19953 Zhiyu Zhang

DS565 Week 7 Homework 1

Nov 1st, 2024

SFBU Customer Support System - Document Loading

Note: In the most updated LangChain v0.3.x (by 09/16/24), LangChain has migrated many modules from `langchain` to `langchain_community`, including all the document loaders. You can check the newest code snippets and other components here.

Step 1: Set up API

```
import os
import openai
import sys
sys.path.append('../..')

from dotenv import load_dotenv, find_dotenv
_ = load_dotenv(find_dotenv()) # read local .env file
openai.api_key  = os.environ['OPENAI_API_KEY']
```

Step 2: Load PDFs

Make sure to import from `langchain_community` instead of `langchain` for all code snippets.

```
# !pip install langchain_community pypdf
from langchain_community.document_loaders import PyPDFLoader

loader = PyPDFLoader(
     "docs/sfbu-2024-2025-university-catalog-8-20-2024.pdf")
pages = loader.load_and_split()
```

Result:

[5]:
```
from langchain_community.document_loaders import PyPDFLoader
loader = PyPDFLoader(
     "docs/sfbu-2024-2025-university-catalog-8-20-2024.pdf")
pages = loader.load_and_split()
```

[6]:
```
len(pages)
```

[6]:
```
190
```

[7]:
```
page = pages[0]
```

[8]:
```
print(page.page_content[0:500])
```

```
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 sup
```

[9]:
```
page.metadata
```

[9]:
```
{'source': 'docs/sfbu-2024-2025-university-catalog-8-20-2024.pdf', 'page': 1}
```

19953 Zhiyu Zhang

DS565 Week 7 Homework 1

Nov 1st, 2024

Step 3: Load Youtube videos

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

# !pip install yt_dlp pydub torch transformers

# You can insert more URLs into the list. Using only one here to save cost.
urls=["https://www.youtube.com/watch?v=kuZNIvdwnMc"]

save_dir="docs/youtube/"

loader = GenericLoader(
    YoutubeAudioLoader(urls, save_dir), OpenAIWhisperParser())

docs = loader.load()
```

Result:

[10]:
```
# Load Youtube videos
```

[ ]:
```
# !pip install yt_dlp pydub
```

[11]:
```
from langchain_community.document_loaders.blob_loaders.youtube_audio import YoutubeAudioLoader
from langchain_community.document_loaders.generic import GenericLoader
from langchain_community.document_loaders.parsers.audio import OpenAIWhisperParser
```

[12]:
```
urls=["https://www.youtube.com/watch?v=kuZNIvdwnMc"]

save_dir="docs/youtube/"

loader = GenericLoader(
    YoutubeAudioLoader(urls, save_dir), OpenAIWhisperParser())

docs = loader.load()
```

```
[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   10.20MiB
[ExtractAudio] Not converting audio docs/youtube//San Francisco Bay University MBA Student Spotlig
ht:  John Odebode.m4a; file is already in target format m4a
Transcribing part 1!
```

[13]:
```
docs[0].page_content[0:500]
```

[13]:
```
"My name is John, John Odebode. I am studying for an MBA program here at SFBU. It's my final trime
ster at SFBU and I will be graduating in two weeks. I am from Nigeria. I studied at the University
of Lagos for my first degree in philosophy. I also studied for my first master's degree in philoso
phy as well at the same university. I have been practicing within the supply chain industry for th
e past six years. I have spent the most part of my career at ExxonMobil and I recently completed a
six-month"
```

19953 Zhiyu Zhang

DS565 Week 7 Homework 1

Nov 1st, 2024

Step 4: Load URLs

I modified the code so that it will print out cleaner content.

```
from langchain_community.document_loaders import WebBaseLoader
import re

urls = ["https://www.sfbu.edu/student-health-insurance",
    "https://www.sfbu.edu/why-we-are-here",
    "https://www.sfbu.edu/admissions",
    "https://www.sfbu.edu/learning-teaching",
    "https://www.sfbu.edu/student-life-support",
    "https://www.sfbu.edu/contact-us"]

for url in urls:
    loader = WebBaseLoader(url)
    docs = loader.load()
    raw_content = docs[0].page_content
    cleaned_content = re.sub(r'\n\s*\n', '\n', raw_content).strip()
```

Result:

[62]:
```
# Load URLs
```

[64]:
```
from langchain_community.document_loaders import WebBaseLoader
import re

urls = ["https://www.sfbu.edu/student-health-insurance",
    "https://www.sfbu.edu/why-we-are-here",
    "https://www.sfbu.edu/admissions",
    "https://www.sfbu.edu/learning-teaching",
    "https://www.sfbu.edu/student-life-support",
    "https://www.sfbu.edu/contact-us"]

for url in urls:
    loader = WebBaseLoader(url)
    docs = loader.load()
    raw_content = docs[0].page_content
    cleaned_content = re.sub(r'\n\s*\n', '\n', raw_content).strip()
```

[66]:
```
print(docs[0].metadata)
```

```
{'source': 'https://www.sfbu.edu/contact-us', 'title': 'Contact Us | San Francisco Bay University', 'description': 'Hav
e questions and need to connect with an SFBU team member? Call, email, or fill out a form. We are here to help you!', '
language': 'en'}
```

[68]:
```
print(cleaned_content[:500])
```

```
Contact Us | 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
Culti
```

19953 Zhiyu Zhang

DS565 Week 7 Homework 1

Nov 1st, 2024

Step 5: Load Notion database

Export a Notion database in 'Markdown/CSV' format, then unzip the file under a folder named 'Notion_DB' inside the directory you're working in.

```
from langchain_community.document_loaders import NotionDirectoryLoader

loader = NotionDirectoryLoader("Notion_DB")
docs = loader.load()
```

Result:

```
[54]: # Load Notion Database

[56]: from langchain_community.document_loaders import NotionDirectoryLoader

      loader = NotionDirectoryLoader("Notion_DB")
      docs = loader.load()

[58]: print(docs[0].page_content[0:500])

      # Blendle's Employee Handbook

      This is a living document with everything we've learned working with people while running a startup. And, of course, we
      continue to learn. Therefore it's a document that will continue to change.

      **Everything related to working at Blendle and the people of Blendle, made public.**

      These are the lessons from three years of working with the people of Blendle. It contains everything from [how our lead
      ers lead](https://www.notion.so/ecfb7e647136468a9a0a32f1771a8f52?pv

[60]: docs[0].metadata

[60]: {'source': "Notion_DB/Blendle's Employee Handbook 13082cad8cd680a4af71e02dfcdcfde2.md"}
```