Python program for analysis log file.

**[description]**

This program read data from zipped log files and then analysis those and save the result to mysql database.

**[HOW to USE]**

Copy file to that folder where is specified zip file. And run this program. This program read all files and performs function.

Type: console program

Language: python

If you want to use this program, first of all, install packages below.

zipfile, pathlib, dateutil, datetime, calendar, numpy, mysql, ip2geotools

First part of code, there are global variables to control the program.

So if you set variable **g\_dir\_name** to specified directory, this program will check that directory and analysis zipped log files.

Variable **g\_mysql\_db\_name** is db name.

Variable **g\_tbl\_\*\*\*\***s are table names.

Caculation will take some times.

Code explain

**Global functions**

Function is\_date is one to check string is date.

Function leap\_year is one to check year is leap year.

Function days\_in\_month is one to get dates in month.

**Global variables**

g\_timesteps is variable of time step you choose in doc file.

users is variable to save users who connect to the server.

dates is variable to get requests of every day.

dates\_pages is variable to get pages of every day.

**Class**

Class node is a class to explorer zipfile and read the logs.

Function set\_sub\_dirs is one to get subdirectories.

Function explorer is one to explore.

Function analysis is one to unzipped file and read lines one by one.

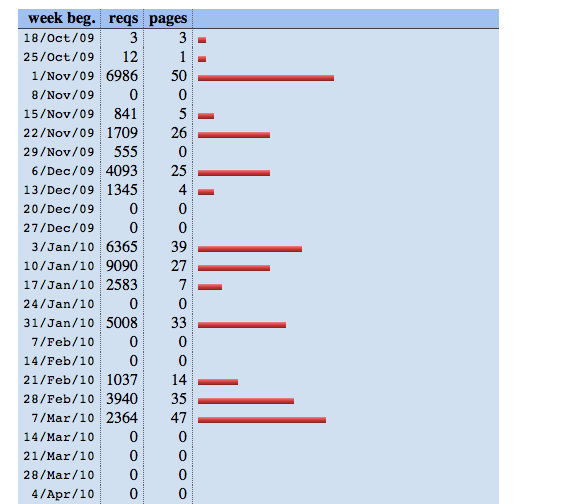
By class node, reading zipped log files is possible.

Class result is a class to save the read data to structured variable.

Class’s member variable csv\_data is variable to store all the logs except comments.

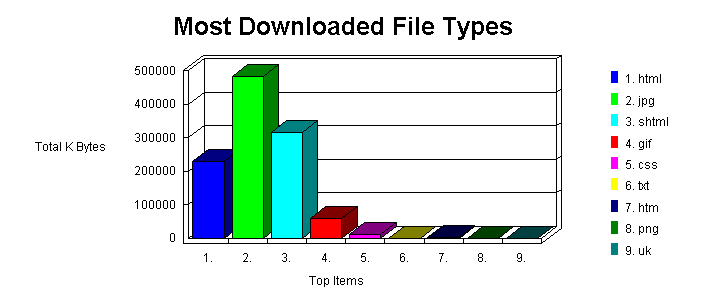
Codes of line 84 and 86 is function to get users list from log text.

Codes of line 88 to 109 is function for picture below



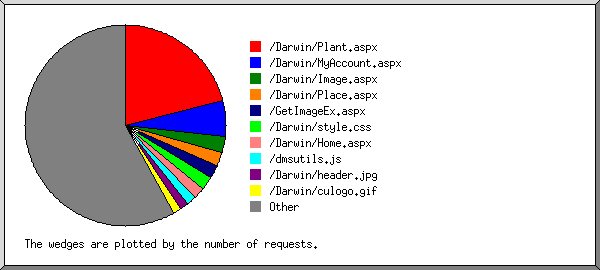
Function get\_cnt\_pages is to get download file types.

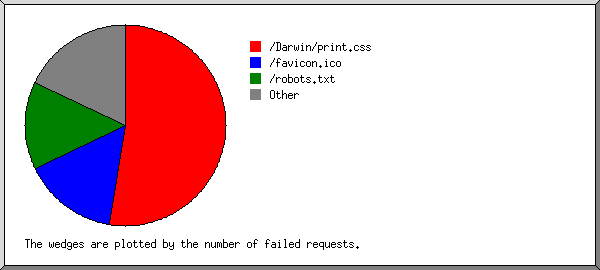
Image below and table are relative to this function. They are relative to each other.

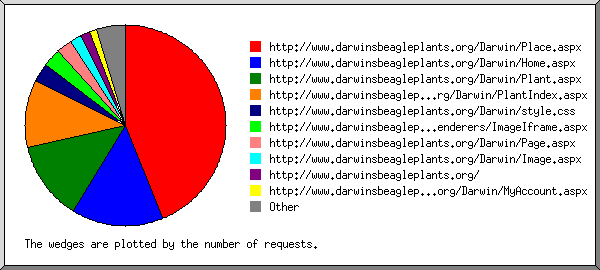


| **reqs** | **%bytes** | **extension** |
| --- | --- | --- |
| 35705 | 81.15% | .aspx |
| 30654 | 14.06% | .jpeg [JPEG graphics] |
| 2526 | 1.93% | .jpg [JPEG graphics] |
| 2170 | 1.43% | .css [Cascading Style Sheets] |
| 8336 | 0.49% | .gif [GIF graphics] |
| 2001 | 0.40% | .js [JavaScript code] |
| 1906 | 0.29% | [directories] |
| 2562 | 0.24% | .png [PNG graphics] |
| 2 |  | [not listed: 1 extension] |

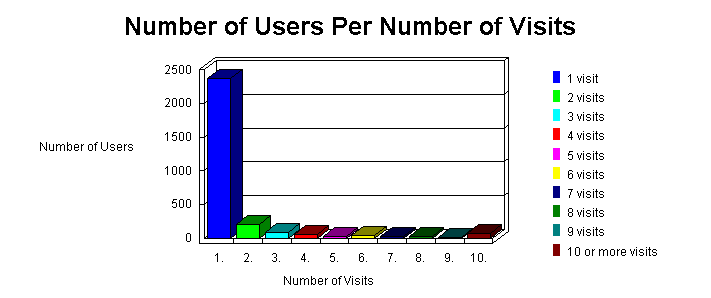
Function get\_cnt\_special\_pages is to count special pages.





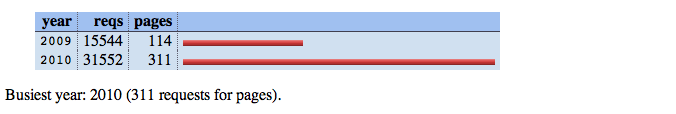


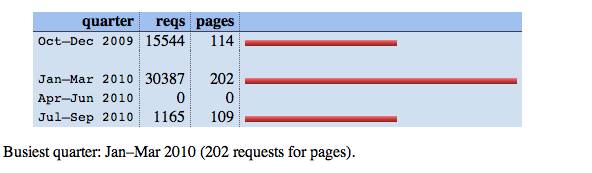
Function get\_user\_per\_numbers is to get number of users to visit

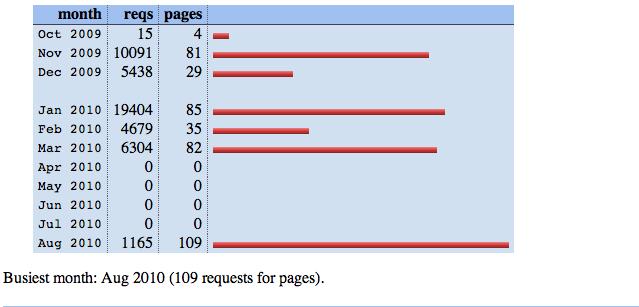


Function get\_year is one to get availiable years of those zipped log files.

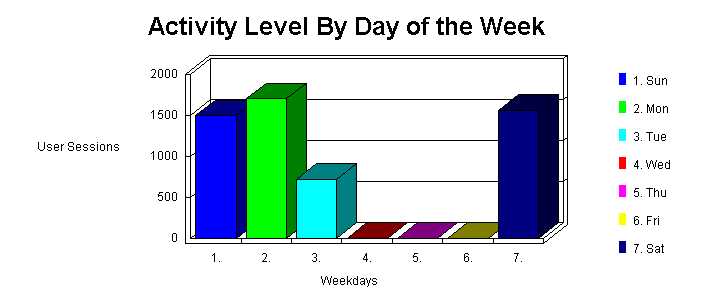
Function get\_req\_pages\_cnt is one to get requests and pages in definite period.



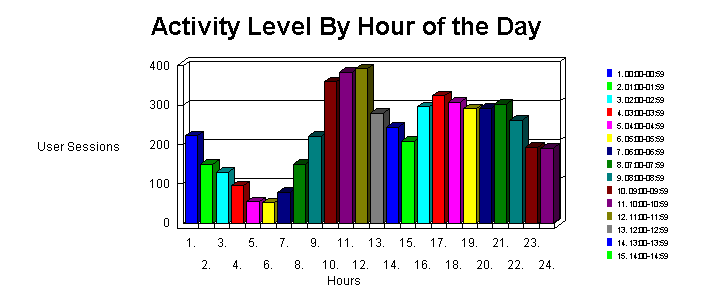




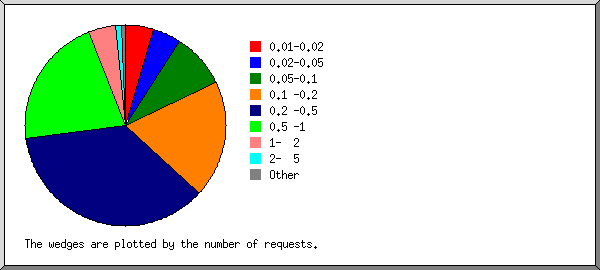
Function get\_activities is one to calculate activities.



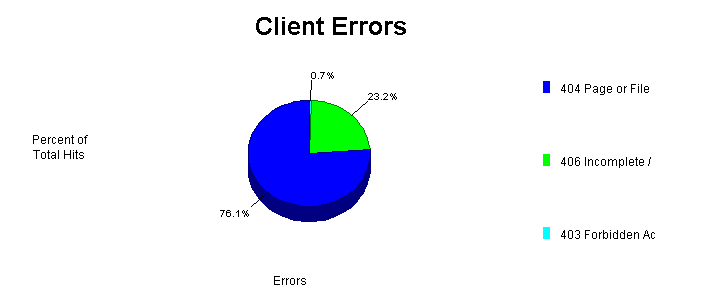
Function get\_activity\_level\_hours is one to calc activities of hours a day.

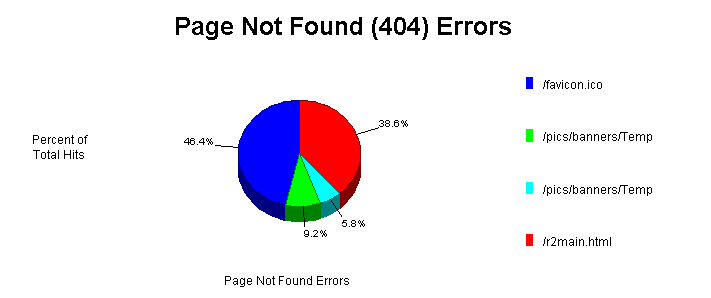


Function get\_time\_taken is one for image below.

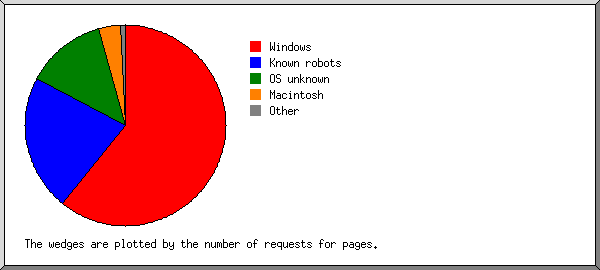


Function get\_clients\_errors is one to count errors of request.

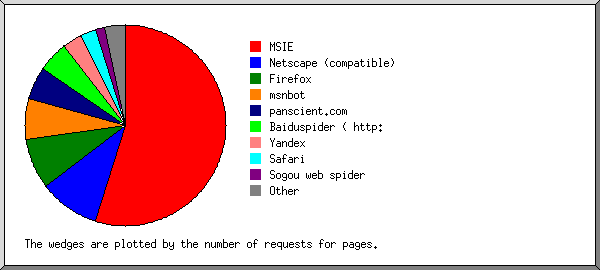




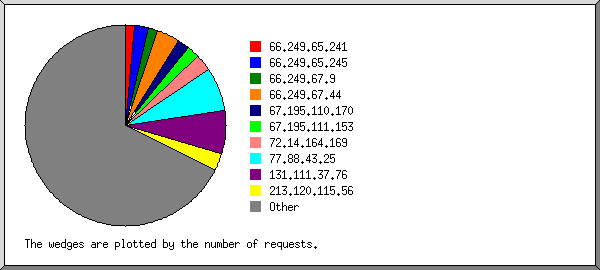
Function get\_os\_info is for below



Function get\_browser\_info is for image below



Function get\_host\_report is for image below



**My sql description**

If you run this code, db will be created and add the analysised data.

Dbname: pydb\_log\_analysis

Username: root

|  |  |  |
| --- | --- | --- |
| table | fields | desctiption |
| log\_download\_file\_types | filetype | Field for filetypes like .aspx, .php, jpg, css etc |
|  | counts | Count for those files |
| log\_users\_per\_visits | visits | Count of visits |
|  | users | Users fo visits |
| log\_access\_pages | urlss | Special urls |
|  | counts | Count for urls |
| log\_extension\_percent | extension | Extensions like aspx,jpg etc |
|  | reqs | Request of those extension |
|  | percent | Percent of this extension |
| log\_year\_request | year | Specified year |
|  | reqs | Count of request in this year |
|  | pages | Count of pages in this year |
| log\_month\_request | Month | Specified month |
|  | reqs | Count for requests in month |
|  | Pages | Count for pages in month |
| log\_day\_request\_pages | date | Specified date |
|  | reqs | Count for reqs on this date |
|  | pages | Count for pages |
| log\_weekday\_request | weekday | One of weekdays |
|  | users | Number of users in this weekday |
| log\_hour\_request | hours | Hour in a day |
|  | users | Number of users in this hour |
| log\_time\_taken | Timestep | Period of time step of time taken |
|  | request | Number of requests within this timestep |
| log\_client\_error | errorcode | Specified errorcode like 404,403, etc |
|  | counts | Count errors of this status |
| log\_os\_info | ostype | Type of client’s os |
|  | counts | Number of request from that os |
| log\_browser\_info | browsertype | Type of client’s browser |
|  | counts | Number of request from that browser |
| log\_host\_report | hostip | Host IP |
|  | counts | Number of that ips |