HTML

> The HyperText Markup Language or HTML is the standard markup language for documents designed to be displayed in a web browser.

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

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

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

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

```
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`.

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

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

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