Oboseeselebor@gmail.com   @Obose Eselebor
UNIVERSITY OF CAMBRIDGE   NIHR | BioResource   #10000 BLACK INTERNS   UK Health Data Research Alliance   HDRUK Health Data Research UK

AUTOMATED EXTRACTION OF STRUCTURED TABLES FROM TEXT FILES FOR ENHANCED HEALTHCARE INSIGHTS

OBOSEKOKHUNE ESELEBOR

INTRODUCTION

Extracting structured data from unstructured text files, particularly tables, is crucial for efficient data analysis across various domains. Health data often comes in forms that are not particularly accessible and are not fully utilized due to time-consuming manual extraction.

This project focuses on automating the extraction of flat tables from PDF collections, ensuring that critical data within test results, prescription sub-tables and other sections is readily available and actionable.

KEY FINDINGS

Effective Extraction
Successful extraction of flat tables from diverse text files

Challenges in Prescription Sub-Tables
Dealing with prescription sub-tables has presented challenges due to their non-standard formats

Refinement in Progress
As the project continues, ongoing efforts are being directed toward refining the handling of prescription sub-tables

METHODOLOGY

1. Browse through PDF Files

2. Use software to detect flat tables

3. Extract and save to any file format

4. Use python for further analysis.
i.e. using regex expression to detect unique identifiers

5. Make available for research

SCAN ME

BENEFIT TO HEALTHCARE

Improved Access for Research:
Extraction improves access for researchers to a wide range of medical data and in turn supporting innovation benefitting patient care.

Benefits for Existing Patient Care:
Timely access to precise data can also positively impact patient outcomes through improving speed of decision-making and overall quality of care in a clinical setting.

Reduced Data Entry Burden and Error:
Automation reduces the need for manual data entry and helps accurately capture and communicate patient information across medical records.
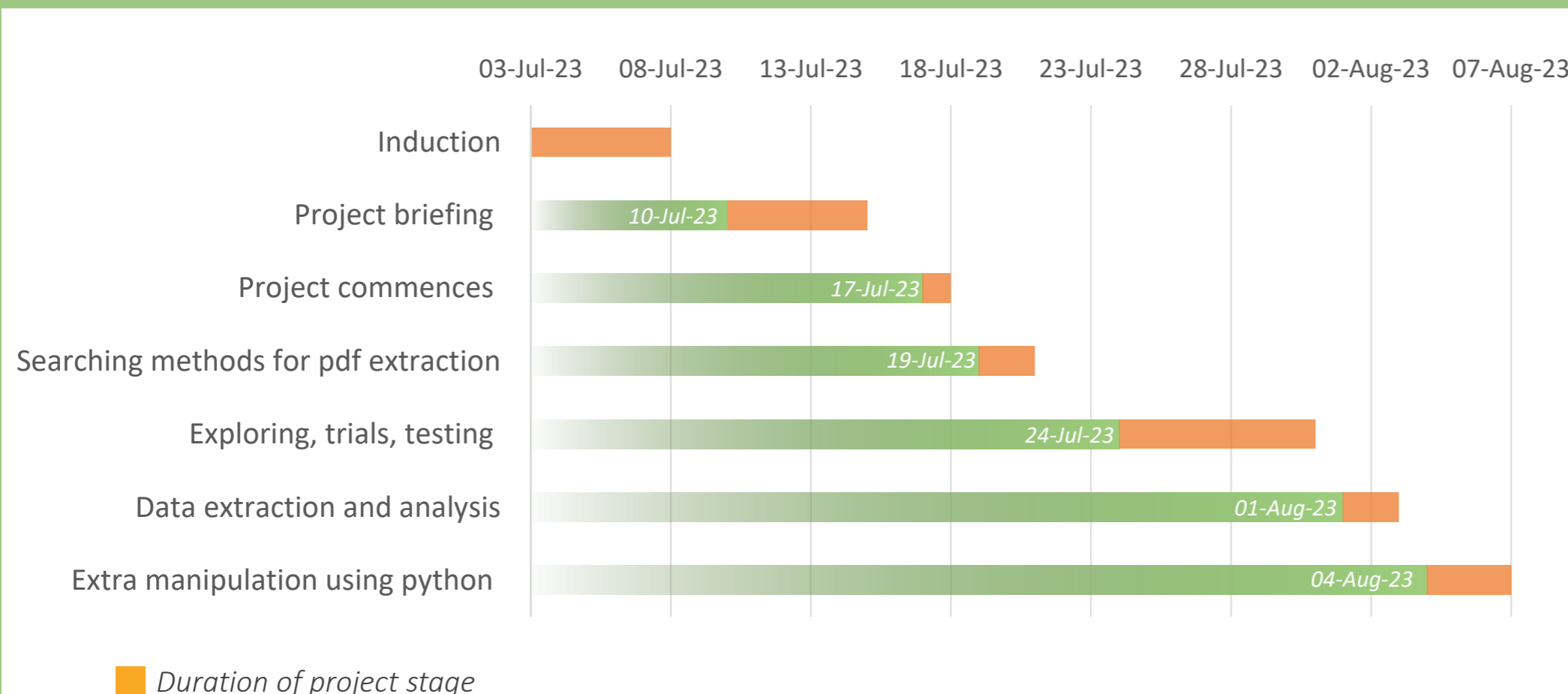
ACKNOWLEDGEMENT

- I extend my sincere appreciation to HDRUK and Black Internship Program for giving me this platform to learn and grow
- A special thank you to Dona, my supervisors Alvaro and Tom. Additionally, I would like to acknowledge my manager, Laura.
- Your combined efforts have played a pivotal role in the accomplishment of this project

SCHEDULE

Gantt chart columns: 03-Jul-23, 08-Jul-23, 13-Jul-23, 18-Jul-23, 23-Jul-23, 28-Jul-23, 02-Aug-23, 07-Aug-23

- Induction
- Project briefing — 10-Jul-23
- Project commences — 17-Jul-23
- Searching methods for pdf extraction — 19-Jul-23
- Exploring, trials, testing — 24-Jul-23
- Data extraction and analysis — 01-Aug-23
- Extra manipulation using python — 04-Aug-23

Duration of project stage

REFERENCES

- https://undraw.co/illustrations
- https://iconify.design/
- https://upload.wikimedia.org/wikipedia/commons/thumb/8/87/PDF_file_icon.svg/1200px-PDF_file_icon.svg.png
- https://www.wikipedia.org/

https://bioresource.nihr.ac.uk/