Analytics Vidhya

Document Loaders

Instructor

Dipanjan Sarkar

Head of Community & Principal AI Scientist at Analytics Vidhya

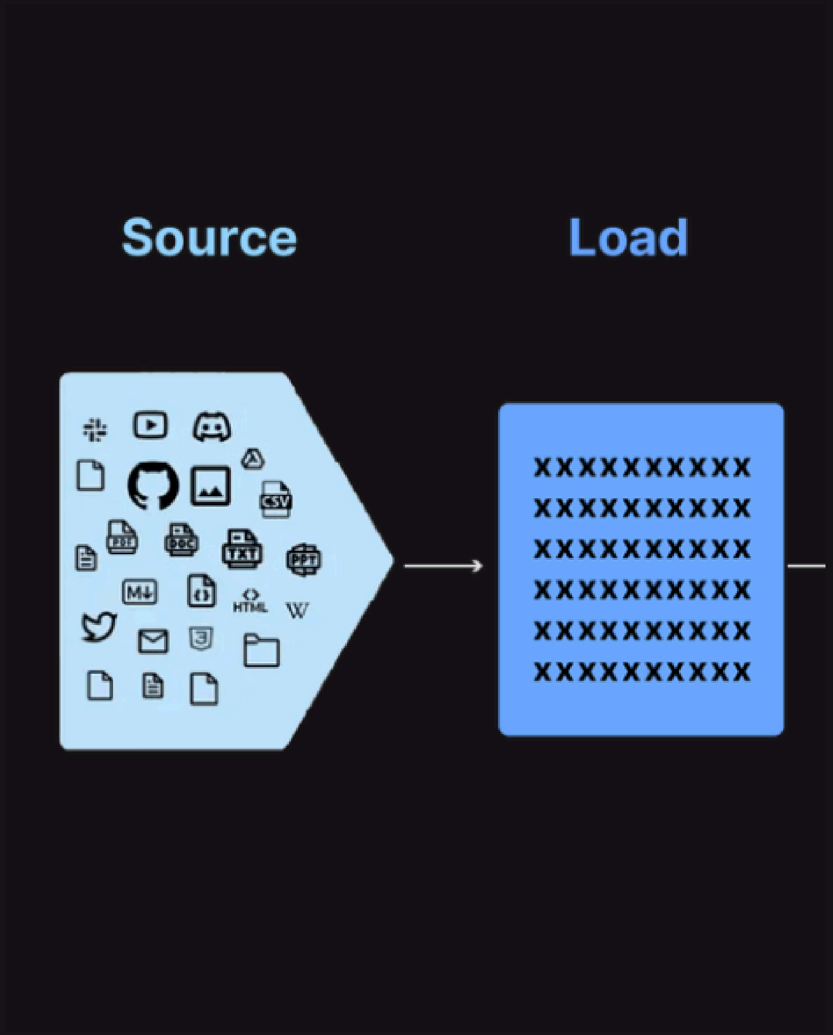
Google Developer Expert - ML & Cloud Champion Innovator

Published Author

Outline

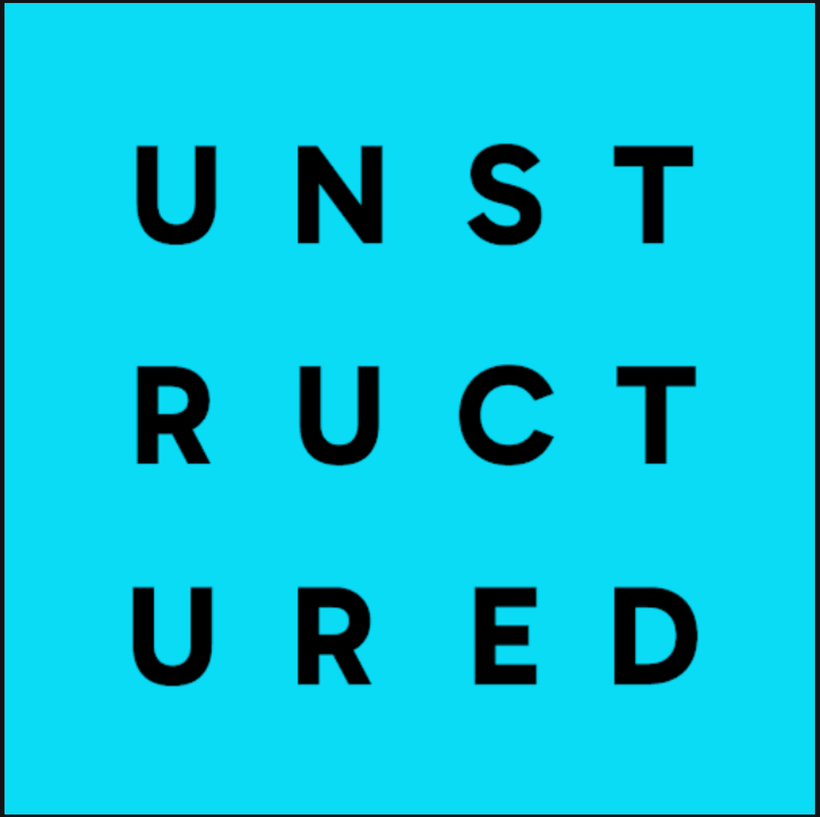
- Document Loaders
- Document Loaders from Unstructured.io
- Popular Document Loaders in LangChain

Document Loaders



- LangChain provides document loaders to load data from a source as a Document data type object
- A Document has a piece of text content and associated metadata
- LangChain has loaders for almost every possible document format
- Document loaders provide a load method for loading data as documents from a configured source

Document Loaders from Unstructured.io



- The unstructured library provides open-source components for ingesting and pre-processing documents, such as:
  - PDFs
  - HTML
  - Microsoft Word and many more
- LangChain has bindings to the Unstructured library to access and use various data loaders

Popular Document Loaders in LangChain

Loader Type	Description
CSV	LangChain implements a CSV Loader that will load CSV files into a sequence of Document objects
Markdown	LangChain implements an UnstructuredMarkdownLoader object which uses the Unstructured library to load data from Markdown files
Text	LangChain has a simple text loader to load data from text or markdown files
JSON	LangChain implements a JSONLoader to convert JSON and JSONL data into LangChain Document objects. It uses the jq package to enable extraction of specific fields from the data
PDF	LangChain integrates with a host of PDF parsers including PyPDF, PyMuPDF, PDFMiner and Unstructured PDF Loader which supports OCR, image and table extraction also.

Thank You