667, 96.50666667)
W-T_{Lw} =(43.13666667, 96.18333333)
X-T_{LX} =(45.66, 95.31333333)
Y-T_{LY} =(37.99, 95.70333333)
Z-T_{LZ} =(23.23666667, 93.86)