Name : Reddi Mohana Krishna | Scholar ID : 2322111 | Subject : AI | M.Tech AI Sem I

REPORT

Assignment-I: Search Engine using Apache Nutch

Abstract : -

Our task in this assignment is to develop search engine using Apache Nutch and Tomcat as primary technologies.

The objective of this assignment was to design and implement a web-based search engine capable of crawling and indexing web content from diverse sources, providing users with efficient and relevant search results. The project successfully demonstrates the creation of a functional search engine capable of crawling and indexing web pages, providing relevant search results, and demonstrating the feasibility of building custom search engines using Apache Nutch and Tomcat.

Tools used :-

- Operating System : MacOS Sonoma 14.1(Unix Based)
- Apache Nutch 0.9
- Apache Tomcat 9.0.82
- Java 21.0.1

While it was recommended to use a Linux distribution to perform this task. I choose to use MacOS for sake of convenience because by default it comes with Linux.

❖ I used Apache Tomcat 9.0.82 on my system since it is required to view and browse the crawled data using a browser in local environment.I downloaded the "Apache-Tomcat-9.0.82.tar" from their official website and extracted to my 'Downloads' folder. Apache Tomcat is commonly used for hosting web applications, especially those built with Java technologies like Java Servlets and JavaServer Pages (JSP). It was also required to set up the "JAVA_HOME" environment variables since it is needed by Apache Tomcat to perform its operations. I started the server by executing the following command:

```
"/Users/krishna/Downloads/apache-tomcat-9.0.82/bin/startup.sh"
```

❖ For this task, Apache Nutch 0.9 was used. Apache Nutch is an open-source, extensible web crawling and indexing framework developed by the Apache Software Foundation. It is designed to help developers build web search engines, web data extraction, and data mining applications.The download link for this version was provided by our faculty or we can download it from the Apache Software Foundation.

❖ A packaged file named “`nutch-0.9.tar`” was downloaded and extracted the files that are kept inside a folder named “`nutch-0.9`” and moved into the directory “`/Users/krishna/Downloads/nutch-0.9`” . Inside the folder ‘`nutch-0.9/bin`’ created a directory named ‘urls’ and a text file ‘seed.txt’ was created inside it, which contained list of URLs that are needed to be crawl. In our case there was only one URL “http://www.nits.ac.in” which we are required to crawl.

Inside the directory named “conf” located at “`/Users/krishna/Downloads/nutch-0.9/conf`” several files were modified to configure the web crawler. All changes that were made listed below :

- In the file named ‘`crawl-urlfilter.txt`’ added a additional line

  “`+^http://([a-z0-9]*\.)*www.nits.ac.in/`”
- In the file named ‘`regex-urlfilter.txt`’ added a additional line

  “`+^http://([a-z0-9]*\.)*www.nits.ac.in`”
- In the file named ‘nutch-site.xml’ added a property ‘`http.agent.name`’ copied from ‘`nutc-default.xml`’ and value changed to ‘ My Nutch Spider’.

1. Once the configurations are finished, Apache Nutch is ready for first crawl. In the Terminal change the directory to “`/Users/krishna/Downloads/nutch-0.9/bin`” and following command is used to begin the crawl :

   “`./nutch crawl urls -dir Crawled_Data -depth 3 -topN 50`”
2. Where, ‘Crawled_Data’ is a folder where the crawled data to be stored and -depth and -topN determine the depth and number of pages to be crawled respectively.

3. After successfully crawling the data, a file from the ‘nutch-0.9’ named ‘nutch-0.9.war’ is copied and pasted into to “/Users/krishna/Downloads/apache-tomcat-9.0.82/webapps” and run the Apache Tomcat server. Open the file ‘nutch-site.xml’ from

“`/Users/krishna/Downloads/apache-tomcat-9.0.82/webapps/nutch-0.9/WEB-INF/classes/nutch-site.xml`”.

4. Modify the property ‘search.dir’ and set its value to- “`/Users/krishna/Downloads/nutch-0.9/bin/Crawled_Data`” which is a path to the directory that contains the crawled data from the Apache Nutch. Save the changes and open a browser.

Enter the url ‘`localhost:8080/nutch-0.9/`’ to see the nutch homepage with a search bar. Enter the text you want to search and then click on search, if it displays error that the file ‘Search.jsp’ located at ““`/Users/krishna/Downloads/apache-tomcat-9.0.82/webapps/nutch-0.9`”needs to be modified at line 151, after adding escape sequence it will look like this “`<jsp:include page="<%= language + \"/include/header.html\"%>"/>`” restart the Apache Tomcat Server.

Result :-

After Successful execution of all required tasks, a homepage with Apache Nutch Logo and a search bar is displayed. Enter a search query relevant to the data that was crawled to see the results. In my case I entered the query “**m.tech**” and received about 61 results from the crawled data and successfully finished the task

All the results are uploaded to my github : https://github.com/mohanreddy91/search_engine

Conclusion :-

Throughout the course of this assignment, we achieved several key milestones:

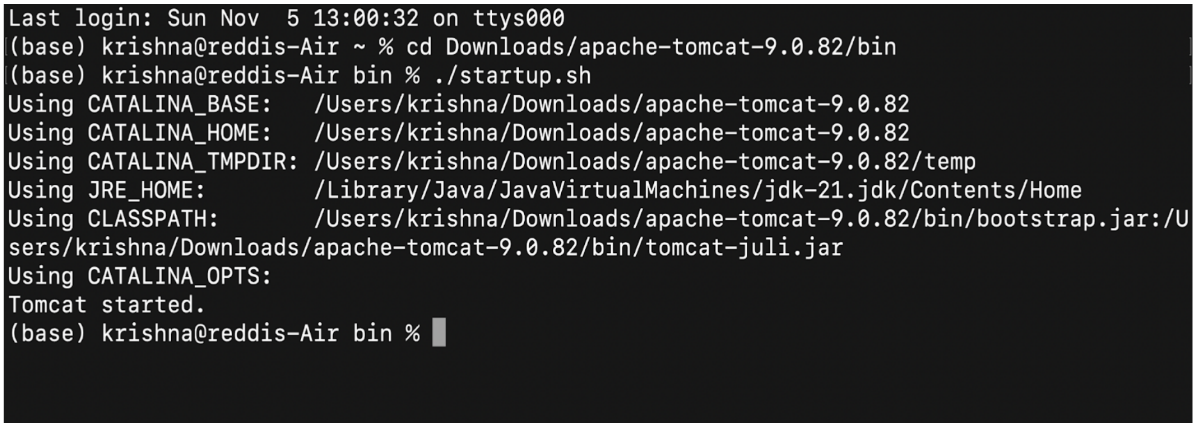
- We successfully configured and integrated Apache Nutch and Tomcat, laying the foundation for our search engine’s infrastructure.
- The web crawling and indexing processes proved effective in collecting data from various websites and demonstrating the search engine’s ability.
- Challenges faced during the project, such as managing web diversity and optimizing crawling performance, were addressed through iterative refinement of our techniques and strategies.

References :-

1. https://cwiki.apache.org/confluence/display/NUTCH/NutchTutorial
2. https://youtube.com/playlist?list=PL_RrEj88onS_-T5zBnkkm07suqQtCLFpT&si=O-vofl62X0qu0kan
3. https://nutchinstall.blogspot.com/2007/07/setting-up-cygwin-and-nutch.html

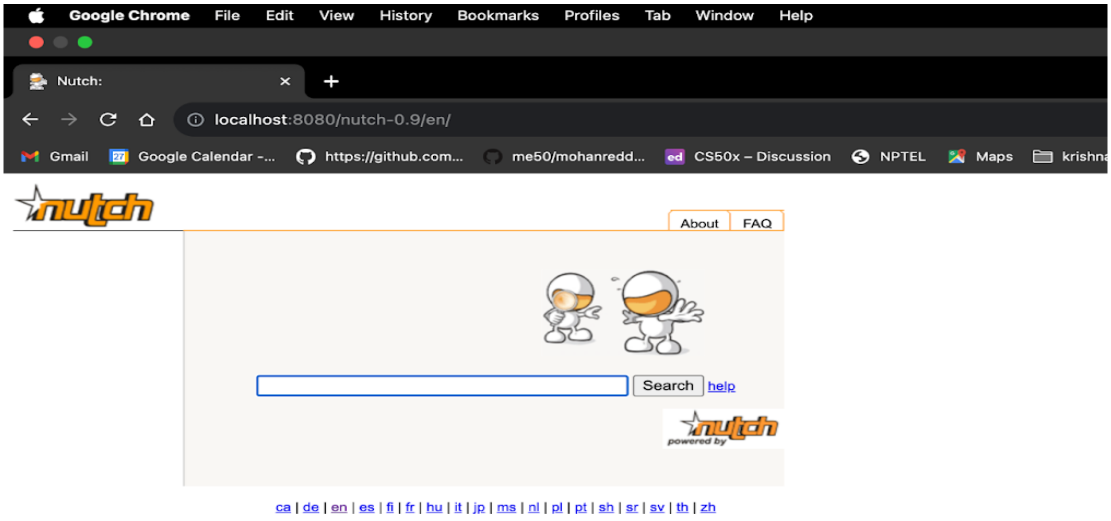
Screenshots :-

1.Starting Apache Tomcat Server :-



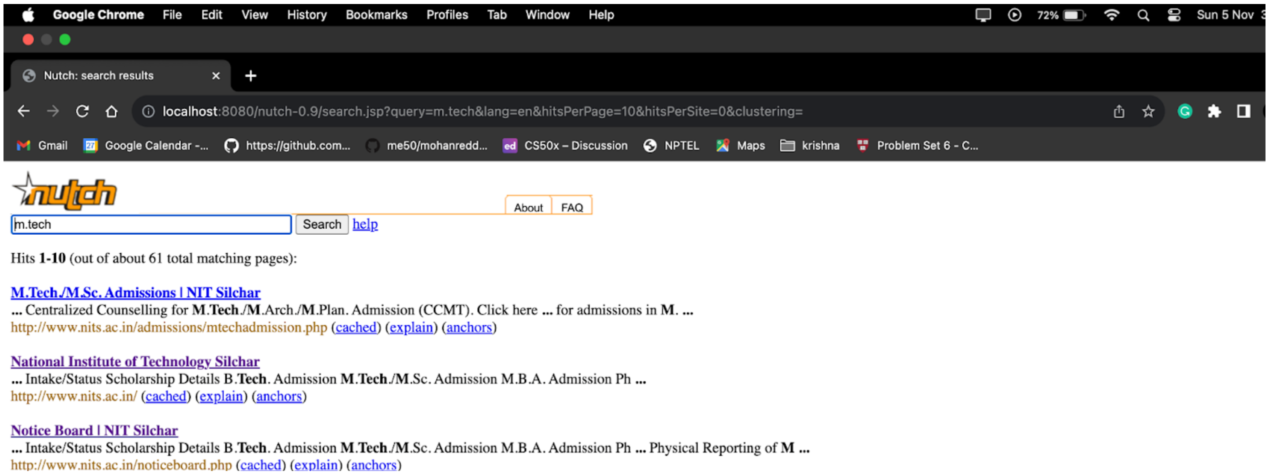
```
Last login: Sun Nov  5 13:00:32 on ttys000
(base) krishna@reddis-Air ~ % cd Downloads/apache-tomcat-9.0.82/bin
(base) krishna@reddis-Air bin % ./startup.sh
Using CATALINA_BASE:   /Users/krishna/Downloads/apache-tomcat-9.0.82
Using CATALINA_HOME:   /Users/krishna/Downloads/apache-tomcat-9.0.82
Using CATALINA_TMPDIR: /Users/krishna/Downloads/apache-tomcat-9.0.82/temp
Using JRE_HOME:        /Library/Java/JavaVirtualMachines/jdk-21.jdk/Contents/Home
Using CLASSPATH:       /Users/krishna/Downloads/apache-tomcat-9.0.82/bin/bootstrap.jar:/U
sers/krishna/Downloads/apache-tomcat-9.0.82/bin/tomcat-juli.jar
Using CATALINA_OPTS:
Tomcat started.
(base) krishna@reddis-Air bin %
```

2. After opening url “`localhost:8080/nutch-0.9/`” chrome browser :-



localhost:8080/nutch-0.9/en/

3. Search results after entering search query “m.tech”



localhost:8080/nutch-0.9/search.jsp?query=m.tech&lang=en&hitsPerPage=10&hitsPerSite=0&clustering=

m.tech

Hits **1-10** (out of about 61 total matching pages):

M.Tech./M.Sc. Admissions | NIT Silchar
... Centralized Counselling for **M**.**Tech**./**M**.Arch./**M**.Plan. Admission (CCMT). Click here ... for admissions in **M**. ...
http://www.nits.ac.in/admissions/mtechadmission.php (cached) (explain) (anchors)

National Institute of Technology Silchar
... Intake/Status Scholarship Details B.**Tech**. Admission **M**.**Tech**./**M**.Sc. Admission M.B.A. Admission Ph ...
http://www.nits.ac.in/ (cached) (explain) (anchors)

Notice Board | NIT Silchar
... Intake/Status Scholarship Details B.**Tech**. Admission **M**.**Tech**./**M**.Sc. Admission M.B.A. Admission Ph ... Physical Reporting of **M** ...
http://www.nits.ac.in/noticeboard.php (cached) (explain) (anchors)