Carl Zaff
Carl.zaff@outlook.com

April 11, 2023

Method instructions – Tesseract OCR Oryx

This is a very rough first draft for the method instructions. The full process is not described in detail due to lack of time completing the instruction, but still manages to cover roughly 90% of the entire process.

Repository:

https://github.com/CarlZaff/Tesseract-Date-Extraction_Oryxspioenkop

Part 1:

Initial setup:

1. Create a source folder to save all data in, for example C:\Users\Username\Documents*EquipmentAnalysis*

Installation:

Install the following software according to your Operative System. It is recommended to install all software in English.

1. Microsoft Excel Desktop version. Note. This is not free software. However, using a school/work account may work, or a 30day trial.
2. Tesseract. **Important! When installing the software, choose alternative "Install for just me" this allows the framework to automatically find the installation folder** : https://github.com/UB-Mannheim/tesseract/wiki
3. Notepad++: https://notepad-plus-plus.org/downloads/
4. ImageMagick: https://imagemagick.org/script/download.php
5. Python: https://www.python.org/downloads/

Download:

Download the following files to you source folder:

1. Excel Framework: https://github.com/CarlZaff/Tesseract-Date-Extraction_Oryxspioenkop/blob/main/Framework_Oryx_Automatic.xlsm
2. Oryx Sorting Algorithm: https://github.com/CarlZaff/Tesseract-Date-Extraction_Oryxspioenkop/blob/main/Algorithm1_Oryxs_HTML.py
3. Tesseract Algorithm: https://github.com/CarlZaff/Tesseract-Date-Extraction_Oryxspioenkop/blob/main/Algorithm2_Tesseract.py

Carl Zaff
Carl.zaff@outlook.com
April 11, 2023

Installing Algorithms:

1. Open Notepad++.
2. Navigate to tab: Plugins > Plugins admin…
3. Search for: PythonScript
4. Check the box and click Install
5. Navigate to tab: Plugins > Python Script > New Script
6. When the folder opens. Copy the two algorithms earlier downloaded from your source folder into the script folder.
7. Restart Notepad++
8. Algorithms should now be available when navigating to tab: Plugins > Python Script > Scripts

Part 2:

Download the Oryx data:

1. Open the Excel Framework
2. **Note. The source will be counted as untrusted by Microsoft.** In order to activate Macros, follow this link: https://support.microsoft.com/en-us/topic/a-potentially-dangerous-macro-has-been-blocked-0952faa0-37e7-4316-b61d-5b5ed6024216
3. Open Oryxspioenkop image collection: example, https://www.oryxspioenkop.com/2022/02/attack-on-europe-documenting-equipment.html
4. Save the website to the source folder by using CTRL+S.

Sorting and cleaning the Oryx data

1. Navigate to your source folder, right click the newly downloaded Oryx HTML data, choose “Edit with Notepad++”
2. Run the Oryx cleaning algorithm by navigating to Plugins > Python Script > Scripts > Algorithm1_Oryxs_HTML **Note. The give the algorithm time to process. Recommended is to give it at least 60 seconds, or until the document stops moving data.**
3. Copy all data by using CTRL+A then CTRL+C.

Formatting the data in Excel

1. Navigate to Excel, Sheet ‘Setup’. Click the button ‘SPECIFY SOURCE FOLDER’ and navigate to your source folder created earlier.
2. Navigate to Sheet ‘Database_Oryx_Automated’.

3. Paste the copied data from Oryx into cell B4. **Note. The correct cell contains a note within the Framework. This process will also require some time.**
4. Press the 'INPUT FORMULAS' button. Allow the program to process.
5. Press the 'REFRESH' button. Allow the program to process.
6. Control if the equipment count in cell B3 corresponds to the amount specified by Oryxspioenkop. **Note. The value can differ by 0-30 without anything being wrong. If the error is over 5%, the setup has not been completed successfully.**
7. Now you should be able to navigate the Oryx data. By using the Hyperlinks included in Sheet 'Overview' you can quickly jump to whichever unit you want to.

Part 3:

Download the images:

1. Navigate to sheet 'Image_Download'
2. Click the button 'CHOOSE DESTINATION FOLDER', the prompt will allow you to choose your download folder. **Note. By right clicking within the prompt – you can create a new folder for better organization.** Example, C:\Users\Username\Documents\EquipmentAnalysis***Downloads***
3. Click the button 'FETCH DOWNLOADABLE IMAGES'. This will list all downloadable images. **Note. Control how many images will be downloaded by scrolling down the list until the end.**
4. Click the button 'DOWNLOAD IMAGES'. This will start the download. **Note. Depending on your internet speed, this process will take more or less time. The program might stop responding during the process, this is normal. You can control that the images are being downloaded by navigating to your download folder. You will know the status by comparing the number of images in the folder and how many should be downloaded.**

Correct the images

1. Navigate to Excel sheet 'MagickFix'.
2. Click the button 'CREATE IMAGE-CORRECTION SCRIPT'.
3. Navigate to the download folder, click on the image correction script called '0_ImageFix.bat'.
4. Allow the software to correct all images.

Part 4:

Processing the first Batch

1. Navigate to Excel sheet 'ImageMagick_Pre-Processing'.
2. Specify the Threshold value. For example 15.
3. Press shift+F9 to update the sheet. Control that the download folder is correct.

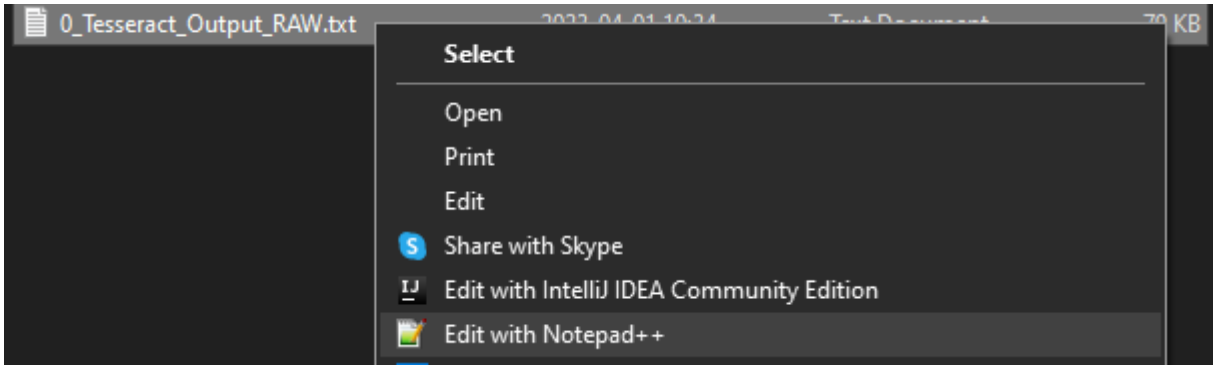
4. Click the button ‘CREATE NEW FOLDER AND LAUNCH PRE-PROCESSING’.
5. Choose a name for your list name in the prompt. **Note. It is important to change the file extension to a text document, on windows: Text (MS-DOS) (*.txt)**
6. Allow the program to process the request. When the process is done, a new window will open.
7. Run the binarization script in the folder. Example ‘0_ImageMagick_Threshold_15.bat’
8. **Note. This process is the most time-consuming part.** Depending on how many images are available, the binarization process will be longer or shorter. To speed up the process, the script can be manually split into 2-6 other instances, allowing the program to use more CPU power. This is achieved by editing the script, and cutting parts of the code and pasting them into a new ‘.bat’ script file. Speeding up the process by roughly 200%.

Tesseract OCR

1. Navigate to sheet ‘Tesseract_Commands’.
2. Click the button ‘CHOOSE FOLDER FOR IMAGES TO OCR’. And find the Threshold folder. Example, C:\Users\Username\Documents\EquipmentAnalysis\Downloads**Images_Threshold_15**
3. Control that the folder name in cell C1 is correct. If not redo the previous step.
4. Tesseract_Results_New_BatchSpecify the Tesseract setting to use, then click the button ‘REFRESH AND LAUNCH TESSERACT’ to analyse the images. This will launch Tesseract OCR.
5. When Tesseract is done with analysing, progress to the next step.

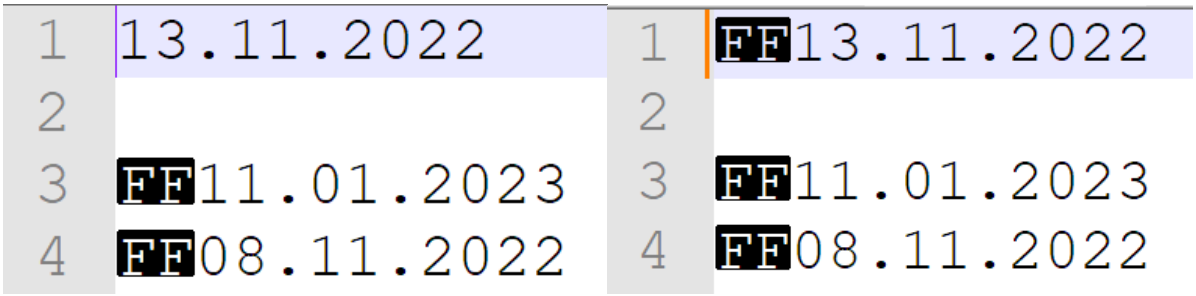
Input the data

1. Navigate to Excel sheet ‘Tesseract_Results’.
2. Click the button ‘IMPORT DATA FROM TESSERACT’.
3. Navigate to the Threshold folder. Right-click the Tesseract output file. Choose ‘Edit with Notepad++’. Example, 0_Tesseract_Output_RAW.txt



4. **Important! Always add an extra ‘FF’ character to the first line, even if the line already has one.**

Carl Zaff
Carl.zaff@outlook.com
April 11, 2023



5. Run the Tesseract algorithm by navigating to Plugins > Python Script > Scripts > Algorithm2_Tesseract
6. When the algorithm is completed, save the document by using ctrl+S
7. Go back to Excel and open Tesseract output file.
8. Click the button 'FIX INCORRECT ROWS' to patch the placeholder IDs.

Part 5, feedback batches:

Processing Batch 2

1. Navigate to Excel sheet 'ImageMagick_Pre-Processing'.
2. Specify the Threshold value. For example 215.
3. Press shift+F9 to update the sheet. Control that the download folder is correct.
4. Click the button 'CREATE NEW FOLDER AND LAUNCH PRE-PROCESSING'.
5. Choose a name for your list name in the prompt. **Note. It is important to change the file extension to a text document, on windows: Text (MS-DOS) (*.txt)**
6. Allow the program to process the request. When the process is done, a new window will open.
7. Run the binarization script in the folder. Example '0_ImageMagick_Threshold_15.bat'
8. **Note. This process is the most time-consuming part.** Depending on how many images are available, the binarization process will be longer or shorter. To speed up the process, the script can be manually split into 2-6 other instances, allowing the program to use more CPU power. This is achieved by editing the script, and cutting parts of the code and pasting them into a new '.bat' script file. Speeding up the process by roughly 200%.

Tesseract OCR

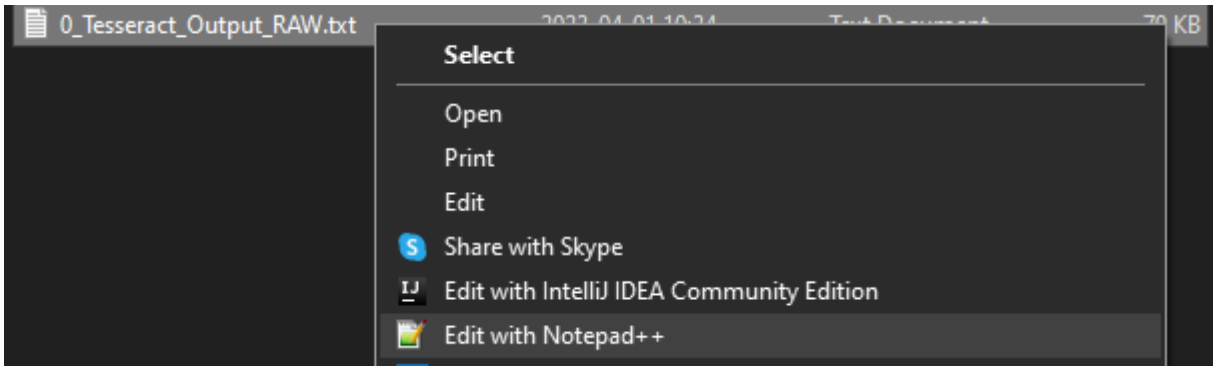
1. Navigate to sheet 'Tesseract_Commands'.
2. Click the button 'CHOOSE FOLDER FOR IMAGES TO OCR'. And find the Threshold folder. Example, C:\Users\Username\Documents\EquipmentAnalysis\Downloads**Images_Threshold_215**
3. Control that the folder name in cell C1 is correct. If not redo the previous step.
4. Specify the Tesseract setting to use, then click the button 'REFRESH AND LAUNCH TESSERACT' to analyse the images. This will launch Tesseract OCR.
5. When Tesseract is done with analysing, progress to the next step.

Carl Zaff
Carl.zaff@outlook.com

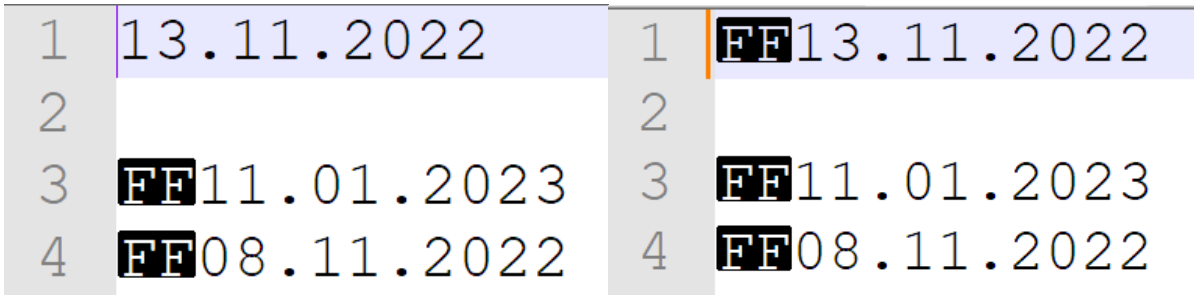
April 11, 2023

Input the data

1. Navigate to Excel sheet 'Tesseract_Results_New_Batch'. **Note. When adding data from a new batch. Always input the data into 'Tesseract_Results_New_Batch'. The new batch will then be included to the 'Tesseract_Results' Sheet.**
2. Click the button 'IMPORT DATA FROM TESSERACT'.
3. Navigate to the latest Threshold folder. Right-click the Tesseract output file. Choose 'Edit with Notepad++'. Example, 0_Tesseract_Output_RAW.txt



4. **Important! Always add an extra 'FF' character to the first line, even if the line already has one.**



5. Run the Tesseract algorithm by navigating to Plugins > Python Script > Scripts > Algorithm2_Tesseract
6. When the algorithm is completed, save the document by using ctrl+S
7. Go back to Excel and open Tesseract output file.
8. Click the button 'FIX INCORRECT ROWS' to patch the placeholder IDs.
9. Click the button 'TRANSFER BATCH WITH MISSING DATES' to transfer the new dates to 'Tesseract_Results'.

Feedback

1. Repeat Part 5, feedback batches for the number of batches wanted.