GE Web Crawler Assignment: User's Guide and Design

How to use WebCrawler.exe

Simple validation check: Just run the WebCrawler.exe as you would any other exe, it will automatically test the Internet1 and Internet2 .json files.

Additional test cases: A single argument may be given to WebCrawler.exe in the form of a filepath. If you input runtests as an argument, the tests on the Internet1 and Internet2 .json files will be ran.

Design

I set out to make this crawler as fast and simple to read as possible. A concurrent dictionary (hashmap) was used to accomplish both of these goals. The concurrent dictionary allowed hashtable based search to rule out previously visited pages in log(n) time (instead of up to n time with arrays), the uniqueness of a page, and synchronization handling in one data structure.

Additional methods and members are present for decomposing the dictionary to success, skip and failure arrays, should this be a necessary operation in the broader scope of the system. For n pages, filling the arrays will take O(n) time. For the purposes of this assignment such arrays were unnecessary, and would have provided complications in performing concurrent crawls.