Scholarship Search Engine

University of Windsor

Submitted By:-

Karan Vishavjit (110099867)

Neel Pandya (110095825)

Shreyansh Dalwadi (110097536)

Manjinder Singh (110097177)

Harbhajan Singh (110100089)

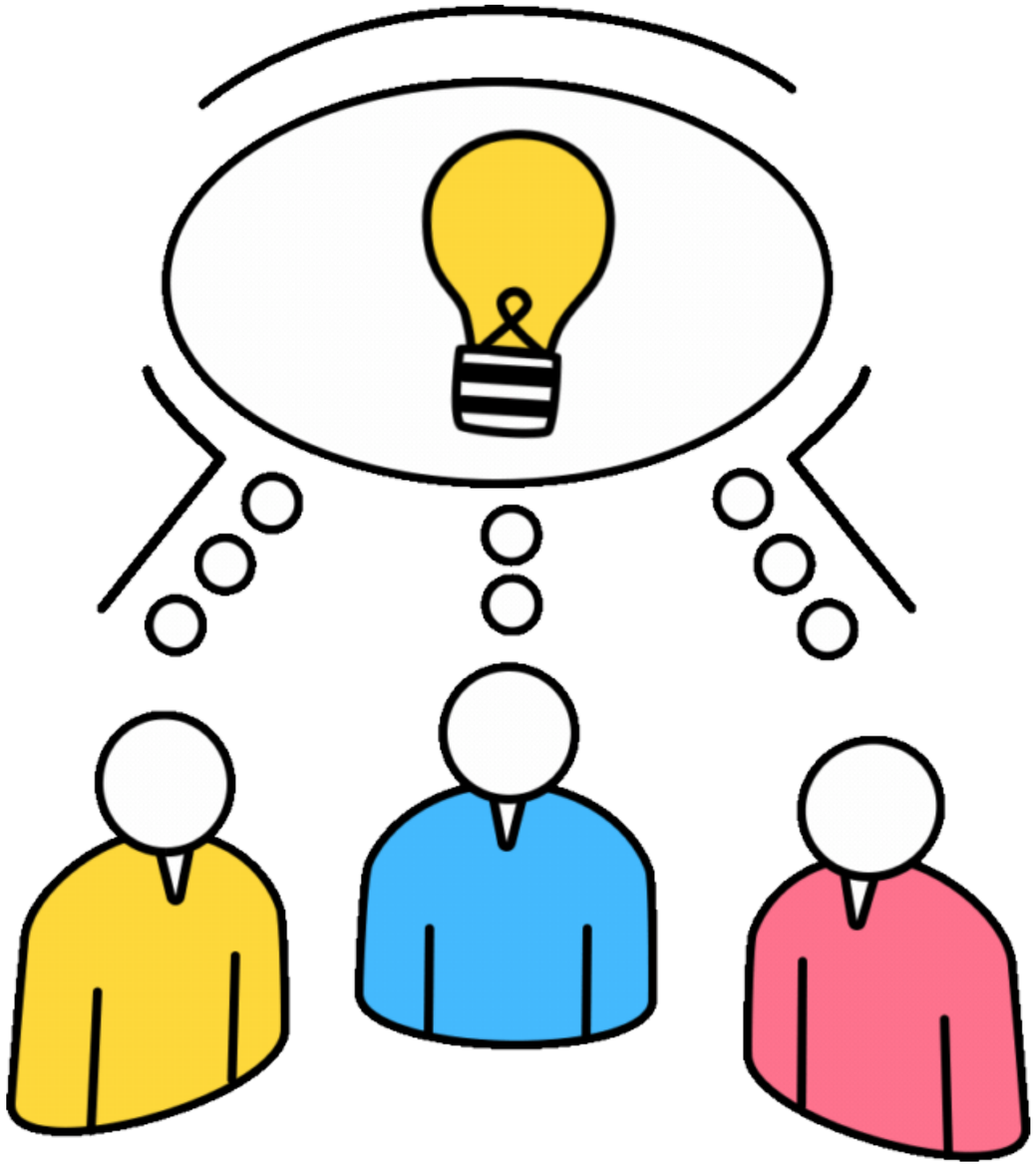
Presented To:-

Dr. Mahdi Firoozjaei

Graduate Assistant:

Roisul Islam Rumi

TABLE OF CONTENTS



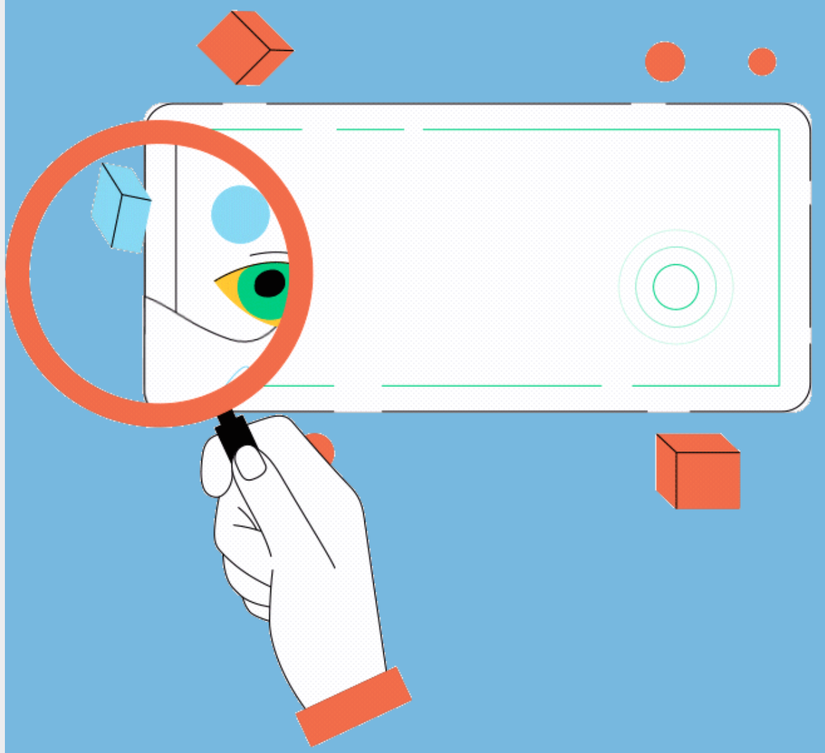
TEAM ROLES

Manjinder Singh	Web Crawling
Neel Pandya	Html Parser
Karan Vishavjit	Pattern Matching
Harbhajan Singh	Auto Correct
Shreyansh Dalwadi	Sorting and deletion

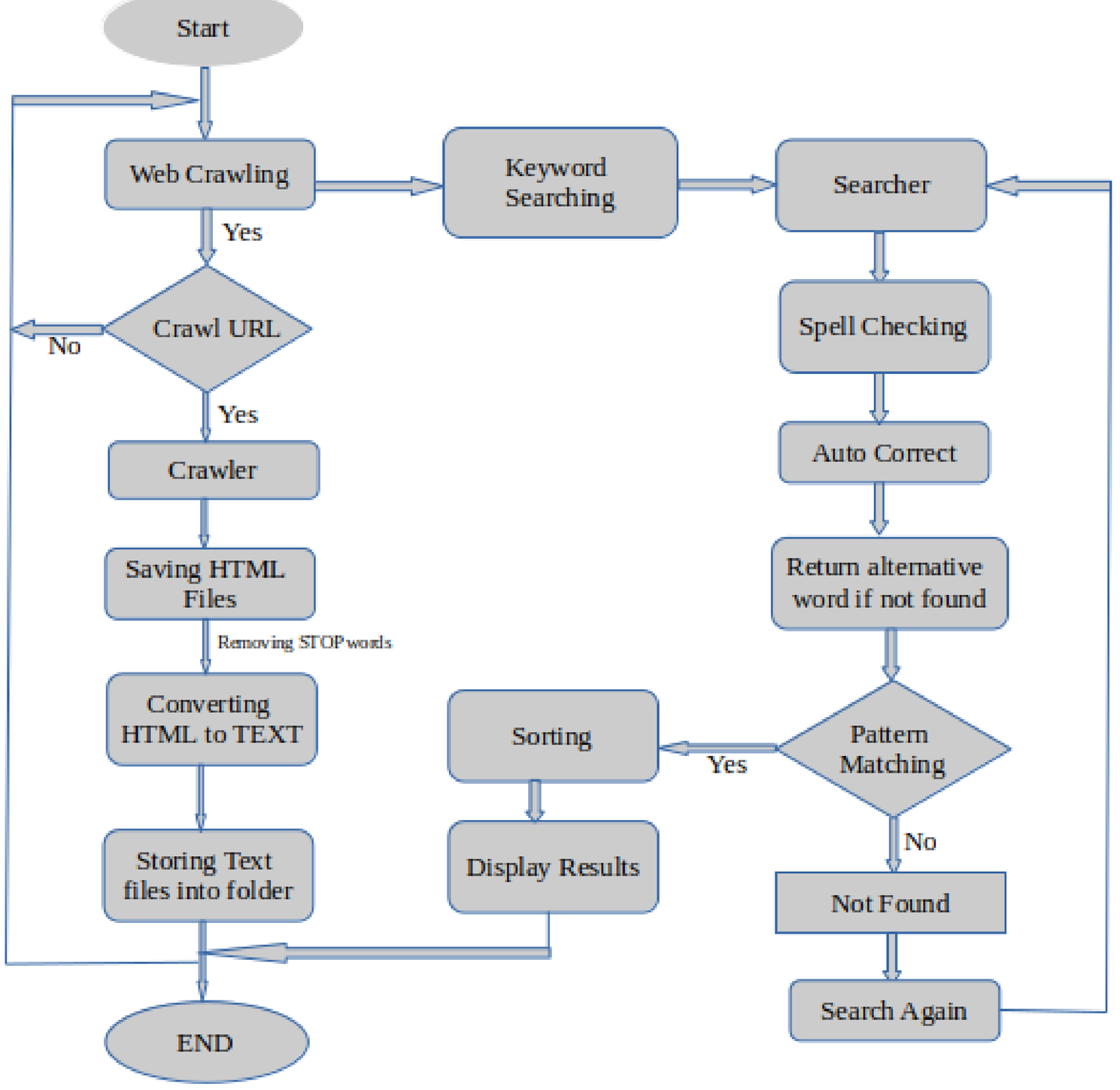
SCHOLARSHIP SEARCH ENGINE

- Scholarship opportunities are a great way to finance one's education. However, finding the right scholarship can be a daunting task. With the vast amount of information available online, it can be overwhelming to sift through all the data and find the most relevant scholarships.

- The scholarship web search engine uses jsoup to crawl web pages and download HTML files. It then reads these files, filters out stop words, and converts them to text files. Once the text files are created, the user inputs a keyword, which is autocorrected using edit distance. The engine then searches for patterns in the downloaded text files using the Boyer-Moore algorithm to find occurrences of the keyword.



PROJECT FLOWCHART



PROGRAM FEATURES

WEB CRAWLING

PATTERN MATCHING,
PAGE RANKING

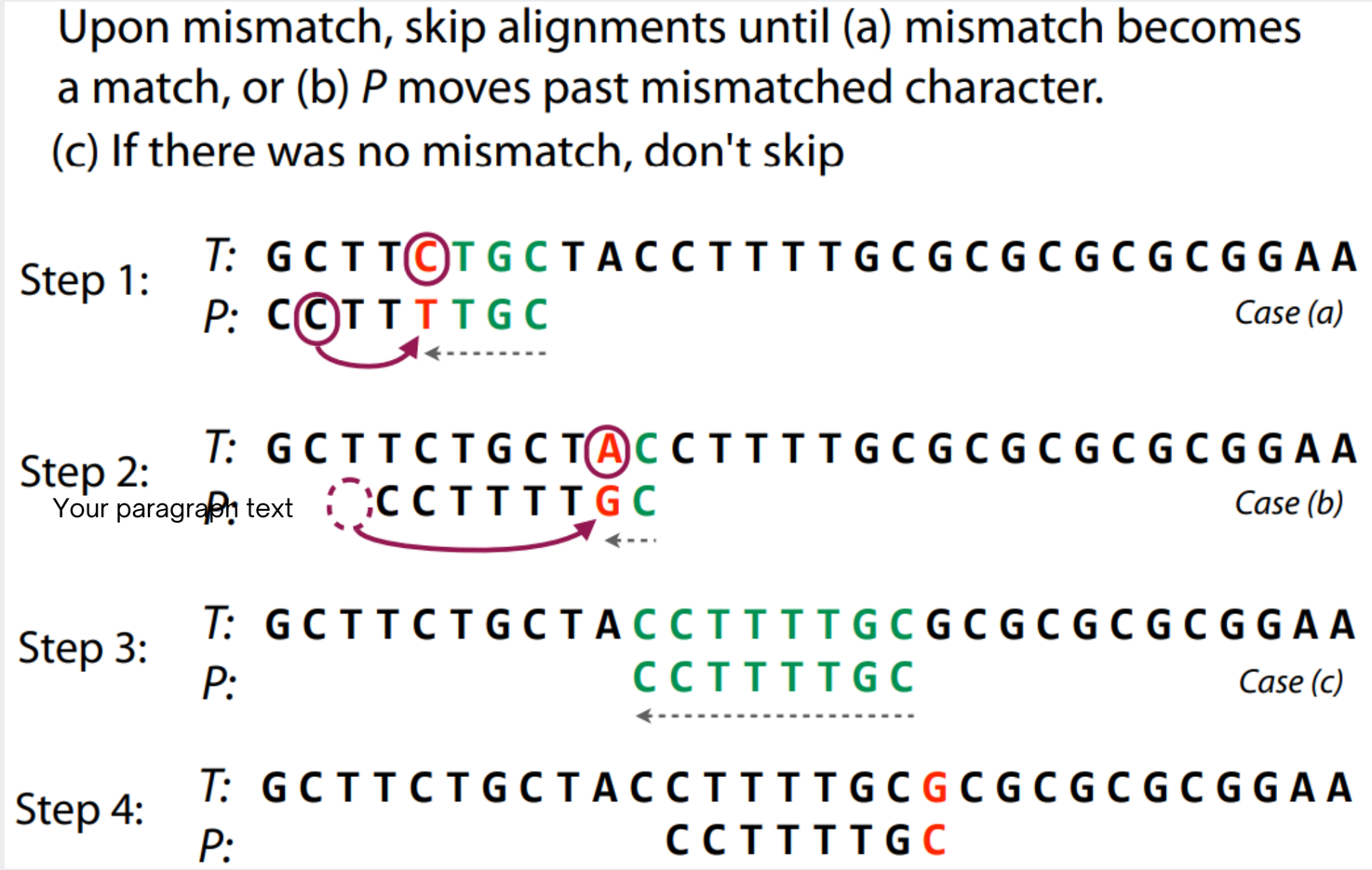
DICTIONARY
CREATION,
SPELL CHECKING,
AUTO CORRECT

HTML PARSER,
FILTERING STOP
WORDS

SORTING , DELETION

PATTERN MATCHING

- Boyer-Moore algorithm (Bad Character) to find occurrences of the keyword in the text files.
- Find all the occurrences in the file in single method call.
- Return list of indices searched pattern is present.
- Optimized result due to filtered data and bad character implementation.



Your paragraph text

Filtering Stop Words while converting HTML to Text

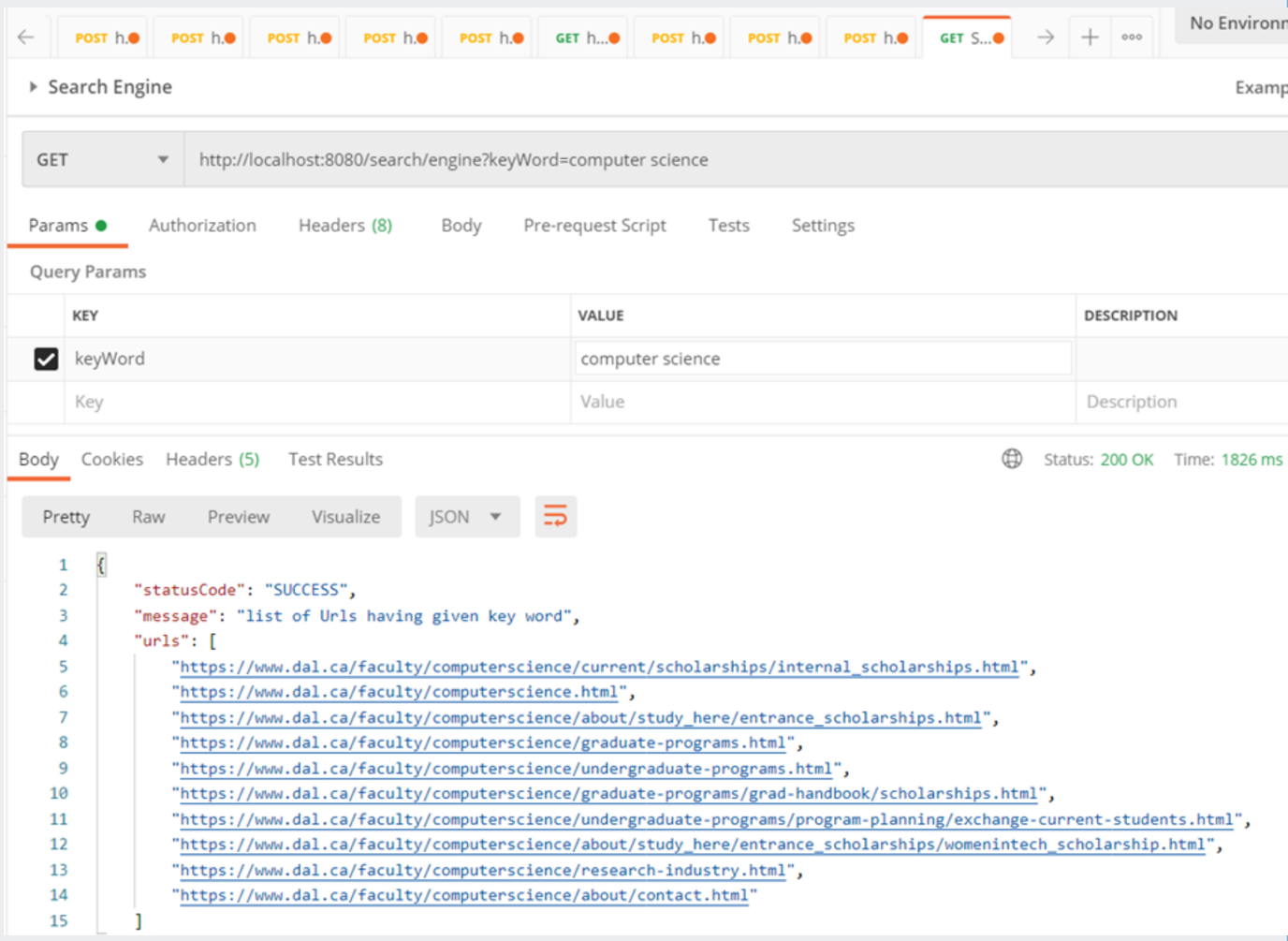
- Maintain a set of all stop words .
- Return filtered text without stop words.
- Approximately halved pattern search time.
- Reduced storage size on disk.

Size:	3.96 MB (41,60,593 bytes)
Size on disk:	4.50 MB (47,26,784 bytes)
Contains:	286 Files, 0 Folders

Size:	3.08 MB (32,39,914 bytes)
Size on disk:	3.62 MB (38,01,088 bytes)
Contains:	286 Files, 0 Folders

FUTURE IMPLEMENTATION

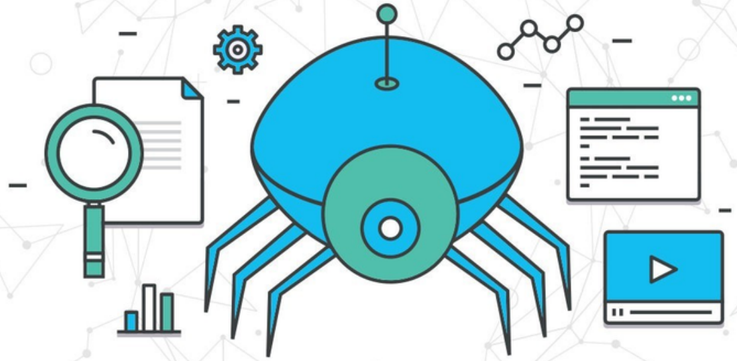
- Search Engine can be provided as an API service.
- Return list of url's based on search keywords.
- Can be customized to send other details as well.
- Light weighted json request and response no need of web UI to display results.
- Can be easily integrated with any existing application.



WEB CRAWLING

- The first step in creating a scholarship web search engine is to crawl web pages. This is done using a Java library called jsoup. Jsoup allows us to connect to a URL and retrieve the HTML content of the page. Once we have the HTML content, we can extract the relevant information such as the title, body, and links.

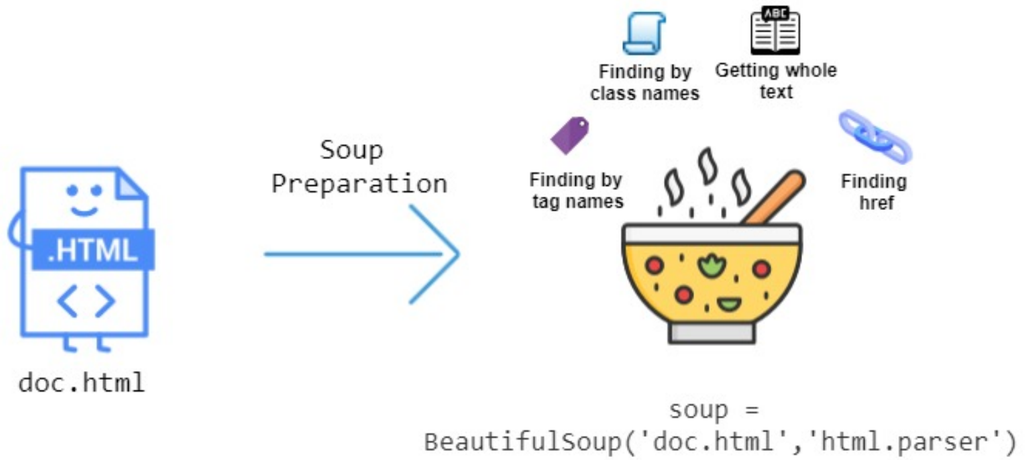
- After the HTML files are downloaded, the engine reads the files and converts them to text files. During this process, stop words are filtered out. Stop words are common words such as 'the', 'and', and 'a' that do not carry much meaning and can be safely removed from the text. Removing stop words reduces noise and makes it easier to find relevant information.



HTML PARSER

- Reading HTML code and producing a structured representation of the content are the functions of a Java HTML parser. It makes use of a library or framework, such as Jsoup, HTML Parser, or JSF, which offers an API for accessing and modifying the document's components, attributes, and content. online scraping, data extraction, and the development of online applications that produce dynamic HTML content all make use of HTML parsers.

- It our system it fetches the HTML files, fetches the directory path, parse file using Jsoup library and converts it into text files.

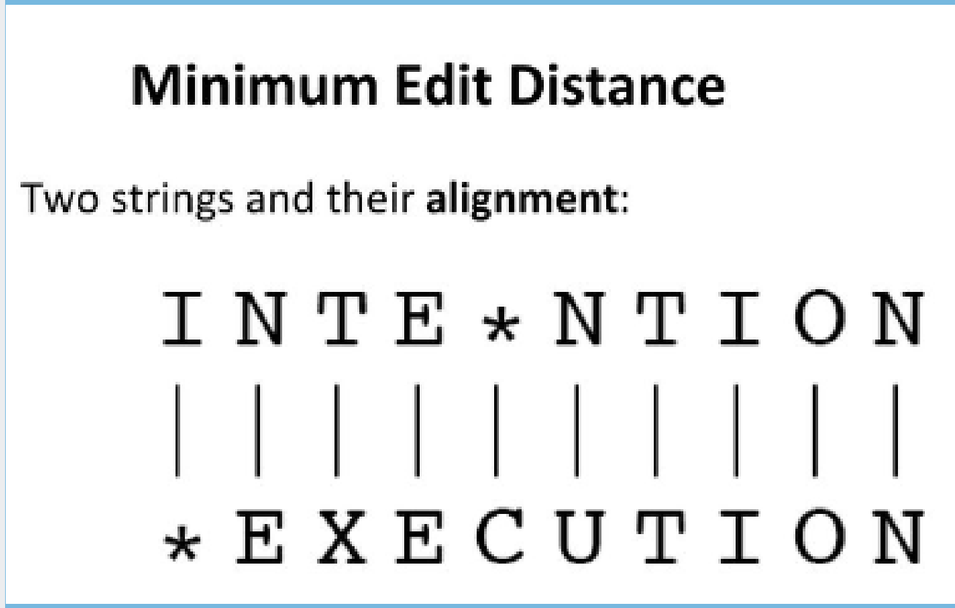


AUTOCORRECT

- Once the text files are created, the user inputs a keyword. The engine then autocorrects the keyword using edit distance.

- Edit distance is a measure of how many edits (insertions, deletions, or substitutions) are needed to transform one word into another. Autocorrecting the keyword ensures that even if the user misspells the keyword, the engine can still find relevant information.

- After autocorrecting the keyword, the engine searches for patterns in the downloaded text files using the Boyer-Moore algorithm.

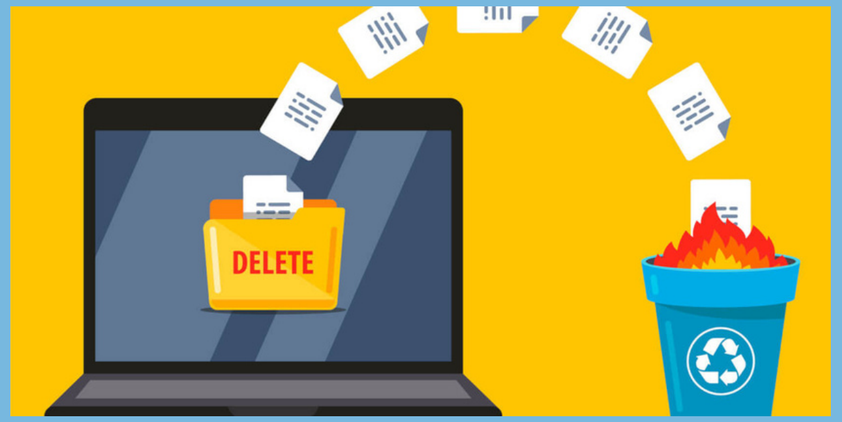


SORTING AND DELETION

- Sorting refers to ordering data in an increasing or decreasing manner according to some linear relationship among the data items.

- In our project, it will arrange the top 10 files on descending order based on the word which has occured in the particular file the most to the least.

- Deletion of text and html file is done.



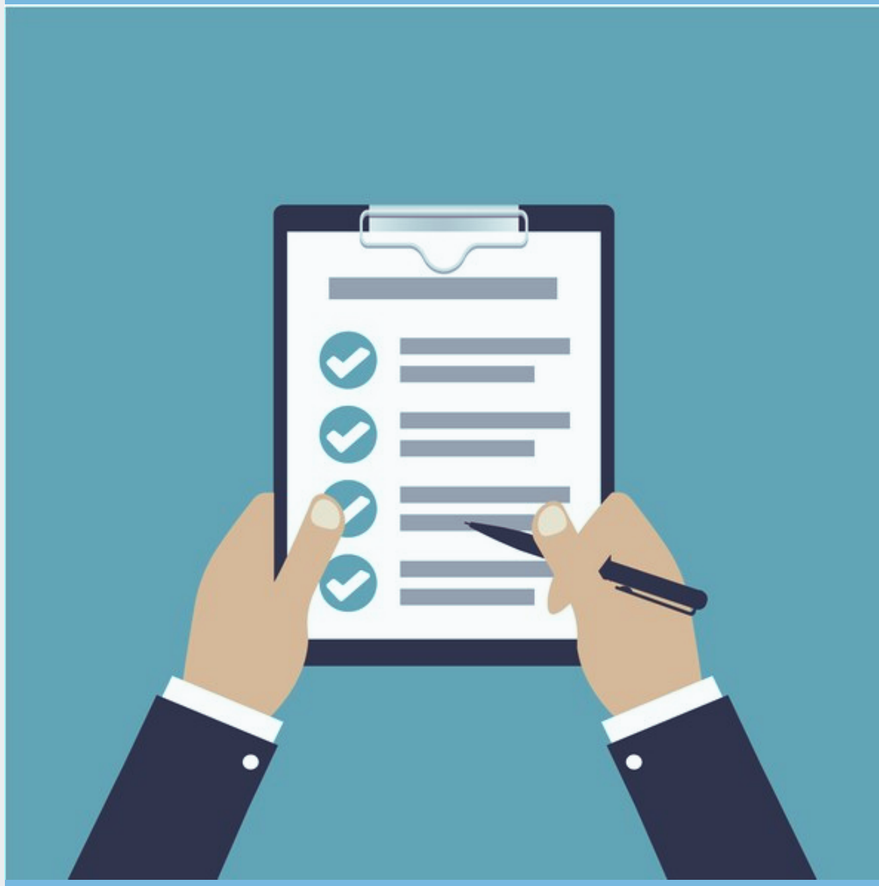
ACKNOWLEDGEMENT

- We appreciate Prof. Mahdi's instruction in algorithm design and the practicals he provided to assist us better comprehend the ideas. We appreciate your leading us through the programme.

- We appreciate Roisul Rumi's help with the labs, the project, and answering our questions.

REFERENCES

1. Lecture notes and lab practice files provided by Dr. Mahdi Firoozjaei
2.https://www.cs.jhu.edu/~langmea/resources/lecture_notes/04_boyer_moore_v2.pdf
3. http://www.mieliestronk.com/wordlist.html
4. https://countwordsfree.com/stopwords
5. https://developer.mozilla.org/en-US/docs/Web/HTML/Element/a
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/04_boyer_moore_v2.pdf
6.https://www.crio.do/blog/top-10-sorting-algorithms/
7.https://www.revouninstaller.com/blog/how-to-permanently-delete-files-from-your-computer/
8.https://slideplayer.com/slide/12877072/




Thank you!