Name: Hrishikesh Gogoi

Scholar ID: 2322222

Branch: CSE M.tech 2nd Sem

Subject: AI

REPORT ON

Assignment 1:- Search Engine using Apache Nutch and Apache Tomcat

Abstract:-

In this assignment, my objective is to develop a robust web search engine using Apache Nutch and Apache Tomcat as the primary technologies. The aim is to design and implement a web-based search engine capable of crawling and indexing content from diverse sources, providing users with efficient and relevant search results. Through this project, the creation of a functional search engine capable of crawling and indexing web pages, delivering relevant search results, and showcasing the feasibility of custom search engine development using Apache Nutch and Apache Tomcat is demonstrated.

Tools Used:-

- Operating System: MacOS Sonoma 14.3.1
- Apache Nutch 0.9
- Apache Tomcat 9.0.82
- Java 21.0.1

Apache Nutch 0.9 served as the primary web crawling and indexing framework, allowing for the extraction of data from various web sources. Java 21.0.1 was employed for its compatibility and robustness in implementing the necessary functionalities within the search engine. Despite the recommendation for a Linux distribution, MacOS Sonoma 14.1 was chosen for its inherent Unix-based environment, ensuring seamless integration and development processes. Apache Tomcat 9.0.82 facilitated the local browsing and viewing of the crawled data, with essential environment variables like JAVA_HOME configured to enable its operations.

Procedure:-

- *Setting up the environment*: I downloaded "`apache-tomcat-9.0.82.tar`" and "`nutch-0.9.tar`" from their official websites and extracted them to a folder named "`ai`" inside my "`Downloads`" directory. Additionally, I installed Java 21.0.1 on my system and ensured that the JAVA_HOME environment variable was set up to enable Apache Tomcat operations by executing the following command:

```
"echo 'export JAVA_HOME=$(/usr/libexec/java_home)' | sudo tee -a ~/.zshrc"
```

● *Configuring Apache Tomcat*: I started the Apache Tomcat server by executing the following command in the terminal:

```
“/Users/hrishikeshgogoi/Downloads/ai/apache-tomcat-9.0.82/bin/startup.sh”
```

● *Configuring Apache Nutch*: I navigated to the directory “`/Users/hrishikeshgogoi/Downloads/ai/nutch-0.9/bin`” and created a directory named “`urls`”. Within this directory, I created a text file named “`seed.txt`”, which contained the list of URLs that are needed to be crawled. In this case, the file included only one URL: “`http://www.nits.ac.in`”. I then navigated to the directory “`/Users/hrishikeshgogoi/Downloads/ai/nutch-0.9/conf`” and modified three files to configure the web crawler as follows:

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

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

In the file named “`nutch-site.xml`” I added a property “`http.agent.name`” which was copied from “`nutch-default.xml`” with it’s value changed to “`My Nutch Spider`”.

● *Crawling the data*: In the terminal, I changed my directory to “`/Users/hrishikeshgogoi/Downloads/ai/nutch-0.9/bin`” and started crawling by executing the following command in the terminal:

```
“./nutch crawl urls -dir Crawled_Data -depth 3 -topN 50”
```

In this context, “`Crawled_Data`" refers to the folder designated for storing the crawled data. The parameter “`-depth`” specifies the crawling depth, determining how many levels of pages the crawler will traverse from the seed URL. On the other hand, “`-topN`” specifies the maximum number of URLs to fetch from the frontier at each iteration of the crawling process. In this case, crawling will be conducted to a depth of 3 levels, with the first 50 URLs being retrieved at each depth level.

● *Deploying the crawled data to Apache Tomcat*: After successfully crawling the data, the file “`nutch-0.9.war`” from the directory “`/Users/hrishikeshgogoi/Downloads/ai/nutch-0.9`” is copied and pasted into the directory “`/Users/hrishikeshgogoi/Downloads/apache-tomcat-9.0.82/webapps`”. Additionally, I opened the file named “`nutch-site.xml`” located in the directory “`/Users/hrishikeshgogoi/Downloads/apache-tomcat-9.0.82/webapps/nutch-0.9/WEB-INF/classes`” and modified the property “`searcher.dir`”. I changed it’s value to “`/Users/hrishikeshgogoi/Downloads/ai/nutch-0.9/bin/Crawled_Data`” which is the path to the directory containing the crawled data from Apache Nutch.

● *Testing and viewing results*: The Apache Tomcat server is restarted, and a web browser is opened. The url “`http://localhost:8080/nutch-0.9/`” is accessed to view the Nutch homepage with a search bar. The desired text is entered into the search bar and a search is performed. In my

case, an error was encountered. To address this, the file “search.jsp” located at “/Users/hrishikeshgogoi/Downloads/ai/apache-tomcat-9.0.82/webapps/nutch-0.9” was modified at line 151. An escape sequence was added to the code, resulting in the following modification:
“<jsp:include page="<%= language + \”/include/header.html\"%>"/>”
After modifying the code, the Apache Tomcat server was restarted.

Result:-

Following the completion of all essential tasks, a homepage showcasing the Apache Nutch logo and an interactive search bar becomes visible. Input a search query aligned with the crawled data to view the results. For my assessment, I inputted the query "nits" yielding around 84 results sourced from the crawled data, thereby marking the successful completion of the task.

The results are uploaded in my GitHub:

https://github.com/hrishikeshgogoi/search-engine

Conclusion:-

Throughout the completion of this project, several significant milestones were achieved. Firstly, the successful configuration and integration of Apache Nutch and Tomcat laid a robust foundation for the search engine's infrastructure. Secondly, the effectiveness of the web crawling and indexing processes was demonstrated through the collection of data from various websites, showcasing the search engine's capability to retrieve relevant information. Lastly, challenges encountered during the project, such as managing web diversity and optimizing crawling performance, were addressed through iterative refinement of techniques and strategies. Overall, this project not only showcased the practical application of Apache Nutch and Tomcat but also provided valuable insights into the complexities of developing custom search engines.

References:-

https://nutch.apache.org/

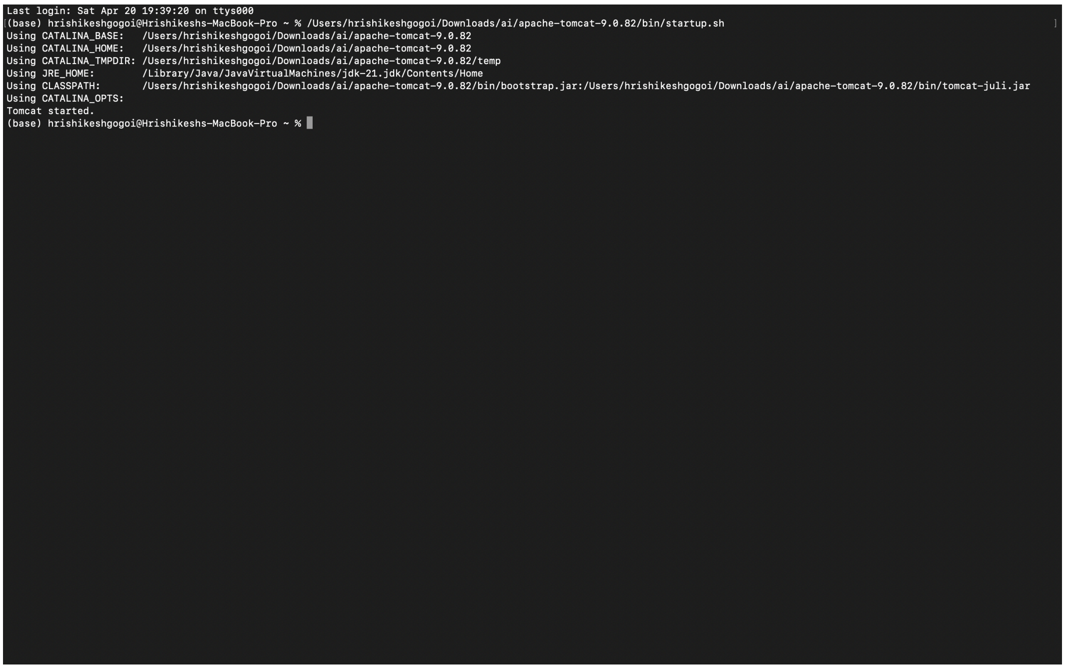
https://tomcat.apache.org/

https://cwiki.apache.org/confluence/display/NUTCH/NutchTutorial

https://www.youtube.com/playlist?list=PL_RrEj88onS_-T5zBnkkm07suqQtCLFpT

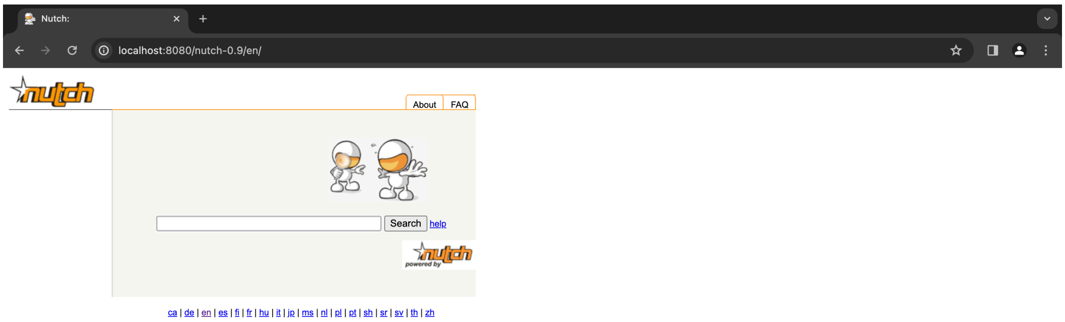
Screenshots:-

- Starting the Apache Tomcat server:



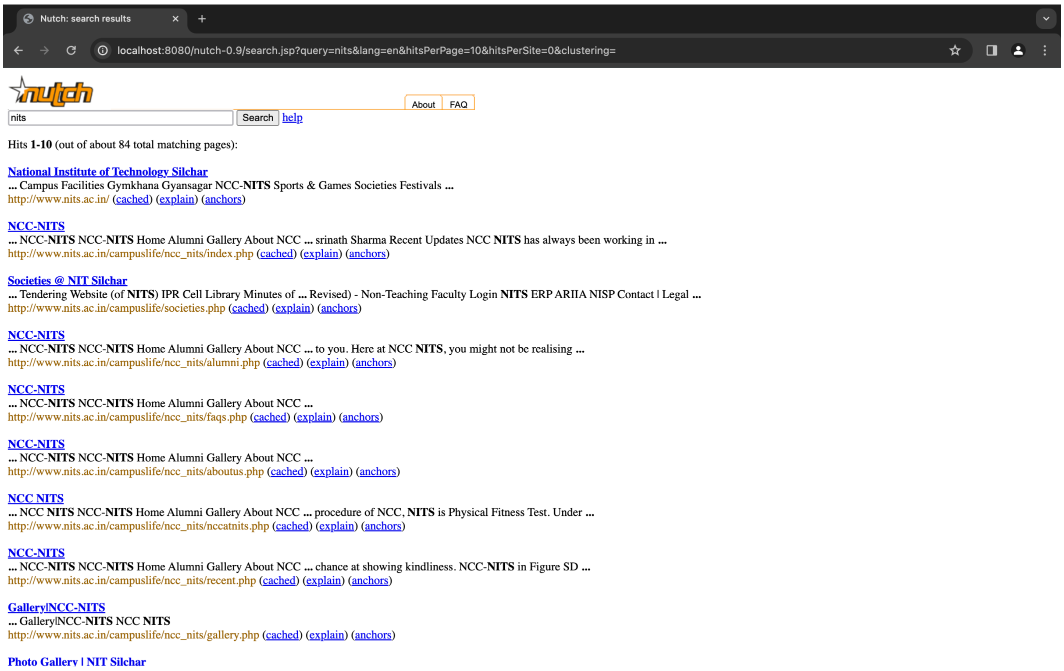
```
Last login: Sat Apr 20 19:39:20 on ttys000
(base) hrishikeshgogoi@Hrishikeshs-MacBook-Pro ~ % /Users/hrishikeshgogoi/Downloads/ai/apache-tomcat-9.0.82/bin/startup.sh
Using CATALINA_BASE:   /Users/hrishikeshgogoi/Downloads/ai/apache-tomcat-9.0.82
Using CATALINA_HOME:   /Users/hrishikeshgogoi/Downloads/ai/apache-tomcat-9.0.82
Using CATALINA_TMPDIR: /Users/hrishikeshgogoi/Downloads/ai/apache-tomcat-9.0.82/temp
Using JRE_HOME:        /Library/Java/JavaVirtualMachines/jdk-21.jdk/Contents/Home
Using CLASSPATH:       /Users/hrishikeshgogoi/Downloads/ai/apache-tomcat-9.0.82/bin/bootstrap.jar:/Users/hrishikeshgogoi/Downloads/ai/apache-tomcat-9.0.82/bin/tomcat-juli.jar
Using CATALINA_OPTS:
Tomcat started.
(base) hrishikeshgogoi@Hrishikeshs-MacBook-Pro ~ %
```

- Opening the url "`http://localhost:8080/nutch-0.9/`" with a web browser:



Nutch:
localhost:8080/nutch-0.9/en/
nutch
About
FAQ
Search
help
powered by nutch
ca | de | en | es | fi | fr | hu | it | jp | ms | nl | pl | pt | sh | sr | sv | th | zh

- Search results after entering the query “nits”:



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

nits

Hits **1-10** (out of about 84 total matching pages):

National Institute of Technology Silchar
... Campus Facilities Gymkhana Gyansagar NCC-**NITS** Sports & Games Societies Festivals ...
http://www.nits.ac.in/ (cached) (explain) (anchors)

NCC-NITS
... NCC-**NITS** NCC-**NITS** Home Alumni Gallery About NCC ... srinath Sharma Recent Updates NCC **NITS** has always been working in ...
http://www.nits.ac.in/campuslife/ncc_nits/index.php (cached) (explain) (anchors)

Societies @ NIT Silchar
... Tendering Website (of **NITS**) IPR Cell Library Minutes of ... Revised) - Non-Teaching Faculty Login **NITS** ERP ARIIA NISP Contact | Legal ...
http://www.nits.ac.in/campuslife/societies.php (cached) (explain) (anchors)

NCC-NITS
... NCC-**NITS** NCC-**NITS** Home Alumni Gallery About NCC ... to you. Here at NCC **NITS**, you might not be realising ...
http://www.nits.ac.in/campuslife/ncc_nits/alumni.php (cached) (explain) (anchors)

NCC-NITS
... NCC-**NITS** NCC-**NITS** Home Alumni Gallery About NCC ...
http://www.nits.ac.in/campuslife/ncc_nits/faqs.php (cached) (explain) (anchors)

NCC-NITS
... NCC-**NITS** NCC-**NITS** Home Alumni Gallery About NCC ...
http://www.nits.ac.in/campuslife/ncc_nits/aboutus.php (cached) (explain) (anchors)

NCC NITS
... NCC **NITS** NCC-**NITS** Home Alumni Gallery About NCC ... procedure of NCC, **NITS** is Physical Fitness Test. Under ...
http://www.nits.ac.in/campuslife/ncc_nits/nccatnits.php (cached) (explain) (anchors)

NCC-NITS
... NCC-**NITS** NCC-**NITS** Home Alumni Gallery About NCC ... chance at showing kindliness. NCC-**NITS** in Figure SD ...
http://www.nits.ac.in/campuslife/ncc_nits/recent.php (cached) (explain) (anchors)

Gallery|NCC-NITS
... Gallery|NCC-**NITS** NCC **NITS**
http://www.nits.ac.in/campuslife/ncc_nits/gallery.php (cached) (explain) (anchors)

Photo Gallery | NIT Silchar