```
!pip install pandas==2.1.3 fsspec==2023.6.0 pillow==10.0.0
wrapt==1.14.0 protobuf==4.20.0 torch==2.1.0

Collecting pandas==2.1.3
  Using cached pandas-2.1.3-cp310-cp310-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl (12.3 MB)
Collecting fsspec==2023.6.0
  Downloading fsspec-2023.6.0-py3-none-any.whl (163 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 163.8/163.8 kB 3.3 MB/s eta
0:00:00
anylinux_2_28_x86_64.whl (3.4 MB)
Collecting wrapt==1.14.0
  Using cached wrapt-1.14.0-cp310-cp310-
manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_17_x86_64.manylinux
2014_x86_64.whl (77 kB)
ERROR: Could not find a version that satisfies the requirement
protobuf==4.20.0 (from versions: 2.0.0b0, 2.0.3, 2.3.0, 2.4.1, 2.5.0,
2.6.0, 2.6.1, 3.0.0a2, 3.0.0a3, 3.0.0b1, 3.0.0b1.post1, 3.0.0b1.post2,
3.0.0b2, 3.0.0b2.post1, 3.0.0b2.post2, 3.0.0b3, 3.0.0b4, 3.0.0, 3.1.0,
3.1.0.post1, 3.2.0rc1, 3.2.0rc1.post1, 3.2.0rc2, 3.2.0, 3.3.0, 3.4.0,
3.5.0.post1, 3.5.1, 3.5.2, 3.5.2.post1, 3.6.0, 3.6.1, 3.7.0rc2,
3.7.0rc3, 3.7.0, 3.7.1, 3.8.0rc1, 3.8.0, 3.9.0rc1, 3.9.0, 3.9.1,
3.9.2, 3.10.0rc1, 3.10.0, 3.11.0rc1, 3.11.0rc2, 3.11.0, 3.11.1,
3.11.2, 3.11.3, 3.12.2, 3.12.4, 3.13.0rc3, 3.13.0, 3.14.0rc1,
3.14.0rc2, 3.14.0rc3, 3.14.0, 3.15.0rc1, 3.15.0rc2, 3.15.0, 3.15.1,
3.15.2, 3.15.3, 3.15.4, 3.15.5, 3.15.6, 3.15.7, 3.15.8, 3.16.0rc1,
3.16.0rc2, 3.16.0, 3.17.0rc1, 3.17.0rc2, 3.17.0, 3.17.1, 3.17.2,
3.17.3, 3.18.0rc1, 3.18.0rc2, 3.18.0, 3.18.1, 3.18.3, 3.19.0rc1,
3.19.0rc2, 3.19.0, 3.19.1, 3.19.2, 3.19.3, 3.19.4, 3.19.5, 3.19.6,
3.20.0rc1, 3.20.0rc2, 3.20.0, 3.20.1rc1, 3.20.1, 3.20.2, 3.20.3,
4.0.0rc1, 4.0.0rc2, 4.21.0rc1, 4.21.0rc2, 4.21.0, 4.21.1, 4.21.2,
4.21.3, 4.21.4, 4.21.5, 4.21.6, 4.21.7, 4.21.8, 4.21.9, 4.21.10,
4.21.11, 4.21.12, 4.22.0rc2, 4.22.0rc3, 4.22.0, 4.22.1, 4.22.3,
4.22.4, 4.22.5, 4.23.0rc2, 4.23.0rc3, 4.23.0, 4.23.1, 4.23.2, 4.23.3,
4.23.4, 4.24.0rc1, 4.24.0rc2, 4.24.0rc3, 4.24.0, 4.24.1, 4.24.2,
4.24.3, 4.24.4, 4.25.0rc1, 4.25.0rc2, 4.25.0, 4.25.1, 4.25.2, 4.25.3,
5.26.0rc1, 5.26.0rc2, 5.26.0rc3)
ERROR: No matching distribution found for protobuf==4.20.0

!pip install langchain unstructured[all-docs] pydantic lxml openai
chromadb tiktoken

Requirement already satisfied: langchain in
/usr/local/lib/python3.10/dist-packages (0.1.11)
Requirement already satisfied: unstructured[all-docs] in
/usr/local/lib/python3.10/dist-packages (0.12.6)
Requirement already satisfied: pydantic in
/usr/local/lib/python3.10/dist-packages (2.6.3)
Requirement already satisfied: lxml in /usr/local/lib/python3.10/dist-
packages (5.1.0)
```

```
Requirement already satisfied: openai in
/usr/local/lib/python3.10/dist-packages (1.13.3)
Requirement already satisfied: chromadb in
/usr/local/lib/python3.10/dist-packages (0.4.17)
Requirement already satisfied: tiktoken in
/usr/local/lib/python3.10/dist-packages (0.6.0)
Requirement already satisfied: PyYAML>=5.3 in
/usr/local/lib/python3.10/dist-packages (from langchain) (6.0.1)
Requirement already satisfied: SQLAlchemy<3,>=1.4 in
/usr/local/lib/python3.10/dist-packages (from langchain) (2.0.28)
Requirement already satisfied: aiohttp<4.0.0,>=3.8.3 in
/usr/local/lib/python3.10/dist-packages (from langchain) (3.9.3)
Requirement already satisfied: async-timeout<5.0.0,>=4.0.0 in
/usr/local/lib/python3.10/dist-packages (from langchain) (4.0.3)
Requirement already satisfied: dataclasses-json<0.7,>=0.5.7 in
/usr/local/lib/python3.10/dist-packages (from langchain) (0.6.4)
Requirement already satisfied: jsonpatch<2.0,>=1.33 in
/usr/local/lib/python3.10/dist-packages (from langchain) (1.33)
Requirement already satisfied: langchain-community<0.1,>=0.0.25 in
/usr/local/lib/python3.10/dist-packages (from langchain) (0.0.27)
Requirement already satisfied: langchain-core<0.2,>=0.1.29 in
/usr/local/lib/python3.10/dist-packages (from langchain) (0.1.30)
Requirement already satisfied: langchain-text-splitters<0.1,>=0.0.1 in
/usr/local/lib/python3.10/dist-packages (from langchain) (0.0.1)
Requirement already satisfied: langsmith<0.2.0,>=0.1.17 in
/usr/local/lib/python3.10/dist-packages (from langchain) (0.1.23)
Requirement already satisfied: numpy<2,>=1 in
/usr/local/lib/python3.10/dist-packages (from langchain) (1.26.4)
Requirement already satisfied: requests<3,>=2 in
/usr/local/lib/python3.10/dist-packages (from langchain) (2.31.0)
Requirement already satisfied: tenacity<9.0.0,>=8.1.0 in
/usr/local/lib/python3.10/dist-packages (from langchain) (8.2.3)
Requirement already satisfied: backoff==2.2.1 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(2.2.1)
Requirement already satisfied: beautifulsoup4==4.12.3 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(4.12.3)
Requirement already satisfied: certifi==2024.2.2 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(2024.2.2)
Requirement already satisfied: chardet==5.2.0 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(5.2.0)
Requirement already satisfied: charset-normalizer==3.3.2 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(3.3.2)
Requirement already satisfied: click==8.1.7 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
```

```
(8.1.7)
Requirement already satisfied: dataclasses-json-speakeasy==0.5.11
in /usr/local/lib/python3.10/dist-packages (from unstructured[all-
docs]) (0.5.11)
Requirement already satisfied: emoji==2.10.1 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(2.10.1)
Requirement already satisfied: filetype==1.2.0 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(1.2.0)
Requirement already satisfied: idna==3.6 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(3.6)
Requirement already satisfied: joblib==1.3.2 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(1.3.2)
Requirement already satisfied: jsonpath-python==1.0.6 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(1.0.6)
Requirement already satisfied: langdetect==1.0.9 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(1.0.9)
Requirement already satisfied: marshmallow==3.20.2 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(3.20.2)
Requirement already satisfied: mypy-extensions==1.0.0 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(1.0.0)
Requirement already satisfied: nltk==3.8.1 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(3.8.1)
Requirement already satisfied: packaging==23.2 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(23.2)
Requirement already satisfied: python-dateutil==2.8.2 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(2.8.2)
Requirement already satisfied: python-iso639==2024.2.7 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(2024.2.7)
Requirement already satisfied: python-magic==0.4.27 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(0.4.27)
Requirement already satisfied: rapidfuzz==3.6.1 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(3.6.1)
Requirement already satisfied: regex==2023.12.25 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(2023.12.25)
```

```
Requirement already satisfied: six==1.16.0 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(1.16.0)
Requirement already satisfied: soupsieve==2.5 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(2.5)
Requirement already satisfied: tabulate==0.9.0 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(0.9.0)
Requirement already satisfied: tqdm==4.66.2 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(4.66.2)
Requirement already satisfied: typing-extensions==4.9.0 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(4.9.0)
Requirement already satisfied: typing-inspect==0.9.0 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(0.9.0)
Requirement already satisfied: unstructured-client==0.18.0 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(0.18.0)
Requirement already satisfied: urllib3==1.26.18 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(1.26.18)
Requirement already satisfied: wrapt==1.16.0 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(1.16.0)
Requirement already satisfied: filelock==3.13.1 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(3.13.1)
Requirement already satisfied: pycocotools==2.0.7 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(2.0.7)
Requirement already satisfied: pypandoc==1.12 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(1.12)
Requirement already satisfied: humanfriendly==10.0 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(10.0)
Requirement already satisfied: pypdf==4.0.1 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(4.0.1)
Requirement already satisfied: pdfplumber==0.10.4 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(0.10.4)
Requirement already satisfied: pdfminer-six==20221105 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(20221105)
Requirement already satisfied: safetensors==0.3.2 in
```

```
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(0.3.2)
Requirement already satisfied: msg-parser==1.2.0 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(1.2.0)
Requirement already satisfied: unstructured-inference==0.7.23 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(0.7.23)
Requirement already satisfied: cycler==0.12.1 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(0.12.1)
Requirement already satisfied: antlr4-python3-runtime==4.9.3 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(4.9.3)
Requirement already satisfied: openpyxl==3.1.2 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(3.1.2)
Requirement already satisfied: olefile==0.47 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(0.47)
Requirement already satisfied: torch==2.2.0 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(2.2.0)
Requirement already satisfied: cryptography==42.0.2 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(42.0.2)
Requirement already satisfied: pikepdf==8.11.0 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(8.11.0)
Requirement already satisfied: tokenizers==0.15.2 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(0.15.2)
Requirement already satisfied: deprecated==1.2.14 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(1.2.14)
Requirement already satisfied: zipp==3.17.0 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(3.17.0)
Requirement already satisfied: pyparsing==3.0.9 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(3.0.9)
Requirement already satisfied: torchvision==0.17.0 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(0.17.0)
Requirement already satisfied: pycparser==2.21 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(2.21)
Requirement already satisfied: pytesseract==0.3.10 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
```

```
(0.3.10)
Requirement already satisfied: timm==0.9.12 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(0.9.12)
Requirement already satisfied: python-pptx==0.6.23 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(0.6.23)
Requirement already satisfied: transformers==4.37.1 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(4.37.1)
Requirement already satisfied: mpmath==1.3.0 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(1.3.0)
Requirement already satisfied: effdet==0.4.1 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(0.4.1)
Requirement already satisfied: markupsafe==2.1.5 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(2.1.5)
Requirement already satisfied: et-xmlfile==1.1.0 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(1.1.0)
Requirement already satisfied: omegaconf==2.3.0 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(2.3.0)
Requirement already satisfied: xlrd==2.0.1 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(2.0.1)
Requirement already satisfied: portalocker==2.8.2 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(2.8.2)
Requirement already satisfied: python-docx==1.1.0 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(1.1.0)
Requirement already satisfied: pdf2image==1.17.0 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(1.17.0)
Requirement already satisfied: pillow==10.2.0 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(10.2.0)
Requirement already satisfied: pillow-heif==0.15.0 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(0.15.0)
Requirement already satisfied: fonttools==4.49.0 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(4.49.0)
Requirement already satisfied: onnxruntime==1.15.1 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(1.15.1)
```

```
Requirement already satisfied: iopath==0.1.10 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(0.1.10)
Requirement already satisfied: xlsxwriter==3.1.9 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(3.1.9)
Requirement already satisfied: onnx==1.15.0 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(1.15.0)
Requirement already satisfied: tzdata==2024.1 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(2024.1)
Requirement already satisfied: scipy==1.10.1 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(1.10.1)
Requirement already satisfied: networkx==3.2.1 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(3.2.1)
Requirement already satisfied: protobuf==4.23.4 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(4.23.4)
Requirement already satisfied: fsspec==2024.2.0 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(2024.2.0)
Requirement already satisfied: huggingface-hub==0.20.3 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(0.20.3)
Requirement already satisfied: cffi==1.16.0 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(1.16.0)
Requirement already satisfied: kiwisolver==1.4.5 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(1.4.5)
Requirement already satisfied: matplotlib==3.7.2 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(3.7.2)
Requirement already satisfied: pandas==2.2.0 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(2.2.0)
Requirement already satisfied: importlib-metadata==7.0.1 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(7.0.1)
Requirement already satisfied:
layoutparser[layoutmodels,tesseract]==0.3.4 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(0.3.4)
Requirement already satisfied: contourpy==1.2.0 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(1.2.0)
```

```
Requirement already satisfied: pypdfium2==4.27.0 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(4.27.0)
Requirement already satisfied: opencv-python==4.8.0.76 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(4.8.0.76)
Requirement already satisfied: pytz==2024.1 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(2024.1)
Requirement already satisfied: flatbuffers==23.5.26 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(23.5.26)
Requirement already satisfied: importlib-resources==6.1.1 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(6.1.1)
Requirement already satisfied: jinja2==3.1.3 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(3.1.3)
Requirement already satisfied: python-multipart==0.0.9 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(0.0.9)
Requirement already satisfied: sympy==1.12 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(1.12)
Requirement already satisfied: markdown==3.5.2 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(3.5.2)
Requirement already satisfied: coloredlogs==15.0.1 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(15.0.1)
Requirement already satisfied: unstructured-pytesseract==0.3.12 in
/usr/local/lib/python3.10/dist-packages (from unstructured[all-docs])
(0.3.12)
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.1.105 in
/usr/local/lib/python3.10/dist-packages (from torch==2.2.0-
>unstructured[all-docs]) (12.1.105)
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.1.105
in /usr/local/lib/python3.10/dist-packages (from torch==2.2.0-
>unstructured[all-docs]) (12.1.105)
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.1.105 in
/usr/local/lib/python3.10/dist-packages (from torch==2.2.0-
>unstructured[all-docs]) (12.1.105)
Requirement already satisfied: nvidia-cudnn-cu12==8.9.2.26 in
/usr/local/lib/python3.10/dist-packages (from torch==2.2.0-
>unstructured[all-docs]) (8.9.2.26)
Requirement already satisfied: nvidia-cublas-cu12==12.1.3.1 in
/usr/local/lib/python3.10/dist-packages (from torch==2.2.0-
>unstructured[all-docs]) (12.1.3.1)
Requirement already satisfied: nvidia-cufft-cu12==11.0.2.54 in
```

```
/usr/local/lib/python3.10/dist-packages (from torch==2.2.0-
>unstructured[all-docs]) (11.0.2.54)
Requirement already satisfied: nvidia-curand-cu12==10.3.2.106 in
/usr/local/lib/python3.10/dist-packages (from torch==2.2.0-
>unstructured[all-docs]) (10.3.2.106)
Requirement already satisfied: nvidia-cusolver-cu12==11.4.5.107 in
/usr/local/lib/python3.10/dist-packages (from torch==2.2.0-
>unstructured[all-docs]) (11.4.5.107)
Requirement already satisfied: nvidia-cusparse-cu12==12.1.0.106 in
/usr/local/lib/python3.10/dist-packages (from torch==2.2.0-
>unstructured[all-docs]) (12.1.0.106)
Requirement already satisfied: nvidia-nccl-cu12==2.19.3 in
/usr/local/lib/python3.10/dist-packages (from torch==2.2.0-
>unstructured[all-docs]) (2.19.3)
Requirement already satisfied: nvidia-nvtx-cu12==12.1.105 in
/usr/local/lib/python3.10/dist-packages (from torch==2.2.0-
>unstructured[all-docs]) (12.1.105)
Requirement already satisfied: triton==2.2.0 in
/usr/local/lib/python3.10/dist-packages (from torch==2.2.0-
>unstructured[all-docs]) (2.2.0)
Requirement already satisfied: nvidia-nvjitlink-cu12 in
/usr/local/lib/python3.10/dist-packages (from nvidia-cusolver-
cu12==11.4.5.107->torch==2.2.0->unstructured[all-docs]) (12.4.99)
Requirement already satisfied: annotated-types>=0.4.0 in
/usr/local/lib/python3.10/dist-packages (from pydantic) (0.6.0)
Requirement already satisfied: pydantic-core==2.16.3 in
/usr/local/lib/python3.10/dist-packages (from pydantic) (2.16.3)
Requirement already satisfied: anyio<5,>=3.5.0 in
/usr/local/lib/python3.10/dist-packages (from openai) (3.7.1)
Requirement already satisfied: distro<2,>=1.7.0 in
/usr/lib/python3/dist-packages (from openai) (1.7.0)
Requirement already satisfied: httpx<1,>=0.23.0 in
/usr/local/lib/python3.10/dist-packages (from openai) (0.27.0)
Requirement already satisfied: sniffio in
/usr/local/lib/python3.10/dist-packages (from openai) (1.3.1)
Requirement already satisfied: chroma-hnswlib==0.7.3 in
/usr/local/lib/python3.10/dist-packages (from chromadb) (0.7.3)
Requirement already satisfied: fastapi>=0.95.2 in
/usr/local/lib/python3.10/dist-packages (from chromadb) (0.110.0)
Requirement already satisfied: uvicorn[standard]>=0.18.3 in
/usr/local/lib/python3.10/dist-packages (from chromadb) (0.28.0)
Requirement already satisfied: posthog>=2.4.0 in
/usr/local/lib/python3.10/dist-packages (from chromadb) (3.5.0)
Requirement already satisfied: pulsar-client>=3.1.0 in
/usr/local/lib/python3.10/dist-packages (from chromadb) (3.4.0)
Requirement already satisfied: opentelemetry-api>=1.2.0 in
/usr/local/lib/python3.10/dist-packages (from chromadb) (1.16.0)
Requirement already satisfied: opentelemetry-exporter-otlp-proto-
grpc>=1.2.0 in /usr/local/lib/python3.10/dist-packages (from chromadb)
```

```
(1.16.0)
Requirement already satisfied: opentelemetry-sdk>=1.2.0 in
/usr/local/lib/python3.10/dist-packages (from chromadb) (1.16.0)
Requirement already satisfied: pypika>=0.48.9 in
/usr/local/lib/python3.10/dist-packages (from chromadb) (0.48.9)
Requirement already satisfied: overrides>=7.3.1 in
/usr/local/lib/python3.10/dist-packages (from chromadb) (7.7.0)
Requirement already satisfied: grpcio>=1.58.0 in
/usr/local/lib/python3.10/dist-packages (from chromadb) (1.62.0)
Requirement already satisfied: bcrypt>=4.0.1 in
/usr/local/lib/python3.10/dist-packages (from chromadb) (4.1.2)
Requirement already satisfied: typer>=0.9.0 in
/usr/local/lib/python3.10/dist-packages (from chromadb) (0.9.0)
Requirement already satisfied: kubernetes>=28.1.0 in
/usr/local/lib/python3.10/dist-packages (from chromadb) (29.0.0)
Requirement already satisfied: aiosignal>=1.1.2 in
/usr/local/lib/python3.10/dist-packages (from aiohttp<4.0.0,>=3.8.3-
>langchain) (1.3.1)
Requirement already satisfied: attrs>=17.3.0 in
/usr/local/lib/python3.10/dist-packages (from aiohttp<4.0.0,>=3.8.3-
>langchain) (23.2.0)
Requirement already satisfied: frozenlist>=1.1.1 in
/usr/local/lib/python3.10/dist-packages (from aiohttp<4.0.0,>=3.8.3-
>langchain) (1.4.1)
Requirement already satisfied: multidict<7.0,>=4.5 in
/usr/local/lib/python3.10/dist-packages (from aiohttp<4.0.0,>=3.8.3-
>langchain) (6.0.5)
Requirement already satisfied: yarl<2.0,>=1.0 in
/usr/local/lib/python3.10/dist-packages (from aiohttp<4.0.0,>=3.8.3-
>langchain) (1.9.4)
Requirement already satisfied: exceptiongroup in
/usr/local/lib/python3.10/dist-packages (from anyio<5,>=3.5.0->openai)
(1.2.0)
Requirement already satisfied: starlette<0.37.0,>=0.36.3 in
/usr/local/lib/python3.10/dist-packages (from fastapi>=0.95.2-
>chromadb) (0.36.3)
Requirement already satisfied: httpcore==1.* in
/usr/local/lib/python3.10/dist-packages (from httpx<1,>=0.23.0-
>openai) (1.0.4)
Requirement already satisfied: h11<0.15,>=0.13 in
/usr/local/lib/python3.10/dist-packages (from httpcore==1.*-
>httpx<1,>=0.23.0->openai) (0.14.0)
Requirement already satisfied: jsonpointer>=1.9 in
/usr/local/lib/python3.10/dist-packages (from jsonpatch<2.0,>=1.33-
>langchain) (2.4)
Requirement already satisfied: google-auth>=1.0.1 in
/usr/local/lib/python3.10/dist-packages (from kubernetes>=28.1.0-
>chromadb) (2.27.0)
Requirement already satisfied: websocket-client!=0.40.0,!=0.41.*,!
```

```
=0.42.*,>=0.32.0 in /usr/local/lib/python3.10/dist-packages (from
kubernetes>=28.1.0->chromadb) (1.7.0)
Requirement already satisfied: requests-oauthlib in
/usr/local/lib/python3.10/dist-packages (from kubernetes>=28.1.0-
>chromadb) (1.3.1)
Requirement already satisfied: oauthlib>=3.2.2 in
/usr/local/lib/python3.10/dist-packages (from kubernetes>=28.1.0-
>chromadb) (3.2.2)
Requirement already satisfied: orjson<4.0.0,>=3.9.14 in
/usr/local/lib/python3.10/dist-packages (from
langsmith<0.2.0,>=0.1.17->langchain) (3.9.15)
Requirement already satisfied: setuptools>=16.0 in
/usr/local/lib/python3.10/dist-packages (from opentelemetry-
api>=1.2.0->chromadb) (67.7.2)
Requirement already satisfied: googleapis-common-protos~=1.52 in
/usr/local/lib/python3.10/dist-packages (from opentelemetry-exporter-
otlp-proto-grpc>=1.2.0->chromadb) (1.62.0)
Requirement already satisfied: opentelemetry-proto==1.16.0 in
/usr/local/lib/python3.10/dist-packages (from opentelemetry-exporter-
otlp-proto-grpc>=1.2.0->chromadb) (1.16.0)
Requirement already satisfied: opentelemetry-semantic-
conventions==0.37b0 in /usr/local/lib/python3.10/dist-packages (from
opentelemetry-sdk>=1.2.0->chromadb) (0.37b0)
Requirement already satisfied: monotonic>=1.5 in
/usr/local/lib/python3.10/dist-packages (from posthog>=2.4.0-
>chromadb) (1.6)
Requirement already satisfied: greenlet!=0.4.17 in
/usr/local/lib/python3.10/dist-packages (from SQLAlchemy<3,>=1.4-
>langchain) (3.0.3)
Requirement already satisfied: httptools>=0.5.0 in
/usr/local/lib/python3.10/dist-packages (from
uvicorn[standard]>=0.18.3->chromadb) (0.6.1)
Requirement already satisfied: python-dotenv>=0.13 in
/usr/local/lib/python3.10/dist-packages (from
uvicorn[standard]>=0.18.3->chromadb) (1.0.1)
Requirement already satisfied: uvloop!=0.15.0,!=0.15.1,>=0.14.0 in
/usr/local/lib/python3.10/dist-packages (from
uvicorn[standard]>=0.18.3->chromadb) (0.19.0)
Requirement already satisfied: watchfiles>=0.13 in
/usr/local/lib/python3.10/dist-packages (from
uvicorn[standard]>=0.18.3->chromadb) (0.21.0)
Requirement already satisfied: websockets>=10.4 in
/usr/local/lib/python3.10/dist-packages (from
uvicorn[standard]>=0.18.3->chromadb) (12.0)
Requirement already satisfied: cachetools<6.0,>=2.0.0 in
/usr/local/lib/python3.10/dist-packages (from google-auth>=1.0.1-
>kubernetes>=28.1.0->chromadb) (5.3.3)
Requirement already satisfied: pyasn1-modules>=0.2.1 in
/usr/local/lib/python3.10/dist-packages (from google-auth>=1.0.1-
>kubernetes>=28.1.0->chromadb) (0.3.0)
```

```
Requirement already satisfied: rsa<5,>=3.1.4 in
/usr/local/lib/python3.10/dist-packages (from google-auth>=1.0.1-
>kubernetes>=28.1.0->chromadb) (4.9)
Requirement already satisfied: pyasn1<0.6.0,>=0.4.6 in
/usr/local/lib/python3.10/dist-packages (from pyasn1-modules>=0.2.1-
>google-auth>=1.0.1->kubernetes>=28.1.0->chromadb) (0.5.1)
```

```
from typing import Any
import os
from unstructured.partition.pdf import partition_pdf
import pytesseract
import os

pytesseract.pytesseract.tesseract_cmd = r'C:\Program Files\Tesseract-
OCR\tesseract.exe'

input_path = os.getcwd()
output_path = os.path.join(os.getcwd(), "output")

# Print the value of tesseract_cmd for debugging
print("Tesseract executable path:",
pytesseract.pytesseract.tesseract_cmd)
```

```
Tesseract executable path: C:\Program Files\Tesseract-OCR\
tesseract.exe
```

```
# Get elements
print("Input path:", input_path)
print("Output path:", output_path)
```

```
Input path: /content
Output path: /content/output
```

```
# Install pdftoppm
!apt-get install poppler-utils
```

```
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following NEW packages will be installed:
  poppler-utils
0 upgraded, 1 newly installed, 0 to remove and 35 not upgraded.
Need to get 186 kB of archives.
After this operation, 696 kB of additional disk space will be used.
Get:1 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64
poppler-utils amd64 22.02.0-2ubuntu0.3 [186 kB]
Fetched 186 kB in 0s (1,020 kB/s)
Selecting previously unselected package poppler-utils.
(Reading database ... 121749 files and directories currently
installed.)
Preparing to unpack .../poppler-utils_22.02.0-2ubuntu0.3_amd64.deb ...
```

```
Unpacking poppler-utils (22.02.0-2ubuntu0.3) ...
Setting up poppler-utils (22.02.0-2ubuntu0.3) ...
Processing triggers for man-db (2.10.2-1) ...
```

```
# Import required libraries
import subprocess

!apt-get install tesseract-ocr
```

```
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  tesseract-ocr-eng tesseract-ocr-osd
The following NEW packages will be installed:
  tesseract-ocr tesseract-ocr-eng tesseract-ocr-osd
0 upgraded, 3 newly installed, 0 to remove and 35 not upgraded.
Need to get 4,816 kB of archives.
After this operation, 15.6 MB of additional disk space will be used.
Get:1 http://archive.ubuntu.com/ubuntu jammy/universe amd64 tesseract-
ocr-eng all 1:4.00~git30-7274cfa-1.1 [1,591 kB]
Get:2 http://archive.ubuntu.com/ubuntu jammy/universe amd64 tesseract-
ocr-osd all 1:4.00~git30-7274cfa-1.1 [2,990 kB]
Get:3 http://archive.ubuntu.com/ubuntu jammy/universe amd64 tesseract-
ocr amd64 4.1.1-2.1build1 [236 kB]
Fetched 4,816 kB in 0s (26.5 MB/s)
Selecting previously unselected package tesseract-ocr-eng.
(Reading database ... 121779 files and directories currently
installed.)
Preparing to unpack .../tesseract-ocr-eng_1%3a4.00~git30-7274cfa-
1.1_all.deb ...
Unpacking tesseract-ocr-eng (1:4.00~git30-7274cfa-1.1) ...
Selecting previously unselected package tesseract-ocr-osd.
Preparing to unpack .../tesseract-ocr-osd_1%3a4.00~git30-7274cfa-
1.1_all.deb ...
Unpacking tesseract-ocr-osd (1:4.00~git30-7274cfa-1.1) ...
Selecting previously unselected package tesseract-ocr.
Preparing to unpack .../tesseract-ocr_4.1.1-2.1build1_amd64.deb ...
Unpacking tesseract-ocr (4.1.1-2.1build1) ...
Setting up tesseract-ocr-eng (1:4.00~git30-7274cfa-1.1) ...
Setting up tesseract-ocr-osd (1:4.00~git30-7274cfa-1.1) ...
Setting up tesseract-ocr (4.1.1-2.1build1) ...
Processing triggers for man-db (2.10.2-1) ...
```

```
!pip install pytesseract
```

```
Requirement already satisfied: pytesseract in
/usr/local/lib/python3.10/dist-packages (0.3.10)
Requirement already satisfied: packaging>=21.3 in
/usr/local/lib/python3.10/dist-packages (from pytesseract) (23.2)
```

```
Requirement already satisfied: Pillow>=8.0.0 in
/usr/local/lib/python3.10/dist-packages (from pytesseract) (10.2.0)
```

```
# Get elements
raw_pdf_elements = partition_pdf(
    filename=os.path.join(input_path, "test.pdf"),
    extract_images_in_pdf=True,
    infer_table_structure=True,
    chunking_strategy="by_title",
    max_characters=4000,
    new_after_n_chars=3800,
    combine_text_under_n_chars=2000,
    image_output_dir_path=output_path,
)
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     /root/nltk_data...
[nltk_data]   Unzipping taggers/averaged_perceptron_tagger.zip.
WARNING:unstructured:This function will be deprecated in a future
release and `unstructured` will simply use the DEFAULT_MODEL from
`unstructured_inference.model.base` to set default model name
```

```
{"model_id":"7f738d3005bd46859ad4d4d09e781b8c","version_major":2,"vers
ion_minor":0}
```

```
{"model_id":"3b6d974f3b374325a65a6d48c8b22a6e","version_major":2,"vers
ion_minor":0}
```

```
{"model_id":"ab64e50af50a4c1d9ccb833b28d91318","version_major":2,"vers
ion_minor":0}
```

```
Some weights of the model checkpoint at microsoft/table-transformer-
structure-recognition were not used when initializing
TableTransformerForObjectDetection:
['model.backbone.conv_encoder.model.layer2.0.downsample.1.num_batches_
tracked',
'model.backbone.conv_encoder.model.layer3.0.downsample.1.num_batches_t
racked',
'model.backbone.conv_encoder.model.layer4.0.downsample.1.num_batches_t
racked']
- This IS expected if you are initializing
TableTransformerForObjectDetection from the checkpoint of a model
trained on another task or with another architecture (e.g.
initializing a BertForSequenceClassification model from a
BertForPreTraining model).
- This IS NOT expected if you are initializing
TableTransformerForObjectDetection from the checkpoint of a model that
you expect to be exactly identical (initializing a
```

```
BertForSequenceClassification model from a
BertForSequenceClassification model).

import base64

text_elements = []
table_elements = []
image_elements = []

# Function to encode images
def encode_image(image_path):
    with open(image_path, "rb") as image_file:
        return base64.b64encode(image_file.read()).decode('utf-8')

for element in raw_pdf_elements:
    if 'CompositeElement' in str(type(element)):
        text_elements.append(element)
    elif 'Table' in str(type(element)):
        table_elements.append(element)

table_elements = [i.text for i in table_elements]
text_elements = [i.text for i in text_elements]

# Tables
print(len(table_elements))

# Text
print(len(text_elements))

3
24

import os

# Specify the path to the "content" folder
content_path = "/content"  # Update this with the actual path if
different

# Construct the path to the "figures" folder inside the "content"
folder
figures_path = os.path.join(content_path, "figures")

# Print the list of JPEG files in the "figures" folder
print("JPEG files found in the 'figures' folder:")
for file in os.listdir(figures_path):
    if file.endswith('.jpg'):
        print(file)


JPEG files found in the 'figures' folder:
figure-18-7.jpg
```

```
figure-7-3.jpg
figure-4-2.jpg
figure-19-9.jpg
figure-8-4.jpg
figure-14-5.jpg
figure-18-8.jpg
figure-15-6.jpg
figure-3-1.jpg
figure-19-10.jpg
```

```
import os

# Create the output directory
output_path = '/content/output'
os.makedirs(output_path, exist_ok=True)

for image_file in os.listdir(figures_path):
    if image_file.endswith(('.png', '.jpg', '.jpeg')):
        image_path = os.path.join(figures_path, image_file)
        encoded_image = encode_image(image_path)
        image_elements.append(encoded_image)
print(len(image_elements))
```

```
10
```

```
# Print the list of image files
print("Image files found:")
for image_file in os.listdir(figures_path):
    if image_file.endswith(('.png', '.jpg', '.jpeg')):
        print(image_file)
```

```
Image files found:
figure-18-7.jpg
figure-7-3.jpg
figure-4-2.jpg
figure-19-9.jpg
figure-8-4.jpg
figure-14-5.jpg
figure-18-8.jpg
figure-15-6.jpg
figure-3-1.jpg
figure-19-10.jpg
```

```
!pip install -U langchain-openai
```

```
Collecting langchain-openai
  Downloading langchain_openai-0.0.8-py3-none-any.whl (32 kB)
Requirement already satisfied: langchain-core<0.2.0,>=0.1.27 in
/usr/local/lib/python3.10/dist-packages (from langchain-openai)
(0.1.30)
Requirement already satisfied: openai<2.0.0,>=1.10.0 in
```

```
/usr/local/lib/python3.10/dist-packages (from langchain-openai)
(1.13.3)
Requirement already satisfied: tiktoken<1,>=0.5.2 in
/usr/local/lib/python3.10/dist-packages (from langchain-openai)
(0.6.0)
Requirement already satisfied: PyYAML>=5.3 in
/usr/local/lib/python3.10/dist-packages (from langchain-
core<0.2.0,>=0.1.27->langchain-openai) (6.0.1)
Requirement already satisfied: anyio<5,>=3 in
/usr/local/lib/python3.10/dist-packages (from langchain-
core<0.2.0,>=0.1.27->langchain-openai) (3.7.1)
Requirement already satisfied: jsonpatch<2.0,>=1.33 in
/usr/local/lib/python3.10/dist-packages (from langchain-
core<0.2.0,>=0.1.27->langchain-openai) (1.33)
Requirement already satisfied: langsmith<0.2.0,>=0.1.0 in
/usr/local/lib/python3.10/dist-packages (from langchain-
core<0.2.0,>=0.1.27->langchain-openai) (0.1.23)
Requirement already satisfied: packaging<24.0,>=23.2 in
/usr/local/lib/python3.10/dist-packages (from langchain-
core<0.2.0,>=0.1.27->langchain-openai) (23.2)
Requirement already satisfied: pydantic<3,>=1 in
/usr/local/lib/python3.10/dist-packages (from langchain-
core<0.2.0,>=0.1.27->langchain-openai) (2.6.3)
Requirement already satisfied: requests<3,>=2 in
/usr/local/lib/python3.10/dist-packages (from langchain-
core<0.2.0,>=0.1.27->langchain-openai) (2.31.0)
Requirement already satisfied: tenacity<9.0.0,>=8.1.0 in
/usr/local/lib/python3.10/dist-packages (from langchain-
core<0.2.0,>=0.1.27->langchain-openai) (8.2.3)
Requirement already satisfied: distro<2,>=1.7.0 in
/usr/lib/python3/dist-packages (from openai<2.0.0,>=1.10.0->langchain-
openai) (1.7.0)
Requirement already satisfied: httpx<1,>=0.23.0 in
/usr/local/lib/python3.10/dist-packages (from openai<2.0.0,>=1.10.0-
>langchain-openai) (0.27.0)
Requirement already satisfied: sniffio in
/usr/local/lib/python3.10/dist-packages (from openai<2.0.0,>=1.10.0-
>langchain-openai) (1.3.1)
Requirement already satisfied: tqdm>4 in
/usr/local/lib/python3.10/dist-packages (from openai<2.0.0,>=1.10.0-
>langchain-openai) (4.66.2)
Requirement already satisfied: typing-extensions<5,>=4.7 in
/usr/local/lib/python3.10/dist-packages (from openai<2.0.0,>=1.10.0-
>langchain-openai) (4.9.0)
Requirement already satisfied: regex>=2022.1.18 in
/usr/local/lib/python3.10/dist-packages (from tiktoken<1,>=0.5.2-
>langchain-openai) (2023.12.25)
Requirement already satisfied: idna>=2.8 in
/usr/local/lib/python3.10/dist-packages (from anyio<5,>=3->langchain-
```

```
core<0.2.0,>=0.1.27->langchain-openai) (3.6)
Requirement already satisfied: exceptiongroup in
/usr/local/lib/python3.10/dist-packages (from anyio<5,>=3->langchain-
core<0.2.0,>=0.1.27->langchain-openai) (1.2.0)
Requirement already satisfied: certifi in
/usr/local/lib/python3.10/dist-packages (from httpx<1,>=0.23.0-
>openai<2.0.0,>=1.10.0->langchain-openai) (2024.2.2)
Requirement already satisfied: httpcore==1.* in
/usr/local/lib/python3.10/dist-packages (from httpx<1,>=0.23.0-
>openai<2.0.0,>=1.10.0->langchain-openai) (1.0.4)
Requirement already satisfied: h11<0.15,>=0.13 in
/usr/local/lib/python3.10/dist-packages (from httpcore==1.*-
>httpx<1,>=0.23.0->openai<2.0.0,>=1.10.0->langchain-openai) (0.14.0)
Requirement already satisfied: jsonpointer>=1.9 in
/usr/local/lib/python3.10/dist-packages (from jsonpatch<2.0,>=1.33-
>langchain-core<0.2.0,>=0.1.27->langchain-openai) (2.4)
Requirement already satisfied: orjson<4.0.0,>=3.9.14 in
/usr/local/lib/python3.10/dist-packages (from langsmith<0.2.0,>=0.1.0-
>langchain-core<0.2.0,>=0.1.27->langchain-openai) (3.9.15)
Requirement already satisfied: annotated-types>=0.4.0 in
/usr/local/lib/python3.10/dist-packages (from pydantic<3,>=1-
>langchain-core<0.2.0,>=0.1.27->langchain-openai) (0.6.0)
Requirement already satisfied: pydantic-core==2.16.3 in
/usr/local/lib/python3.10/dist-packages (from pydantic<3,>=1-
>langchain-core<0.2.0,>=0.1.27->langchain-openai) (2.16.3)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.10/dist-packages (from requests<3,>=2-
>langchain-core<0.2.0,>=0.1.27->langchain-openai) (3.3.2)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.10/dist-packages (from requests<3,>=2-
>langchain-core<0.2.0,>=0.1.27->langchain-openai) (1.26.18)
Installing collected packages: langchain-openai
Successfully installed langchain-openai-0.0.8
```

```
import os

# Set the OpenAI API key as an environment variable
os.environ["OPENAI_API_KEY"] = "sk-
G9wwOLFL4bHZ2j9zvaAZT3BlbkFJoWtuQZITynnykw6u1a8Z"

from langchain.chat_models import ChatOpenAI
from langchain.schema.messages import HumanMessage, AIMessage

chain_gpt_35 = ChatOpenAI(model="gpt-3.5-turbo", max_tokens=1024)
chain_gpt_4_vision = ChatOpenAI(model="gpt-4-vision-preview",
max_tokens=1024)

# Function for text summaries
def summarize_text(text_element):
    prompt = f"Summarize the following text:\n\n{text_element}\n\
```

```
nSummary:"
    response = chain_gpt_35.invoke([HumanMessage(content=prompt)])
    return response.content

# Function for table summaries
def summarize_table(table_element):
    prompt = f"Summarize the following table:\n\n{table_element}\n\
nSummary:"
    response = chain_gpt_35.invoke([HumanMessage(content=prompt)])
    return response.content

# Function for image summaries
def summarize_image(encoded_image):
    prompt = [
        AIMessage(content="You are a bot that is good at analyzing
images."),
        HumanMessage(content=[
            {"type": "text", "text": "Describe the contents of this
image."},
            {
                "type": "image_url",
                "image_url": {
                    "url": f"{encoded_image}"
                },
            },
        ])
    ]
    response = chain_gpt_4_vision.invoke(prompt)
    return response.content
```

```
/usr/local/lib/python3.10/dist-packages/langchain_core/_api/
deprecation.py:117: LangChainDeprecationWarning: The class
`langchain_community.chat_models.openai.ChatOpenAI` was deprecated in
langchain-community 0.0.10 and will be removed in 0.2.0. An updated
version of the class exists in the langchain-openai package and should
be used instead. To use it run `pip install -U langchain-openai` and
import as `from langchain_openai import ChatOpenAI`.
  warn_deprecated(
```

```
# Processing table elements with feedback and sleep
table_summaries = []
for i, te in enumerate(table_elements[0:2]):
    summary = summarize_table(te)
    table_summaries.append(summary)
    print(f"{i + 1}th element of tables processed.")
```

```
1th element of tables processed.
2th element of tables processed.
```

```
# Processing text elements with feedback and sleep
text_summaries = []
for i, te in enumerate(text_elements[0:2]):
    summary = summarize_text(te)
    text_summaries.append(summary)
    print(f"{i + 1}th element of texts processed.")
```

```
1th element of texts processed.
2th element of texts processed.
```

```
# Processing image elements with feedback and sleep
image_summaries = []
for i, ie in enumerate(image_elements[0:2]):
    summary = summarize_image(ie)
    image_summaries.append(summary)
    print(f"{i + 1}th element of images processed.")
```

```
1th element of images processed.
2th element of images processed.
```

```
import uuid

from langchain.embeddings import OpenAIEmbeddings
from langchain.retrievers.multi_vector import MultiVectorRetriever
from langchain.schema.document import Document
from langchain.storage import InMemoryStore
from langchain.vectorstores import Chroma

# Initialize the vector store and storage layer
vectorstore = Chroma(collection_name="summaries",
embedding_function=OpenAIEmbeddings())
store = InMemoryStore()
id_key = "doc_id"

# Initialize the retriever
retriever = MultiVectorRetriever(vectorstore=vectorstore,
docstore=store, id_key=id_key)

# Function to add documents to the retriever
def add_documents_to_retriever(summaries, original_contents):
    doc_ids = [str(uuid.uuid4()) for _ in summaries]
    summary_docs = [
        Document(page_content=s, metadata={id_key: doc_ids[i]})
        for i, s in enumerate(summaries)
    ]
    retriever.vectorstore.add_documents(summary_docs)
    retriever.docstore.mset(list(zip(doc_ids, original_contents)))
```

```
/usr/local/lib/python3.10/dist-packages/langchain_core/_api/
deprecation.py:117: LangChainDeprecationWarning: The class
```

```
`langchain_community.embeddings.openai.OpenAIEmbeddings` was
deprecated in langchain-community 0.0.9 and will be removed in 0.2.0.
An updated version of the class exists in the langchain-openai package
and should be used instead. To use it run `pip install -U langchain-
openai` and import as `from langchain_openai import OpenAIEmbeddings`.
  warn_deprecated(
```

```
# Add text summaries
add_documents_to_retriever(text_summaries, text_elements)

# Add table summaries
add_documents_to_retriever(table_summaries, table_elements)

# Add image summaries
add_documents_to_retriever(image_summaries, image_summaries) #
hopefully real images soon

# We can retrieve this table
retriever.get_relevant_documents(
    "What do you see on the images in the database?"
)
```

```
['Visual features Before Last Best variant Predict answer first
Training from scratch 7B model size 90.92 - 85.81 (-5.11) 89.84 (-
1.08) 89.96 (-0.96) 89.77 (-1.15) - -',
 'The image depicts a screenshot of a conversation in a messaging
interface. The conversation appears to be between a user and an AI
assistant. The user is asking for meal recommendations based on the
contents of their refrigerator, and the AI assistant responds with a
suggestion to make a fruit salad including a detailed recipe. The
recipe provided by the AI includes ingredients like strawberries,
blueberries, carrots, lemon juice, parsley or mint, and optional honey
or maple syrup. There are instructions for preparing and serving the
fruit salad as well. The screenshot includes a view of the inside of a
refrigerator stocked with various food items such as milk, eggs, and a
selection of fruits and vegetables, which seem to correspond to the
ingredients mentioned in the recipe. The conversation is focused on
cooking and recipes, and there is no display of personal or sensitive
information.',
 "The image shows a person ironing clothes on an ironing board that is
set up on the back of a yellow taxi cab. The individual is wearing a
yellow shirt, which matches the color of the taxi, and appears to be
focused on the task of ironing. There is another yellow taxi in
motion, blurred due to its speed, in the background, indicating that
this scene is likely taking place in a busy urban area, possibly a
city known for its yellow cabs like New York City. The setting looks
like a street lined with buildings and there are pink decorations
visible, suggesting some kind of festive or seasonal event. The scene
is unusual and somewhat humorous because it's not common to see
someone ironing clothes on the street, let alone on the back of a
```

```
taxi.",
 'Conversation Detail description Complex reasoning All Full data
Detail + Complex Conv + 5% Detail + 10% Complex Conversation No
Instruction Tuning 83.1 81.5 (-1.6) 81.0 (-2.1) 76.5 (-6.6) 22.0 ( -
61.1 ) 75.3 73.3 (-2.0) 68.4 (-7.1) 59.8 (-16.2 ) 24.0 ( -51.3 ) 96.5
90.8 (-5.7) 91.5 (-5.0) 84.9 (-12.4 ) 18.5 ( -78.0 ) 85.1 81.9 (-3.2)
80.5 (-4.4) 73.8 (-11.3 ) 21.5 ( -63.6 )']
```

```
from langchain.schema.runnable import RunnablePassthrough
from langchain.prompts import ChatPromptTemplate
from langchain.schema.output_parser import StrOutputParser

template = """Answer the question based only on the following context,
which can include text, images and tables:
{context}
Question: {question}
"""
prompt = ChatPromptTemplate.from_template(template)

model = ChatOpenAI(temperature=0, model="gpt-3.5-turbo")

chain = (
    {"context": retriever, "question": RunnablePassthrough()}
    | prompt
    | model
    | StrOutputParser()
)

chain.invoke(
     "What do you see on the images in the database?"
)
```

```
{"type":"string"}
```

```
!pip install gradio
```

```
Collecting gradio
  Downloading gradio-4.21.0-py3-none-any.whl (17.0 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 17.0/17.0 MB 57.1 MB/s eta
0:00:00
 gradio)
  Downloading aiofiles-23.2.1-py3-none-any.whl (15 kB)
Requirement already satisfied: altair<6.0,>=4.2.0 in
/usr/local/lib/python3.10/dist-packages (from gradio) (4.2.2)
Requirement already satisfied: fastapi in
/usr/local/lib/python3.10/dist-packages (from gradio) (0.110.0)
Collecting ffmpy (from gradio)
  Downloading ffmpy-0.3.2.tar.gz (5.5 kB)
  Preparing metadata (setup.py) ...  gradio)
  Downloading gradio_client-0.12.0-py3-none-any.whl (310 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 310.7/310.7 kB 35.2 MB/s eta
```

```
0:00:00
ent already satisfied: httpx>=0.24.1 in
/usr/local/lib/python3.10/dist-packages (from gradio) (0.27.0)
Requirement already satisfied: huggingface-hub>=0.19.3 in
/usr/local/lib/python3.10/dist-packages (from gradio) (0.20.3)
Requirement already satisfied: importlib-resources<7.0,>=1.3 in
/usr/local/lib/python3.10/dist-packages (from gradio) (6.1.1)
Requirement already satisfied: jinja2<4.0 in
/usr/local/lib/python3.10/dist-packages (from gradio) (3.1.3)
Requirement already satisfied: markupsafe~=2.0 in
/usr/local/lib/python3.10/dist-packages (from gradio) (2.1.5)
Requirement already satisfied: matplotlib~=3.0 in
/usr/local/lib/python3.10/dist-packages (from gradio) (3.7.2)
Requirement already satisfied: numpy~=1.0 in
/usr/local/lib/python3.10/dist-packages (from gradio) (1.26.4)
Requirement already satisfied: orjson~=3.0 in
/usr/local/lib/python3.10/dist-packages (from gradio) (3.9.15)
Requirement already satisfied: packaging in
/usr/local/lib/python3.10/dist-packages (from gradio) (23.2)
Requirement already satisfied: pandas<3.0,>=1.0 in
/usr/local/lib/python3.10/dist-packages (from gradio) (2.2.0)
Requirement already satisfied: pillow<11.0,>=8.0 in
/usr/local/lib/python3.10/dist-packages (from gradio) (10.2.0)
Requirement already satisfied: pydantic>=2.0 in
/usr/local/lib/python3.10/dist-packages (from gradio) (2.6.3)
Collecting pydub (from gradio)
  Downloading pydub-0.25.1-py2.py3-none-any.whl (32 kB)
Requirement already satisfied: python-multipart>=0.0.9 in
/usr/local/lib/python3.10/dist-packages (from gradio) (0.0.9)
Requirement already satisfied: pyyaml<7.0,>=5.0 in
/usr/local/lib/python3.10/dist-packages (from gradio) (6.0.1)
Collecting ruff>=0.2.2 (from gradio)
  Downloading ruff-0.3.2-py3-none-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl (7.9 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 7.9/7.9 MB 97.9 MB/s eta
0:00:00
antic-version~=2.0 (from gradio)
  Downloading semantic_version-2.10.0-py2.py3-none-any.whl (15 kB)
Collecting tomlkit==0.12.0 (from gradio)
  Downloading tomlkit-0.12.0-py3-none-any.whl (37 kB)
Requirement already satisfied: typer[all]<1.0,>=0.9 in
/usr/local/lib/python3.10/dist-packages (from gradio) (0.9.0)
Requirement already satisfied: typing-extensions~=4.0 in
/usr/local/lib/python3.10/dist-packages (from gradio) (4.9.0)
Requirement already satisfied: uvicorn>=0.14.0 in
/usr/local/lib/python3.10/dist-packages (from gradio) (0.28.0)
Requirement already satisfied: fsspec in
/usr/local/lib/python3.10/dist-packages (from gradio-client==0.12.0-
>gradio) (2024.2.0)
```

```
Collecting websockets<12.0,>=10.0 (from gradio-client==0.12.0->gradio)
  Downloading websockets-11.0.3-cp310-cp310-
manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_17_x86_64.manylinux
2014_x86_64.whl (129 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 129.9/129.9 kB 19.1 MB/s eta
0:00:00
ent already satisfied: entrypoints in /usr/local/lib/python3.10/dist-
packages (from altair<6.0,>=4.2.0->gradio) (0.4)
Requirement already satisfied: jsonschema>=3.0 in
/usr/local/lib/python3.10/dist-packages (from altair<6.0,>=4.2.0-
>gradio) (4.19.2)
Requirement already satisfied: toolz in
/usr/local/lib/python3.10/dist-packages (from altair<6.0,>=4.2.0-
>gradio) (0.12.1)
Requirement already satisfied: anyio in
/usr/local/lib/python3.10/dist-packages (from httpx>=0.24.1->gradio)
(3.7.1)
Requirement already satisfied: certifi in
/usr/local/lib/python3.10/dist-packages (from httpx>=0.24.1->gradio)
(2024.2.2)
Requirement already satisfied: httpcore==1.* in
/usr/local/lib/python3.10/dist-packages (from httpx>=0.24.1->gradio)
(1.0.4)
Requirement already satisfied: idna in /usr/local/lib/python3.10/dist-
packages (from httpx>=0.24.1->gradio) (3.6)
Requirement already satisfied: sniffio in
/usr/local/lib/python3.10/dist-packages (from httpx>=0.24.1->gradio)
(1.3.1)
Requirement already satisfied: h11<0.15,>=0.13 in
/usr/local/lib/python3.10/dist-packages (from httpcore==1.*-
>httpx>=0.24.1->gradio) (0.14.0)
Requirement already satisfied: filelock in
/usr/local/lib/python3.10/dist-packages (from huggingface-hub>=0.19.3-
>gradio) (3.13.1)
Requirement already satisfied: requests in
/usr/local/lib/python3.10/dist-packages (from huggingface-hub>=0.19.3-
>gradio) (2.31.0)
Requirement already satisfied: tqdm>=4.42.1 in
/usr/local/lib/python3.10/dist-packages (from huggingface-hub>=0.19.3-
>gradio) (4.66.2)
Requirement already satisfied: contourpy>=1.0.1 in
/usr/local/lib/python3.10/dist-packages (from matplotlib~=3.0->gradio)
(1.2.0)
Requirement already satisfied: cycler>=0.10 in
/usr/local/lib/python3.10/dist-packages (from matplotlib~=3.0->gradio)
(0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in
/usr/local/lib/python3.10/dist-packages (from matplotlib~=3.0->gradio)
(4.49.0)
```

```
Requirement already satisfied: kiwisolver>=1.0.1 in
/usr/local/lib/python3.10/dist-packages (from matplotlib~=3.0->gradio)
(1.4.5)
Requirement already satisfied: pyparsing<3.1,>=2.3.1 in
/usr/local/lib/python3.10/dist-packages (from matplotlib~=3.0->gradio)
(3.0.9)
Requirement already satisfied: python-dateutil>=2.7 in
/usr/local/lib/python3.10/dist-packages (from matplotlib~=3.0->gradio)
(2.8.2)
Requirement already satisfied: pytz>=2020.1 in
/usr/local/lib/python3.10/dist-packages (from pandas<3.0,>=1.0-
>gradio) (2024.1)
Requirement already satisfied: tzdata>=2022.7 in
/usr/local/lib/python3.10/dist-packages (from pandas<3.0,>=1.0-
>gradio) (2024.1)
Requirement already satisfied: annotated-types>=0.4.0 in
/usr/local/lib/python3.10/dist-packages (from pydantic>=2.0->gradio)
(0.6.0)
Requirement already satisfied: pydantic-core==2.16.3 in
/usr/local/lib/python3.10/dist-packages (from pydantic>=2.0->gradio)
(2.16.3)
Requirement already satisfied: click<9.0.0,>=7.1.1 in
/usr/local/lib/python3.10/dist-packages (from typer[all]<1.0,>=0.9-
>gradio) (8.1.7)
Collecting colorama<0.5.0,>=0.4.3 (from typer[all]<1.0,>=0.9->gradio)
  Downloading colorama-0.4.6-py2.py3-none-any.whl (25 kB)
Collecting shellingham<2.0.0,>=1.3.0 (from typer[all]<1.0,>=0.9-
>gradio)
  Downloading shellingham-1.5.4-py2.py3-none-any.whl (9.8 kB)
Requirement already satisfied: rich<14.0.0,>=10.11.0 in
/usr/local/lib/python3.10/dist-packages (from typer[all]<1.0,>=0.9-
>gradio) (13.7.1)
Requirement already satisfied: starlette<0.37.0,>=0.36.3 in
/usr/local/lib/python3.10/dist-packages (from fastapi->gradio)
(0.36.3)
Requirement already satisfied: attrs>=22.2.0 in
/usr/local/lib/python3.10/dist-packages (from jsonschema>=3.0-
>altair<6.0,>=4.2.0->gradio) (23.2.0)
Requirement already satisfied: jsonschema-specifications>=2023.03.6 in
/usr/local/lib/python3.10/dist-packages (from jsonschema>=3.0-
>altair<6.0,>=4.2.0->gradio) (2023.12.1)
Requirement already satisfied: referencing>=0.28.4 in
/usr/local/lib/python3.10/dist-packages (from jsonschema>=3.0-
>altair<6.0,>=4.2.0->gradio) (0.33.0)
Requirement already satisfied: rpds-py>=0.7.1 in
/usr/local/lib/python3.10/dist-packages (from jsonschema>=3.0-
>altair<6.0,>=4.2.0->gradio) (0.18.0)
Requirement already satisfied: six>=1.5 in
/usr/local/lib/python3.10/dist-packages (from python-dateutil>=2.7-
```

```
>matplotlib~=3.0->gradio) (1.16.0)
Requirement already satisfied: markdown-it-py>=2.2.0 in
/usr/local/lib/python3.10/dist-packages (from rich<14.0.0,>=10.11.0-
>typer[all]<1.0,>=0.9->gradio) (3.0.0)
Requirement already satisfied: pygments<3.0.0,>=2.13.0 in
/usr/local/lib/python3.10/dist-packages (from rich<14.0.0,>=10.11.0-
>typer[all]<1.0,>=0.9->gradio) (2.16.1)
Requirement already satisfied: exceptiongroup in
/usr/local/lib/python3.10/dist-packages (from anyio->httpx>=0.24.1-
>gradio) (1.2.0)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.10/dist-packages (from requests->huggingface-
hub>=0.19.3->gradio) (3.3.2)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.10/dist-packages (from requests->huggingface-
hub>=0.19.3->gradio) (1.26.18)
Requirement already satisfied: mdurl~=0.1 in
/usr/local/lib/python3.10/dist-packages (from markdown-it-py>=2.2.0-
>rich<14.0.0,>=10.11.0->typer[all]<1.0,>=0.9->gradio) (0.1.2)
Building wheels for collected packages: ffmpy
  Building wheel for ffmpy (setup.py) ... py: filename=ffmpy-0.3.2-
py3-none-any.whl size=5584
sha256=b0d322fff7ebc0703f80f01d8c86852c7981962bc91f277d003f994094284d3
c
  Stored in directory:
/root/.cache/pip/wheels/bd/65/9a/671fc6dcde07d4418df0c592f8df512b26d7a
0029c2a23dd81
Successfully built ffmpy
Installing collected packages: pydub, ffmpy, websockets, tomlkit,
shellingham, semantic-version, ruff, colorama, aiofiles, gradio-
client, gradio
  Attempting uninstall: websockets
    Found existing installation: websockets 12.0
    Uninstalling websockets-12.0:
      Successfully uninstalled websockets-12.0
Successfully installed aiofiles-23.2.1 colorama-0.4.6 ffmpy-0.3.2
gradio-4.21.0 gradio-client-0.12.0 pydub-0.25.1 ruff-0.3.2 semantic-
version-2.10.0 shellingham-1.5.4 tomlkit-0.12.0 websockets-11.0.3
```

```python
import gradio as gr
from unstructured.partition.pdf import partition_pdf
import base64

# Function to encode images
def encode_image(image_path):
    with open(image_path, "rb") as image_file:
        return base64.b64encode(image_file.read()).decode('utf-8')

# Function for summarizing images
def summarize_image(encoded_image):
```

```
    prompt = [
        AIMessage(content="You are a bot that is good at analyzing
images."),
        HumanMessage(content=[
            {"type": "text", "text": "Describe the contents of this
image."},
            {
                "type": "image_url",
                "image_url": {
                    "url": f"{encoded_image}"
                },
            },
        ])
    ]
    response = chain_gpt_4_vision.invoke(prompt)
    return response.content

# Define your Gradio interface
def pdf_chat_interface(pdf_file):
    # Process the PDF file
    raw_pdf_elements = partition_pdf(
        filename=pdf_file.name,
        extract_images_in_pdf=True,
        infer_table_structure=True,
        chunking_strategy="by_title",
        max_characters=4000,
        new_after_n_chars=3800,
        combine_text_under_n_chars=2000,
        image_output_dir_path=output_path,
    )

    text_elements = []
    table_elements = []
    image_elements = []

    for element in raw_pdf_elements:
        if 'CompositeElement' in str(type(element)):
            text_elements.append(element)
        elif 'Table' in str(type(element)):
            table_elements.append(element)

    table_elements = [i.text for i in table_elements]
    text_elements = [i.text for i in text_elements]

    # Encode and summarize images
    for image_file in os.listdir(figures_path):
      if image_file.endswith(('.png', '.jpg', '.jpeg')):
          image_path = os.path.join(figures_path, image_file)
          encoded_image = encode_image(image_path)
          image_elements.append(encoded_image)
```

```
    # Process text and table elements
    text_summaries = [summarize_text(te) for te in text_elements]
    table_summaries = [summarize_table(te) for te in table_elements]

    # Processing image elements with feedback and sleep
    image_summaries = []
    for ie in image_elements:
        summary = summarize_image(ie)
        image_summaries.append(summary)

    response = {
        "text_summaries": text_summaries,
        "table_summaries": table_summaries,
        "image_summaries": image_elements
    }

    return response

# Deploy your Gradio app
gr.Interface(
    fn=pdf_chat_interface,
    inputs="file",
    outputs="json",
    title="PDF Chat Summarizer",
    description="Upload a PDF file and get separate summaries for
text, tables, and image descriptions.",
).launch()
```

```
Setting queue=True in a Colab notebook requires sharing enabled.
Setting `share=True` (you can turn this off by setting `share=False`
in `launch()` explicitly).

Colab notebook detected. To show errors in colab notebook, set
debug=True in launch()
Running on public URL: https://8191cdc2415a509bf7.gradio.live

This share link expires in 72 hours. For free permanent hosting and
GPU upgrades, run `gradio deploy` from Terminal to deploy to Spaces
(https://huggingface.co/spaces)

<IPython.core.display.HTML object>
```

```
import gradio as gr
from unstructured.partition.pdf import partition_pdf
import base64

# Function to encode images
def encode_image(image_path):
```

```
    with open(image_path, "rb") as image_file:
        return base64.b64encode(image_file.read()).decode('utf-8')

# Function for summarizing images
def summarize_image(encoded_image):
    prompt = [
        AIMessage(content="You are a bot that is good at analyzing
images."),
        HumanMessage(content=[
            {"type": "text", "text": "Describe the contents of this
image."},
            {
                "type": "image_url",
                "image_url": {
                    "url": f"{encoded_image}"
                },
            },
        ])
    ]
    response = chain_gpt_4_vision.invoke(prompt)
    return response.content

# Define your Gradio interface
def pdf_chat_interface(pdf_file):
    # Process the PDF file
    raw_pdf_elements = partition_pdf(
        filename=pdf_file.name,
        extract_images_in_pdf=True,
        infer_table_structure=True,
        chunking_strategy="by_title",
        max_characters=4000,
        new_after_n_chars=3800,
        combine_text_under_n_chars=2000,
        image_output_dir_path=output_path,
    )

    text_elements = []
    table_elements = []
    image_elements = []

    for element in raw_pdf_elements:
        if 'CompositeElement' in str(type(element)):
            text_elements.append(element)
        elif 'Table' in str(type(element)):
            table_elements.append(element)

    table_elements = [i.text for i in table_elements]
    text_elements = [i.text for i in text_elements]

    # Encode and summarize images
```

```
    for image_file in os.listdir(output_path):
        if image_file.endswith(('.png', '.jpg', '.jpeg')):
            image_path = os.path.join(output_path, image_file)
            encoded_image = encode_image(image_path)
            summary = summarize_image(encoded_image)
            image_elements.append(summary)

    # Process text and table elements
    text_summaries = [summarize_text(te) for te in text_elements]
    table_summaries = [summarize_table(te) for te in table_elements]

    # Construct final response
    response = "Text Summaries:\n" + "\n".join(text_summaries) + "\n\
nTable Summaries:\n" + "\n".join(table_summaries) + "\n\nImage
Summaries:\n" + "\n".join(image_elements)

    return response

# Deploy your Gradio app
gr.Interface(
    fn=pdf_chat_interface,
    inputs="file",
    outputs="text",
    title="PDF Chat Summarizer",
    description="Upload a PDF file and get summarized text, tables,
and image descriptions.",
).launch()
```

```
Setting queue=True in a Colab notebook requires sharing enabled.
Setting `share=True` (you can turn this off by setting `share=False`
in `launch()` explicitly).

Colab notebook detected. To show errors in colab notebook, set
debug=True in launch()
Running on public URL: https://8a38aae24820fdaac5.gradio.live

This share link expires in 72 hours. For free permanent hosting and
GPU upgrades, run `gradio deploy` from Terminal to deploy to Spaces
(https://huggingface.co/spaces)

<IPython.core.display.HTML object>
```