My crawler uses the root(http://www.ccs.neu.edu) as the start URL. And the delay time is 5 seconds , the domain name is “neu.edu/northeastern.edu”.

Firstly, the crawler will respect robots.txt, then detect whether the page has been visited is an valid link and whether the link is in the required domains. Secondly, the crawler will check whether the page is html of PDF using the appropriate content-type http header. In my crawler, after fetching pages it uses set to filter the pages has been visited automatic. I define an argument “count” to calculate the quantity of pages , the it reach 100, the crawler will stop and output the file named”result.txt” to save the crawling links.

The strength is my crawler can exactly fetch the links in the required domains and ignore the outgoing links, In addition, it can also filter the repeated links. In my crawler, I haven’t realized converting links to a canonical form. If the robots.txt denies the crawler to fetch its pages then the crawler won’t work in required domains.