Portfolio for mini-project 1: digitization

In the first mini-project, you digitised a book. Starting from scanning the book using a library scanner and collecting its metadata, you learned how to turn the scanned images into searchable text using open-source OCR software and how to train the software to recognize the lines, main text and other parts of a page. In the process, you also learned about how training data for machine learning is created, and how to use git, GitHub and Python basics (up to the point where you can use it to compare different transcriptions of a text).

Portfolio overview

To finalize this project, you will hand in a portfolio that consists of the following :

1. [individual] A reviewed version of all blog posts on your website, based on the feedback you received from the instructors
2. [individual] A new website section that presents an overview of the entire digitization process
3. [individual] A short new blog post that alerts the readers of your website that there is a new website section
4. [group] Metadata of the work you scanned.
5. [class] The segmentation guidelines the whole class produced
6. [group] A script that compares the average CAR / CER of different transcription models
7. [group] A zip file containing the images and (uncorrected) OCR transcription of the book, in ALTO xml format, exported from eScriptorium.

The items marked [group] / [class] are group work; you should divide the work among group members ([group]) or the entire class ([class]). Each group member is responsible for the quality of the group work. Make sure to review each other's work and respectfully offer feedback if you notice a problem with the quality of another group member's work.

The items marked [individual] are individual work. This work should be done independently from other students.

Step by step

1. [group] **Create a project plan / log**:
    - create a Google sheet and share it with the members of your group
    - copy the project's tasks into the sheet

    - Sort the tasks in the order they should be executed (some tasks are dependent on other tasks being finished beforehand - e.g., segmentation has to be done before transcription)
    - Assign a due date for each task
    - Assign a group member to each task (make sure to distribute fairly)
    - Keep this project log updated: add the time and date when a task was finalized as you're going.
2. [group] **Clean the scanned images:** Create a folder containing the final version of the images that can be uploaded to the Aga Khan Library website:
    - All pages of the book must be scanned.
    - Each scan must be sharp and free from blur or shadow.
    - Pages must be in correct sequential order.
    - All pages must be scanned right-side up (no rotated images).
    - Images should not include desk backgrounds. Crop if necessary.
    - Remove any pages that were scanned incorrectly, such as blurred, duplicated, or partial scans.
    - If necessary, re-scan an image (ask Pedro for the hardcopy book)
    - Ensure all pages are numbered sequentially (you could ask ChatGPT to write a Python script to renumber the pages if necessary - always keep a backup copy before running the script!)
    - Don't change the image format (jpeg, png, tiff) of the scanned files.
3. [group] **Create the metadata and table of contents**:
    - Use the templates you find on Moodle (and below)
    - Make sure image file names in the table of contents agree with the image file names in the cleaned images folder
4. [class] **Finalize the segmentation guidelines**:
    - Implement the feedback given by the instructors; remember to write them in such a way that they can be reused by external annotators
    - Export the document as PDF and share it with the entire class
5. [group] **Evaluate which transcription model** works best for your group's book
    - **Write a Python script** that compares different transcription models' output to the ground truth of the three corrected pages, based on the Python script we created in session 7. Make sure that
        - the script calculates the mean/average/mode for the CAR / CER of each model across the pages.
        - the code is well-documented using code comments
    - Choose which model works best based on CER / CAR, and write up a short paragraph to justify your choice. Each group member should include this paragraph in the project description page on their portfolio website.
6. [group] **OCR the cleaned images:**
    - Create a new document in your group's eScriptorium project (give it a fitting name)
    - Upload the final versions of the scanned images to this new document

    - Run the finetuned segmentation model on all pages. No need to correct the segmentation at this point.
    - Run the transcription model that worked best for your book on all pages
    - Export the images and transcription (in ALTO XML format)

7. [individual] **Create a folder** in your Github portfolio repository called "project1-digitization"
    - Add to it the following:
        - A text file containing the metadata of the book your group scanned (see the template below)
        - A CSV file containing the table of contents of your group's book (see the template below)
        - A PDF of the segmentation guidelines the class produced
        - The zip file exported from eScriptorium
        - A folder for the CAR / CER comparison, containing:
            - The Python script comparing the CAR /CER of the different transcription models
            - A folder containing the exported transcription layer for the ground truth of the three test pages
            - A folder containing the exported transcription layer for the different transcription models of the three test pages
8. [individual] **Create the project description** on your portfolio website:
    - Create a new website sub-page called "Digitization" (you can find instructions on how to do this in the README file in your portfolio repository on GitHub)
    - Write up an overview of the entire digitization project, aimed at an audience unfamiliar with the course, in which you showcase what you have learned in the first part of the course:
        - At least one paragraph on the goal of the mini-project
        - At least one paragraph on the book you digitized
        - At least one paragraph on the scanning process
        - A couple of paragraphs on the OCR process
        - At least one paragraph of reflections on the entire project.
        - Include in your description links to the relevant files in your GitHub folder and your earlier blog posts
9. [individual] **Create a new blog post** about the finalization of the first project
    - A very short blog post alerting the reader to the finalization of the first mini-project, with a link to your new project description page.
10. [individual] Upload on Moodle a link to your portfolio website's project description page
11. [individual] upload a link to your group's project work plan/log to Moodle

Important remarks

1. If you have any questions, post them to the discussion forum.
2. Always follow **good practices for naming your files and folders**:
    a. File and folder names should be **meaningful**: the name should make it immediately clear to a user what the file/folder contains
    b. File and folder names should be **short**: max. 25 characters for files, 15 for folders
    c. **Don't use spaces or special characters**: stick to ASCII letters, digits, underscores and hyphens
    d. **Use zero-padded numbers** for numbered file names: this ensures that files will always be processed/displayed in the correct order: e.g., Shelley_Frankenstein_001.jpg, Shelley_Frankenstein_002.jpg, instead of Shelley_Frankenstein_1.jpg, Shelley_Frankenstein_2.jpg
3. Use meaningful names for any commits you make to git/GitHub

Template for metadata file:

Title (in original script):
Title (transliterated):
Author(s) (in original script, comma-separated):
Author(s) (transliterated, comma-separated):
Place of publication (in original script):
Place of publication (transliterated):
Publisher (in original script):
Publisher (transliterated):
Year of publication (hijri era):
Year of publication (common era):
Total number of pages:
Keywords (English):

NB:

- Author(s): if the book has more than one author, separate their names using a comma. If the author is unknown, use "Anonymous".
- Use Library of Congress romanization guidelines for the transcription

Template for table of contents:

section_title,start_page_number,start_image_filename
Front matter,NA,digitization_manual_001.png
Chapter 1. Introduction,7,digitization_manual_010.png
Chapter 2. What is digitization,10,digitization_manual_013.png
Back matter,22,digitization_manual_025.png

NB:

- Section titles should be in the original script
- if you're not sure, look up what is considered front and back matter!
- The start page number is the number that is shown on the page