Data Acquisition Report
Differentially Private Location-based Histogram Publication

Name: CHAN, Lo Yuet

1. General Purpose

The dataset will be used in the data science project for implementing and reviewing the differential privacy policy suggested by Zhang. As a result, the dataset will be processed into a 2-variable dataset.

2. Description of Dataset

The acquired dataset consists of 3 different part, dataset_TIST2015, dataset_TSMC2014, dataset_ubicomp2013 including check-in records and other data over a a few years of time in a global scale.

a. dataset_TIST2015

It contains 33,278,683 check-in records by 266,909 users at 3,690,126 venues located at 415 cities in 77 countries in tsv format. The set contains 3 files, dataset_TIST2015_Checkins containing check-in records; dataset_TIST2015_POIs containing data of 3,690,126 venues; dataset_TIST2015_Cities containing 415 cities' data.

dataset_TIST2015_Checkins			
Column	Variable	Type	Details
0	User ID	Integer	Anonymous ID
1	Venue ID	String	Anonymous ID
2	UTC Time	String	
3	Timezone Offset	Integer	In minutes
dataset_TIST2015_POIs			
Column	Variable	Type	Details
0	Venue ID	String	Anonymous ID
1	Latitude	Float	
2	Longitude	Float	
3	Venue Category Name	String	From Foursquare
4	Country Code	String	
dataset_TIST2015_Cities			
0	City Name	String	
1	Latitude	Float	
2	Longitude	Float	
3	Country Code	String	
4	Country Name	String	
5	City Type	String	

b. dataset_TSMC2014

It contains check-in at New York City and Tokyo in tsv format. The set contains 2 files, dataset_ TSMC2014_NYC with 227,428 check-in records at New York City; dataset_TSMC2014_TKY with 573,703 check-in records at Tokyo.

dataset_TSMC2014_NYC			
Column	Variable	Type	Details
0	User ID	Integer	Anonymous ID
1	Venue ID	String	Anonymous ID
2	Venue Category ID	String	From Foursquare
3	Venue Category Name	String	From Foursquare
4	Latitude	Float	
5	Longitude	Float	
6	Timezone Offset	Integer	In minutes
7	UTC Time	String	
dataset_TSMC2014_TKY			

Column	Variable	Type	Details
0	User ID	Integer	Anonymous ID
1	Venue ID	String	Anonymous ID
2	Venue Category ID	String	From Foursquare
3	Venue Category Name	String	From Foursquare
4	Latitude	Float	
5	Longitude	Float	
6	Timezone Offset	Integer	In minutes
7	UTC Time	String	

c. dataset_ubicomp2013

The dataset contains 27,149 check-in records by 3,112 users at 3,298 venues and 10,377 tips in tsv format. The set contains 3 files, dataset_ubicomp2013_checkins containing check-in records; dataset_ubicomp2013_tips containing 10,377 tips records; dataset_ubicomp2013_tags containing the optional tags for venues left by users.

dataset_ubicomp2013_checkins			
Column	Variable	Type	Details
0	User ID	Integer	Anonymous ID
1	Venue ID	String	Anonymous ID
dataset_ubicomp2013_tips			
Column	Variable	Type	Details
0	User ID	Integer	Anonymous ID
1	Venue ID	String	Anonymous ID
2	Tip Text	String	
dataset_ubicomp2013_tags			
0	Venue ID	String	
1	Tags	String	

3. Curation of Dataset

dataset_ubicomp2013 files has been converted to UTF-8 Encoding such that it can be. Then, dataset_TIST2015 _Checkins, dataset_TSMC2014_NYC, dataset_TSMC2014_TKY, and dataset_ubicomp2013_checkins are processed to extract User ID and Venue ID from the tsv files into separate csv format according to the source dataset name for easy access in the future.

dataset_TIST2015_Checkins.csv (33,263,633 records)			
Column	Variable	Type	Details
0	User ID	Integer	Anonymous ID
1	Venue ID	String	Anonymous ID
dataset_TSMC2014_NYC.csv (227,428 records)			
Column	Variable	Type	Details
0	User ID	Integer	Anonymous ID
1	Venue ID	String	Anonymous ID
dataset_TSMC2014_TKY.csv (573,703 records)			
Column	Variable	Type	Details
0	User ID	Integer	Anonymous ID
1	Venue ID	String	Anonymous ID
dataset_ubicomp2013_checkins.csv (27,149 records)			
Column	Variable	Type	Details
0	User ID	Integer	Anonymous ID
1	Venue ID	String	Anonymous ID

4. Reference

All the raw datasets are from https://sites.google.com/site/yangdingqi/home/foursquare-dataset