Document Loading

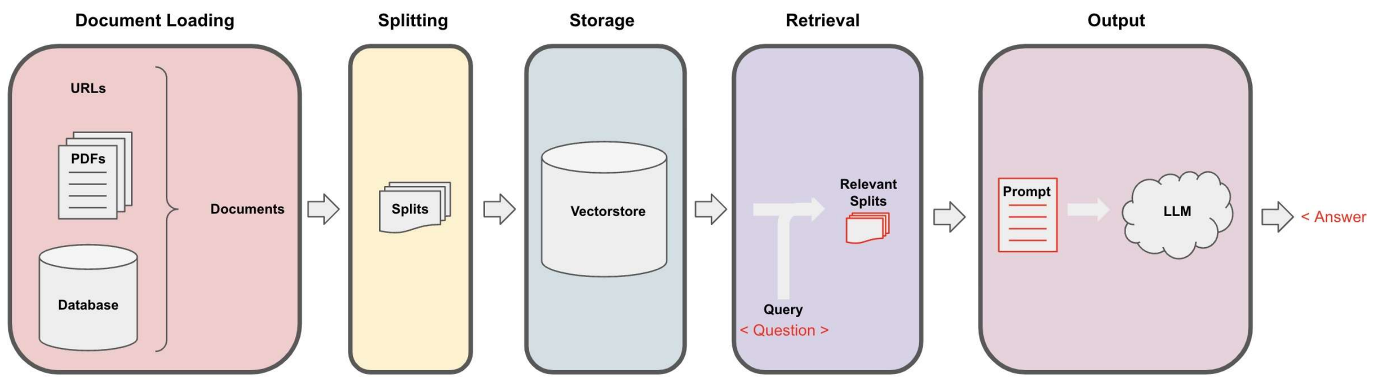
Note to students.

During periods of high load you may find the notebook unresponsive. It may appear to execute a cell, update the completion number in brackets [#] at the left of the cell but you may find the cell has not executed. This is particularly obvious on print statements when there is no output. If this happens, restart the kernel using the command under the Kernel tab.

Retrieval augmented generation

In retrieval augmented generation (RAG), an LLM retrieves contextual documents from an external dataset as part of its execution.

This is useful if we want to ask question about specific documents (e.g., our PDFs, a set of videos, etc).



```
In [1]: #! pip install langchain
```

```
In [2]: 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']
```

PDFs

Let's load a PDF transcript (https://see.stanford.edu/materials/aimlcs229/transcripts/MachineLearning-Lecture01.pdf) from Andrew Ng's famous CS229 course! These documents are the result of automated transcription so words and sentences are sometimes split unexpectedly.

In [3]:
```
# The course will show the pip installs you would need to install packages c
# These packages are already installed on this platform and should not be ru
#! pip install pypdf
```

In [4]:
```
from langchain.document_loaders import PyPDFLoader
loader = PyPDFLoader("docs/cs229_lectures/MachineLearning-Lecture01.pdf")
pages = loader.load()
```

Each page is a `Document`.

A `Document` contains text (`page_content`) and `metadata`.

In [5]:
```
len(pages)
```

```
22
```

In [6]:
```
page = pages[0]
```

In [7]:
```
print(page.page_content[0:500])
```

```
MachineLearning-Lecture01
Instructor (Andrew Ng):  Okay. Good morning. Welcome to CS229, the machine
learning class. So what I wanna do today is ju st spend a little time going o
ver the logistics
of the class, and then we'll start to  talk a bit about machine learning.
By way of introduction, my name's  Andrew Ng and I'll be instru ctor for this
class. And so
I personally work in machine learning, and I' ve worked on it for about 15 ye
ars now, and
I actually think that machine learning i
```

In [8]:
```
page.metadata
```

```
{'source': 'docs/cs229_lectures/MachineLearning-Lecture01.pdf', 'page': 0}
```

YouTube

In [9]:
```
from langchain.document_loaders.generic import GenericLoader
from langchain.document_loaders.parsers import OpenAIWhisperParser
from langchain.document_loaders.blob_loaders.youtube_audio import YoutubeAuc
```

```
In [10]: # ! pip install yt_dlp
         # ! pip install pydub
```

Note: This can take several minutes to complete.

```
In [11]: url="https://www.youtube.com/watch?v=jGwO_UgTS7I"
         save_dir="docs/youtube/"
         loader = GenericLoader(
             YoutubeAudioLoader([url],save_dir),
             OpenAIWhisperParser()
         )
         docs = loader.load()
```

```
    225     )
--> 226     resp, got_stream = self._interpret_response(result, stream)
    227     return resp, got_stream, self.api_key

File /usr/local/lib/python3.9/site-packages/openai/api_requestor.py:619, i
n APIRequestor._interpret_response(self, result, stream)
    611     return (
    612         self._interpret_response_line(
    613             line, result.status_code, result.headers, stream=True
    614         )
    615         for line in parse_stream(result.iter_lines())
    616     ), True
    617 else:
    618     return (
--> 619         self._interpret_response_line(
    620             result.content.decode("utf-8"),
    621             result.status_code,
    622             result.headers,
    623             stream=False,
    624         ),
```

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

URLs

```
In [ ]: from langchain.document_loaders import WebBaseLoader

        loader = WebBaseLoader("https://github.com/basecamp/handbook/blob/master/37s
```

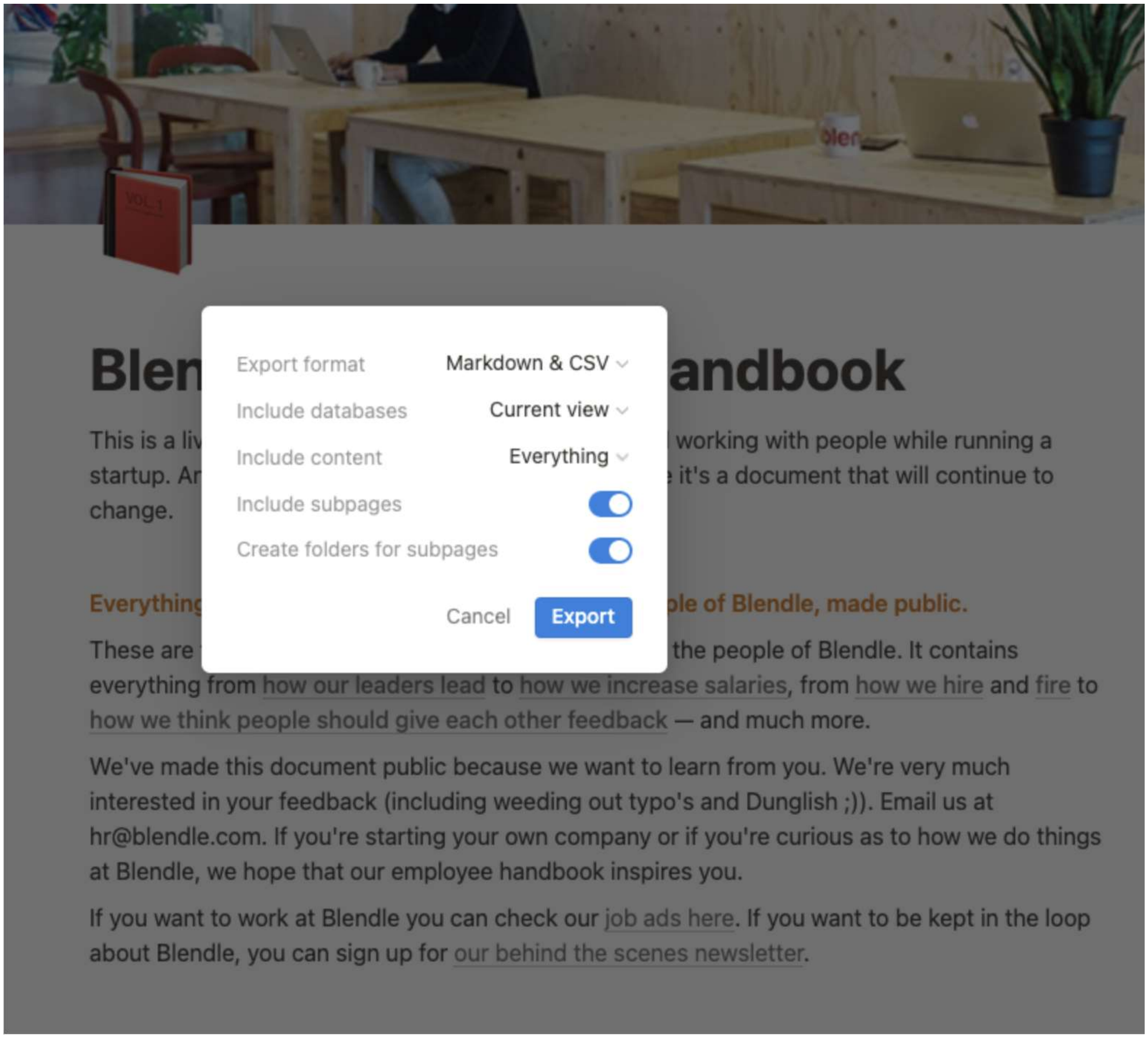
```
In [ ]: docs = loader.load()
```

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

Notion

Follow steps [here (https://python.langchain.com/docs/modules/data_connection/document_loaders/integrations/notion)](https://python.langchain.com/docs/modules/data_connection/document_loaders/integrations/notion) for an example Notion site such as [this one (https://yolospace.notion.site/Blendle-s-Employee-Handbook-e31bff7da17346ee99f531087d8b133f)](https://yolospace.notion.site/Blendle-s-Employee-Handbook-e31bff7da17346ee99f531087d8b133f):

- Duplicate the page into your own Notion space and export as `Markdown / CSV`.
- Unzip it and save it as a folder that contains the markdown file for the Notion page.



Blen andbook

This is a liv working with people while running a startup. Ar it's a document that will continue to change.

Everything ple of Blendle, made public.

These are the people of Blendle. It contains everything from how our leaders lead to how we increase salaries, from how we hire and fire to how we think people should give each other feedback — and much more.

We've made this document public because we want to learn from you. We're very much interested in your feedback (including weeding out typo's and Dunglish ;)). Email us at hr@blendle.com. If you're starting your own company or if you're curious as to how we do things at Blendle, we hope that our employee handbook inspires you.

If you want to work at Blendle you can check our job ads here. If you want to be kept in the loop about Blendle, you can sign up for our behind the scenes newsletter.

Export format: Markdown & CSV
Include databases: Current view
Include content: Everything
Include subpages: on
Create folders for subpages: on

```
In [12]: from langchain.document_loaders import NotionDirectoryLoader
         loader = NotionDirectoryLoader("docs/Notion_DB")
         docs = loader.load()
```

In [13]:
```
print(docs[0].page_content[0:200])
```

```
# Blendle's Employee Handbook

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

In [14]:
```
docs[0].metadata
```

```
{'source': "docs/Notion_DB/Blendle's Employee Handbook e367aa77e225482c849111
687e114a56.md"}
```

In [ ]: