# ĐỒ ÁN GIỮA KỲ

## Ngữ liệu: Tổng hợp văn học việt nam 06 (Trang 57-96)

* **Cấu trúc thư mục của đồ án**

A screenshot of a computer

Description automatically generated

* **Chức năng của từng thư mục**
  + **./data:** Chứa những dữ liệu cần xử lý và dùng để lưu những dữ liệu sẽ xử lý.
  + **./clean:** Chứa dữ liệu được xử lý và dữ liệu này được dùng để dóng hàng
  + **./images:** chứa các ảnh được chiết xuất từ ngữ liệu thô
  + **./ocr:** Chứa các dữ liệu được ocr (hán nôm và chữ quốc ngữ)
  + **./resources:** Chứa ngữ liệu thô và từ điển hán nôm thầy đã cung cấp
    - A screenshot of a computer

      Description automatically generated
  + **./result:** Chứa kết quả khi đã tiến hành dóng hàng và tô màu ngữ liệu.
* **Các bước xử lý ngữ liệu**
  + **Bước 1:** Chạy file *extract\_pdf.py* để tiến hành chiết xuất những hình ảnh trong ngữ liệu lưu vào thư mục **./images**
  + **Bước 2:** Chạy file *extract\_ocr.py* để tiến hành ocr văn bản trong file ảnh và lưu kết quả vào thư mục **./ocr**
  + **Bước 3:** Sử dụng colab để gọi api của Gemini AI để tiến hành sửa lỗi chính tả của kết quả (*ocr\_results.quocngu.json*) vừa được ocr. Lưu lại kết quả trong thư mục **./ocr** có tên là *gemini\_api\_results.quocngu.json*
  + **Bước 4:** Chạy file *preprocess\_quocngu\_ocr.py* để làm sạch dữ liệu và định dạng lại dữ liệu lưu vào thư mục **./clean** để chuẩn bị cho việc dóng hàng
  + **Bước 5:** Chạy file *request.py* để tiến hành gọi api từ tool kandianguji và lưu kết quả ở thư mục **./ocr** có tên *ocr\_results.hannom.json*
  + **Bước 6:** Chạy file *preprocess\_hannom\_ocr.py* để tiến hành làm sạch dữ liệu hán nôm được ocr
  + **Bước 7:** Chạy file *concat\_hannom.py* để tiến hành nối lại những dòng hán nôm bị tách ra. Bằng cách kiểm tra tọa độ các dòng xem nó có phải thuộc một dòng không với sai số nhất định nào đó. Sau đó xử lý nối box lại với nhau, cập nhật tọa độ và xóa box thừa.
  + **Bước 8**: Chạy file *convert\_output.py* để tiến hành dóng hàng và tô màu ngữ liệu lưu lại vào file *result.xlsx*.