# LangChain如何加载不同格式的数据

LangChain的数据加载能力并不限于单一的数据源或格式，它可以处理各种常见的数据格式，例如CSV、文件目录、HTML、JSON、Markdown以及PDF等。下面，我们将分别解析一下这些不同格式数据的加载方法。

#### 1. CSV文件的加载

逗号分隔值（Comma-Separated Values，简称CSV）文件是一种使用逗号来分隔值的文本文件。每一行都是一条数据记录，每条记录包含一个或多个用逗号分隔的字段。LangChain可以加载CSV数据，其中每一行都被视为一个独立的文档。

#### 2. 文件目录的加载

对于文件目录，LangChain提供了一种方法来加载目录中的所有文档。在底层，它默认使用UnstructuredLoader来实现这个功能。这意味着，只要将文档存放在同一目录下，无论数量多少，LangChain都能够将它们全部加载进来。

#### 3. HTML文件的加载

HTML是用于设计在Web浏览器中显示的文档的标准标记语言。LangChain可以将HTML文档加载为我们后续使用的文档格式。这就意味着，我们可以直接从网页上提取并处理数据。

#### 4. JSON文件的加载

JSON是一种使用人类可读文本来存储和传输数据对象的开放标准文件格式和数据交换格式，这些对象由属性-值对和数组（或其他可序列化值）组成。LangChain的JSONLoader使用指定的jq模式来解析JSON文件。jq是一种适用于Python的软件包。JSON文件的每一行都被视为一个独立的文档。

#### 5. Markdown文件的加载

Markdown是一种使用纯文本编辑器创建格式化文本的轻量级标记语言。LangChain可以将Markdown文档加载为我们后续使用的文档格式。

#### 6. PDF文件的加载

PDF（Portable Document Format）是Adobe在1992年开发的一种文件格式，用于以独立于应用软件、硬件和操作系统的方式呈现文档，包括文本格式化和图像。LangChain可以将PDF文档加载为我们后续使用的文档格式。

总的来说，通过对各种不同数据格式的加载能力，LangChain为大规模、多样性的数据处理提供了强大的支持。