MULTITHREADED CRAWLER

By,

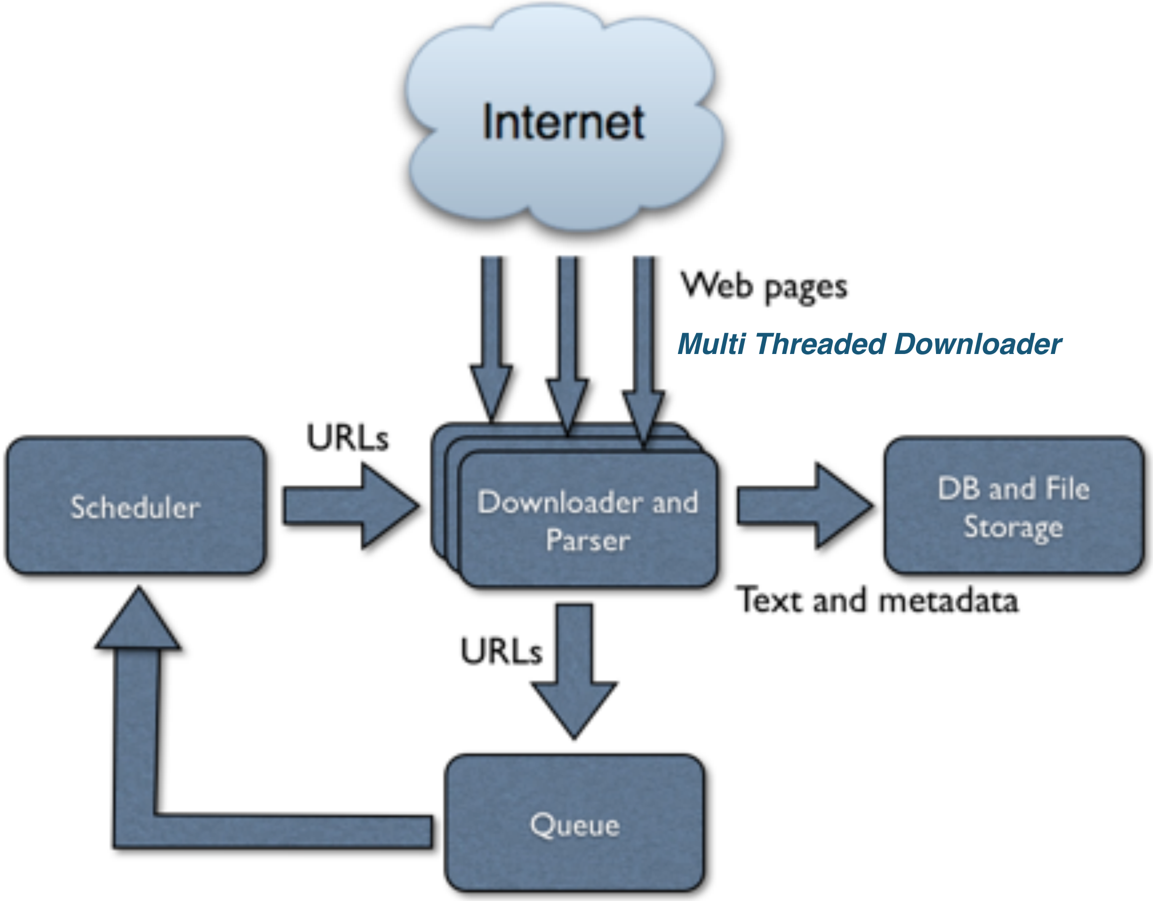
Piyush Mangrulkar - 861245524
Vikash Kumar - 861245614

INTRODUCTION

A **Web crawler** is an Internet bot which systematically browses the World Wide Web, typically for the purpose of Web indexing. A Web crawler may also be called a **Web spider**, an **ant**, an **automatic indexer**, or (in the FOAF software context) a **Web scutter**.

OVERVIEW

A web crawler finds and downloads all the pages on a websites, perhaps to archive or index them. Beginning with root URL (seed URL), it fetches each page, parse it for links to pages it has not seen, and adds the new link to queue (called frontier). When it fetches a page with no unseen links and the queue (frontier) is empty, it stops.



ARCHITECTURE

We have hasten the Crawling process by downloading many pages concurrently. As the crawler finds new link, it launches simultaneous fetch operations for the new pages on separate sockets. It parses response as they arrive, adding new link to the *frontier.* There may come some point of diminishing returns where too much concurrency degrades performance, so we cap the number of concurrent request (that is number of the threads - which is specified by the user), and leave the remaining links in the queue until some-in fight requests compete.

THE CRAWLING / DATA COLLECTION STRATEGY

-Saving A Page

Below are the steps followed
The two main tasks of Web Crawler

i) Saving data from URL to a file
ii) Extracting hyperlinks (Here, we extracting only hyper links of .edu domains)

-While crawling, below steps are followed:

i) Getting seed URLs from the specified text file
ii) Storing these URLs on the frontier (queue)
iii) Taking one of these URLs and using Java classes like URL, InputStreamReader, BufferedStreamReader we download the page and save the pages on disks.
iv) Parsing the robots.txt file of each URL
    i) Check whether particular site has permission to crawl or not.
    ii) If Site permit crawling then extracting all the Directories, Sub directories, pages which the site has not allowed for crawling.
    iii) Parsing Robot.txt & Meta tag and check whether the crawler has permission for indexing and Follow / Unfollow the links in that web page.
v) Parsing the contents of page
    i) Extracting the hyperlinks of .edu domain
    ii) Handling the relative URLs in the page
    iii) Handling the improper URLs in the page.
vi) After parsing putting all the URLs (which has permission to crawl as specified by the robots.txt and Robot Meta tags) into the frontier.
vii) Calculating check sum of the file and stored in the checksum map. If any new file has same checksum than it is a duplicate page and thus will not be stored.

-Parsing & Extracting Hyper links

Use the static Jsoup.parse(String html) method, or Jsoup.parse(String html, String baseUri) if the page came from the web to get the absolute URLs.

COLLABORATION

i) Download Page Contents - Piyush
ii) Check for Page Duplicates (MD5) on disk - Vikash
iii) Parse downloaded file - Piyush
iv) Crawler Ethics (Robots.txt, META) - Vikash
v) Clean extracted URLs - Piyush
vi) Multithreading - Vikash
vii) Normalize recursive Links - Piyush
viii) Batch File creation - Vikash
ix) Report - Piyush

LIMITATIONS

i) Non static pages are not handled
ii) Crawler can only crawl the web pages which follow “http” protocols.
iii) FTP, HTTPS protocols are not handled.
iv) PDF, Images links are discarded from crawling.

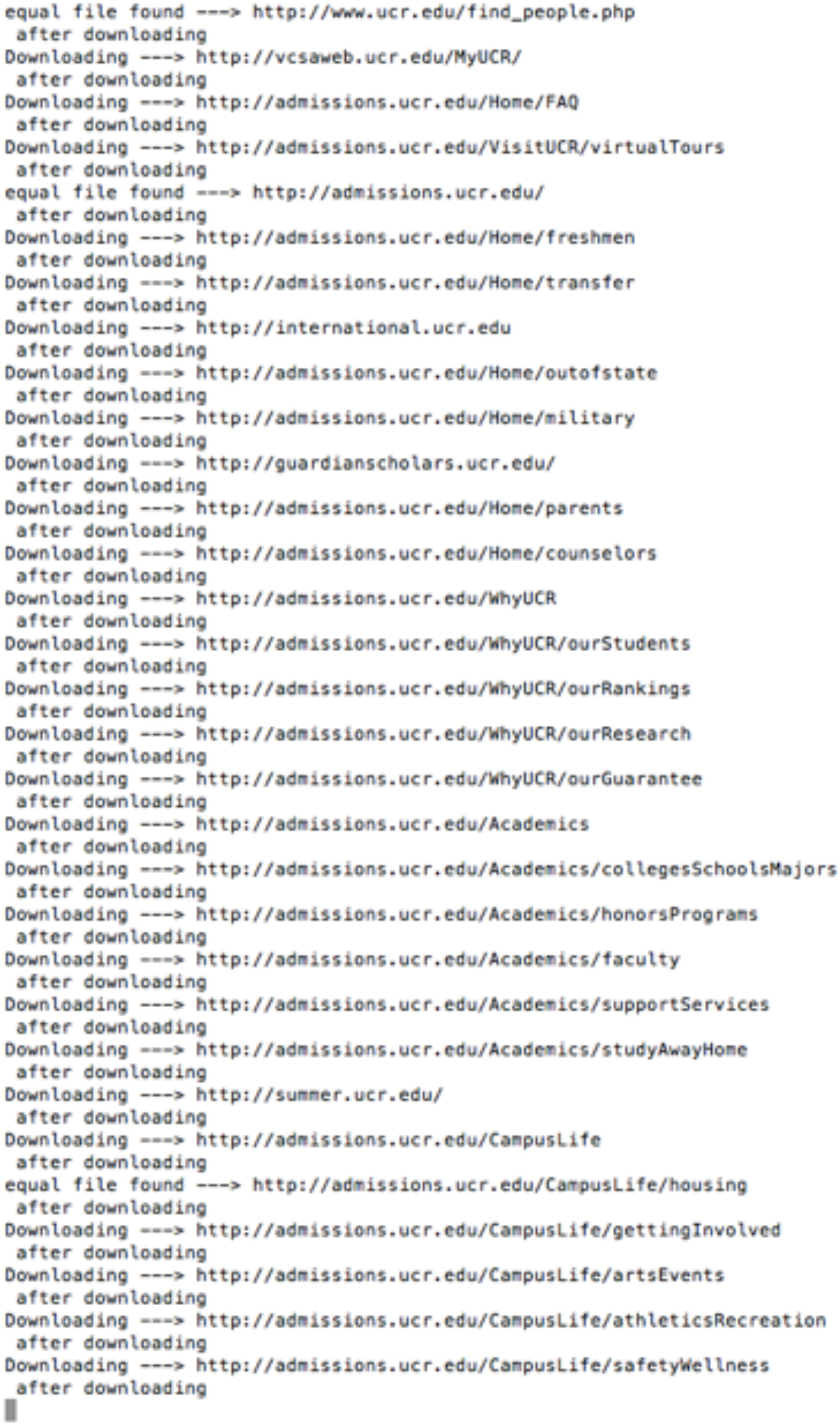
EXTRA FOR BONUS POINTS

i) Crawler is Multi threaded.
ii) Duplicated page handling is there. Used MD5 for calculating the checksum and store on the map.
iii) Connection timeout of 3 seconds, if page is not able to connect

DEPLOYMENT INSTRUCTIONS

i) Make these changes in ‘crawler.sh’ batch file.
ii) Provide the JAR path of multiThreadedCrawler.jar
iii) Provide 4 parameters - No of hops, No of pages, Document Store path, Seeds.txt path
iv) A 5th parameter is optional - Number of Threads, by default Max Thread Count is 100
v) On Terminal run this command - $ sh crawler.sh

SCREENSHOT



```
equal file found ---> http://www.ucr.edu/find_people.php
 after downloading
Downloading ---> http://vcsaweb.ucr.edu/MyUCR/
 after downloading
Downloading ---> http://admissions.ucr.edu/Home/FAQ
 after downloading
Downloading ---> http://admissions.ucr.edu/VisitUCR/virtualTours
 after downloading
equal file found ---> http://admissions.ucr.edu/
 after downloading
Downloading ---> http://admissions.ucr.edu/Home/freshmen
 after downloading
Downloading ---> http://admissions.ucr.edu/Home/transfer
 after downloading
Downloading ---> http://international.ucr.edu
 after downloading
Downloading ---> http://admissions.ucr.edu/Home/outofstate
 after downloading
Downloading ---> http://admissions.ucr.edu/Home/military
 after downloading
Downloading ---> http://guardianscholars.ucr.edu/
 after downloading
Downloading ---> http://admissions.ucr.edu/Home/parents
 after downloading
Downloading ---> http://admissions.ucr.edu/Home/counselors
 after downloading
Downloading ---> http://admissions.ucr.edu/WhyUCR
 after downloading
Downloading ---> http://admissions.ucr.edu/WhyUCR/ourStudents
 after downloading
Downloading ---> http://admissions.ucr.edu/WhyUCR/ourRankings
 after downloading
Downloading ---> http://admissions.ucr.edu/WhyUCR/ourResearch
 after downloading
Downloading ---> http://admissions.ucr.edu/WhyUCR/ourGuarantee
 after downloading
Downloading ---> http://admissions.ucr.edu/Academics
 after downloading
Downloading ---> http://admissions.ucr.edu/Academics/collegesSchoolsMajors
 after downloading
Downloading ---> http://admissions.ucr.edu/Academics/honorsPrograms
 after downloading
Downloading ---> http://admissions.ucr.edu/Academics/faculty
 after downloading
Downloading ---> http://admissions.ucr.edu/Academics/supportServices
 after downloading
Downloading ---> http://admissions.ucr.edu/Academics/studyAwayHome
 after downloading
Downloading ---> http://summer.ucr.edu/
 after downloading
Downloading ---> http://admissions.ucr.edu/CampusLife
 after downloading
equal file found ---> http://admissions.ucr.edu/CampusLife/housing
 after downloading
Downloading ---> http://admissions.ucr.edu/CampusLife/gettingInvolved
 after downloading
Downloading ---> http://admissions.ucr.edu/CampusLife/artsEvents
 after downloading
Downloading ---> http://admissions.ucr.edu/CampusLife/athleticsRecreation
 after downloading
Downloading ---> http://admissions.ucr.edu/CampusLife/safetyWellness
 after downloading
```