Team\_16GB MSA8010 Final Project Documentation

Overview

The project is essentially a log file analysis tool. A log in this instance is an automatically produced and time-stamped record of events that occurred on a particular system. The purpose of the project is to analyze the log files and serve as part of a report generation process. The report is expected to assist the client in reviewing how much continuing application support to give its customers on certain browser types and to also access application support requirements for particular parts of the country where it’s website is accessed from.

To demonstrate its execution, a PostgreSQL database was created to warehouse 16GB of actual log-files from client. The log files are parsed and details of unique users is added into the database. To extract certain location details from the data, a web scraper is used to determine the states where the users are accessing from. Due to the restriction on making unlimited calls to the website, some of the states were determined with the aid of a free database. Browser details are extracted from the data. Finally, some analysis is done on the data and plots are generated for visualization.

Key Files

The key files include the following;

1. ***Team\_16GB.ipynb*** – This is the ipython notebook file made up of the different source files that are included in the code.
2. ***DatabaseClass.py*** – contains a class to handle all of the database access. It gets used by the other python files.
3. ***project.py*** – the first step in the project implementation. It parses the log files and adds unique users into the database.
4. ***webscraper.py*** – scrapes a web site to determine the user’s State from the IP address.
5. ***ip\_update\_from\_db.py*** - uses a database to determine user’s State from the IP address.
6. ***extract\_browser\_info***.py – parses the user agent string to determine the type of browser and operating system that was used.
7. ***plot.py*** – does some analysis and generates the plots based on the data in the database.

Python Libraries

Python libraries utilized include; psycopg2, os, sys, datetime, re, urllib2, pandas, numpy, matplotlib, scipy.stats, collections and json.