HTML CSS Selector 工具用户使用手册

罗毅凡 (2024202715)

2025 年 11 月 30 日

1 概述

`HtmlSelectorTool` 是一个基于 C/C++ 实现的高效 HTML 解析和数据抽取工具。它将 HTML 文档结构建模为一棵树，并支持常用的 CSS 选择器（Selector）进行节点定位，以及方便地抽取节点的文本内容、完整 HTML 代码和链接信息。本手册将指导用户如何导入 HTML 文档、执行 CSS 查询以及对查询结果进行操作。

2 核心功能

功能	描述
HTML 建模	将 HTML 文件或 URL 内容解析为内存中的树形数据结构（DOM）。
CSS 选择器	支持实现列表中的基本 CSS 选择器，快速定位目标节点。
数据抽取	支持抽取节点的内部文本、外部 HTML 代码以及链接（`href`）属性。
链式查询	支持对已选中的节点集合再次执行 CSS 查询，实现局部精确抽取。

3 命令行交互操作

程序提供一个交互式的命令行界面，用户可以通过输入以下命令进行操作：

3.1 文档加载：read()

该命令用于加载 HTML 内容到程序中，构建 DOM 树。

命令格式	描述
`read(file_name)`	从本地文件加载 HTML 内容。
`read(url)`	从指定的 URL 地址下载并加载 HTML 内容。

示例:

```
> read(实验03\examples\example.html)
文档加载成功。
```

3.2 全局查询：query()

该命令用于在当前加载的整个 HTML 文档中，根据给定的 CSS 选择器检索所有符合条件的节点。

命令格式	描述
`query(selector)`	执行查询，结果将存储为一个有序列表，供后续操作使用。

支持的基本 CSS 选择器 (Implemented Selectors):

选择器	例子	例子描述
.class	.intro	选择 class=”intro” 的所有元素。
.class1.class2	.name1.name2	选择 class 属性中同时有 name1 和 name2 的所有元素。
.class1 .class2	.name1 .name2	选择作为类名 name1 元素后代的所有类名 name2 元素。
#id	#firstname	选择 id=”firstname” 的元素。
	*	选择所有元素。
element	p	选择所有 <p> 元素。
element.class	p.intro	选择 class=”intro” 的所有 <p> 元素。
element,element	div, p	选择所有 元素和所有 <p> 元素。
element element	div p	选择 元素内的所有 <p> 元素。
element>element	div > p	选择父元素是 的所有 <p> 元素。
element+element	div + p	选择紧跟 元素的首个 <p> 元素。
element1~element2	p ~ ul	选择前面有 <p> 元素的每个 元素。

查询结果输出: 程序将按文档顺序打印每个匹配节点的简要信息，并将其保存到 **结果列表** (`Out`) 中。

示例:

```
> query(.class2.class3)
[div.class2.class3, span.class3.class2, img.class2.class0.class3]
> query(#id4.class2)
[div#id4.class2]
```

3.3 XPath 查询：xpath()

该命令允许用户使用 XPath 语法在当前文档中进行灵活的节点查找和定位。

表达式	描述	示例
nodename	选取此节点的所有子节点。	body
/	从根节点选取（绝对路径）。	/body/div
//	从匹配选择的当前节点选择文档中的节点，而不考虑它们的位置（递归查找）。	//div
.	选取当前节点。	.
..	选取当前节点的父节点。	..
@	选取属性（常用于谓语条件中）。	//div[@id='id1']
*	通配符，匹配任何元素节点。	/body/*

示例:

```
> xpath(/body/div)
[div.class0.class1, div, div#id1, div, ...]
> xpath(//div[@id='id1'])
[div#id1]
```

3.4 结果列表操作：Out[k].operation

查询完成后，用户可以通过 Out[k] 访问结果列表中的第 k 个节点 (索引 k 从 0 开始)。

A. 获取内部文本：Out[k].innerText 返回节点及其所有后代节点的**纯文本内容**，去除所有 HTML 标签。

B. 获取完整 HTML：Out[k].outerHTML 返回节点自身的**完整 HTML 代码**，包括其自身标签及其所有子节点。

C. 获取链接属性：Out[k].href 仅适用于 <a> 标签节点。返回该节点的 href 属性值。

D. 链式查询/二次查询：Out[k].query(selector) 以 Out[k] 节点为根，在其内部进行一次新的 CSS 查询。新的查询结果将**覆盖**原有的结果列表 (Out)。

E. 链式 XPath 查询：Out[k].xpath(path) 以 Out[k] 节点为上下文环境（当前节点），执行 XPath 查询。

3.5 退出

输入 q 或 quit 或 exit 退出程序。