**20100117**

**Muhammad Abu Bakar Aziz**

**ANALYSING WEB PAGES OVER DIFFERENT NETWORKS**

**Websites Analyzed:**

I analyzed the following websites:

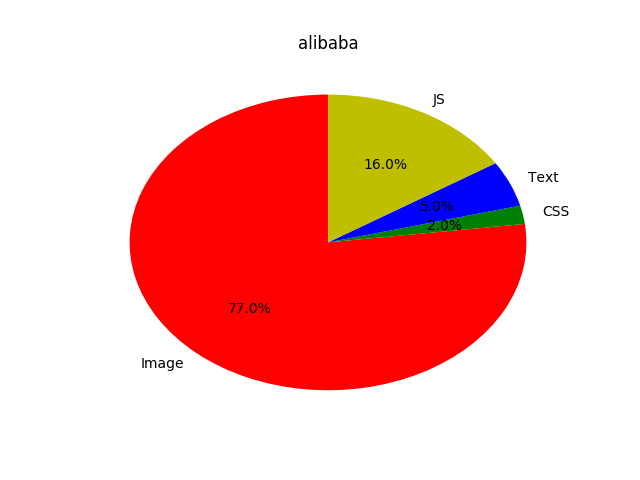
* [www.washingtonpost.com](http://www.washingtonpost.com/)
* [www.reddit.com](http://www.reddit.com/)
* [www.espncricinfo.com](http://www.espncricinfo.com/)
* [www.alibaba.com](http://www.alibaba.com/)
* [www.imdb.com](http://www.imdb.com/)
* [www.cnn.com](http://www.cnn.com/)

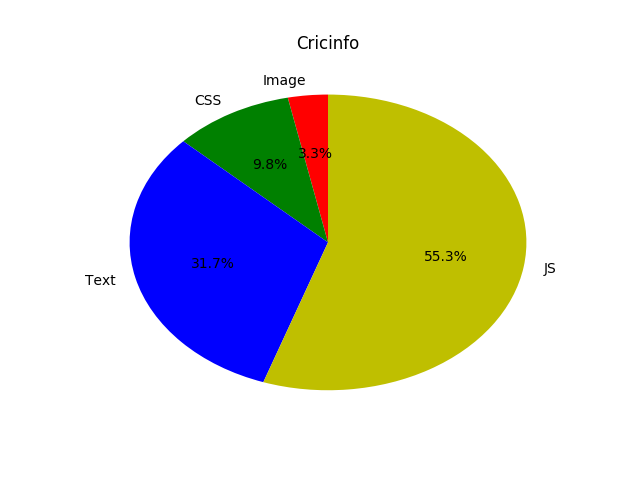
**Goal:**

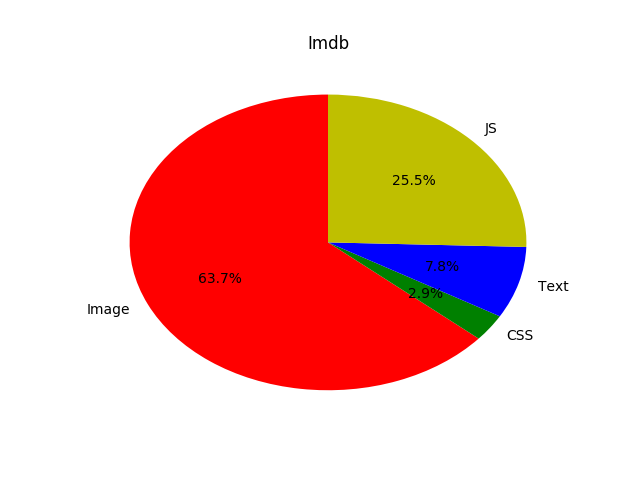
My primary aim was to find out which of the web page feature or type of object impact the load time the most and how we can improve the load time. I have only tested the features which I think will impact the most. Moreover, not all the graphs are shown here. There is a separate folder which contain all the graphs and pie charts.

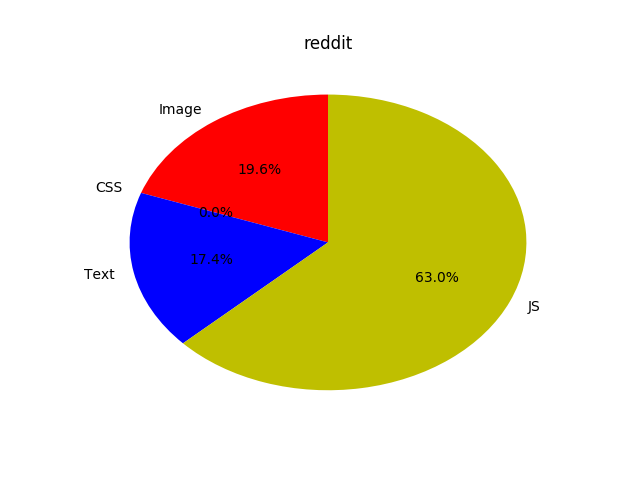
**1. Web Page Complexity.**

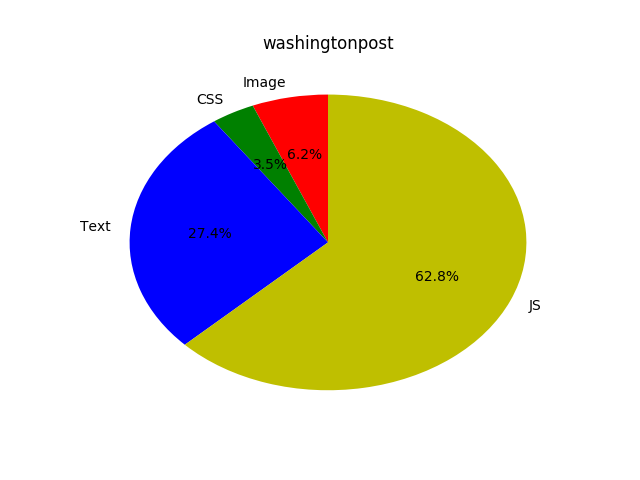
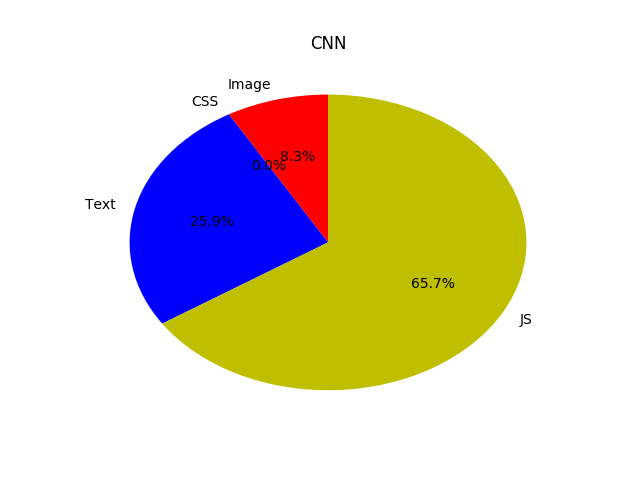
The websites consists of mainly four objects-Text, Images, CSS, and JS. Basically, majority of the websites content was dominated by Images and JavaScript. Below are the pie charts of all the websites showing different content and their respective percentages.







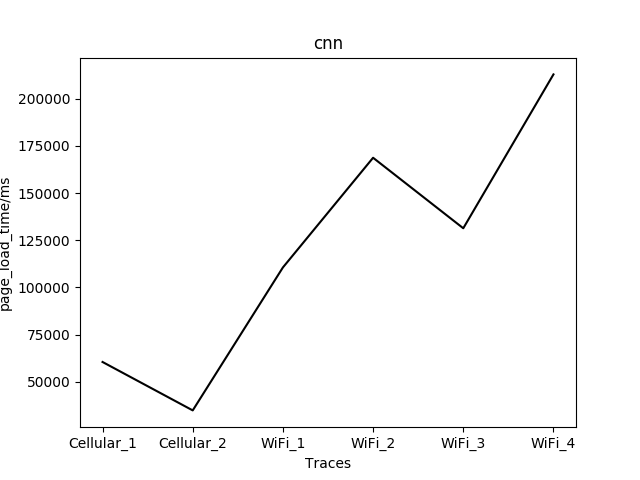
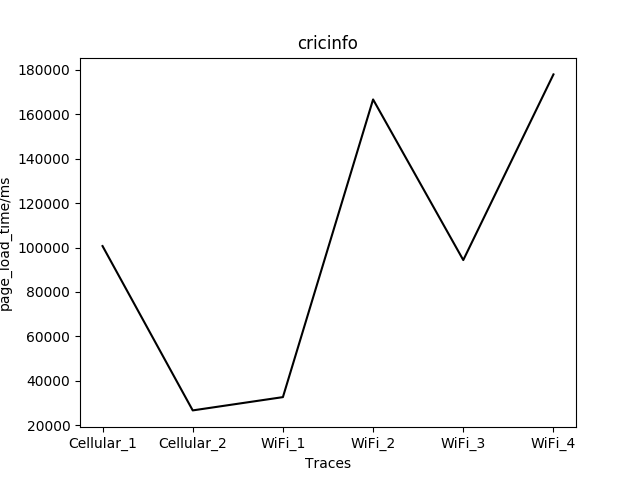


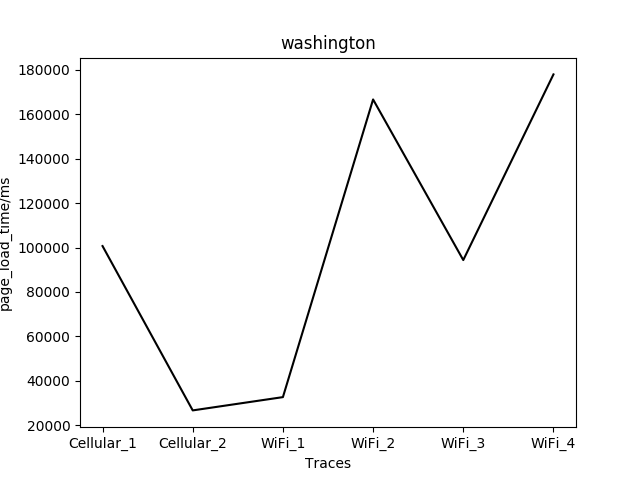
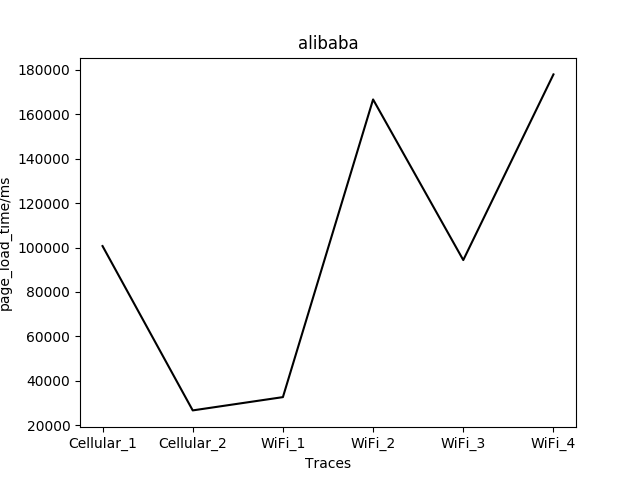


1. **Network Throughput and Page Load Time:**

**Method Employed:**

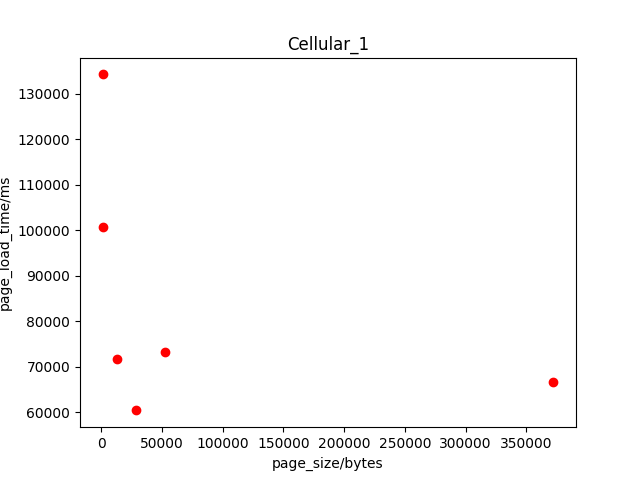
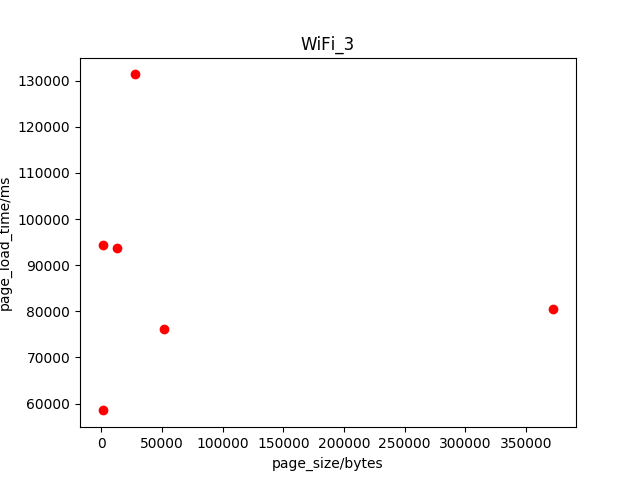
Each website load\_time was compared with different traces. For every website, I plotted graph of Page\_Load\_Time against different traces. Following are some of the graphs.

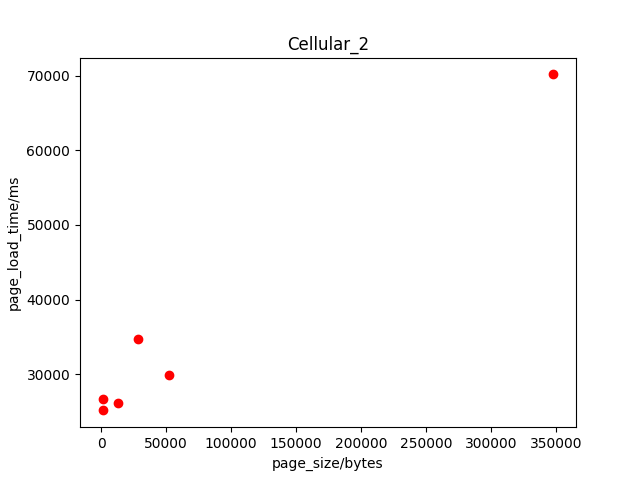
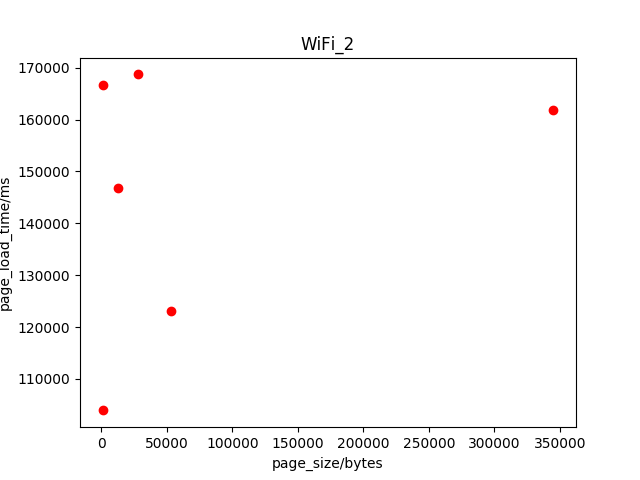
****

****

**Conclusion:** Results show that Cellular\_2 is the fastest as for every website load\_time is minimum. Moreover, WiFi\_4 and WiFi\_2 are slowest. For some websites , load\_page\_time was greatest on WiFi\_4 and for some WiFi\_2 was the slowest.

**II: Page\_Load\_Time vs Page\_Size**





On every trace, I compared Page\_Sizes of 6 websites with their Page\_Load Times. I tried to plot graphs of Page\_Load\_time vs Page\_Size. Following are some of the graphs.

**Conclusion:**

There is no direct relationship between Page\_Load\_Time and Page\_Size. Same Page\_Size for the same trace can have significantly very different Load\_Page\_Time values.

**Object with greatest Impact:**

Mainly Images and Javascript takes up most of the content on the websites. Following is the list of what type of content to remove from each website to decrease load time.

Websites : ObjectRemove:

[www.washingtonpost.com](http://www.washingtonpost.com/) JS

[www.reddit.com](http://www.reddit.com/) JS

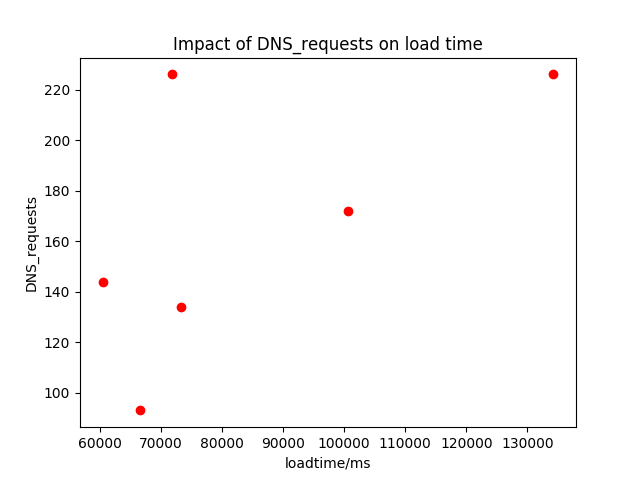
[www.espncricinfo.com](http://www.espncricinfo.com/) JS

[www.alibaba.com](http://www.alibaba.com/) Images

[www.imdb.com](http://www.imdb.com/) Images

[www.cnn.com](http://www.cnn.com/) JS

**I- DNS Requests:**

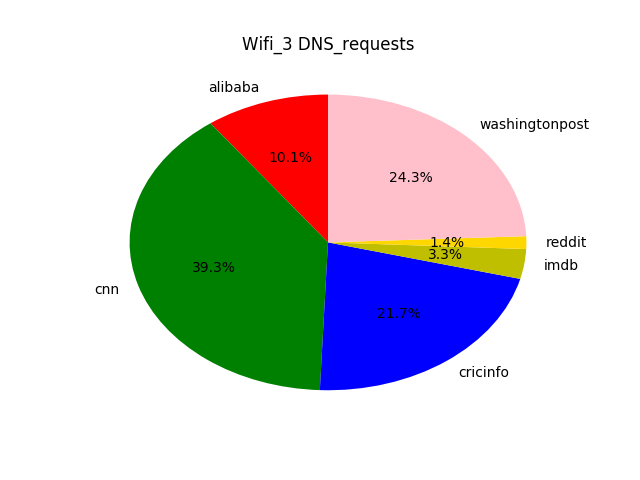
****

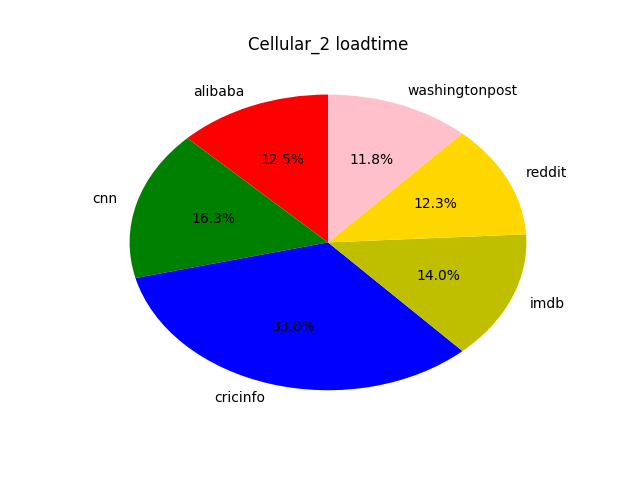
For some websites, the number of DNS requests made from external servers impact the load time. For example, DNS requests made by CNN is 226 and it has the highest load\_time. However, reddit also has 226 DNS requests but it doesn’t have highest load time. That means web load time is not entirely dependent on the DNS \_requests made by a website.

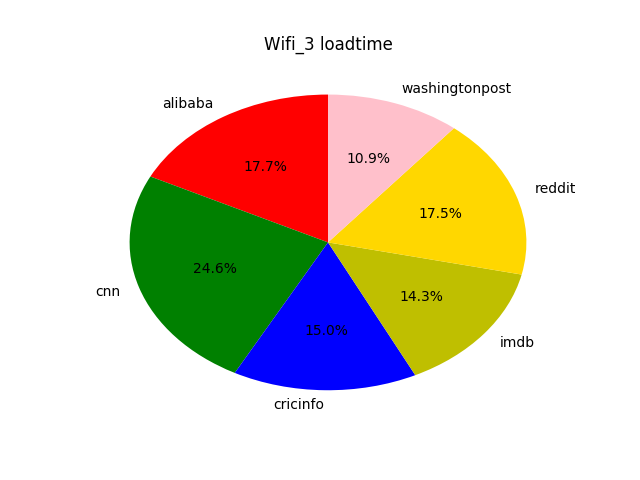
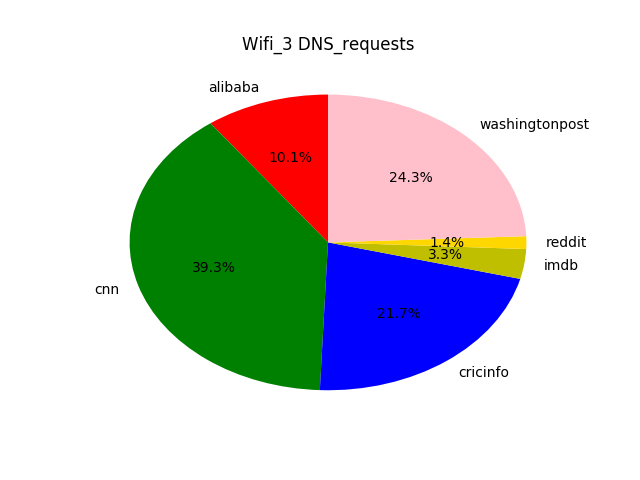
**DNS Look up time vs load time:**

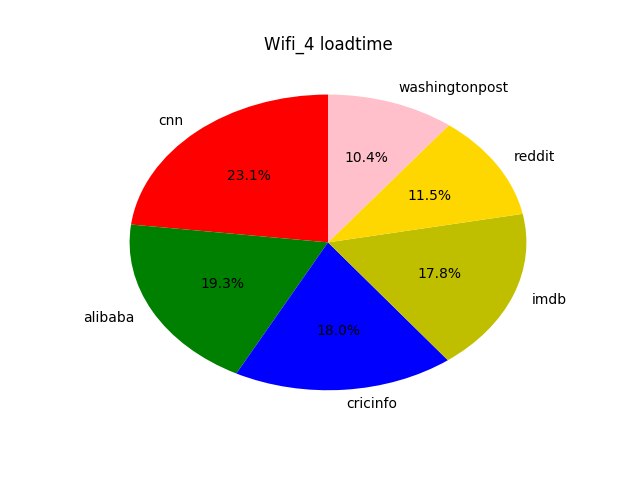
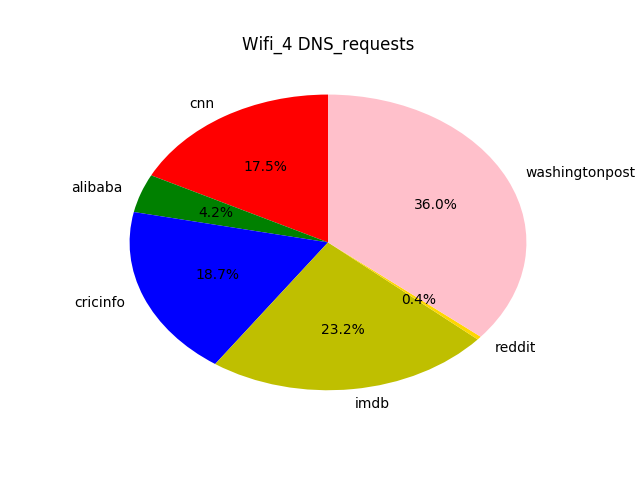
From the last experiment, we noticed that load\_time doesn’t depend entirely on DNS\_requests made by the web\_pages . In this experiment, I analyzed DNS look up time of each website on various networks and compared them with load times. Below pie charts contain results for three networks.

**CORRECTION: In below piecharts, the heading is DNS\_look\_up time not DNS\_requests.**





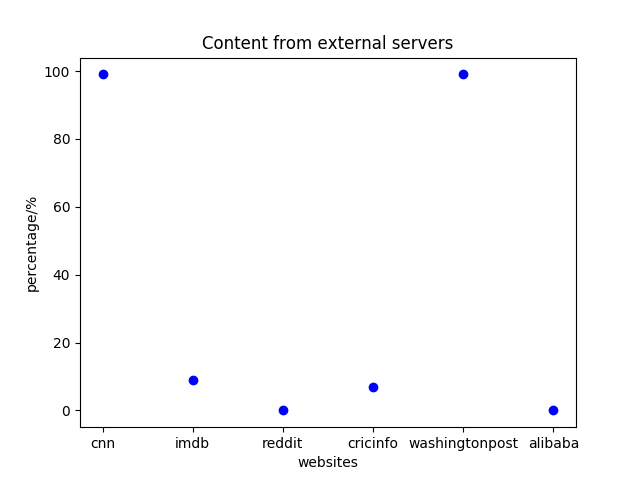


****

**Analysis:**

It can be seen that for some websites such as CNN load time does depends on the DNS lookup time made by the website. For example, for CNN and WashingtonPost, as DNS look up time increases, the load time also increases. Moreover, although increase in DNS load\_time increases the load time for many websites, the lower DNS lookup time doesn’t mean that load time will be lower. Moreover, in the last experiment we saw that DNS requests made by the CNN and reddit was 226 but their load times were different. Reddit has a less load time. This can be observed from above pie charts as reddit DNS look up time is very less compared to CNN even though both has the same number of requests. Hence, DNS look up time is more important factor in determining load time compared to number of DNS requests made by the web

**Role of External Servers**

I analyzed the HAR file and extracted the content individual websites requests from the external server. Following graph shows that mainly CNN and Washington Post requests all their content (nearly 99 %) from external servers. The content was mainly Images from external sources and text . Moreover, websites like CNN and WashingtonPost may have their internal servers very near to the requesting IP\_addresses. However, since they have majority of content from external servers, the closeness of the internal servers to may not matter much on the load time.

**Possible Improvements:**

More efforts should be made to make web pages lighter on traces which are comparatively slower than others. For example, Wifi\_2 and Wifi\_4 are the slowest. On these traces, the heaviest websites such as CNN and Cricinfo should be made lighter. Websites can be made lighter in many ways. For example, CNN spend much time on DNS look up. DNS look up time can be reduced by using the services of better DNS provider. Moreover, in these websites majority of the content is requested from the external servers. So, for slower traces we can can reduce the content from external servers by requesting only necessary information. Moreover, image load time takes most of the time of the websites like imdb. In these webites, if the throughput is very less, then there could be some mechanism by which number of pictures uploaded can be decreased and only download the necessary images. Images can also be further compressed.Moreover, on slower trace networks such as wireless sometimes not all the javascript functionality is needed. Some websites contain much Javascript functionality such as CNN and Cricinfo. This js can be minimized to show only basic functionality for slower networks.

**Note:** I also took some help from the Internet sources in understanding different parts of HAR file and parsing different content and how to improve DNS look up times. Moreover, basically there is one main script starter\_code.py that I used for different files. Moreover, not all the graphs that I made are shown in the report.