Test Manual

Historic Census Scanning Project

CITS3200 - Professional Computing

Winter 2024

University of Western Australia

Objectives

The purpose of this document is to specify tests that will indicate whether important aspects and requirements for the project are met. The testing strategy will include conducting frequent testing of the software and automated testing through a CI/CD pipeline. This document will provide information about the unit test specifications.

This document refers to the criteria that must be met for the project to be accepted as successful. The tests discussed in this document are general and part of a greater, lower level testing suite which will be implemented.

These tests should be seen as criteria for important aspects of the project which have been identified in the scope of work document and audits such as the $100 test. Therefore the tests described in this document should reflect the overall needs of the project.

Document References

Major documents produced during project development:

1. Scope of work/Requirements analysis document: statement of the scope of the project, requirements to be met and general statement of what the project is to achieve.
2. Skills and resources audit: audit of the skills and resources that the project will require to be successful and identification of gaps in the skills of the team members.
3. Risk register: description of the risks that may impact the success of the project and how these risks will be mitigated.
4. User stories & acceptance criteria: list of potential user stories and acceptance criteria that is required to be met to successfully satisfy the associated story.

Test Summary

The functions that are immediately being tested are:

1. OCR processing and export: Functions used for extracting text from PDF documents, includes functions for processing directories, image preprocessing as well.
2. Image preprocessing and enhancement: image enhancement functions for contrast, sharpness and noise reduction to improve image quality for OCR extraction.
3. PDF conversion to images: function converts PDFs into images using the pdf2image library and applies image enhancements.

Testing Strategy

The overall system's goal is to accurately convert a PDF containing data in tables to a usable format. Sub systems that make up the overall system that require testing include the PDF to image, image preprocessing and OCR processing. The immediate testing strategy will be conducted using unit tests that will test the functionality of these sub systems. Integration and system tests are important to ensure that the system will function as expected too. These tests will occur during development as functions of the sub system are created, initially on developer's devices using the unit tests that have been written by one of our team members. A system test will occur when the sub systems have been integrated together so the entire system can be comprehensively tested. Developers of the individual systems can use the unit tests to test the functions they have developed in their specific system they are responsible for. System tests should be conducted by all team members ensuring comprehensive testing of the overall flow of the system.

Diagrams of the system flow can be found in the requirements analysis document, these diagrams include mock-ups of the user flow, data flow and decision-making flow.

Test OCR processing (output accuracy)

This Unit Test is part of a testing suite for the OCR processing and export of data into an appropriate CSV format. It will ensure that the OCR processing is functioning correctly and the extracted text from provided images matches the expected output and is 1:1 with the original data uploaded.

Test Specification

The requirements whose satisfaction will be demonstrated by the test are that the OCR processing functions extract text to a high degree of accuracy and can process images provided to it for processing. It will run the OCR functions and check that the output of specified files matches the expected output that we have pre-determined through manual verification.

Test Description

- Location of test: Code/unitTests/TestOCRToCSV.py in the function: test_ocr_processing
- Means of Control: python unit tests are used in a testing script
- Data
    - Input Data: specified files which have already been extracted by another sub system for processing such as file 'page_iiicgko_0_0.png'.
    - Input Commands: the script automatically invokes the unit test
    - Expected Output Data:

      ['Geschlecht', 'Wien', 'Sonstiges Nieder- Oesterreich', 'Nieder-Oesterreich iiberhampt', 'Ober- Oesterreich'],

      ['m.', '291.183', '70.739', '361.922', '10.010']
    - System Messages: the framework will output an error message if the test fails

- Procedures:
    - Specify files to apply OCR function to
    - Run OCR on the specified files
    - Store the output in a CSV format
    - Check the actual output with the expected output

Test Analysis Report

- Function: the OCR must extract the text from the provided images with accuracy so that the output of the OCR functions matches the expected output from the user.
- Performance: the OCR function must extract the text within a reasonable amount of time, although, this unit test uses a small data set and performance likely won't be an issue.
- Data measures, including whether target requirements have been met: the resulting output data should match the expected output data from the user, target requirements will need to be tested as development continues, however basic tests are passing now.

If the test passes, then the actual output of the OCR matches the expected output, and we can state that for the provided test case the OCR software accurately extracts text from PDFs.

An error or deficiency is indicated by the actual output not matching the expected output. This impacts the requirement for 99% accuracy from the OCR text extraction functions and will have a major effect on the success of the project.

Test OCR processing (Batch document performance)

This Unit Test is part of a testing suite for the OCR processing and export of data into an appropriate CSV format. It will ensure that the OCR processing is functioning correctly and the extracted text from provided images matches the expected output and is 1:1 with the original data uploaded when provided with multiple images to process. The aim of this test is to evaluate the performance of the OCR processing when a user wishes to batch process many PDF documents at a time.

Test Specification

The requirements whose satisfaction will be demonstrated by the test are that the OCR processing functions extract text from many images in a reasonable amount of time to ensure the function meets the requirement for 10 pages per minute. It will run the OCR functions and check that the output of specified files matches the expected output when provided with many images. The conditions are that the output matches.

Test Description

- Location of test: Code/unitTests/TestOCRToCSV.py in the func: test_ocr_performance
- Means of Control: the script automatically invokes the unit test and generates the required images to conduct the test.
- Data
    - Input Data: images generated by the test with specific text to be extracted
    - Input Commands: the script automatically invokes the OCR functions using the images generated
    - Expected Output Data: output from the OCR functions should match the generated files by the test
    - System Messages: the framework will output an error message if the test fails
- Procedures: The test procedure is often specified in form of a test script.
    - Generates random strings, creates images with the strings and saves them
    - Run the OCR on the files
    - Validate the actual output with the expected output generated by the test.

Test Analysis Report

- Function: the OCR must extract the text from the provided images with accuracy so that the output of the OCR functions matches the expected output from the user.
- Performance: the OCR function must be able to meet performance requirements when provided many images, this include time constraints and being able to accurately process text when provided many images. The required output speed is 10 pages per minute.
- Data measures, including whether target requirements have been met: the resulting output data should match the expected output data from the user, target requirements will need to be tested as development continues, however basic tests are passing now within a reasonable amount of time, although the number of images should be variable.

If the test passes, then the actual output of the OCR matches the expected output, and we can state that for the provided test case the OCR software accurately extracts text from PDFs when provided multiple images and therefore the OCR is performing to requirements.

An error or deficiency is indicated by the actual output not matching the expected output. This impacts the requirement for 99% accuracy from the OCR text extraction functions and will have a major effect on the success of the project.

Test PDF to Image Conversion

This Unit Test is part of a testing suite for the OCR processing and export of data into an appropriate CSV format. It will ensure that the conversion of PDF documents to images functions as expected so that the images can later be processed by OCR and have text extracted.

Test Specification

The requirements whose satisfaction will be demonstrated by the test are that the PDF documents are correctly converted to images and enhancement can occur at this step. It is required that the software returns images for the corresponding PDF uploaded.

Test Description

- Location of test: /Code/unitTests/TestPDFToImages.py
- Means of Control: the script automatically invokes the unit test and generates the required images to conduct the test.
- Data
    - Input Data: set of PDF documents that are to be converted to images
    - Input Commands: each PDF is processed by pdf_to_images
    - Output Data: images corresponding to the PDF documents input
    - System Messages: the framework will output an error message if the test fails
- Procedures:
    - Iterate over PDF files and call the pdf_to_images function to convert to images
    - Check the validity of the output list, i.e. it is not empty, and items are an image

Test Analysis Report

- Function: testing the function of the conversion of PDFs to images. PDFs must be correctly converted to images for the OCR software to extract text from them.
- Performance: this test focuses on functionality rather than performance; however, this process should be completed within an appropriate time frame.
- Data measures, including whether target requirements have been met: the output data that is measured to check if the test passed are: the output must be a list containing images corresponding to the PDFs and be an Image.Image object.

If the test passes, then the function successfully converts PDF documents into images so that the OCR tools can correctly extract text from them and output it in the required format.

If the test fails then the function does not convert PDF documents into images correctly which will have a significant impact on the success of the project as the OCR tools require images to extract text from.

Automated Table Detection

This acceptance test refers to the automatic table detection functionality that is required to improve the accuracy of the OCR and remove manual work required by the user. This will ensure that the system can accurately extract text from the tables provided in the PDFs as images.

Test Specification

Testing that the automated table detection works correctly will be demonstrated by the software detecting a table, rows, columns and the data in cells without requiring end user intervention. The structure of the table should be maintained during this process.

Tables should accurately be detected, with rows and columns being detected. Multiple table styles should be detected such as those without specific borders. Therefore, multiple tables from PDFs should be tested to ensure this functionality works.

Test Description

- Location of test (hyperlink to test)
- Means of Control: Describes how data are entered (manually or automatically with a test driver)
- Data
    - Input Data: PDF documents
    - Input Commands: manual workflow/system tests
    - Output Data: tables in images being automatically detected
    - System Messages: report error message if tables cannot be automatically detected
- Procedures: The test procedure is often specified in form of a test script.
    - Submit PDF document
    - Convert PDF to image
    - Detect tables automatically
    - Check automatic table detection

Test Analysis Report

- Function: tests the function of the automatic table detection tools
- Performance: performance is not a concern regarding this acceptance criteria
- Data measures, including whether target requirements have been met: manual review of the automatic table detection may be required.

If tests pass, then the automatic table detection tools work as expected and client requirements will be met for this aspect of the system.

If the test fails, then the acceptance criteria has not been met and this aspect of the system would not have met the requirements of the client for automatic table detection.

Test Materials

Requires example PDF documents and expected outputs after OCR extraction has been used on the PDF documents. This will ensure the software is tested using real examples of documents that may be processed.