实验02.HTML结构解析和校验 实验报告

罗毅凡 2024202715

1.实现思路

简述整体实现框架，具体细节见 section2

1.1 读取并存储HTML文件

- 对html文件采用线性表的形式存储，线性表中每一个元素都是字符串，根据内容分为两种：
  - 带 `< >` 的tag
  - 普通文本
- 读取采用单字符 `getc` 循环读取：
  - 若 `current` 字符为 `<` ，累计读取直到遇到 `>` ，将这段内容存入线性表
  - 若 `current` 字符不为 `<` 时，累计读取直到遇到 `<` ，将这段内容存入线性表
  - 这样整个html被切分为块存入线性表

1.2 checkHtml

- 区分不同的 `tag` 块及其嵌套关系，使用一个 `enum` 枚举，同时恰好实现不同的 `tag` 块的嵌套优先级关系
- 依次读取线性表：
  - 若为文本块：直接跳过
  - 若为自闭和 `tag` 块：跳过
  - 若为开始 `tag` 块：判断嵌套优先级，提取 `tagname` 并入栈
  - 若为结束 `tag` 块：对比栈顶元素，出栈
- 若正确读完整个线性表还需判断栈是否为空（读取中途遇到问题可直接终止）

1.3 OuterHtml & Text

- 将路径读取并存储在线性表中，这里用一个 `index` 指示当前路径
- 依次读取线性表匹配路径：
  - 若 `tag` 块匹配：入栈， `index++`
  - 若 `tag` 块不匹配：跳过到闭合块
- 当读取到当前路径，输出内容

2.遇到问题与思考收获

问题一：读取时(getc)读取到的字符串中会有大量空格和换行符

- 对最后读取到的字符串进行处理，去除换行符并将多个空格替换为一个

问题二：Html的tag嵌套问题

- 定义了一个 `enum` 区分不同的tag块，并定义嵌套优先级，采用类似运算表达式转换的出入栈形式

问题三：注释、css块等影响判断html合法性

- 在读取阶段读取到注释、css等内容直接跳过不存储

3.实现结果

- 正确实现功能，通过助教老师测试