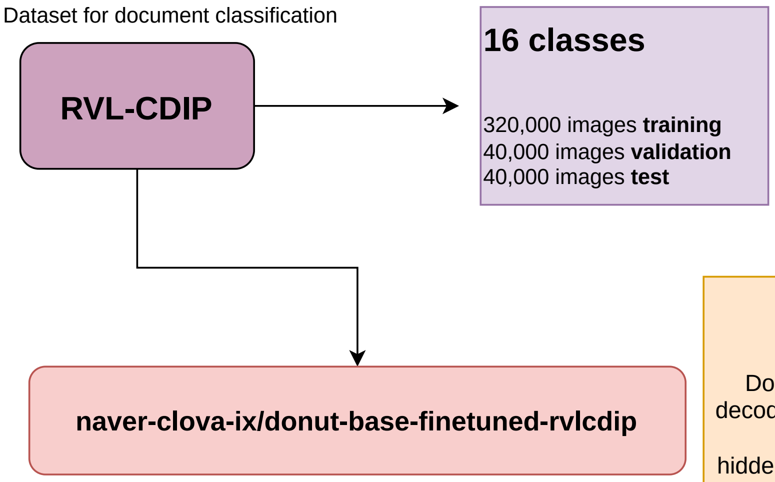


Dataset for document classification

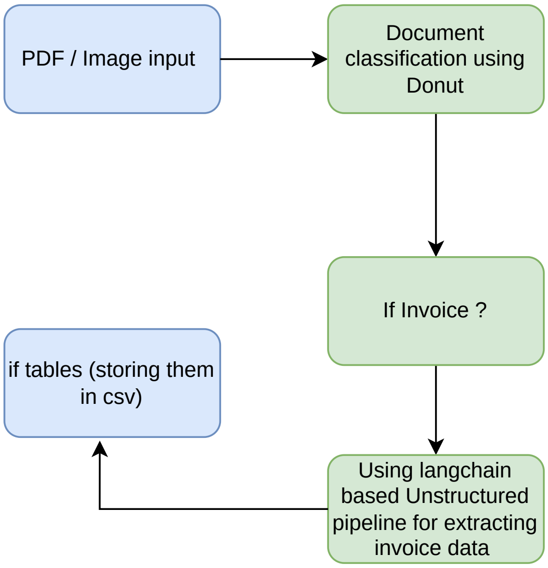
RVL-CDIP

16 classes

320,000 images **training**
40,000 images **validation**
40,000 images **test**

naver-clova-ix/donut-base-finetuned-rvlcdip

Donut consists of a vision encoder (Swin Transformer) and a text decoder (BART). Given an image, the encoder first encodes the image into a tensor of embeddings (of shape batch_size, seq_len, hidden_size), after which the decoder autoregressively generates text, conditioned on the encoding of the encoder.



PDF / Image input

Document classification using Donut

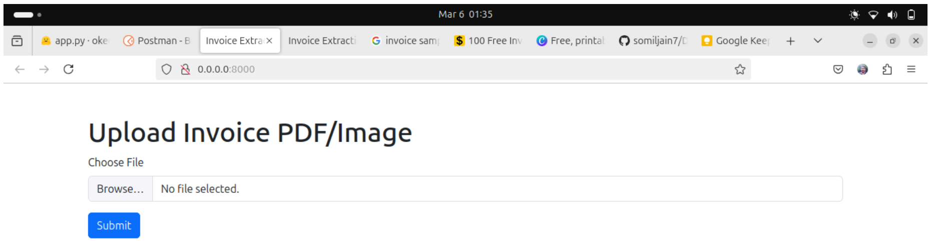
If Invoice ?

Using langchain based Unstructured pipeline for extracting invoice data

if tables (storing them in csv)

FUTURE OPTIONS:

- a yolo v11 finetune for rvlclip dataset
- benchmarking yolov11 vs donut (yolo takes less cpu and ram utilisation compared to a transformer)
- Unstructured can also be benchmarked with docling.

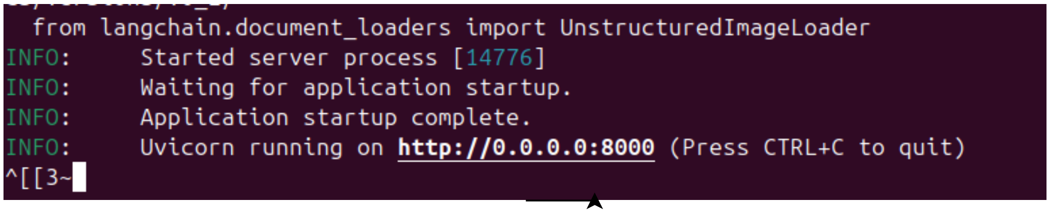


Upload Invoice PDF/Image

Choose File

Browse... No file selected.

Submit



```
  from langchain.document_loaders import UnstructuredImageLoader
INFO:     Started server process [14776]
INFO:     Waiting for application startup.
INFO:     Application startup complete.
INFO:     Uvicorn running on http://0.0.0.0:8000 (Press CTRL+C to quit)
^[[3~
```