Assignment 2

Due Date: Jan 12th, 23:59pm, BR time.

This is a followup on the first assignment. Modify your program such that it receives an input file containing a list of "seed URLS", one URL per line. Each seed URL is a homepage and is considered at level 0. URLs that appear directly in the homepage are the URLs at level 1. For each one of the seed URLs, your crawler should collect all pages at level 1. It is not necessary to collect URLs at levels 2 or higher, but URLs of Web sites other than the seed URLs should be included in the crawling process. That is, your crawler will start with the seed URLs and will also crawl pages of sites other than the seed URLs. For each site crawled, your crawler should respect the Web protocol and do a delay of at least 100ms between consecutive requests to a same Web server. Your crawler should stop when it has collected 100,000 pages. Only pages in HTML need to be collected, that is, if you find a PDF file, for instance, skip over it and do not include it in your collection.

The architecture of your crawler should include a long-term scheduler and a short-term scheduler. The long-term scheduler should adopt a breadth-first scheduling strategy in order to increase the number of distinct sites crawled. The short-term scheduler should be multi-threaded so as to permit the concurrent crawling of distinct Web servers. Each thread should be associated with a particular Web server to permit properly controlling the time interval between consecutive requests to that Web server. At the end, your crawler should list:

(a) all Web sites crawled;
(b) the number of URLs at level 1 crawled for each Web site crawled;
(c) the average size of each page crawled on a per Web site basis;
(d) the average crawling time per page for each one of the sites crawled.

Each student will be provided with a distinct list of seed URLs to use in their program. The final run of the crawler, after debugging and testing, should be based on this list of seed URLs. Once that run is concluded, you should write a small report explaining what you have done, include your source code at the end, as well as basic statistics on the output of your crawler (number of seed URLs visited, number of URLs at level 1 crawled for each seed URL, average page size crawled per Web site). Most important, your collection of 100,000 pages should be stored in a local disk, it will be used in the next assignment.