**RenAIssance Project Tests for Prospective GSoC 2025 Applicants**

Below are the tests we will use to evaluate prospective GSoC students for the [**RenAIssance**](https://humanai.foundation/gsoc/projects/2025/project_RenAIssance.html) project. Please thoroughly complete the specific test for your project of interest and (optionally) other tests if you would like to also be considered for additional projects at the same time. This may increase your chances of success, but make sure you don't do it at the expense of the specific project you are interested in.

**Note:** please work in your own github branch (i.e. NO PRs should be made). Send us a link to your code when you are finished by following the instructions below, and we will evaluate it. We encourage you to submit your solutions at least 1 week before the GSoC Proposal Submission deadline, or earlier, so that you have enough time to write the proposal.

**General Dataset:** choose any public dataset appropriate for this text recognition task.

**Specific Dataset:**

**Dataset: text scans -** [**PDF sources scans**](https://bama365-my.sharepoint.com/:f:/g/personal/xgranja_ua_edu/EgOq8lyZrspEuja8pEDtUisB5P-LtlNH8MttYlUvtas3Lg?e=FTBHd2)

**Dataset reference -** [**transcribed sources text**](https://bama365-my.sharepoint.com/:f:/g/personal/xgranja_ua_edu/EliXjgL8OBdLoQwOGwW9yGYBA0Tj2r4pUHAzaJgXqgWKLA?e=tqQZWY)

**Description of the Specific Dataset**

The dataset consists of 6 scanned early modern printed sources. The images have a simple recognition applied that reflects the limitations of the OCR already used (missed letters, incorrectly recognized words...), each source is saved as separate PDF file. The PDF is editable and can be saved as JPEGs if that helps with processing. Marginalia can be ignored, applicants should only focus on the main text, which may vary in layout and page organization depending on the source. The dataset also includes a transcription of the first 3 pages of each PDF source – they should be used as reference while training the AI models for the project. The transcriptions also include a few notes that should help manage expectations of slight variability in spelling errors or limitations throughout the text. Each source has more than the first 3 transcribed pages, so that renAIssance can evaluate the degree of accuracy and viability of the test method employed.

**Tasks:**

* Those interested in the Layout Organization Recognition should complete Test I
* Those interested in the Optical Character Recognition should complete Test II.

---------------------------------------------------------------------------------------------------------------------**Specific Test I. Layout Organization Recognition**

**Task:** Build a model based on convolutional-recurrent, transformer, or self-supervised architectures for optically recognizing the layout of each data source. Your model should be able to detect where in each page main text is present, and disregard other embellishments. Pick the most appropriate approach and discuss your strategy.

**Evaluation Metrics:** discuss which evaluation metrics you are using to evaluate your model performance

---------------------------------------------------------------------------------------------------------------------

**Specific Test II. Optical Character Recognition**

**Task:** Build a model based on convolutional-recurrent, transformer, or self-supervised architectures for optically recognizing the text of each data source. Your model should be able to detect the main text in each page, while disregarding other embellishments. Pick the most appropriate approach and discuss your strategy.

**Evaluation Metrics:** discuss which evaluation metrics you are using to evaluate your model performance

--------------------------------------------------------------------------------------------------------------------- **Specific Test III. Synthetic renaissance text generation with generative models**

**Task:** Design a Mid-scale generative model to create synthetic Renaissance-style printed text. The model should introduce realistic printing imperfections such as ink bleed, smudging, or faded text. Use a sample of historical Spanish text (from the 17th century, available as part of other specific tests) and generate images for at least 5 pages from the Word file with visible degradation effects. Document the approach, including model choice (GANs, diffusion models, or another method), dataset preparation, and implementation steps. Also, evaluate how well the generated text mimics historical printing artifacts using a quantitative metric.

**Evaluation Metrics:** discuss which evaluation metrics you are using to evaluate your model performance

--------------------------------------------------------------------------------------------------------------------- **Submission Guidelines:**

Please send us your CV and a link to all your completed work (github repo, Jupyter notebook + pdf of Jupyter notebook with output) to [human-ai@cern.ch](mailto:human-ai@cern.ch) with Evaluation Test: RenAIssance in the title.