Abhishek Sinha CSCE 612

Homework Assignment part 3

1. The architecture of the code is as follows:
   There are two main classes **Crawler** and **Socket**.

   - **Crawler** class contains the parameters that are used to passed while creation and handling of multi-threaded crawling.
   - **Socket** class has the attributed and method used to open, communicae and close a TCP connection.

   There are two important functions - Producer and the Consumer.
   Producer function parses the file the stores the URL from the files into a queue,
   Consumer function picks the front of the queue, and checks the lights of the car. In the main function, user defined number of threads are created and are assigned the Conusmer function to execute. First the Producer thread is created which parses the URL and stores them in a Queue. Then n number of threads are created that crawl the URLs by poping from the queue. Also a status thread is created for printing of status

   Through this exercise I learnt the following:

   - Robots and their usage in the present day internet
   - Socket programming with windows, WSAEvent
   - Multi-threaded architecture of a web crawler, how to manage the producer consumer code, the Critical sections and about the events signaling
   - The homework also helped to practically implement HTTP requests and read their responses and filter them out based on our needs.

2. 4913 pages responded with a 2xx and the total number of links found were 91024. Thus average HTML links per page is 18. Thus on an average, each webpage contains 18 links which is quite reasonable
   Assuming Google crawls 10^{12} pages (1 trillion).Assuming the 11MB came from only the pages that responded with 2xx code, avg size of page is 2347 bytes.
   . On an average there are 18 links per page. Since each would be stored up as 64bit has, total size needed would be :
   $18 * 64/8 = 144 bytes$
   Thus total size needed to store 1 trillion pages as a webgraph would be (2347+144)*10^{12}/(1024^4)TB = 2265 Terabytes in space.
   Total number of edges each node would have is 18. Thus total nodes = (18+1)*10^{12} nodes.

3. The average size in bytes across all HTTP codes is 194 bytes (It seems very less, but this might be due to the fact that majority of the webpages responded with 4xx).
   Assuming the 11MB came from only the pages that responded with 2xx code, avg size

of page is 2347 bytes.
If Bing crawls 10^{10} pages a day, total bytes crawled is 2347*10^{10} bytes which is 21858.14 GB per day.
So bandwidth required by Bing would be $21858.14 * 8/86400 = 2.023$ Gbps (a day has 86400 seconds)

4. Total Links crawled = 999927
   Total unique DNS names looked up = 793687
   Total unique hosts = 98934 So probability that URL would contain unique host = 98934/99927 = 0.0989
   Probability that unique host has a valid DNS = 98934/793687 = 0.12
   Total robots contacted = 2883
   Total 4xx site = 46873
   So robots were 6.15% of the contacted sites.

5. To obtain information about how many crawled pages had tamu.edu domain, use the handle provided by the HTMLWebpageParser to iterate through the weblinks and check for the term ”tamu.edu”. For our case, only 1342 such cases were present.