01 12 2016 **SEWN 2016 labsheet 5 MSc version (assessed coursework)**
Parsing and filtering of web log data and data analysis

This labsheet is **optional** and is intended to demonstrate how a user's actions in a web browser are recorded in web server log files.

Any MSc student attempting the BSc version of this labsheet may do so only at the discretion of the module's lecturer or TA.

The maximum grade achievable for this labsheet is 100%.

1. Conduct some preliminary research on the web, on the attributes included in a standard/extended web log format. You may find the URL below useful:

   http://www.w3.org/TR/WD-logfile

2. Find out about: (a) what information about users' activity can be obtained by analysing web server logs, and (b) how to partition the log files into meaningful sessions. You may find the paper below useful:

   http://arxiv.org/ftp/arxiv/papers/1101/1101.5668.pdf

3. Download the web log file:

   http://www.dcs.bbk.ac.uk/~sewn_ta_2016/sewn/ls2/sewn_2016_labsheet_2_web_server_log_data.zip

   This log data, which is from a few years ago, is for a single day and has been *cleansed* to exclude images and other multimedia files. The log data has the followings attributes (in this order):

   ```
   date, time, c-ip, cs-username, sc-method, cs-uri-stem, cs-uri-query, sc-status, cs-host, cs-(User-Agent), cs(Cookie), cs(Referrer)
   ```

4. Import the web log data into a DBMS. Create a Microsoft Access, MySQL, Oracle or other database and separate the data into meaningful fields of the table. It is up to you whether you want to create multiple tables or a single table.

5. Using the information found during your research for (1) above, design suitable queries to extract users' activities and behaviours for the following:

   a. The top five most active hours (most requests per hour).
   b. The number of requests made per status.
   c. The top twenty files/pages requested.
   d. The top twenty IP addresses (or users) who requested the most URLs.

6. Write two additional queries of your own choice.

   Extra credit will be given to those who manage to extract *unusual* or *interesting* observations of user activities from the log data for the two original queries, and also for any visualisation(s) created displaying the data for the original queries.

7. Produce and submit a summary report of your web data analysis. Plan and think carefully about how to organise your report and its presentation. The contents of the report are listed below.

What to hand in:

1. A summary report (no more than four A4 pages) of your web data analysis as specified above.
2. A screen shot or text of each query's results.
3. The definitions of each query, submitted as text format (either as `.txt`, `.doc(x)` or `.pdf` file).
4. Any references to books or on-line material for this labsheet should be included in your report.

All of the above should be archived in a single `.zip` file submitted via the **Labsheet 5 Parsing and filtering of web log data and data analysis** drop box in Moodle.

Submission deadline: 15 12 2016.

Late assignments: No extensions are available for this lab and any late submissions will be graded as per the guidelines of the relevant course being studied.