MAULANA AZAD NATIONAL INSTITUTE OF TECHNOLOG, BHOPAL
Department of Computer Science & Engineering
B.Tech. (CSE) VI th Sem.
Lab Assignment - 1

Subject: **Data Warehouse & Date Mining** **Sub Code: CSE-322** **Due Date –**

Data Preprocessing

Consider the sample data set (log file) provided in class. Review the following attributes of the data set.

1) **2004-11-13**: Date at which the entry is recorded.
2) **00:00:32**: Time at which the entry is recorded
3) **W3SVC195**: Name of the server
4) **68.142.250.151**: IP address of the server
5) **Get**: Method of HTTP request
6) **sharedoutandabout/InAndOut/Categories.aspx:** The resource requested
7) **insid=2&langid=1**: Parameters associated with the resource requested
8) **80**: Port number
9) -: This is the user name if the site require user authentication. If not the hyphen is placed
10) **93.186.23.240**: Client IP address
11)**Mozilla/4.0+ (compatible;+MSIE+4.01;+Windows+NT**): Browser name and version and the operating system.
12) **200**: It is the status code returned to the user
13) **3223**: The bytes sent from the server to the client in response to the user request.

Answer the following questions with respect to the sample data set.

Q.1) The purpose of data cleansing process is to remove noisy and unnecessary data that may affect the mining process.

a) Remove log Entry nodes that contain in uri-stem child node extensions like .jpg, .gif, and .css.

b) Remove the records having status code above 299 and below 200.

Q.2) The goal of user identification is to identify who access web site and which pages are accessed.

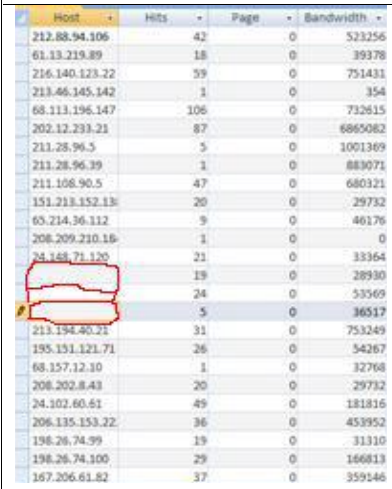
a) Indentify unique users in given log

[Hint: If new IP address then there is a new user. If IP address is same but browser version or operating system is different then it represents different users.]

Q.3) Generate a new data set using the missing data strategies discussed in class.

For Instance

Before Preprocessing	After Preprocessing



Host	Hits	Page	Bandwidth
212.88.94.106	42	0	523256
61.13.219.89	18	0	39378
216.140.123.22	59	0	751431
213.46.145.142	1	0	354
68.113.196.147	106	0	732615
202.12.233.21	87	0	6865082
211.28.96.5	5	0	1001369
211.28.96.39	1	0	883071
211.108.90.5	47	0	680321
151.213.152.13	20	0	29732
65.214.36.112	9	0	46176
208.209.210.18	1	0	0
24.148.71.120	21	0	33364
211.28.96.39	19	0	28930
211.108.90.5	24	0	53569
67.29.222.73	5	0	36517
213.194.40.21	31	0	753249
195.151.121.71	26	0	54267
68.157.12.10	1	0	32768
208.202.8.43	20	0	29732
24.102.60.61	49	0	181816
206.135.153.22	36	0	453952
198.26.74.99	19	0	31310
198.26.74.100	29	0	166813
167.206.61.82	37	0	359146

Host	Hits	Page	Bandwidth
212.88.94.106	42	0	523256
61.13.219.89	18	0	39378
216.140.123.22	59	0	751431
213.46.145.142	1	0	354
68.113.196.147	106	0	732615
202.12.233.21	87	0	6865082
211.28.96.5	5	0	1001369
211.28.96.39	1	0	883071
211.108.90.5	47	0	680321
151.213.152.13	20	0	29732
65.214.36.112	9	0	46176
208.209.210.18	1	0	0
24.148.71.120	21	0	33364
211.28.96.39	19	0	28930
211.108.90.5	24	0	53569
67.29.222.73	5	0	36517
213.194.40.21	31	0	753249

Q. 4) Write a utility to determine the association between two attributes "**Client IP address**" and "**Bytes Sent**". Generate a histogram for the attribute Bytes Sent with respect to Client IP address.

Submission:

1. Report explaining the algorithm, description of functions, and any other implementation details that explain your code.

2. Entire project directory including source files, header files, Data Base (Script), Charts, Graphs and the compiled executable files.

Note: **You may use PHP, Visual C++, C#, Java, SQL, R, Python, Octave/MATLAB, or Excel in any combination to complete this assignment.**

****