AI Assignment Report

Apache Nutch Search Engine Assignment Report

Rahul Raj (2322216)

Introduction

Apache Nutch is an open-source web crawler and search engine toolkit. It is used to crawl the web and index the contents of web pages. Nutch can be used to build a variety of web search applications, including general-purpose search engines, vertical search engines, and enterprise search engines.

Apache Tomcat (also known as Tomcat) is a free and open-source web server and Servlet container that implements the Java Servlet, JavaServer Pages (JSP), Java Unified Expression Language, and WebSocket technologies. It provides a "pure Java" HTTP web server environment in which Java code can also run. Tomcat is used to deploy and run Java web applications, such as web services, web stores, and content management systems. It is also used as a development environment for Java web applications.

Java Development Kit (JDK) is a software development environment used for developing Java applications and applets. It includes the Java Runtime Environment (JRE), an interpreter/loader (Java), a compiler (javac), an archiver (jar), a documentation generator (Javadoc), and other tools needed in Java development. Now we need an environment to make a run of our program.

Cygwin a large collection of GNU and Open Source tools which provide functionality similar to a Linux Distribution on Windows.

Nutch Architecture

Nutch is a distributed system that consists of the following components:

- Crawler: The crawler is responsible for downloading web pages and extracting their contents.
- Parser: The parser extracts the text, images, and other resources from web pages.
- Indexer: The indexer indexes the text, images, and other resources extracted by the parser.
- Query processor: The query processor handles user queries and returns the most relevant results.

Building a Search Engine with Nutch and Tomcat

To build a search engine with Nutch, you will need to:

1. Download Nutch-0.9 and its dependencies(Cygwin 8 , Apache Tomcat 8.5.95 , JDK 17, JRE 8).
2. Configure Nutch to crawl the web pages to be indexed.
    - Extract nutch to c:\nutch-0.9\bin and create a folder with URLs and a text file with a website address.
    - Replace the website with MY.DOMAIN.NAME and add the value of agent name as property in nutch-site.xml in the conf folder.
3. Run the Nutch crawler in cygwin to download the web pages and extract their contents.
    - a. open Cygwin. exe
    - b. Type cd c
    - c. Type cd nutch-0. 9
    - d. Type cd bin
    - e. Type . /nutch crawl urls -dir My crawl-depth 5-topN 5.
    - Crawling will start
    - Wait until Crawling is finished
4. Nutch war file is generated in Nutch directory
5. Copy Nutch war file in C:\Program Files\Apache Software Foundation\Tomcat 8.5_Tomcat8_2\webapps
6. Start the Tomcat Server
7. Nutch folder will be generated automatically
8. Edit the search.jsp in location C:\Program Files\Apache Software Foundation\Tomcat 8.5_Tomcat8_2\webapps\nutch-0.9 to cope up with escape sequence error,
9. Go to C:\Program Files\Apache Software Foundation\Tomcat 8.5_Tomcat8_2\webapps\nutch-0.9\WEB-INF\classes
10. Open nutch-default.xml and nutch-site.xml
11. Search searcher word in xml file and copy the entire HTTP property and paste in nutch-site.xml inside configuration.
12. Paste the path C:\nutch-0.9\bin\abc in between <value></value>
13. Restart Tomcat Server
14. Run the Nutch query processor to handle user queries and return the most relevant results.
15. Go to http://localhost:8080
16. Go to Manager help -> /nutch-0.9
17. Enter user_id and password setup during Tomcat Configuration . If forgot check it in C:\Program Files\Apache Software Foundation\Tomcat 8.5_Tomcat8_2\conf

    inside tomcat-users

18. Search for any keyword related to site that is fetched

Nutch Use Cases

Nutch can be used to build a variety of web search applications, including:

- General-purpose search engines: Nutch can be used to build general-purpose search engines that can index and search the entire web.
- Vertical search engines: Nutch can be used to build vertical search engines that focus on a specific topic or industry.
- Enterprise search engines: Nutch can be used to build enterprise search engines that index and search the internal content of an organisation, such as intranet pages, documents, and emails.

Nutch Benefits

Nutch offers a number of benefits, including:

- Cost-effective: Nutch is an open-source project, which means that it is freely available to use and modify. This can save organisations a significant amount of money on search engine development and maintenance costs.
- Scalable: Nutch is highly scalable and can be used to crawl and index the entire web, or it can be configured to crawl and index specific websites or domains.

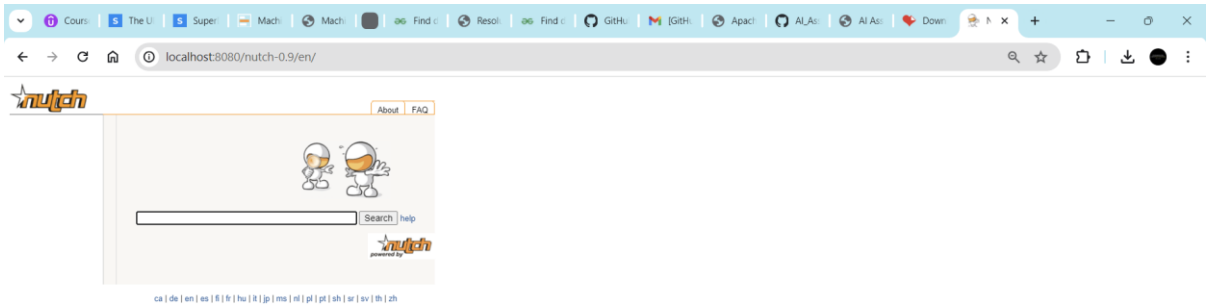
Nutch Challenges

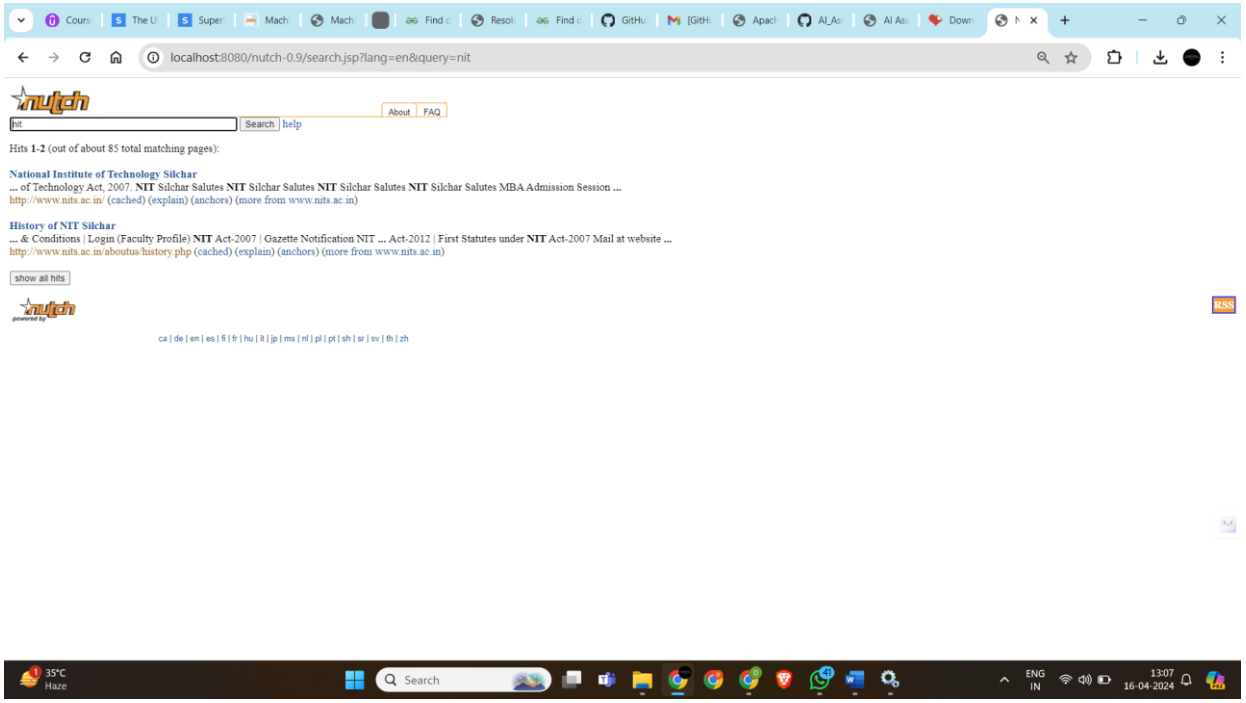
Nutch also presents some challenges, including:

- Complexity: Nutch is a complex system that can be difficult to set up and configure.
- Performance: Nutch can be slow to crawl and index large amounts of data.
- Documentation: Nutch's documentation can be difficult to understand and use.

Results

https://github.com/Rahulrajsit/M.Tech/blob/main/AI%20Assignment%20Results.pdf





Conclusion

Apache Nutch is a powerful and versatile search engine toolkit that can be used to build a variety of web search applications. Nutch is highly scalable, extensible, and open source, making it a cost-effective and powerful solution for many organisations.

References:

1. https://cwiki.apache.org/confluence/display/NUTCH/NutchTutorial
2. https://thecustomizewindows.com/2018/06/install-apache-nutch-web-crawler-on-ubuntu-server/

3. https://nutchinstall.blogspot.com/2007/07/setting-up-cygwin-and-nutch.html
4. https://www.youtube.com/playlist?list=PLBs6yEuoSNkGIbQ4Gk-aTECYox2VP_bEp