Data gathering

1. How did we collect the data?

   We collected the data using a Python script on the CSX servers. We only included two sample data files in our submission, as the all data in the zip format exceeded 90MB.
2. The Format we used to collect the data
   (Number:: Dataset name::Link::Size:: Title)
   1::google.com::http://google.com::206710::Google
   - Delimiter – " :: " -non English characters
3. What type of data that is there :
   Few notable titles:
   a. UNKNOWN – Any site that did not respond within 10 seconds or simply just unreachable, we set the size as -1, and the title as "UNKNOWN". For all sites that returned the page, and allowed us to get the page size but did not have correct tag for the page title, they also got the title as "UNKNOWN".
   b. Loading... – These websites had title as "Loading..." and some redirected to inappropriate sites. We looked at some sites that gave us "Loading" as the tile, and they had a JavaScript for a redirect.

How was the data Collected?

We used a python script (visist2.py). The script takes in two arguments: 1) Begin line# 2) End line#.

Example command to visit sites from line 1001 to 2000: "python visit2.py 1001 2000"

It reads the specified lines from top-1m.csv input file, and visits each site using the user agent of desktop Chrome browser on Windows 10. The use of user agent was done because some sites did not return a content if we did not provide a user agent.

It does not execute JavaScript, so if the site has moved and the page had JavaScript to redirect the visitor, it would not have been redirected. If the http request to the site resulted in http 301 redirect code, the script visited the redirected site. The script tries to visit site by appending "http://" in front of the hostname. If it results in error, it retries visiting the site with "https://". Each visit had maximum timeout of 10 seconds. We divided the job of visiting all million sites into 1000 jobs (one job is responsible of visiting 1000 sites). We ran 100 jobs on each of all 10 CSX servers (csx0 to csx9). We were able to finish visiting all millions sites within 2 hours. Each job generates one output file, so we have total of 1000 output files.

Among all million sites, our script was able to get size of the page from 883725 sites.

How to run on Hadoop ?

- Put those file in the Hadoop cluster and unzip.

Commands used for transfer files form one server to another.

//Transfer directory to another linux server

```
scp -r /home/username/unzip/ hadoop1:. :
```

// Copy directory contaning files to hadoop cluster

```
hadoop fs -copyFromLocal unzip / username /
```

- To run the code –Run the commands.

// Run the jar file

```
hadoop jar Homework2.jar Question1 / username /unzip /username/output1/1
```

// Copy the output file from hadoop cluster to local machine

```
hadoop fs -copyToLocal / username /output1/1/part-r-00000
```

// Open the output file with use of nano editor

```
nano part-r-00000
```

<u>Question 1</u>- List top 10 biggest websites

URL	Size in bytes
http://btc-i.com	313853015
http://cs-skini.me	278040455
http://n-azot.ru	38236923
http://kcsnews.co.kr	21803824
http://live.ci	17764488
http://boswtol.com	12928509
http://unj.ac.id	12733223
http://heeft-nieuwe-jobs.be	12704751
http://kyujin-ascom.com	11727699
http://likealyzer.com	11298813

<u>Question 2</u> – List top 10 most occurring unique titles

Title	# of occurrences
UNKNOWN	150692
Home	2243
Tumblr	1481
Sign in to your account	972
Not Found	931
Index of /	873
Loading...	782
Account Suspended	781
homepage	719
Главная	379