LangChain Chat with Your Data: Document Loading

This script demonstrates how to load and process documents from different sources, such as PDFs, YouTube transcripts, web URLs, and Notion pages, to make them accessible for a large language model (LLM). These documents are chunked and stored, enabling efficient retrieval of specific content for question-answering tasks.

1. **Setup**: Required environment variables

```
# Define your OpenAI API key
OPENAI_API_KEY = "Replace with your actual API key"
USER_AGENT = "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.124 Safari/537.36"
```

```
# Load environment variables for OpenAI API and User-Agent (if required by web requests).
load_dotenv(find_dotenv())
openai.api_key = os.getenv("OPENAI_API_KEY")
user_agent = os.getenv("USER_AGENT")
```

2. **Setup**: Required Libraries

```
import os
import openai
from dotenv import load_dotenv, find_dotenv
from langchain_community.document_loaders import PyPDFLoader, WebBaseLoader, NotionDirectoryLoader # type: ignore
from langchain_community.document_loaders.generic import GenericLoader # type: ignore
from langchain_community.document_loaders.parsers.audio import OpenAIWhisperParser # type: ignore
from langchain_community.document_loaders import YoutubeAudioLoader # type: ignore
```

3. **Loading a PDF Document:**

```
# ------------------------
# Document Sources Loading
# ------------------------

# ### Load PDFs
# Use the `PyPDFLoader` to load PDFs, breaking down the content into separate pages to manage document size and enable efficient retrieval.
pdf_loader = PyPDFLoader("docs/cs229_lectures/sfbu-2024-2025-university-catalog-8-20-2024.pdf")
pages = pdf_loader.load()

# Display page count and a preview of the first page's content.
print(f"PDF loaded with {len(pages)} pages.")
print("Preview of first page content:")
print(pages[1].page_content[:500])  # Display first 500 characters of the first page's content
```

Output:

```
(venv) vaishnavi@DESKTOP-9V8KJG2:/mnt/c/Users/Mohit/Desktop/Gen AI/Week 7$ python3 Document_Loading.py
USER_AGENT environment variable not set, consider setting it to identify your requests.
USER_AGENT loaded successfully.
PDF loaded with 186 pages.
Preview of first page content:


2024 – 2025 University Catalog  1
San Francisco Bay University
2024-2025 University Catalog
      Effective Fall Semester 2024


The 2024-2025 University Catalog is published annually and designed to provide an overview of
general information about San Francisco Bay University and a detailed explanation of the
University’s degree programs, curricular requirements, and Academic Affairs rules and regulations.
Additional information about student life organizations, social and personal
```

4. Loading a YouTube Transcript:

Document_Loading.py > ...

```python
# ### Load YouTube Transcript
# Load audio from a YouTube video, transcribing it with the Whisper API. Make sure `yt-dlp` and `pydub` are installed.
# - This method saves the audio, parses it, and stores the transcription.
# - Install `yt-dlp` and `pydub` locally with: `!pip install yt_dlp pydub`

youtube_url = "https://www.youtube.com/watch?v=kuZNIvdwnMc"
save_dir = "docs/youtube/"
youtube_loader = GenericLoader(
    YoutubeAudioLoader([youtube_url], save_dir),
    OpenAIWhisperParser()
)
docs = youtube_loader.load()

# Display a preview of the YouTube transcript.
print("YouTube Transcript Preview:")
print(docs[0].page_content[:500])  # Display the first 500 characters
```

Output:

```
[youtube] Extracting URL: https://www.youtube.com/watch?v=kuZNIvdwnMc
[youtube] kuZNIvdwnMc: Downloading webpage
[youtube] kuZNIvdwnMc: Downloading ios player API JSON
[youtube] kuZNIvdwnMc: Downloading mweb player API JSON
[youtube] kuZNIvdwnMc: Downloading m3u8 information
[info] kuZNIvdwnMc: Downloading 1 format(s): 140
[download] docs/youtube//San Francisco Bay University MBA Student Spotlight:  John Odebode.m4a has already been downloaded
[download] 100% of   10.19MiB
[ExtractAudio] Not converting audio docs/youtube//San Francisco Bay University MBA Student Spotlight:  John Odebode.m4a; file
is already in target format m4a
Transcribing part 1!
YouTube Transcript Preview:
My name is John, John Odebode. I am studying for an MBA program here at SFBU. It's my final trimester at SFBU and I will be gr
aduating in two weeks. I am from Nigeria. I studied at the University of Lagos for my first degree in philosophy. I also studi
ed for my first master's degree in philosophy as well at the same university. I have been practicing within the supply chain i
ndustry for the past six years. I have spent the most part of my career at ExxonMobil and I recently completed a six-month
```

5. **Loading Webpage Content:**

```python
# ### Load Webpage Content
# Use `WebBaseLoader` to load content from a specified URL. This can be useful for capturing specific content from websites.
# Pass `headers` with USER_AGENT if needed.

# Load content from a specific URL
web_loader = WebBaseLoader("https://www.sfbu.edu/student-health-insurance")
web_docs = web_loader.load()

# Process the webpage content to remove blank lines
web_content = web_docs[0].page_content  # Access the content
cleaned_content = "\n".join([line.strip() for line in web_content.splitlines() if line.strip()])

# Display the cleaned content (first 500 characters as an example)
print("\nWebpage Content Preview:")
print(cleaned_content[:500])

```

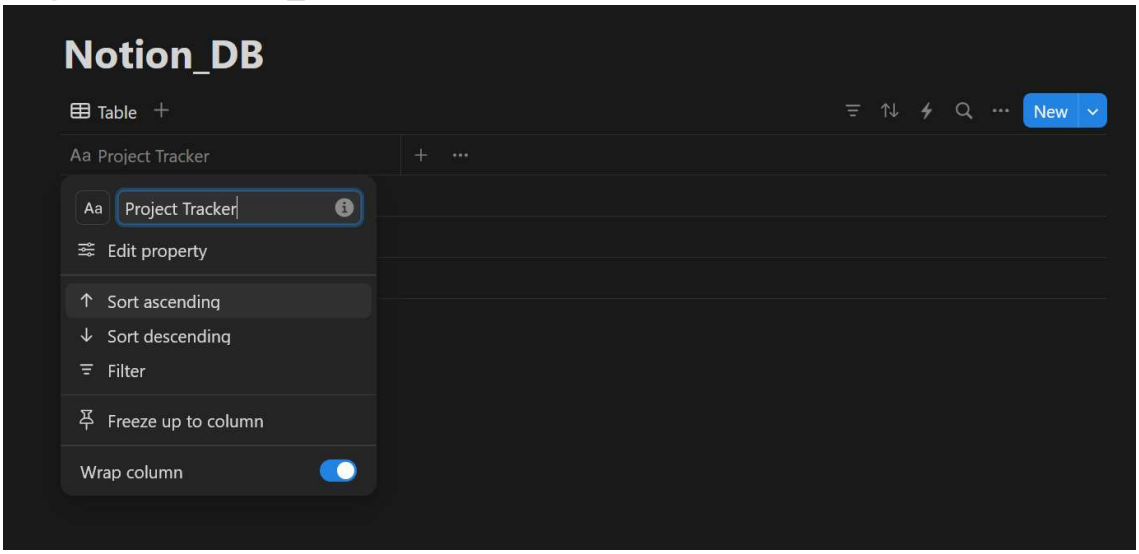
Output:

```
Webpage Content Preview:
Student Health Insurance | San Francisco Bay University
Skip to main content
San Francisco Bay University
Header Action Navigation
Visit
Apply
Online store
Search
Header Action Navigation
Visit
Apply
Online store
Mega Menu
Why We're Here
Our CampusStrategic Plan
Our Leadership
Our Glossary of Terms
Learning & Teaching
Undergraduate ProgramsGraduate ProgramsFaculty
Academic CalendarThe Center for Empowerment and Pedagogical Innovation
Gaining Financial and Life Literacy at SFBULibrary
Cultivating
```

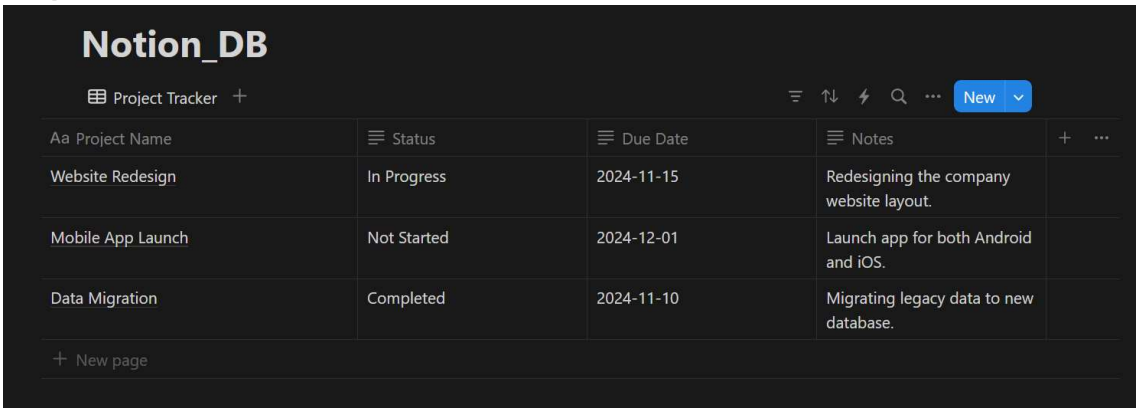
6. **Loading Notion Page Content:**

Step1: Login to https://www.notion.so

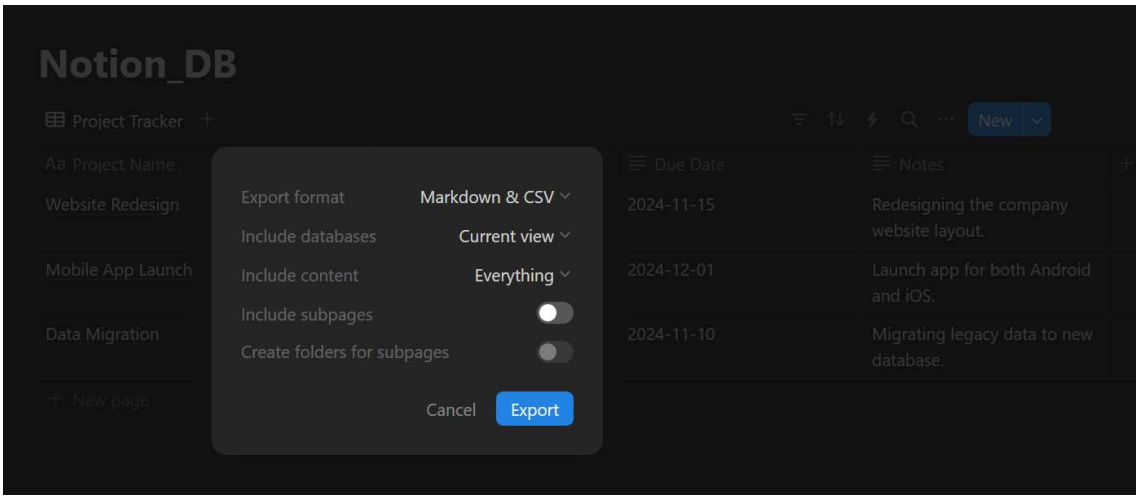
Step2: Create Notion_DB database and add few rows to this.



Step3: Add few rows of content to this Database.



Step4: Export to markdown file as below:



Step5: Load using NotionDirectoryLoader as below:

```python
# ### Load Notion Page Content
# Load documents stored in Notion by specifying the Notion directory path. The content is expected to be in Markdown format.

notion_loader = NotionDirectoryLoader("docs/Notion_DB")
notion_docs = notion_loader.load()

# Display a preview of Notion document content and its metadata.
print("\nLength: ", len(notion_docs))
print("\nNotion Document Content Preview:\n")
print(notion_docs[0].page_content[:200])  # Display the first 200 characters
print("\nNotion Document Metadata:\n")
print(notion_docs[0].metadata)
```

Output:

```
Length:  2

Notion Document Content Preview:

Project Name,Status,Due Date,Notes
Website Redesign,In Progress,2024-11-15,Redesigning the company website layout.
Mobile App Launch,Not Started,2024-12-01,Launch app for both Android and iOS.
Data M

Notion Document Metadata:

{'source': 'docs/Notion_DB/a52c8779-1e0f-40aa-bc38-e94ce9a2ece4_Export-0c2d076b-e39b-4143-86d7-2506dc11c194/Notion_DB 138cf68914cc801f8876cdf8a03c3fb2.md'}
(venv) vaishnavi@DESKTOP-9V8KJG2:/mnt/c/Users/Mohit/Desktop/Gen AI/Week 7$
```

Summary:

This script loads documents from various sources, formats them into manageable chunks, and sets them up for easy retrieval by an LLM. This approach enables specific content to be extracted and used for retrieval-augmented question answering.

GitHub URL:

https://github.com/vaishnavi477/Machine-Learning/tree/main/LangChain%20Chat%20with%20your%20Data/Document%20Loading

Google Slides:

https://docs.google.com/presentation/d/1hHHHeCF_g-HJo7g7f7iW5DbqcpeCgdT0okh3pPrq1BM/edit?usp=sharing