## **实验02、HTML代码解析**

目的

1. 复习巩固线性表数据结构。
2. 练习使用栈数据结构。
3. 熟悉和了解在实际问题中合理设计和使用数据结构

问题描述

1. 实现栈的基本数据结构。
2. 使用线性表存储HTML
3. 基于线性表和栈，完成HTML合法性的校验等任务。

步骤

1. 参考教材上的代码，实现栈。
2. 通过文件读取HTML代码（文本编码），解析HTML代码，识别出其中的各种HTML Tag和文本，采用合适的数据结构进行存储。关于HTML相关知识，请参考： <https://www.w3school.com.cn/tags/index.asp>。假设程序为HtmlParser.exe，html文件为test01.html，执行HtmlParser.exe test01.html后，可等待用户输入CheckHTML，OutHTML，Text, Exit等命令。（遇到注释、css块和javascript块可直接删除）
3. CheckHTML，检验HTML代码的合法性，并输出相关信息。HTML的合法性请参考<https://www.w3school.com.cn/tags/index.asp>。
4. OutHTML(path)：给定单纯由Tag构成的简化的XPATH路径path，输出所有能够匹配该路径的节点的OuterHTML (包含该element的所有html代码段，<https://baike.baidu.com/item/outerHTML/9806570>)，例如，给定path=/HTML/BODY/DIV/P，输出页面html标签下的body标签下的所有div中的所有p标签的代码, 多个elements的html用\n隔开。
5. Text(path)输出页面中符合路径条件的elements中的所有文本内容，多个elements的文本用\n隔开(建议：输出某个element的内部文本时，建议在block元素后面增加换行，inline元素不增加换行，其他连续空白符合并为一个空格)
6. Exit退出程序（不计算分值）
7. 在给定的几个测试文件，以及助教任选的其他测试文件（只考虑英文）基础上，代码能正常运行。
8. 选作：在前面的基础上(满分100分)，考虑实现下面的部分功能。总分如超过100分，只计算100分。
   1. 参考XPATH <https://www.w3school.com.cn/xpath/xpath_syntax.asp> 的语法，实现更复杂的XPath查询，例如部分路径、相对路径、属性等。（5分）
   2. 除本地文件外，支持给定URL进行解析（支持各种编码和语言），通过助教随意输入的URL解析无误后得分 ，请参考Scrapy Shell的功能。（5分）

设备和环境

PC计算机、Windows操作系统、C/C++开发环境

结论

1. 能够熟练掌握线性表和栈的实现方式。
2. 能够结合线性表和栈，实现更复杂的功能完成实际需求。

需要注意的事项以及思考

**示例HTML代码如下（实验02\_case00.html）：**

<html xmlns="http://www.w3.org/1999/xhtml">

<head>

    <meta http-equiv="Content-Type" content="text/html; charset=utf-8" />

    <base href="http://airs2016.ruc.edu.cn/" >

    <title>

        AIRS 2016, Beijing, China, November 30 – December 2, 2016

    </title>

</head>

<body role="document">

    <div class="navbar navbar-default navbar-fixed-top" role="navigation">

        <!--navbar-inverse navbar-fixed-top-->

        <div class="container">

            <div class="navbar-collapse collapse">

                <ul class="nav navbar-nav">

                    <li><a href="/cfp.aspx">CFP</a></li>

                    <li><a href="/dates.aspx">Dates</a></li>

                    <li><a href="/papers.aspx">Program</a></li>

                    <li><a href="/reg.aspx">Registration</a></li>

                    <li><a href="/keynote.aspx">Keynotes</a></li>

                    <li><a href="/committee.aspx">Committee</a></li>

                    <li><a href="/location.aspx">Venue</a></li>

                    <li><a href="/beijing.aspx">Beijing</a></li>

                </ul>

            </div>

            <!--/.navbar-collapse -->

        </div>

    </div>

    <div class="container">

        <div class="row">

            <div class="col-xs-12 col-md-12  col-lg-7">

                <div class="text-center" style="margin:40px 0 20px;">

                    <h2>Welcome to <span class="label label-success">AIRS 2016</span></h2>

                    <h2>

                        The 12th Asia Information Retrieval Societies Conference,

                        <br>

                        Nov. 30 – Dec. 2, 2016 at Tsinghua University, Beijing

                        <br>

                        In cooperation with ACM SIGIR.

                    </h3>

                </div>

                <div>

                    <p>

                        The <a href="/">Asia Information Retrieval Societies Conference (AIRS)</a> aims to bring together researchers and developers

                        to exchange new ideas and latest achievements in the field of information retrieval (IR).

                        The scope of the conference covers applications, systems, technologies and theory aspects of information retrieval

                        in text, audio, image, video and multimedia data. The Twelfth AIRS (AIRS 2016) will be hosted by the

                        Chinese Information Processing Society of China and co-organized by Tsinghua University and Renmin University of China.

                        We welcome submissions of original papers in the broad field of information retrieval.

                        Please find the list of <a href="papers.aspx">Accepted papers</a>.

                    </p>

                    <h2>This sentence is not included in the p element.</h2>

                    <p>

                        Accepted papers will be published as part of the LNCS series from Springer, and will be EI-indexed.

                        Details about relevant topics, publication format and submission deadlines can be found in the

                        <a href="cfp.aspx">Call for Papers</a>.

                    </p>

                </div>

            </div>

        </div>

    </div>

    <script src="/Scripts/ie10-viewport-bug-workaround.js"></script>

    <script src="/Content/bootstrap/js/bootstrap.min.js"></script>

</body>

</html>

**HtmlParser.exe 实验02\_case00.html**

**加载完成后可提示加载成功等相关信息。**

CheckHTML

如果HTML合法，则显示合法，否则提示可能导致不合法的相关信息

（可以根据给定的测试HTML随意修改成不合法的来进行测试）

**Text(/)或者 Text / 输出该HTML的文本大致为：**

CFP

Dates

Program

Registration

Keynotes

Committee

Venue

Beijing

Welcome to AIRS 2016

The 12th Asia Information Retrieval Societies Conference,

Nov. 30 – Dec. 2, 2016 at Tsinghua University, Beijing

In cooperation with ACM SIGIR.

The Asia Information Retrieval Societies Conference (AIRS) aims to bring together researchers and developers to exchange new ideas and latest achievements in the field of information retrieval (IR). The scope of the conference covers applications, systems, technologies and theory aspects of information retrieval in text, audio, image, video and multimedia data. The Twelfth AIRS (AIRS 2016) will be hosted by the Chinese Information Processing Society of China and co-organized by Tsinghua University and Renmin University of China. We welcome submissions of original papers in the broad field of information retrieval. Please find the list of Accepted papers.

This sentence is not included in the p element.

Accepted papers will be published as part of the LNCS series from Springer, and will be EI-indexed. Details about relevant topics, publication format and submission deadlines can be found in the Call for Papers.

**Text /html/body/div/div/div/ul/li/a输出:**

CFP

Dates

Program

Registration

Keynotes

Committee

Venue

Beijing

**OuterHtml /html/body/div/div/div/div/p/a 输出：**

<a href="/">Asia Information Retrieval Societies Conference (AIRS)</a>

<a href="papers.aspx">Accepted papers</a>

<a href="cfp.aspx">Call for Papers</a>

**Text /html/body/div/div/div/div/p 输出:**

The Asia Information Retrieval Societies Conference (AIRS) aims to bring together researchers and developers to exchange new ideas and latest achievements in the field of information retrieval (IR). The scope of the conference covers applications, systems, technologies and theory aspects of information retrieval in text, audio, image, video and multimedia data. The Twelfth AIRS (AIRS 2016) will be hosted by the Chinese Information Processing Society of China and co-organized by Tsinghua University and Renmin University of China.

We welcome submissions of original papers in the broad field of information retrieval. Please find the list of Accepted papers.

Accepted papers will be published as part of the LNCS series from Springer, and will be EI-indexed. Details about relevant topics, publication format and submission deadlines can be found in the Call for Papers.