**Alkiviadis Kariotis 241735**

Homework Delicious Data Set

Acquire some data.

Visit the https://grouplens.org/datasets/hetrec-2011/ and obtain the delicious data set. Store all the files that makeup the data set in a directory called databases.

**1.Create a database named: delicious**

CREATE DATABASE delicious;

USE delicious;

**2. Create a table for tags named: tags, choose the primary key**

CREATE TABLE tags (

tagID INT AUTO\_INCREMENT PRIMARY KEY,

tagName VARCHAR(255)

);

**3. Create a table for bookmarks: named bookmarks, choose the primary key**

CREATE TABLE bookmarks (

bookmarkID INT AUTO\_INCREMENT PRIMARY KEY,

url VARCHAR(255),

title VARCHAR(255)

);

**4. Create a table for user\_taggedbookmarks-timestamps.dat named:**

**tagged bookmarks, choose the primary key. The primary key may contain more than one keys. Set also the foreign keys.**

CREATE TABLE taggedbookmarks (

userID INT,

bookmarkID INT,

tagID INT,

timestamp DATETIME,

PRIMARY KEY (userID, bookmarkID, tagID),

FOREIGN KEY (bookmarkID) REFERENCES bookmarks(bookmarkID),

FOREIGN KEY (tagID) REFERENCES tags(tagID)

);

**5. Import data for bookmarks**

Execute to terminal

mysqlimport --ignore-lines=1 --fields-terminated-by=',' --columns='bookmarkID,url,title' delicious /path/to/databases/bookmarks.dat

**6. Import data for tags**

Execute to terminal

mysqlimport --ignore-lines=1 --fields-terminated-by=',' --columns='tagID,tagName' delicious /path/to/databases/tags.dat

**7. Import data for tagged bookmarks**

Execute to terminal

mysqlimport --ignore-lines=1 --fields-terminated-by=',' --columns='userID,bookmarkID,tagID,timestamp' delicious /path/to/databases/user\_taggedbookmarks-timestamps.dat

**You may need a combination of SQL and python**

**8. Can you display the top ten bookmarks in terms of frequency (in taggedBookmarks)?**

SELECT bookmarks.title, taggedbookmarks.bookmarkID, COUNT(\*) as frequency

FROM taggedbookmarks

JOIN bookmarks ON taggedbookmarks.bookmarkID = bookmarks.bookmarkID

GROUP BY taggedbookmarks.bookmarkID, bookmarks.title

ORDER BY frequency DESC

LIMIT 10;

**9. Can you display the top ten users in terms of frequency of appearance?**

SELECT userID, COUNT(\*) as frequency

FROM taggedbookmarks

GROUP BY userID

ORDER BY frequency DESC

LIMIT 10;

**10. Obtain the user activity per hour or per day or per month.**

**a. Hint-1: place all the taggedbookmarks data in a data frame**

**i. Convert the timestamp to year, month, day, hour (hint: python function:**

**datetime.fromtimestamp(1130845640)**

**ii. Next use: Group-by**

import pandas as pd

from datetime import datetime

df['datetime'] = df['timestamp'].apply(lambda x: datetime.fromtimestamp(x))

# Extract year, month, day, hour

df['year'] = df['datetime'].dt.year

df['month'] = df['datetime'].dt.month

df['day'] = df['datetime'].dt.day

df['hour'] = df['datetime'].dt.hour

# User activity per hour

activity\_per\_hour = df.groupby(['userID', 'year', 'month', 'day', 'hour']).size().reset\_index(name='activity\_count')

# User activity per day

activity\_per\_day = df.groupby(['userID', 'year', 'month', 'day']).size().reset\_index(name='activity\_count')

# User activity per month

activity\_per\_month = df.groupby(['userID', 'year', 'month']).size().reset\_index(name='activity\_count')