Meeting notes

02/22/2024

a) Reviewed the lab assignment comprehensively and discussed the necessary environment setup.

b) Explored the resources mentioned in the article, these tools provide practical guidance for text extraction.

c) Allocated tasks among team members and initiated brainstorming sessions for collaborative problem-solving.

02/23/2024

a) Coded a draft on PDF Extraction, which encompasses creating a database to capture text using PyPDF2 and employing OCR (Optical Character Recognition) technology, pytesseract, to process PDF files containing text in images。

b) Addressed bugs and reviewed initial results, identifying potential issues with the regex patterns.

02/24/2024

a) Optimized our code to enhance efficiency, focusing on resolving issues related to inaccurate information extraction by regularization patterns and improving the accuracy of information extracted via OCR.

b) Completed the first version coding to gather additional information from websites using API numbers. Also conducted an initial review of the information collected.

02/25/2024

a) Completed most part of code, with a continued focus on optimizing regex. Efforts were made to employ various regex patterns to capture as much accurate information as

possible from the PDFs, such as longitude and latitude, improving the data capture rate from an initial 10% to 90%.

b) Completed the data preprocessing tasks, finalizing adjustments to missing values and formatting inconsistencies to align with the database's uniform format requirements.

02/26/2024

a) Reviewed our script, underwent final evaluations and modifications, and submitted the final version.