Data Mining (CS524)

Indian Institute of Technology Ropar

Project: BrightKite **Release Date: 11/03/2019**
Maximum Marks: 100 **Due Date: 21/04/2019**

BrightKite

`Brightkite` was a location-based social networking website, active during the years 2007-2011. The registered users were able to share their locations through "*check in*" at places. Users were also able to see the other nearby users and those who have *checked-in* at that place in the past.

Dataset

Download the `Brightkite` dataset from here (`http://tarique.in/cs524-2019/BrightKite.zip`). The dataset consists of two files, the friendship network (`Brightkite_edges.txt`) and the *check-ins* (`Brightkite_totalCheckins.txt`). The friendship network was collected using their public API, and consists of 58,228 nodes and 214,078 undirected edges, where the nodes represent the users and the edges represent the bi-directional friendship (relationship) between the users. The check-ins file consists of a total of 4,491,143 check-ins of the users over the period of April 2008 - October 2010.

Description of the `Brightkite_edges.txt` file: Each line contains two values separated by a tab character. The two values are node (user) identifiers, and each line represents a friendship between the users corresponding to the pair of node identifiers. For example, the first line
0 1
says that there is a friendship between the users with ids 0 and 1.

Description of the `Brightkite_totalCheckins.txt` file: Each line is a check-in, that contains four values separated by tab characters. The values stand for node id (user who checked-in), check-in time by the user, check-in latitude, check-in longitude, and check-in location id. For example, the first line
0 2010-10-17T01:48:53Z 39.747652 -104.99251 88c46bf20db295831bd2d1718ad7e6f5
says that the user with node id 0 checked-in on 17/10/2010 (date) at 01:48:53 (time) at location (39.747652, -104.99251) whose location id is 88c46bf20db295831bd2d1718ad7e6f5.

Objective

The objective of this project is to understand, mine and analyze the given dataset in order to discover hidden and obscure insights. This can be done through several interesting ways. Some of the tasks that can be performed are given below. Apart from these samples, much more advanced mining tasks can be performed to discover the interesting hidden and obscure insights.

- (Pattern analysis) Discover the pattern of check-ins on a particular day, or how the check-in patterns vary over a period of time. What are the check-in patterns of users, or a group of friends. How their check-ins evolve with respect to the locations.

- (Cluster Analysis) Discover the clusters of check-in. Visualize on a map. Identify the users clusters having common check-in patterns. Identify the different group of users (communities) who are closely related.

- (Regression and Classification) Predict the future check-in of a user. Predict whether a location is going to have a high volume of check-ins at a future point of time. This is a very interesting problem in traffic prediction. Classify users or group of users as *check-in active* or *check-in inactive*.

- (Visualization) Visualize the obtained insights in different ways, so that the mined information can be easily interpreted and understood.

Table 1: Project Timeline (year 2019)

Activity	**11/03**	**11/03 - 24/03**	**24/03**	**24/03 - 07/04**	**07/04**	**07/04 - 21/04**	**21/04**	**22/04**
Project release	■							
Problem identification		■						
Proposal submission			■					
Solution development and implementation				■				
Progress report submission					■			
Implementation and results generation						■		
Final report submission							■	
Demonstration								■

Table 2: Marks distribution

Activity	**11/03**	**11/03 - 24/03**	**24/03**	**24/03 - 07/04**	**07/04**	**07/04 - 21/04**	**21/04**	**22/04**
Proposal submission			10					
Progress report submission					25			
Final report submission							35	
Demonstration								30
Total	**100**							

Project Work

The project is supposed to start immediately. Groups of size 1/2/3 students are to be formed. The first step is to understand the dataset and finalize a problem (tentative) of study. A $\frac{1}{2}$-1 page proposal of the problem to be studied is to be submitted first. Then it will go through the phase of *solution development and implementation*, after which a *progress report* of 2-4 pages is to be submitted. The next phase is *implementation and results generation*, after which the *final report* of 4-6 pages is to be submitted. At the end, each project is to be demonstrated in the lab. The detailed timeline of the progress and submissions is given in Table 1, and the marks distribution is given in Table 2.

End of the Project

Note:

(*i*) *Each project will have 1/2/3 members.*

(*ii*) *All reports are to be written in ACM SIG style and finally converted to pdf for submission.*

(*ii*) *Late submissions will face a penalty of 10% (of the full marks) for each day of delay.*

(*iv*) *Presenting some other person's work as your own without proper citation of the source is an act of plagiarism. It is a serious offence and will be treated strictly.*

(*v*) *Queries, if any, can be directed to our TA Aroof Aimen (2018csz0001@iitrpr.ac.in) through email.*