Langchain Chat with your Data:

Document Loading

By:
Melanie Meby Olisah

Process

Step 1: Study the concepts of Retrieval Rugmented Generation (RAG)

Step 2: Download various data sources

Step 2.1: PDF

Required tasks: SFBU 2023 Catalog

Step 2.2: Youtube

Required tasks: SFBU DeepPiCar

Step 2.3: URL

Required tasks: About Us - SFBU

PDF

```python
from langchain_community.document_loaders import PyPDFLoader
loader = PyPDFLoader("https://www.sfbu.edu/sites/default/files/2022-12/2023Catalog.pdf")
pages: List[Document] = loader.load()
```

✓ 19.7s

```python
len(pages)
```

✓ 0.0s

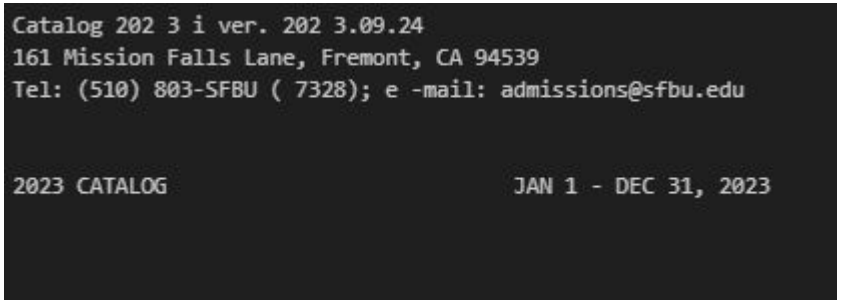
```
197
```

```python
page: Document = pages[0]
```

✓ 0.0s

```python
print(page.page_content[0:500])
```

✓ 0.0s


Catalog 202 3 i ver. 202 3.09.24
161 Mission Falls Lane, Fremont, CA 94539
Tel: (510) 803-SFBU (7328); e -mail: admissions@sfbu.edu
2023 CATALOG
JAN 1 - DEC 31, 2023

Output

YOUTUBE

```python
from langchain_community.document_loaders.generic import GenericLoader
from langchain_community.document_loaders.parsers import OpenAIWhisperParser
from langchain_community.document_loaders.blob_loaders.youtube_audio import YoutubeAudioLoader
```

✓ 0.0s

Note: This can take several minutes to complete.

```python
url="https://youtu.be/AuDodQm7nm8?si=QgtvcsNofH8vqbn0"
save_dir="docs/youtube/"
loader = GenericLoader(
    YoutubeAudioLoader([url],save_dir),
    OpenAIWhisperParser()
)
docs: List[Document] = loader.load()
docs[0].page_content[0:500]
```

✓ 1m 10.2s

```
[youtube] Extracting URL: https://youtu.be/AuDodQm7nm8?si=QgtvcsNofH8vqbn0
[youtube] AuDodQm7nm8: Downloading webpage
[youtube] AuDodQm7nm8: Downloading ios player API JSON
[youtube] AuDodQm7nm8: Downloading android player API JSON
[youtube] AuDodQm7nm8: Downloading m3u8 information
[info] AuDodQm7nm8: Downloading 1 format(s): 140
[download] docs/youtube//SFBU DeepPiCar:  Voice Control.m4a has already been downloaded
[download] 100% of  228.93KiB
[ExtractAudio] Not converting audio docs/youtube//SFBU DeepPiCar:  Voice Control.m4a; file is already in target format m4a
Transcribing part 1!
Transcribing part 1!
Transcribing part 1!
Transcribing part 1!
```

URL

```python
from langchain_community.document_loaders import WebBaseLoader

loader = WebBaseLoader("https://www.sfbu.edu/about-us")
```

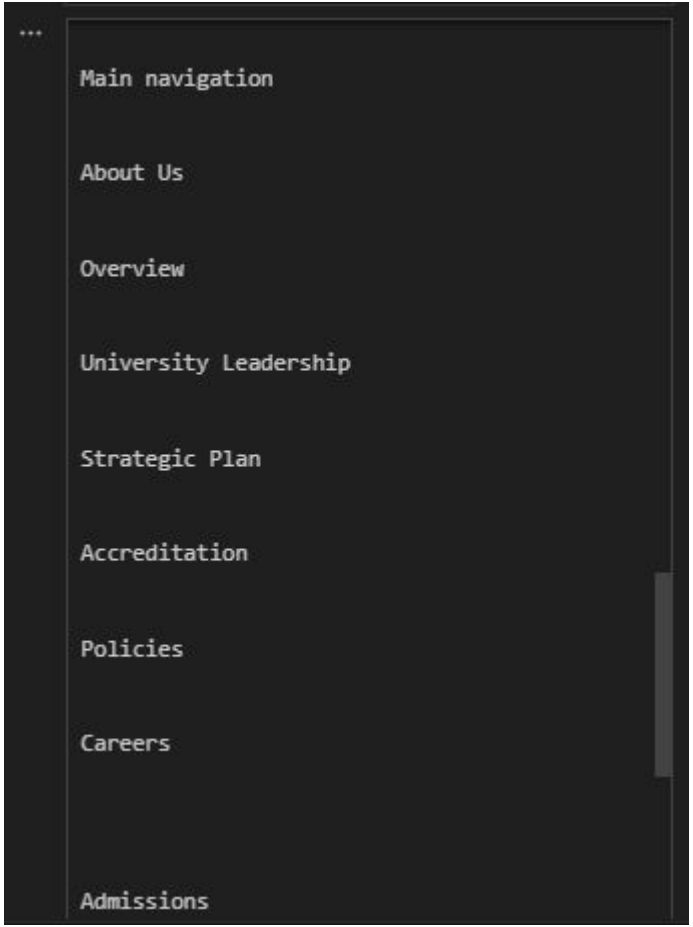
✓ 0.0s

```python
docs: List[Document] = loader.load()
```

✓ 0.7s

```python
print(docs[0].page_content[:500])
```

✓ 0.0s


Main navigation
About Us
Overview
University Leadership
Strategic Plan
Accreditation
Policies
Careers
Admissions

Output