NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR
Cachar, Silchar

B.Tech. VIIth Semester
Artificial Intelligence, CS-401

Subhojit Ghimire, 1912160
Computer Science and Engineering, CSE-B

ABSTRACT:

All the search engines we have seen and used so far use a concept what is known as indexing. There are trillions of websites spread across the Internet, each with its own vast sea of content. It would be heavily time consuming and nearly impossible to go through all of the contents and list the ones searched by the user in the search engine. To minimise this effort, web crawler(s), aka spider(s), are used in these search engines to index the content of websites scattered all across the Web. In this project, we see how a search engine functions with the help of a highly powerful tool called Nutch. Apache Nutch, as annotated by the official website itself, is a highly extensible, highly scalable, matured, production-ready Web crawler which enables fine grained configuration and accommodates a wide variety of data acquisition tasks. We took help of Apache Tomcat to host Nutch as a HTTP web server from our local machine. We crawled and indexed a website and stored the resultant database on our local machine and analysed the search result, all using Apache Nutch.

STEPS INVOLVED:

1. Run and Install JDK's "jdk-8u341-windows-x64.exe" at its default location, with default settings.
2. Run and Install Tomcat's "apache-tomcat-6.0.37.exe".
    a. In "Choose Componenets" setup window, select Host Manager and click Next.
    b. In "Configuration" setup window, set Username and Password as required. DO NOT leave those fields empty. Leaving Username:Password fields empty will call for extra changes later on, which is burdensome. This username:password will be required to log into localhost. Set it as admin:admin, but do not leave it empty, and click Next.
    c. Leave other setup windows as it is, default. Complete the install. Run Configure Tomcat and Start hosting.
    d. In web browser, go to address http://localhost:8080 and if the Apache site opens, the tomcat is running properly.
3. Run and Install Cygwin's "setup-x86_64.exe". This is Cygwin, a clone of linux terminal for windows. An alternative to Cygwin would be Hyper Terminal or even WSL with Ubuntu or other Linux distro.
    a. In "Choose a Download Source" window, select Install from Internet. Click Next and leave everything else default.
    b. In "Select your Internet Connection" window, leave it selected to Use System Proxy Settings, or choose as required. Click Next and wait. If error occurs, re-check the internet connectivity and try again.
    c. In "Choose a Download Site" window, select any one of the provided mirror sites, or recommended "https://cygwin.mirror.constant.com" and click Next and wait for the Download to complete. If error occurs, Retry.
    d. In "Select Packages" window, leave it as default and click Next. Cygwin, by default, doesn't have many of the linux packages like nano, vim etc. so those packages can be manually installed by expanding + All and selecting required packages by searching in the Search bar. These packages can be updated later as well, so leave it as it is and click Next
4. Extract "nutch-0.9.tar.gz" archive file at any location. The location looks something like this:
    D:/Nutch/nutch-0.9/

5. Open Cygwin terminal and change directory to /nutch-0.9/ with the following line of command: (The nutch-0.9 directory may vary in different systems depending on where and how the nutch archive file was unzipped.)

```
$ cd /cygdrive/d/Nutch/nutch-0.9/
```

6. Make a new directory in the nutch-0.9 folder named "urls" and inside urls directory, make a text file "url.txt" with the content "http://nits.ac.in" and/or other websites to crawl. New websites can be added as new lines.
7. In the /nutch-0.9/conf directory, edit the contents of the file named "nutch-site.xml" and add the following property inside <configuration></configuration> tags:

```
<property>
        <name>http.agent.name</name>
        <value>My Nutch Spider</value>
        <description></description>
</property>
```

8. In the same /nutch-0.9/conf directory, edit the contents of the file named "crawl-urlfilter.txt" and replace MY.DOMAIN.NAME with "nits.ac.in" and/or other domain names to be crawled. The change will look something like this:

```
# accept hosts in MY.DOMAIN.NAME
+^http://([a-z0-9]*\.)*nits.ac.in/
```

9. In the already running Cygwin terminal, run the following line to export JAVA_HOME (location may differ in different systems depending upon the installed directory):

```
$ export JAVA_HOME='/cygdrive/c/Program Files/Java/jdk1.8.0_341/'
```

10. Crawl the urls with the follwoing line of command (make sure /nutch-0.9/ directory is open in Cygwin):

```
$ bin/nutch crawl urls -dir crawl -depth 3 -topN 50
```

    a. If execute privilage is not given to nutch file, run: "$ chmod +x bin/nutch" and try the above line of command again.
    b. If there is java.lang.RuntimeException that says crawl already exists, delete the "crawl" folder in /nutch-0.9/ directory and try running the above line of command again.
11. Once the crawl is done successfully, the message "crawl finished: crawl" is displayed.
12. In webbrowser, go to address http://localhost:8080/manager/html. If prompted for Username:Password, enter the username:password set during installation.
13. Under the "WAR file to deploy" section, Choose File named "nutch-0.9.war" located in /Nutch/nutch-0.9/ directory, and Deploy. The page reloads and under the "Applications" list section, "/nutch-0.9" Path is displayed. Stop the hosting of Tomcat.
14. Open directory "Tomcat 6.0" located usually in C:/Program Files/Apache Software Foundation/Tomcat 6.0. From there, go to directory webapps/nutch-0.9/ and open the file named "search.jsp". Remove (do not just comment, but erase entirely) 151th line of code that reads: <jsp:include page="<%=language+\"/include/header.html%>"/>. Save the file.
15. Further, go to /WEB-INF/classes/ and open the file named "nutch-site.xml" and add the following property inside <configuration></configuration> tags (the location in <value> may differ in different systems depending on the Nutch extracted directory and the name given to crawl folder during step 10):

```
<property>
        <name>searcher.dir</name>
        <value>D:\Nutch\nutch-0.9\nutch-0.9\crawl</value>
        <description></description>
</property>
```

16. Start tomcat, goto localhost:8080/manager/html reload /nutch-0.9 Path by pressing "Reload" under Commands column of Applications section.
17. Select /nutch-0.9 Path in Applications section which redirects to nutch webpage. Search any keyword, and the nearest results will be listed.