

Web usage mining lab assignment on 24th October 2018

Q1. Following is a set of web usage logs from an organization. Assume the following rules.

- The time heuristic h1 for a session duration = 30 minutes.
- The time heuristic h2 for the average duration for the user visit of a page = 1 minute.
- Include a page in a session if the page that refers is in that session.
- During a session, the IP address, browser and OS is the same.

You need to printout a table containing session number, IP address, session start time and session end time.

| | | | | | | |
|---------------|---|---|------------------------------|---|-----|-------|
| 172.20.112.25 | - | - | [02/Feb/2000:10:22:02 -0500] | GET /airmedia/images/welcome.gif?GXHC_gx_session_id_airmedia=c7bf39189d98ea9f HTTP/1.0 | 200 | 3081 |
| 172.20.112.25 | - | - | [02/Feb/2000:10:22:02 -0500] | GET /airmedia/images/edit_off.gif?GXHC_gx_session_id_airmedia=c7bf39189d98ea9f HTTP/1.0 | 200 | 268 |
| 172.20.112.25 | - | - | [02/Feb/2000:10:22:02 -0500] | GET /airmedia/images/pubcontent.gif?GXHC_gx_session_id_airmedia=c7bf39189d98ea9f HTTP/1.0 | 200 | 1388 |
| 172.20.112.25 | - | - | [02/Feb/2000:10:22:23 -0500] | GET /airmedia/gateway/1.html HTTP/1.0 | 304 | 252 |
| 172.20.112.25 | - | - | [02/Feb/2000:10:22:23 -0500] | GET /airmedia/gateway/images/66.gif HTTP/1.0 | 304 | 428 |
| 172.20.112.25 | - | - | [02/Feb/2000:10:22:01 -0500] | GET /airmedia/images/topic1_on.gif HTTP/1.0 | 200 | 654 |
| 172.20.112.25 | - | - | [02/Feb/2000:10:22:02 -0500] | GET /airmedia/images/create_on.gif HTTP/1.0 | 200 | 365 |
| 12.3.207.3 | - | - | [02/Feb/2000:10:22:02 -0500] | GET /airmedia/gateway/index.html HTTP/1.0 | 200 | 660 |
| 12.3.207.3 | - | - | [02/Feb/2000:10:22:02 -0500] | GET /airmedia/images/login.gif?HTTP/1.0 | 200 | 660 |
| 12.3.207.3 | - | - | [02/Feb/2000:11:22:02 -0500] | GET /airmedia/images/welcome.gif?GXHC_gx_session_id_airmedia=c7bf39189d98ea9g HTTP/1.0 | 200 | 660 |
| 12.3.207.3 | - | - | [02/Feb/2000:11:22:02 -0500] | GET /airmedia/images/create_off.gif?GXHC_gx_session_id_airmedia=c7bf39189d98ea9g HTTP/1.0 | 200 | 341 |
| 12.3.207.3 | - | - | [02/Feb/2000:11:22:02 -0500] | GET /airmedia/images/w_myh.gif?GXHC_gx_session_id_airmedia=c7bf39189d98ea9g HTTP/1.0 | 200 | 783 |
| 12.3.207.3 | - | - | [02/Feb/2000:11:22:02 -0500] | GET /airmedia/gateway/1.html HTTP/1.0 | 200 | 26814 |
| 12.3.207.3 | - | - | [02/Feb/2000:12:02:13 -0500] | POST /cgi-bin/gx.cgi/AppLogic+airmedia.gateway.Gateway HTTP/1.0 | 200 | - |
| 12.3.207.3 | - | - | [02/Feb/2000:12:02:23 -0500] | GET /airmedia/gateway/images/2.gif HTTP/1.0 | 304 | 150 |
| 172.20.112.25 | - | - | [02/Feb/2000:13:00:07 -0500] | GET /cgi-bin/gx.cgi/AppLogic+airmedia.logout.ShowAsso | 200 | - |

Q2. Use the same program to find do sessionization of a log file that is downloaded from web that is uploaded into moodle today and give it in same format as Q1.