Apify Dataset

Contents

This notebook shows how to load Apify datasets to LangChain.

Apify Dataset is a scaleable append-only storage with sequential access built for storing structured web scraping results, such as a list of products or Google SERPs, and then export them to various formats like JSON, CSV, or Excel. Datasets are mainly used to save results of Apify Actors—serverless cloud programs for varius web scraping, crawling, and data extraction use cases.

Prerequisites

You need to have an existing dataset on the Apify platform. If you don't have one, please first check out this notebook on how to use Apify to extract content from documentation, knowledge bases, help centers, or blogs.

First, import `ApifyDatasetLoader` into your source code:

```
from langchain.document_loaders import ApifyDatasetLoader
from langchain.document_loaders.base import Document
```

Then provide a function that maps Apify dataset record fields to LangChain `Document` format.

For example, if your dataset items are structured like this:

```
{
    "url": "https://apify.com",
```

Skip to main content

```
    "text": "Apify is the best web scraping and automation platform."
}
```

The mapping function in the code below will convert them to LangChain `Document` format, so that you can use them further with any LLM model (e.g. for question answering).

```
loader = ApifyDatasetLoader(
    dataset_id="your-dataset-id",
    dataset_mapping_function=lambda dataset_item: Document(
        page_content=dataset_item["text"], metadata={"source": dataset_item["url"]}
    ),
)
```

```
data = loader.load()
```

An example with question answering

In this example, we use data from a dataset to answer a question.

```
from langchain.docstore.document import Document
from langchain.document_loaders import ApifyDatasetLoader
from langchain.indexes import VectorstoreIndexCreator
```

```
loader = ApifyDatasetLoader(
    dataset_id="your-dataset-id",
    dataset_mapping_function=lambda item: Document(
        page_content=item["text"] or "", metadata={"source": item["url"]}
    ),
)
```

```
index = VectorstoreIndexCreator().from_loaders([loader])
```

```
query = "What is Apify?"
result = index.query_with_sources(query)
```

```
print(result["answer"])
```

Skip to main content

```
 Apify is a platform for developing, running, and sharing serverless cloud
programs. It enables users to create web scraping and automation tools and publish
them on the Apify platform.

https://docs.apify.com/platform/actors,
https://docs.apify.com/platform/actors/running/actors-in-store,
https://docs.apify.com/platform/security,
https://docs.apify.com/platform/actors/examples
```