HTML

Contents

- Loading HTML with BeautifulSoup4

This covers how to load HTML documents into a document format that we can use downstream.

```python
from langchain.document_loaders import UnstructuredHTMLLoader
```

```python
loader = UnstructuredHTMLLoader("example_data/fake-content.html")
```

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

```python
data
```

```
[Document(page_content='My First Heading\n\nMy first paragraph.', lookup_str='',
metadata={'source': 'example_data/fake-content.html'}, lookup_index=0)]
```

Loading HTML with BeautifulSoup4

We can also use BeautifulSoup4 to load HTML documents using the `BSHTMLLoader`. This will extract the text from the html into `page_content`, and the page title as `title` into `metadata`.

```python
from langchain.document_loaders import BSHTMLLoader
```

```python
loader = BSHTMLLoader("example_data/fake-content.html")
```

Skip to main content

```
data
```

```
[Document(page_content='\n\nTest Title\n\n\nMy First Heading\nMy first
paragraph.\n\n\n', lookup_str='', metadata={'source': 'example_data/fake-
content.html', 'title': 'Test Title'}, lookup_index=0)]
```