**DSE I2400 –**

**PROJECT 2 –**

**Prerequisite- completion of Project 1.**

GOAL: Build a web based contents search engine to perform aggregated web search that displays result set ordered by frequency match for relevant search terms

1. Project 1 generated a set of URL obtained through the aggregated search process. Create a list of unique URL that are NOT marked as Ad/Promotional or equivalent.
2. Use Beautiful Soup ([Beautiful Soup: We called him Tortoise because he taught us. (crummy.com)](https://www.crummy.com/software/BeautifulSoup/)) to Scrape data from the URL from step 1 above.
3. Save the output from step 2 in a database object. One possible object structure would be to have columns containing ID, original search engine name, search key words, URL
4. Parse output data to identify the textual content and image/PDF, docs etc.
5. Save the results from step 4 in a database object. One possibility is to create an object with columns- Results\_ID (FK to ID from step 3), info\_type (like PDF, image, text, OCR\_text etc), Content\_details (will have the extracted content).
6. Process the PDF/image from step 5 through a OCR processing engine/API ([Top 5 OCR (Optical Character Recognition) APIs & Software | RapidAPI](https://rapidapi.com/blog/top-5-ocr-apis/))
7. Save the data from step 6 in an existing table of step 5 under info\_type = OCR\_text
8. Scan the text, OCR\_text data in table 6 for search key word (available in step 3) frequency for each of the URL. Save the results in a table having attributes for Stats\_ID (FK to ID from step 3), search term, count of occurrence.
9. Build a web based search engine that allows input of search terms (like done in Google, bing etc search). The result of the search will list the output just like a Google search. The result set will show the list of source URL (clickable link), and search term scoring result. The output will be ordered by data from step 8 based on search term frequency matrix. Note that Google search shows summary of content sometimes. It might be difficult to display similarly in your search engine output.
10. That’s the end of project 2.
11. SUBMIT the code base such that your entire work can be validated.