Computer Networks

Project-2

Group 12

1. Avinash Gogineni
2. Arun Kumar Kalyankar
3. Santosh Rayala

**Multi-Threaded Web Proxy Server**

**Introduction:**

A web proxy is a software which allows to filter or monitor the traffic being sent/received between host machine(s) and the internet. This allows administrators to have a greater control on the access to the internet, thus enforcing an organization's policies.

When a host machine tries to connect to the internet, then the request for the webpage goes to the web proxy. Depending on the policies enforced, the configured web proxy will either allow or deny full access or partial access to the webpage. This is achieved by forcing all the webpage requests to go via the web proxy.

**Problem Analysis:**

To develop a multi-threaded web proxy. This web proxy receives an HTTP request for an object from a browser. It generates a new HTTP request for the same object and sends it to a remote server that is hosting the requested object.

When the proxy receives the corresponding HTTP response with the object from the remote server, it will create a new HTTP response, including the object, and sends it to the client browser.

The web proxy supports multiple requests at a time and has the following features:

1. Block access to the websites if they are found in the black list.
2. Filtering out inappropriate language from requested site.
3. Caching files that have already been requested.

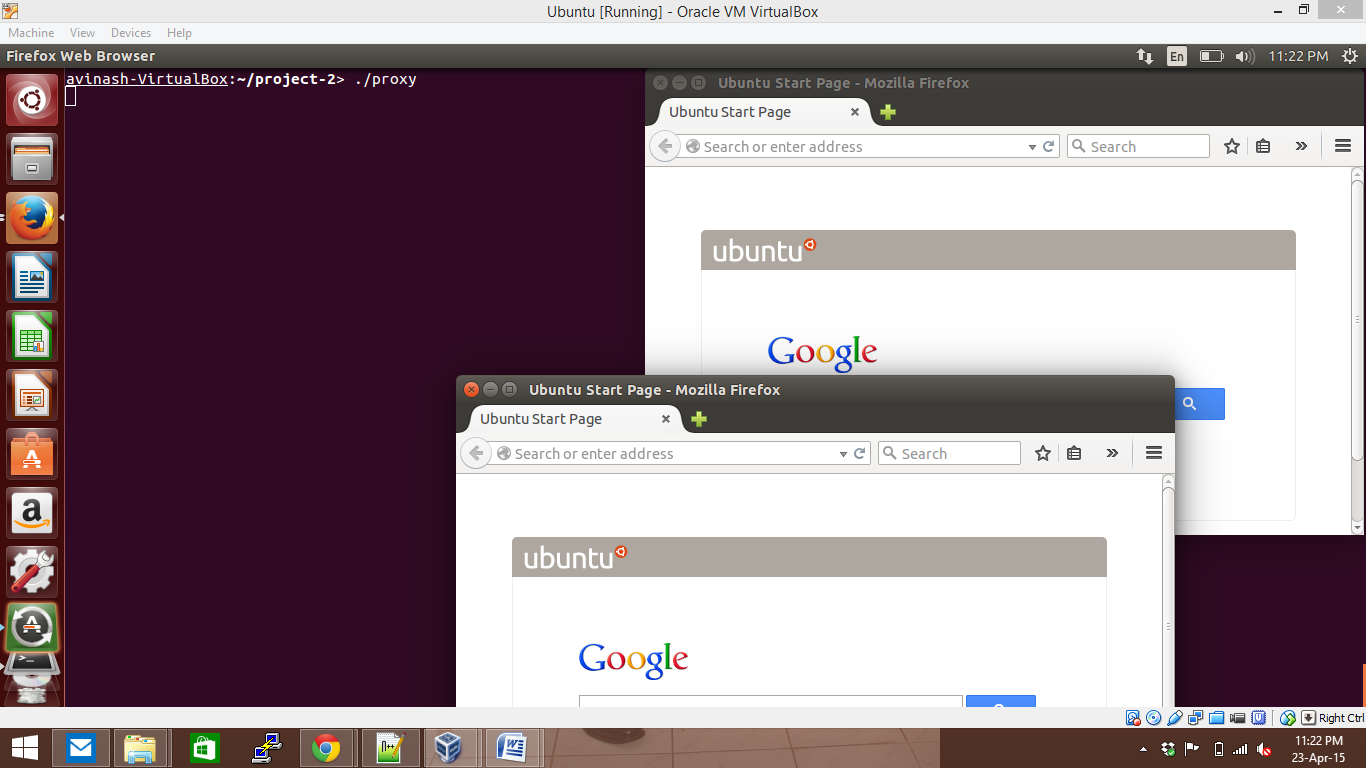
**Problem Approach and Solution:**

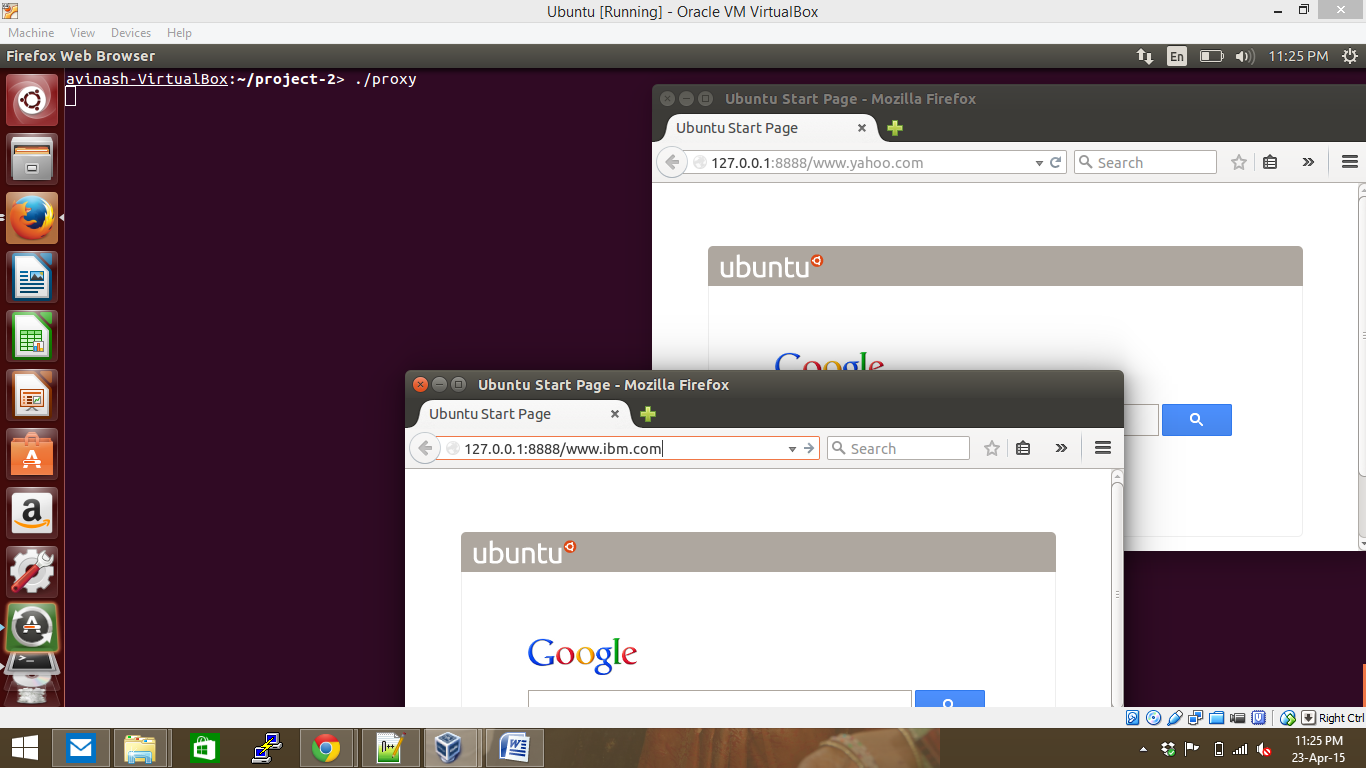
The web proxy "proxy.c" has been developed to provide the following features:

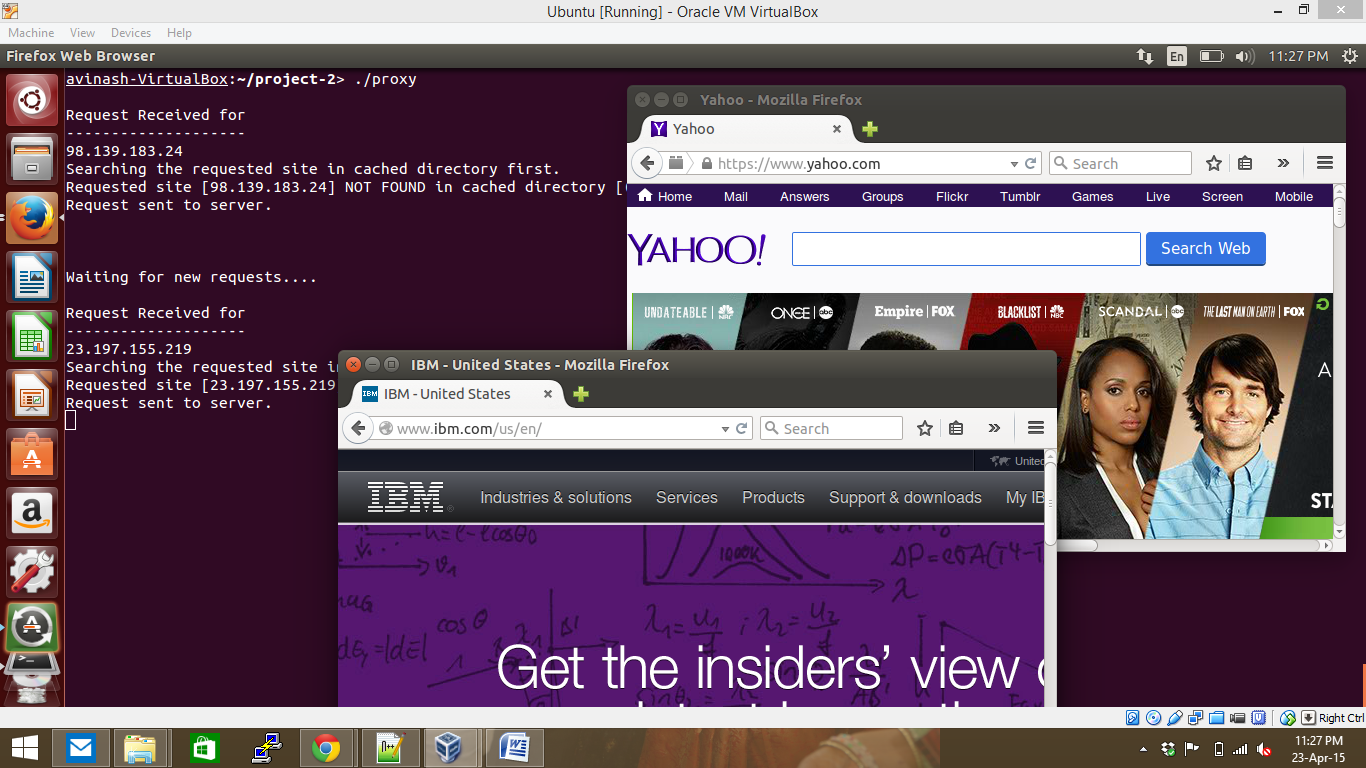
1. Block access to websites found in black list  
   For this, a file "blocked\_sites.txt" is provided. The administrator can add the websites/URLs for which the access should be restricted in this file. If the client tries to access a website/URL which is found in "blocked\_sites.txt" then a message saying "Access to this website is restricted" will be sent back to the client. Thus the client will not be able to access such websites.
2. Filtering out inappropriate language from the requested site  
   For this, a file "blocked\_words.txt" is provided. This file contains the list of swear words from the website "http://www.hyperhero.com/en/insults.htm". However an administrator can add or delete words from this file.  
   When the user accesses any website/URL then the web proxy will check the entire content of the website and if any words are found that are in "blocked\_words.txt" file, then they are replaced with "asterix" (\*) in the website and displayed to the client. Thus the client will not be able to see any words on the website which were found in "blocked\_words.txt" file.
3. Caching files that have already been visited  
   Caching of websites is used for faster access to the website. Some websites do not often change their content. Such websites can be accessed from the local caching copy if they had already been visited.  
   The decision to either request an object from the server where the object is hosted or to send the object that is in the caching folder is done using the HTTP headers.  
   Every website accessed over http protocol has a field called EntityTag or ETag.  
   When a client visits a website or accesses a resource for the first time, then the web proxy stores a copy of the website in "cached\_sites". The next time when the client tries to access the same website, then the web proxy sends a conditional request to the server where the resource is actually hosted.  
   This allows the web proxy to quickly compare the ETag that it has in its "cached\_sites" folder and the ETag from the server where the resource is actually hosted. If the ETags match then it means that the website/resource in the "cached\_sites" folder is up-to-date and the web proxy will then send back the html file from its "cached\_sites" folder to the client.  
   However if the ETags do not match, then the latest website content/resource will be sent back from the server (where the website is actually hosted) to the web proxy. This happens due to the conditional request which was sent earlier. The web proxy then stores this ETag and the website content in its "cached\_sites" folder.  
   In this way, the bandwidth of the internet connection is saved by using the "conditional request" of website content and efficiency of the web proxy is increased by caching visited sites for faster access in future.

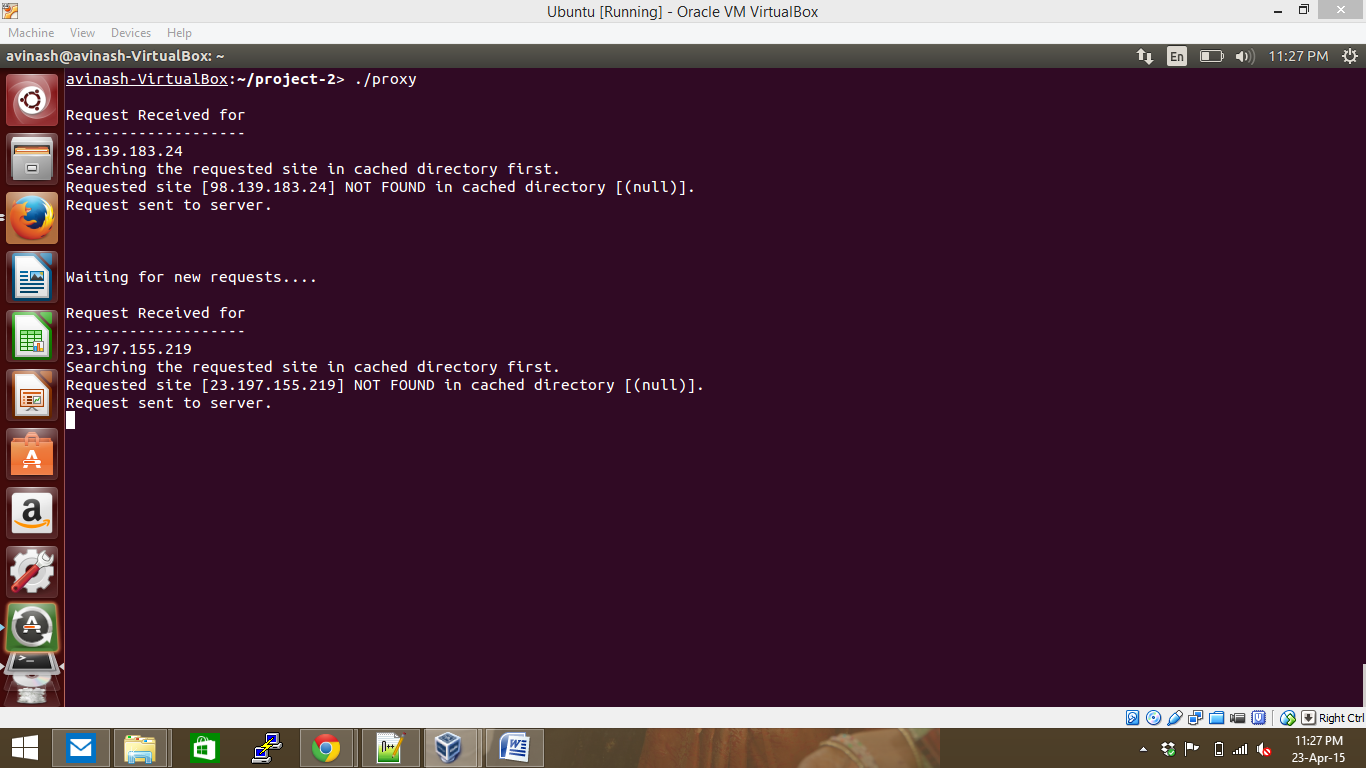
**Execution Screenshots:**

*Handling of Multiple requests*

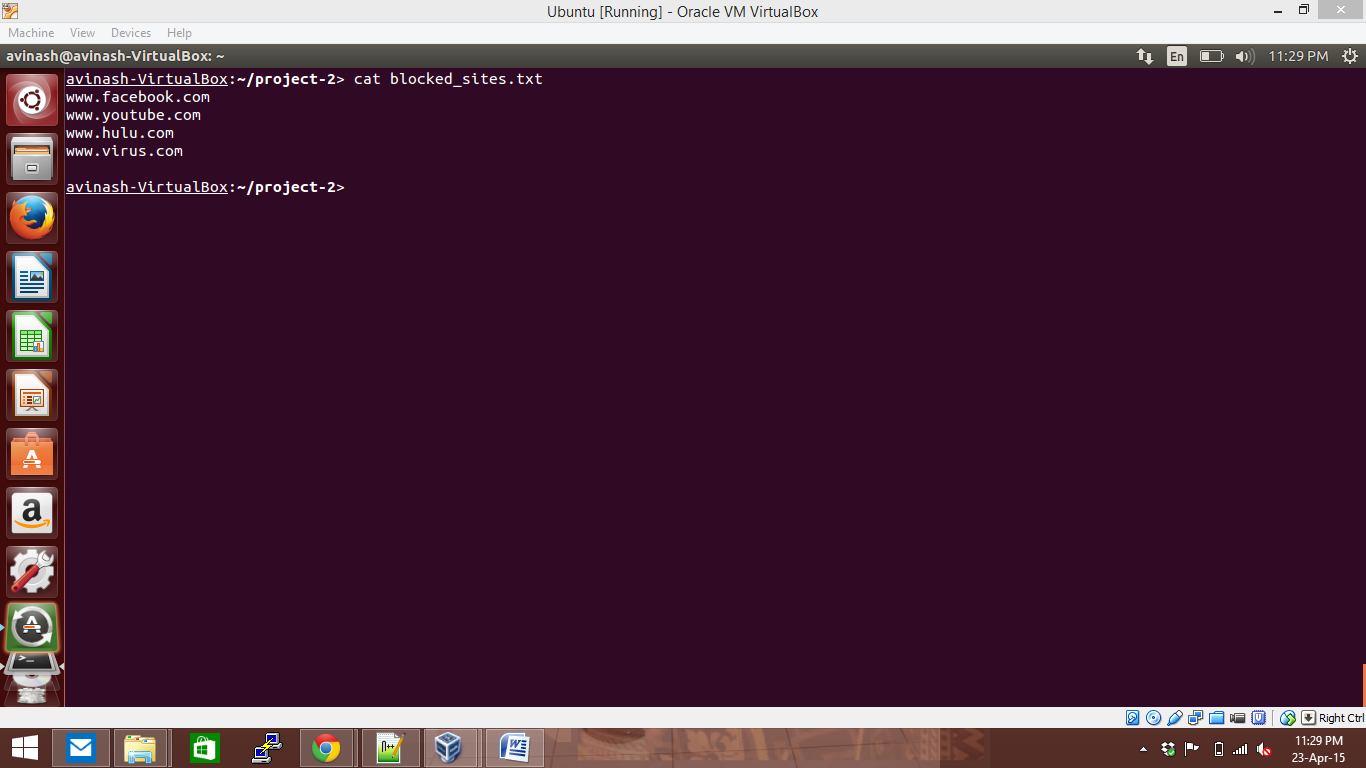


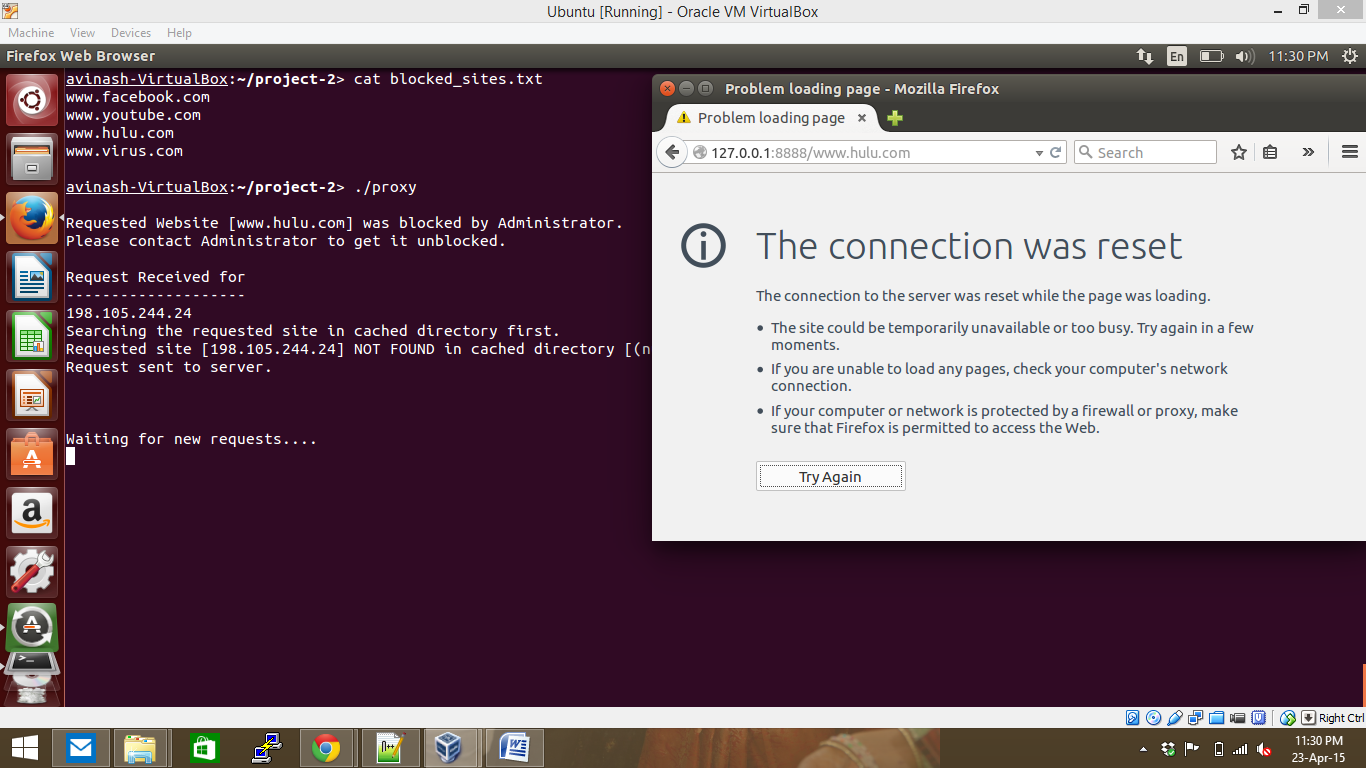




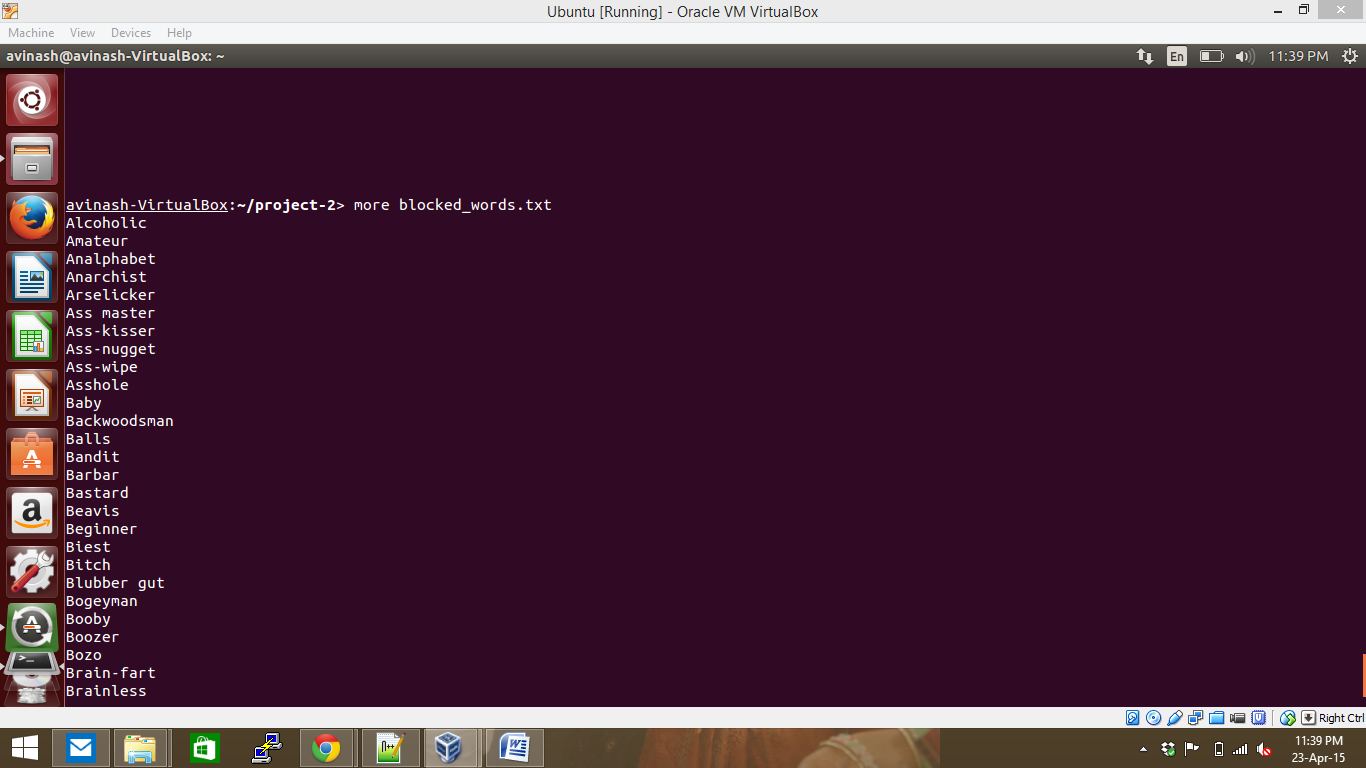


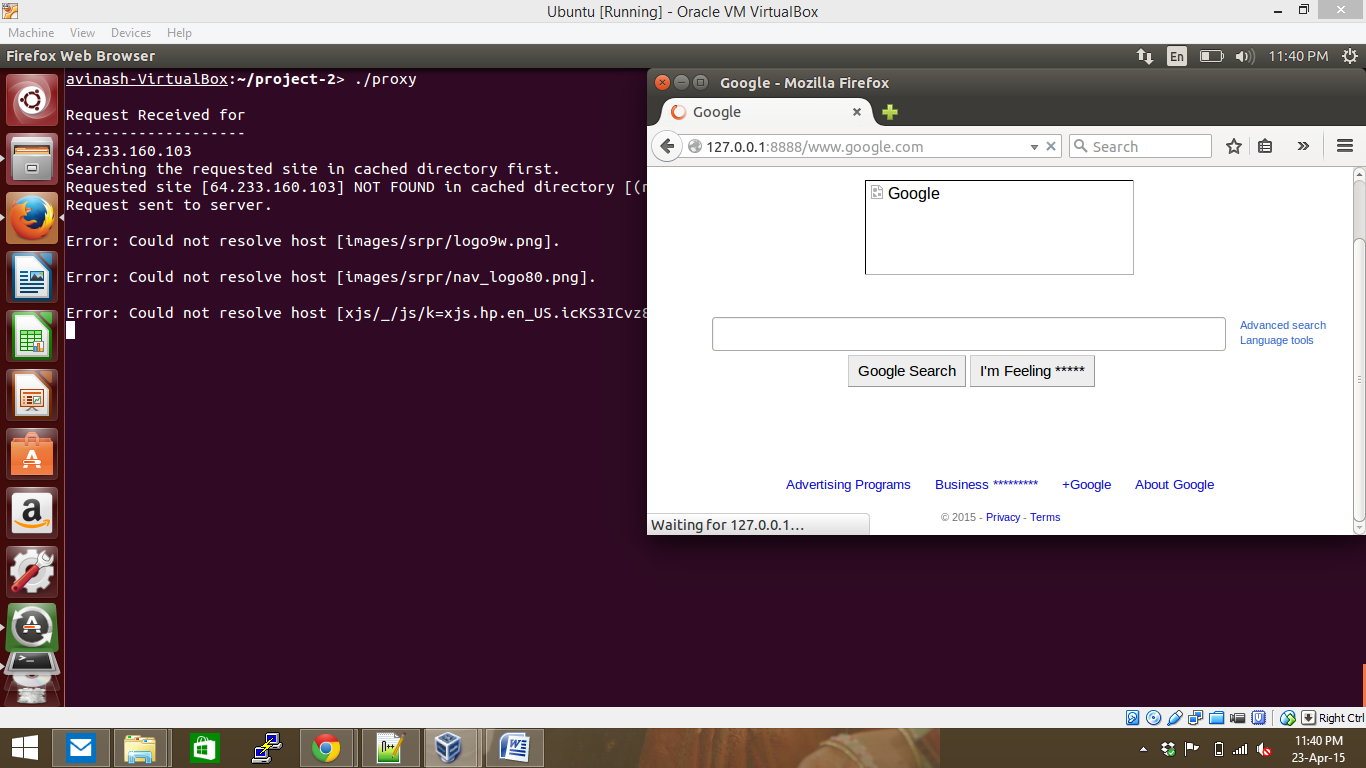
*Handling of Blocked Sites if found in "blocked\_sites.txt" file*

**

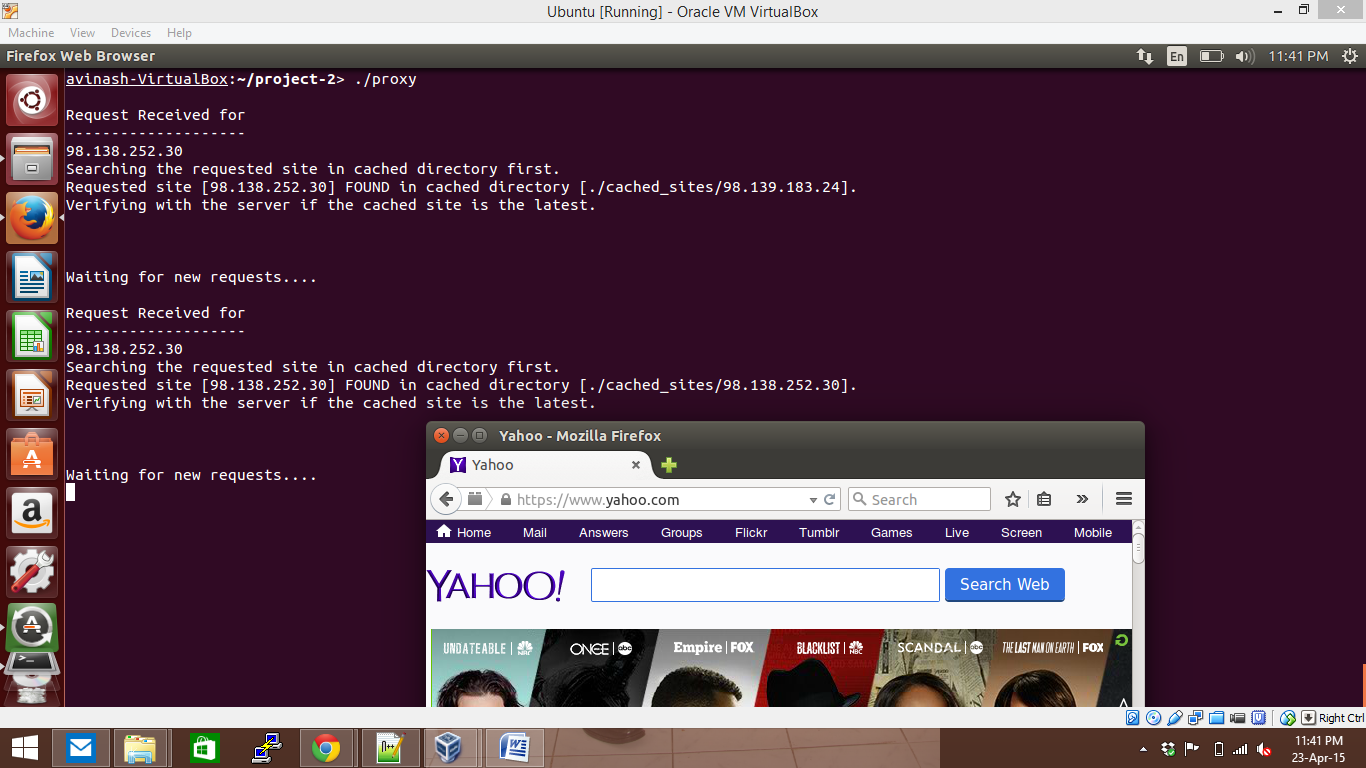
**

*Handling of words if found in "blocked\_words.txt" file  
Here for example, we have added two words "Lucky" and "Solutions" as swear words.  
Those two words are replaced with asterix (\*) in the displayed website.*

**

**

*Caching of already visited sites  
Here we can see that the website "www.yahoo.com" was already visited and hence instead of again requesting from the yahoo server, the website is displayed to the client from the "cached\_sites" folder.*

**