README:

Name: Programming Assignment 3

Purpose: A proxy server with Caching

Date: 8 Nov 2017

Python version: 2.7.14

To run the code:

**$ cd Downloads/proxy/**

**$ python v21.py 8080 100**

The program must get either one or two command line arguments. The first argument is the port no. It must be between 1025 and 65535. The second argument is default timer. It is expressed in seconds.

The program first checks the contents of blocked file. Then it collects the ‘CNAME’ of URLs in the blocked file and adds them to the blocked file.

The program creates a socket to listen to requests from client (browser). If a request arrives then a new thread is spawned. Multi-threading is used to server multiple requests. Target function of threading is “conn\_string”

This function extracts data from the request made by client. The various information it extracts is the name of webserver client wants to connect to, port no to which client wants to connect, method of data retrieval and HTTP version. If version is not HTTP/1.0 or HTTP/1.1 then an error is sent by proxy server to the client browser. If method is any other method than “GET” than a “not implemented” error is sent back to client browser. If the word “pokemon” is present in URL then the URL is blocked. If he requested URL is present in the blocked.txt file, then a HTTP 450 Blocked message is sent back to the client. If none of these errors occurs, then the proxy server function is called.

The proxy creates a hash of the URL, creates a file with this hash name. If cache file for requested URL is present, then it is checked whether cache file is still fresh or not. If it is expired, then tag of that file is sent to webserver to check whether the page contents have changed or if they are still the same. If etag verification fails, then the old cache file is deleted, and new cache is downloaded.

If cache file is not present for requested URL then the request is forwarded to main server, the page is loaded, and a new cache file is created. If “pokemon” is present in data, then the page is blocked. In either case, a new thread is spawned, for prefetching and a save cache function is called.

The save cache function creates new files and writes html data into them. It checks for “must-revalidate”, “etag”, “max-age” headers in data and updates the respective dictionaries accordingly.

Cache validation function is called to check if cache can be refreshed using etag. A HTTP GET message is sent along with the etag. If the server approves, then it sends a 304 message as response. These 304 responses indicate that the same cache can be used.

The prefetch function finds out all the http links present in the html file. All these http links are retrieved and to store these pages, we pass the URL to the save cache function.