Skills and Resources Audit

Project Summary

To develop software that can assist in applying Optical Character Recognition in relation to our curated PDFs which are all data tables - that have been identified from trawling through many millions of pages of archive documentation, to transfer the data from PDF pages into a useable .csv format.

Technical Skills Required

This project will require experience in programming with python and potentially other languages such as HTML depending on the needs of the client for a web interface. Success in this project will rely heavily on the application of optical character recognition (OCR) libraries, therefore it is essential team members research and learn about OCR libraries in python to ensure our software can accurately apply OCR technology. It will be useful for team members to understand how to use APIs and other external libraries which may come in useful such as image enhancement libraries so images can be pre-processed before being processed by OCR. Due to a wide range of quality in the PDF documents that have been curated it will also be important our team to have skills in image enhancement and increasing the quality of images to ensure we can successfully transfer data from the PDFs to csv. Skills in testing frameworks such as PyTest will be important to ensure our project works as expected.

Non-Technical Skills Required

Project management is essential to be able to deliver a successful project to our client, therefore, it is important that team members have an understanding of project management techniques that they can apply throughout the project. Our team will also operate in an agile/scrum methodology and deliver sprints to the client throughout the semester. Many of our team members have experience using an agile methodology from units such as Agile Web Development. It is also important that our project is clearly documented for non-technical and technical users so that we produce a usable and maintainable system for the client.

Tools and Resources Required for the Project

Programming languages: python

Libraries/frameworks: Tesseract, Pillow, OpenCV, CamelotPie, TrOCR, EasyOCR, etc.

Testing tools: PyTest

Version control: Git/Github

Other tools/resources: CITS3200 unit page, CI/CD, Visio, Figma, IDEs.

Personal Evaluations

Skill/Tool	Ciaran Engelbrecht	William Lodge	William vdWB	Shashwat Abrol	Connor Fernie	Oliver Dean
OCR Technology	Intermediate	Beginner	Intermediate	Beginner	Beginner	Beginner
Python Programming	Advanced	Intermediate	Advanced	Advanced	Intremediate	Intermediate
Image enhancement libs	Beginner	Beginner	Beginner	Beginner	Beginner	Beginner
Project management	Intermediate	Beginner	Intermediate	Beginner	Beginner	Beginner
Version control (git)	Advanced	Advanced	Advanced	Advanced	Advanced	Advanced
Testing (i.e. PyTest)	Intermediate	Intermediate	Intermediate	Intermediate	Beginner	Advanced
Conversion of PDF to images	Beginner	Intermediate	Intermediate	Beginner	Beginner	Beginner
Python frontend framework	Beginner	Beginner	Beginner	Intermediate	Beginner	Intermediate
Skill Gap Identified	Learn about OCR tools and AI models, image pre-processing and image enhancement libraries, frontend frameworks and image conversions	Research OCR and image enhancement libraries, watch lecture on project management, frontend frameworks and image conversions	Research documentation on image enhancement, table detection tools, frontend frameworks and image conversions	Research how to automatically detect tables and draw grids in python on images, frontend frameworks and image conversions	Research OCR and image enhancement libraries, watch lecture on project management, frontend frameworks and image conversions	Learn about OCR, and project management techniques, frontend frameworks and image conversions
Resource	Google, OCR documentation, Tesseract, Microsoft TrOCR, Paddle, Additional python libraries, CITS3200 project site, coursera course	OCR documentation, CITS3200 lecture program, Coursera course, look into frontend frameworks for python	Google documentation, Coursera, research into camelotpie for automatic table detection tools.	OCR, image enhancement, table detection python documentation, Coursera course	Documentation OCR, Tesseract, EasyOCR, image enhancement and PyTest, CITS3200 lecture program, Coursera course	OCR documentation and CITS3200 lecture program, Coursera course