Report for Problem No. 2

The following below given steps are involved in crawling.

a) **Algorithm used in implementing the Crawling is Breath First Search.**
b) **User will give the root node from which the crawling has to be start**
c) **Before opening the URL for the root node its hostname has to be fetched and check it should be either neu.edu or northeastern.edu**
d) **After Verification of the correct hostname, it will check for the URL content type, which should be html. For outgoing links it can be either HTML or PDF.**
e) **Now the page will be parse if it is HTML and will extract all the Out links from it and place it into an array**
f) **Now array has to be iterate and will open all the out links one by one after checking the hostname and content-type condition.**
g) **On visiting each Out link URL, all the links which are out link for that visited URL are maintained in a Hash which having Keys as Out Link URL which is parsed and Values are corresponding Links present in that page.**
h) **We do maintain the list of visited URL, check the visited URL condition before opening any new link.**
i) **Robot.txt is handled for the domains. Before visiting any URL it is checked whether is allowed by robot.txt for that domain or not.**
j) **Crawler is visiting only one page in 5 seconds.**
k) **Before putting the links into the text file they are being converted into the canonical form by eliminating # sign from the end.**
l) **The Text file can only maintain the list of 100 visited links.**