Route Choice Model by Text Data Analysis

17. the University of Tokyo B

Aiko Kondo, Kana Masuhashi,
Takuma Murahashi, Yosuke Mochizuki, Muhammad Zeeshan

Introduction

☐ Background

- TwitterなどのSNSの発達により、位置情報と紐付いたテキストデータの取得が容易になっている

  The development of social networking services such as Twitter has made it easier to acquire text data linked to its location data.

- ユーザーの位置情報を介して、テキストデータの特徴量を空間ネットワークに紐付けて時系列的にモデルに組み込める点がTweet形式のテキストデータの強みである

  The strength of Tweet data is that features of the text data can be connected to a spatial network and incorporated into a time-series model through the their location data.

☐ Applications of behavioral models using text data

- **Real time behavior prediction**

  ex. 災害発生時に、Twitterで発せられるテキストデータの内容から人々の行動をリアルタイムに予測するなどの応用が期待される

  It is expected to be applied to real-time prediction of people's behavior based on the content of Twitter text data in the event of a disaster.

- **Extracting the transformation of universal values**

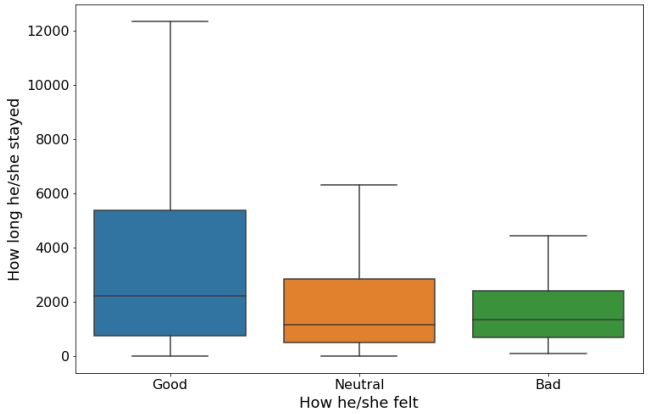
  ex. 「混雑」に関するTweetとその長期的に蓄積されたテキストデータとユーザーの位置情報を通して、COVID-19の流行前後で比較すると「混雑」に対する普遍的な価値観の変容を抽出できる

  Through tweets about "congestion" and their long-term accumulated text data and user location information, we can extract universal changes in values about "congestion" by comparing them before and after the COVID-19 epidemic.

Text data and Feature vector

☐ Relationship between text data and Feature vector

Relationship between time spent in the location and Tweet



① 教師あり機械学習
supervised machine learning

重信川の河川敷にいい感じの公園を見つけました。
春になったら遊びに来てみたいですね。

DNN(Bert)

["Good", "Neutral", "Bad"]

テキストから得た情報を、行動を記述できる特徴量の形にして用いる。

Extract features from text to describe behavior.

重信川の河川敷にいい感じの公園を見つけました。
春になったら遊びに来てみたいですね。

[2.64, -1.5, …, 0.11]

② PCA(主成分分析)
Principal Component Analysis

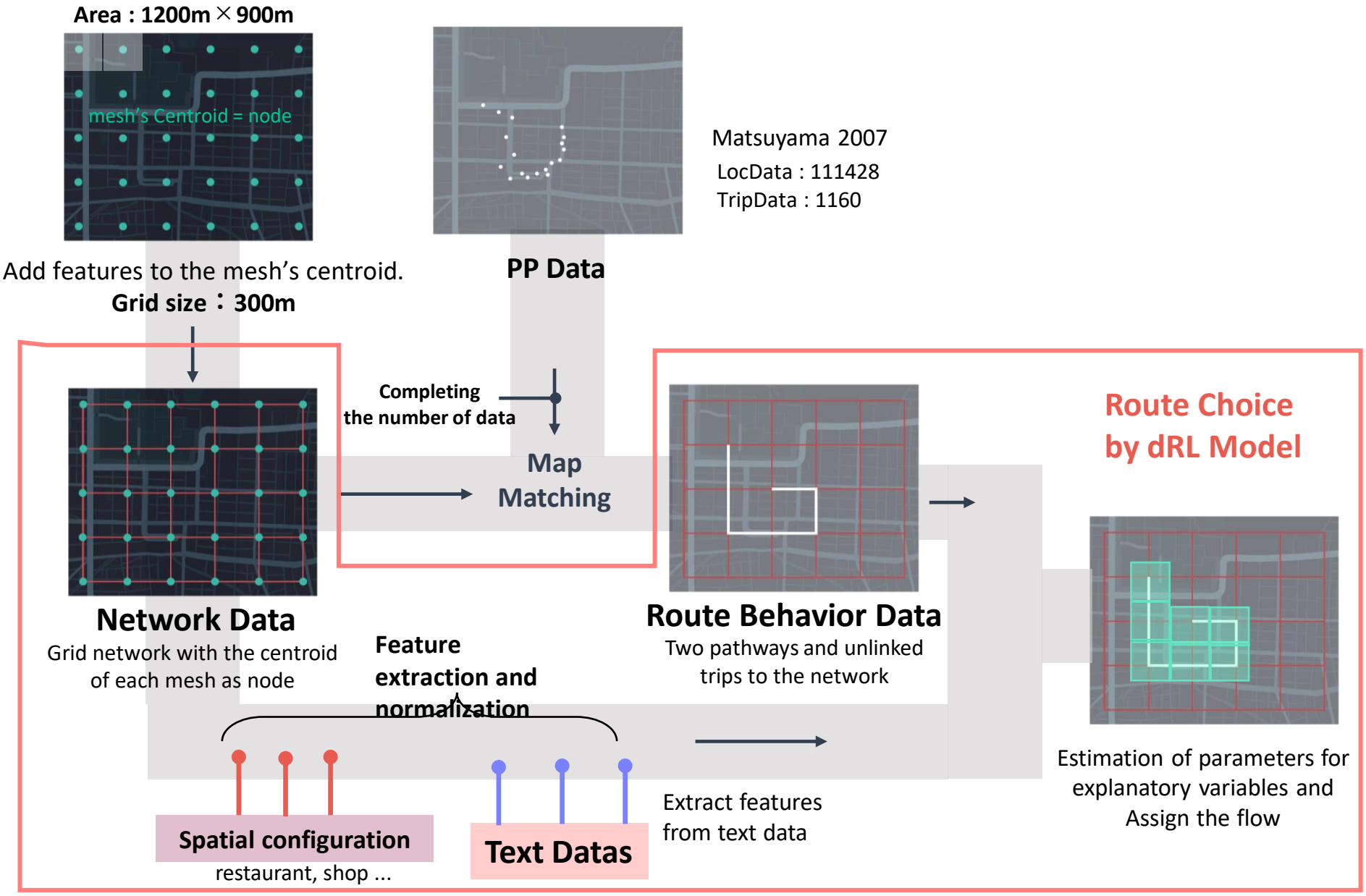
単語ベクトルから分散の大きい軸を抽出

Extract axes with high variance from word vectors

※上図はPCAのものではなく、あくまで単語の特徴を理解してもらうイメージです。

*The image above does not represent PCA, but an image to help you understand characteristics of words.

Framework


Area : 1200m × 900m
mesh's Centroid = node
Add features to the mesh's centroid.
Grid size：300m
Matsuyama 2007
LocData : 111428
TripData : 1160
PP Data
Completing
the number of data
Map
Matching
Network Data
Grid network with the centroid
of each mesh as node
Route Behavior Data
Two pathways and unlinked
trips to the network
Route Choice
by dRL Model
Feature
extraction and
normalization
Spatial configuration
restaurant, shop ...
Text Datas
Extract features
from text data
Estimation of parameters for
explanatory variables and
Assign the flow

RL model and State Quantity

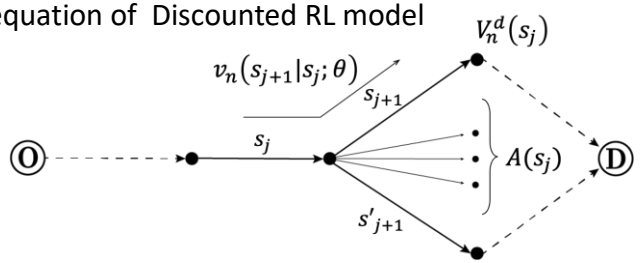
時系列的に逐次的な移動を表すようなモデル → RL(Recursive Logit Model)

dRLモデルの状態価値関数（Bellman方程式） State value function Bellman equation of Discounted RL model

時間割引率
Time discount rate($0 \leq \beta \leq 1$)

$$V^d(s_j) = E\left[\max_{s_{j+1} \in A(s_j)} \left(v(s_{j+1}|s_j;\theta) + \beta V^d(s_{j+1}) + \mu\epsilon(s_{j+1})\right)\right]$$

遷移効用 Transition Utility　将来効用 Future Utility　誤差項 err

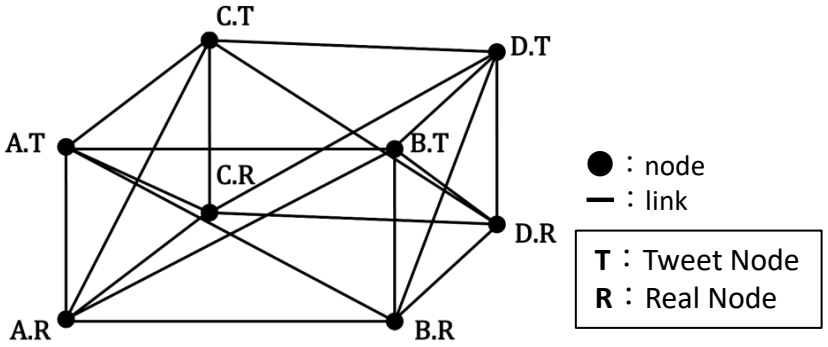


- **Problem：RLモデルは状態量がない** RL model has no state quantity

※ ただし今回は簡単のため$\beta = 0.9$に固定して推定
This time we fixed $\beta = 0.9$

→ ”Tweet 空間ネットワーク”を導入 Introduction of Tweet Space

＝ あるタイムステップにTweetをすることをTweet Nodeに遷移するとして疑似的に状態量を記述する
Tweeting at a certain time step is described as a transition to Tweet Node.



$A.R$ にいるときに遷移先の選択肢は$\{B.R,\ C.R,\ B.T, C.T\}$
When one is at $A.R$, the transition option is

$A.T$ にいるときに遷移先の選択肢は $\{A.T,\ B.R,\ C.R,\ B.T, C.T\}$
When one is at $A.T$, the transition option is

効用関数 Utility Function

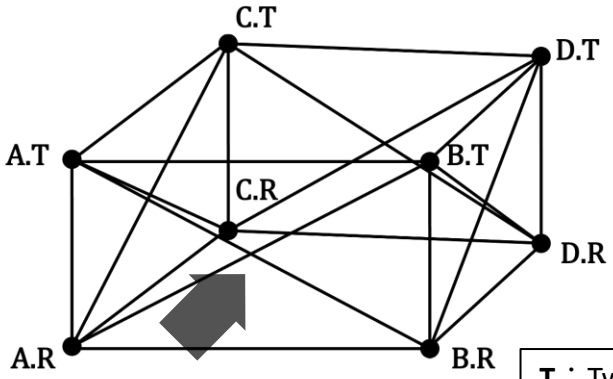
$$U(Z|A) = \beta \cdot X_Z + V_Z + \varepsilon_Z$$

将来効用 Future Utility　誤差 err

このX_Zが遷移のパターンによっていくつかの形に分けられる

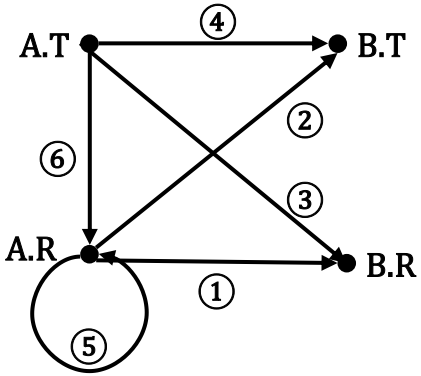
The shape of formulation of X_Z can be divided into several forms depending on the pattern of the transition

Formulation



T：Tweet Node
R：Real Node

O	D	Features of the Grid メッシュ内の特徴量	Features that predict the probability of a Tweet Tweetの確率を予想する特徴量	Emotions extracted from Tweets Tweetテキストの特徴量ベクトル	Constant term for Tweeting 定数項
A.R	B.R	Bでの値	0	0	0
A.R	B.T	Bでの値	Bでの値	0	1
A.T	B.R	Bでの値	0	0	0
A.T	B.T	Bでの値	Bでの値	0	1
A.R	A.R	Aでの値	0	0	0
A.T	A.R	Aでの値	0	Aでの値	0



①③⑤ $A.R \rightarrow B.R / A.T \rightarrow B.R / A.R \rightarrow A.R$

ex. $U(Z|A) = \beta_{shop} \cdot X_{shop} + \beta_{width} \cdot X_{width}$

② $A.R \rightarrow B.T / A.T \rightarrow B.T$

ex. $U(Z|A) = \beta_{shop} \cdot X_{shop} + \beta_{width} \cdot X_{width} + \beta_{nTweet} \cdot X_{nTweet} + \beta_{constant}$

Tweetをする効用
Utility of Tweeting

⑥ $A.T \rightarrow A.R$

ex. $U(Z|A) = \beta_{shop} \cdot X_{shop} + \beta_{width} \cdot X_{width} + \beta_{word\_vec} \cdot X_{word\_vec}$

Tweet後に滞在する効用
The Benefits of Staying After Tweeting

$shop$：店舗数 Number of shops

$width$：車道幅 Lane width

$nTweet$：Tweet数 Number of Tweet

$constant$：Tweetすることに対する定数項 Constant term for Tweeting

$word\_vec$：Tweetテキストから抽出された特徴量ベクトル Feature vector extracted from Tweet

Estimation Result

Basic RL Model

Estimation 1	
Estimated Parameter	t-Test
Evaluation	
the Number of Tweet	
Eval51	
Eval58	
PCA1	
PCA2	
the Width of Road	-6.61 *
the Number of Shop	3.53
the Number of Restaurant	-6.59 *
Constant Term (for Tweet)	
the Number of Sample	103
Initial LL	-835.69
Final LL	-821.72
LL Ratio	0.016
Adjusted LL Ratio	0.013
AIC (赤池情報量基準)	1649
β	0.53

*5%有意 **1%有意

ツイート関連の特徴量がないと、尤度が低い。
widthの推定値も直感に反する。

Without tweet-related features, LL is low. The width estimate is also counterintuitive.

RL Model using estimated evaluation

Estimation 2	
Estimated Parameter	t-Test
Evaluation	
the Number of Tweet	2.74
Eval51	
Eval58	2.92
PCA1	
PCA2	
the Width of Road	
the Number of Shop	
the Number of Restaurant	-3.26 *
Constant Term (for Tweet)	-12.12 **
the Number of Sample	103
Initial LL	-835.69
Final LL	-629.14
LL Ratio	0.24
Adjusted LL Ratio	0.24
AIC (赤池情報量基準)	1270
β	0.97

*5%有意 **1%有意

レストランの数に対してパラメータが負であるのは,
外食に行くトリップが少なかったためと考えられる。

The negative parameter for the number of restaurants is thought to be due to the small number of trips to eat out.

尤度比，t値は安定している。

We succeeded to get high LL ratio & t-value.

Estimation Result

Basic RL Model

Estimation 1		
	Estimated Parameter	t-Test
Evaluation		
the Number of Tweet		
Eval51		
Eval58		
PCA1		
PCA2		
the Width of Road	-6.61	*
the Number of Shop	3.53	
the Number of Restaurant	-6.59	*
Constant Term (for Tweet)		
the Number of Sample	103	
Initial LL	-835.69	
Final LL	-821.72	
LL Ratio	0.016	
Adjusted LL Ratio	0.013	
AIC (赤池情報量基準)	1649	
β	0.53	

*5%有意 **1%有意

ツイート関連の特徴量がないと、尤度が低い。widthの推定値も直感に反する。

Without tweet-related features, LL is low. The width estimate is also counterintuitive.

RL Model using vectorized text data

Estimation 3		
	Estimated Parameter	t-Test
Evaluation	1.79	3.98 *
the Number of Tweet	0.62	3.09 *
Eval51		
Eval58		
PCA1	0.94	2.60 *
PCA2		
the Width of Road		
the Number of Shop		
the Number of Restaurant	-0.11	-3.74 *
Constant Term (for Tweet)	-3.55	-12.14 **
the Number of Sample		103
Initial LL		-835.69
Final LL		-629.14
LL Ratio		0.25
Adjusted LL Ratio		0.24
AIC (赤池情報量基準)		1268
β		0.97

*5%有意 **1%有意

ベクトル化されたテキストデータの第一主成分の寄与率は高くなかったが，尤度比，t値は高かった。

Although the contribution of the first principal component of the vectorized text data was not high, we got high LL ratio and t-values.

Scenario Simulation

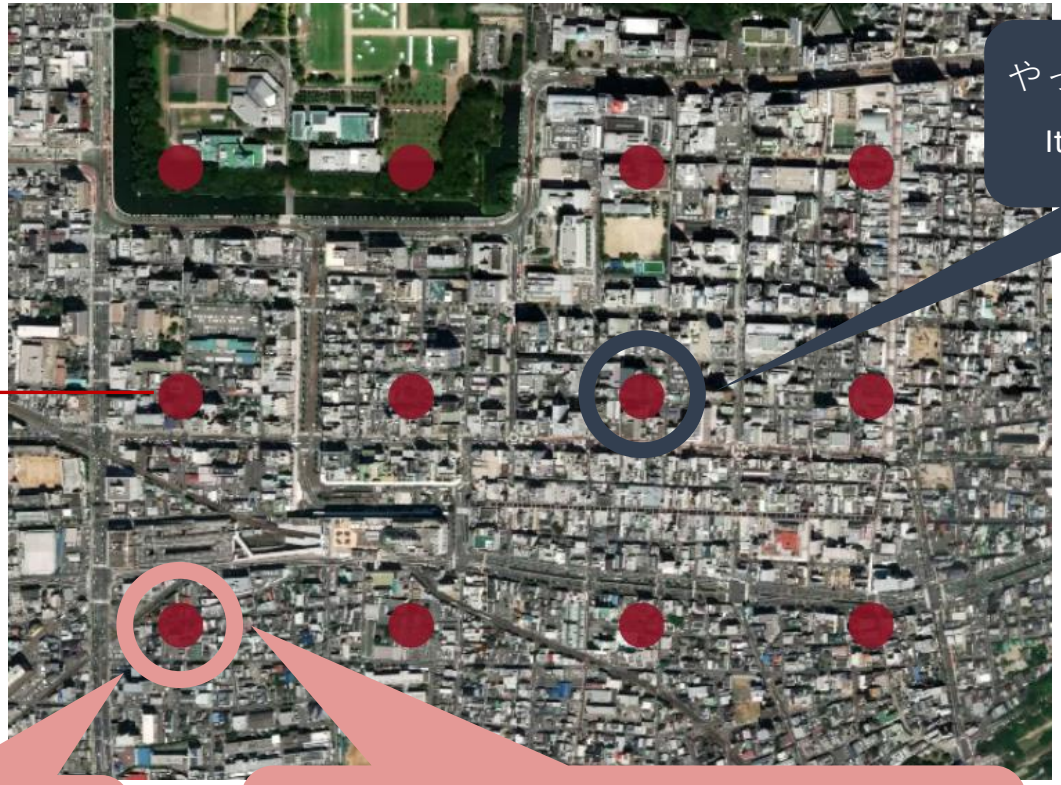
実際の松山中心市街内でのツイート(2021年現在)を、示唆する感情プラスマイナスに基づきノードに付与

We distributed tweets (2021) in Matsuyama-city to nodes, depending on their implying emotion

→ノードでのツイート感情が変化したとき、行動変化は見られるか？

Do you see behavioral changes when the tweet sentiment at the node changes?

Node
= Centroid of each Grid



やっぱ来るの遅過ぎた😫

It was too late to come.

Negative

Positive

良い湯であった☺♨

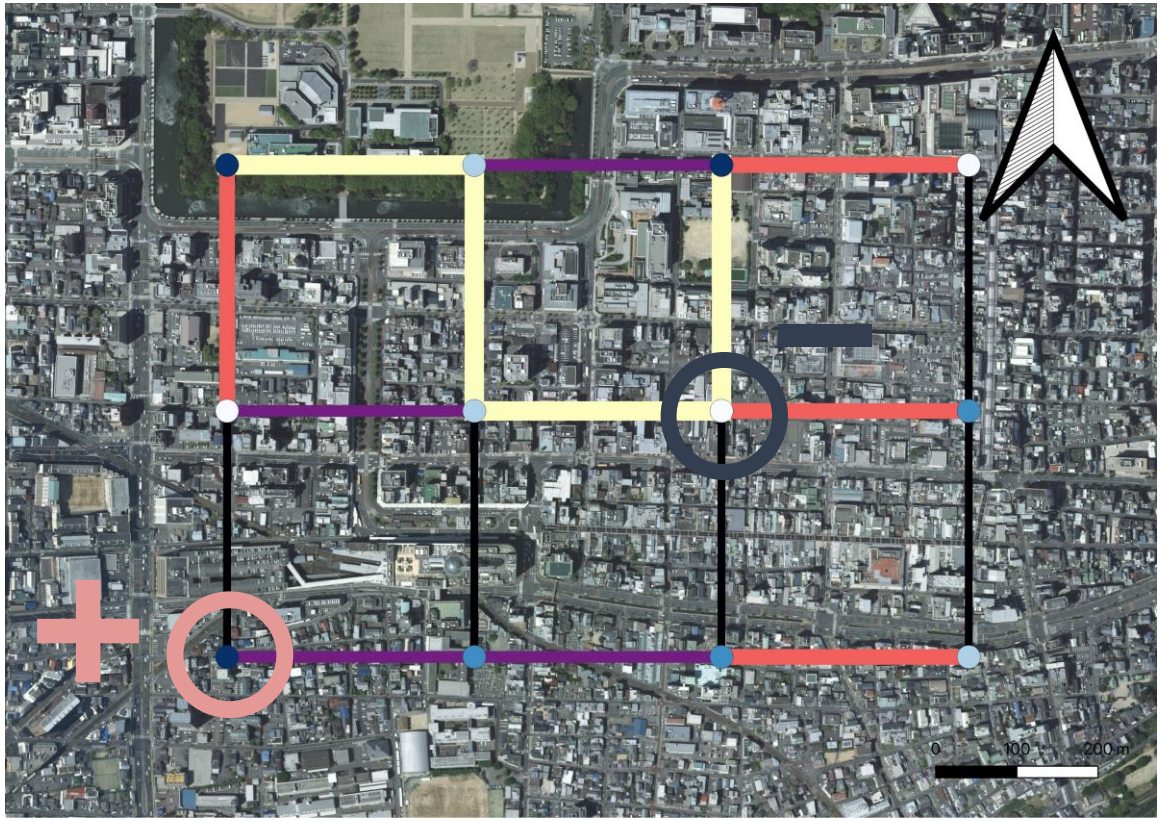
The bath was very comfortable.

油そば周平 醤油久しぶりに食べたけどやっぱり美味い(´艸`)🍴🍜✌

It's been a while since I've had the soy sauce flavored oil soba at Shuhei. It's always good!

Positive

Scenario Simulation Result



Number of Tweet /person (relative)

Least tweeted/person

Most tweeted/person

0.00019 - 0.00037
0.00037 - 0.00094
0.00094 - 0.00181
0.00181 - 0.01132

Traffic Flow
(absolute)

high

low

- 良いツイートを付与したノードの付近では、ツイート数が増加した
  The number of tweets increased in the vicinity of the node that gave good tweets
- 悪いツイートを付与してノード付近では、ツイート行動が減少した
  Tweeting behavior decreased near nodes with bad tweets

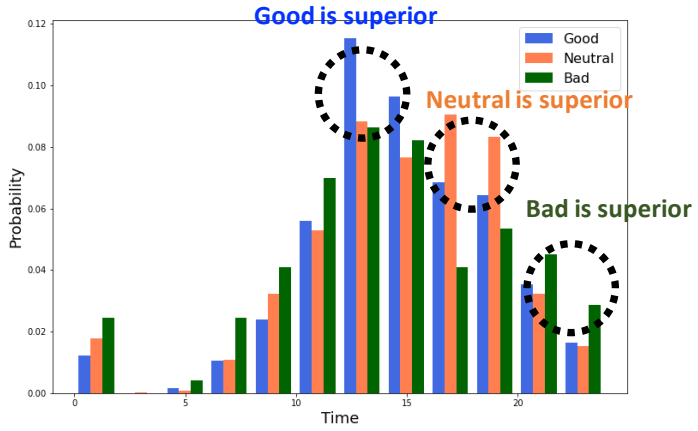
Appendix

☐ Relationship between text data and Emotional evaluation

PPのentryデータの良い・やや良い・どちらでもない・やや悪い・悪いの5段階の指標の回答結果を、良い・どちらでもない・悪いの3段階にして分析を行なった。

The results of responses to the five indicators of PP entry data (good, somewhat good, neither good nor bad, somewhat bad, and bad) were analyzed in three levels: good, neither good nor bad, and bad.

Number of tweets by time zone



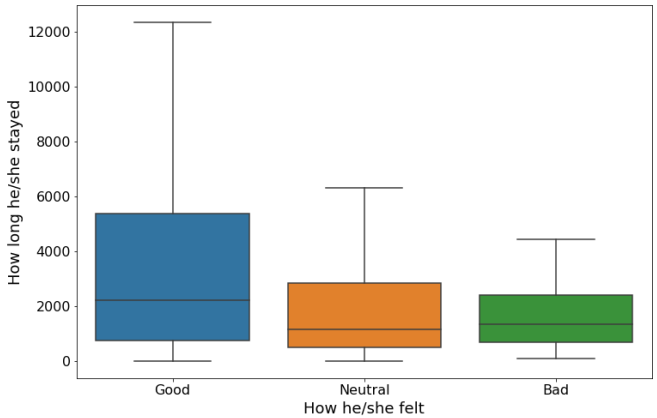
- ✓ Tweet頻度は1日のうちで偏りが見られ、概ね昼間にピークがある

  Tweet frequency is unevenly distributed throughout the day, with a peak generally during the daytime.

- ✓ 1日が終盤になるにつれてbadに分類されるTweetが卓越する→ 疲れてきた?

  Towards the end of the day, the number of Tweets classified as "bad" stands out. People get tired?

Relationship between time spent in the location and Tweet



- ✓ このクラス分類と滞在時間には相関が期待できる

  This Classes in Tweet content can be expected to correlate with time spent.

- ✓ 特に「良い」に分類れさるTweetとTweetされた場所での滞在時間の相関が期待できる

  In particular, we can expect to see a correlation between Tweets that are classified as "good" and the time spent at the location where they were tweeted.

- ✓ 「どちらでもない」「悪い」に分類されるTweetと滞在時間の違いは相対的に小さく見える

  The difference in time spent between Tweets categorized as "neither" or "bad" looks relatively small.

Appendix

☐ Basic Analysis -Text Mining Analysis-

● Time

- ✓ 朝よりも夜の方が購買衝動が強いと思われる

  The urge to buy is greater at night than in the morning.

- ✓ 夜は明日のことを考え始める

  At night, many people think about tomorrow.

● Age

- ✓ 利用施設の世代間の傾向の違いがうかがえる

  The text data shows different trends in the facilities used.

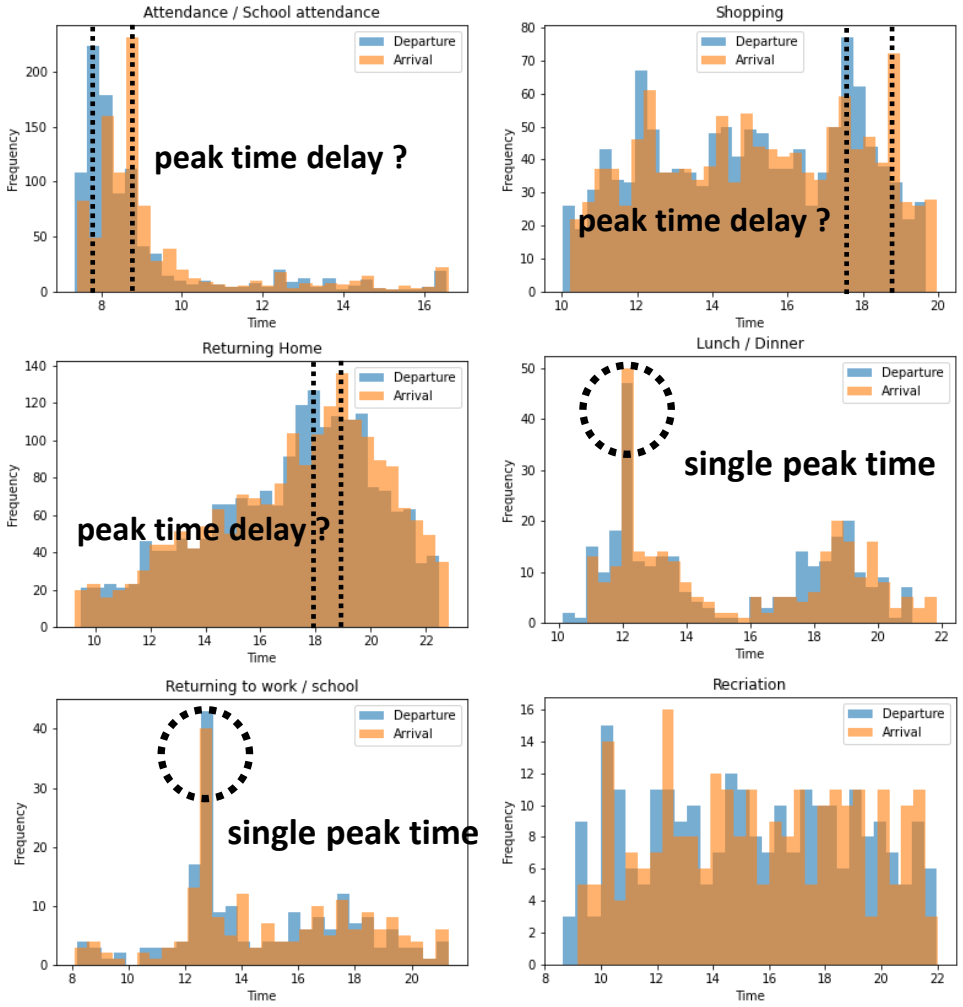
- ✓ 利用施設の世代間の傾向の違いがうかがえる

  Different words are used for the same meaning between generations.

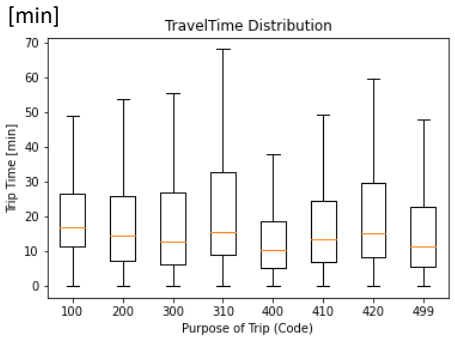
Appendix

☐ Basic Analysis -Differences in trends in behavior occurrence by purpose-

Frequency of Departure and Arrival



cf. Box plot of the length of trip time (5% - 95%)



100：出勤・登校 Going to office/school
200：帰宅 Returning home
300：帰社・帰校 Returning office/school
310：業務 Work
400：買い物 Shopping
410：食事 Eating
420：娯楽 Amusement
499：その他私用 Others

旅行時間に関しては目的ごとにさほど大きな差は見られなかった
There was not much difference in trip time by purpose.

1つの特徴的なピーク時間帯を持つ目的がある一方で、出発時間のピーク時間帯が到着時間のピーク時間帯に遅延しているような集計結果も得られた。

For some purposes there are the single peak time, and for other purposes the peak time of arrival appears later than the peak time for departure.

Appendix

☐ Basic Analysis -Apply tf-idf approach to text data per mesh-