苏州大学实验报告

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 院系 | 计算机学院 | | 年级专业 | | 21计科 | | 姓名 | 方浩楠 | 学号 | 2127405048 |
| 课程名称 | | 编译原理课程实践 | | | | | | | 成绩 |  |
| 指导教师 | | 王中卿 | | 同组实验者 | | 无 | | 实验日期 | 2023.9.22 | |

|  |  |
| --- | --- |
| 实 验 名 称 | 实验二 |

1. 实验目的

将给定的LaTex文件（example.tex）转换为PDF文件

1. 实验内容

 基于正则表达式将LaTex源文件进行解析

 将解析结果构建为HTML字符串

 通过pyfPDF生成pdf

需完成以下标记的解析如下标记：

 document

 title

 abstract

 section

 subsection

 itemize/item

 tabular

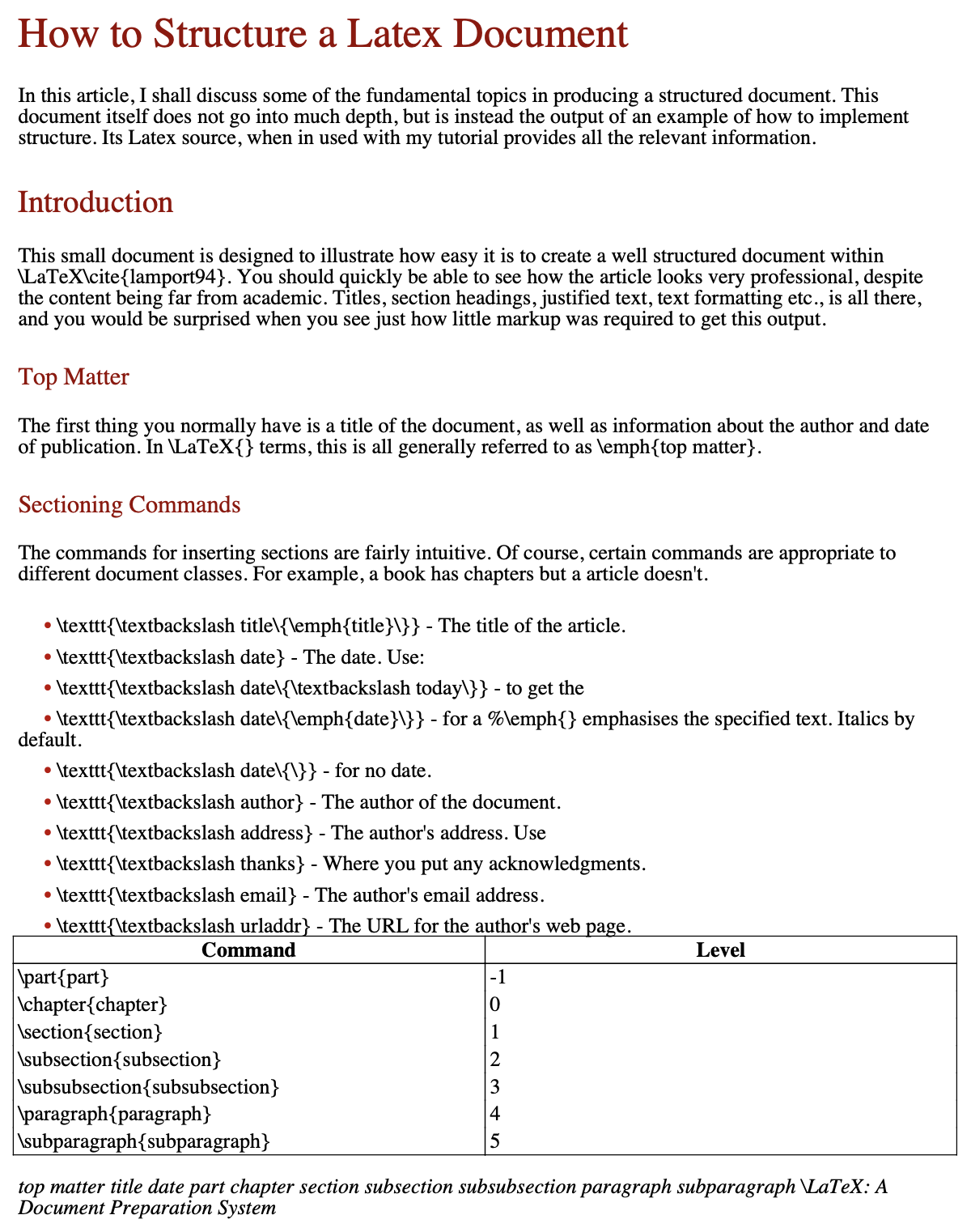
 emph

 textbf

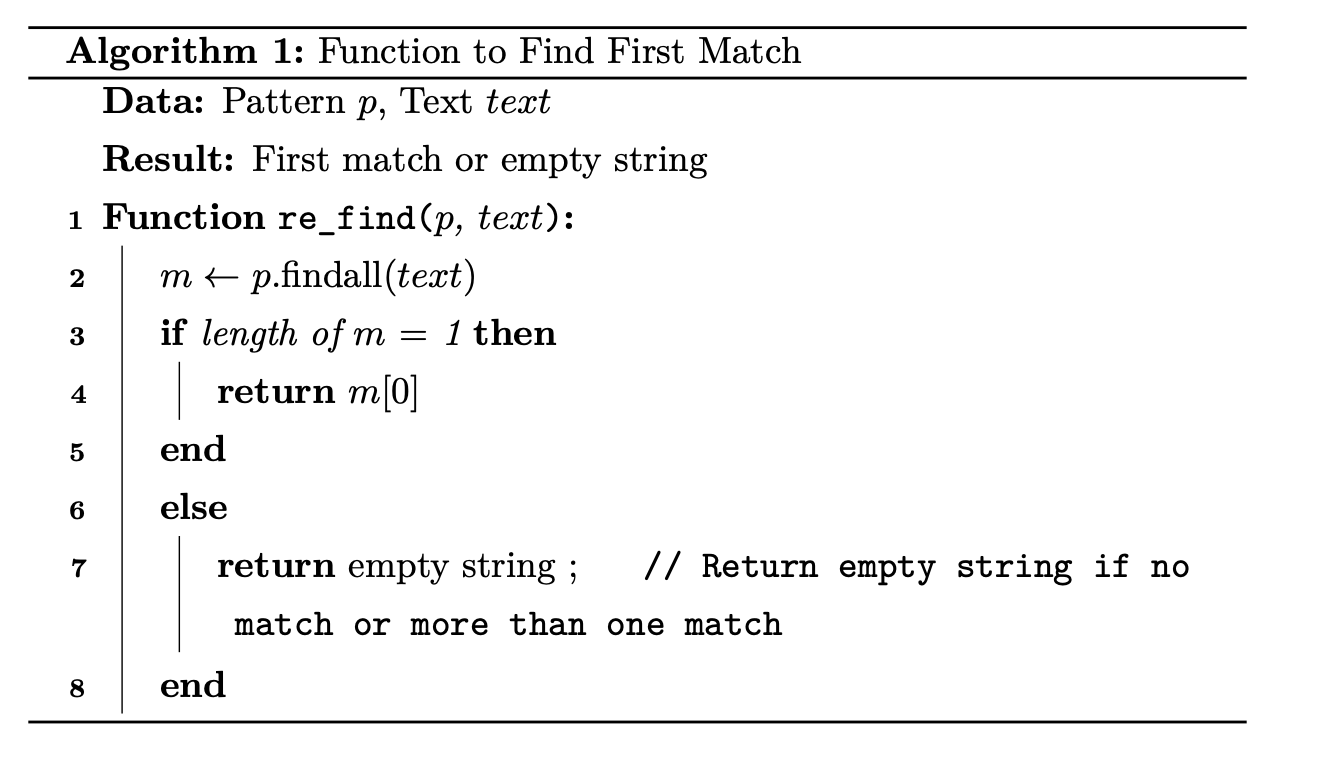
1. 实验步骤和结果

最终运行结果:

运行结果文件位置:实验2/example/example.pdf



函数re\_find伪代码:



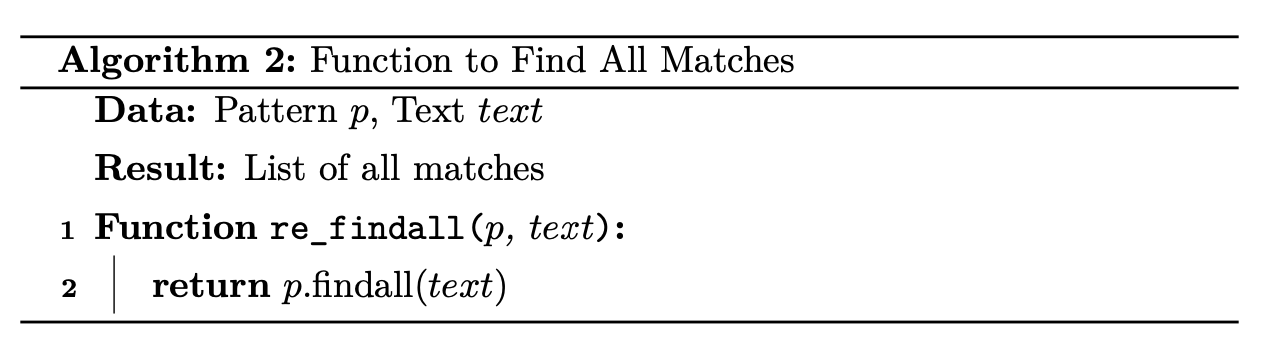
函数re\_find的作用:

此函数接受两个参数：p，是已编译的正则表达式模式，text，是要查找匹配项的输入文本。

它使用正则表达式模式 p 的 findall 方法，在输入文本中查找所有非重叠的匹配项。

如果找到恰好一个匹配项，它将该匹配项作为字符串返回。如果没有找到匹配项或找到多个匹配项，它将返回一个空字符串。

函数re\_findall伪代码:



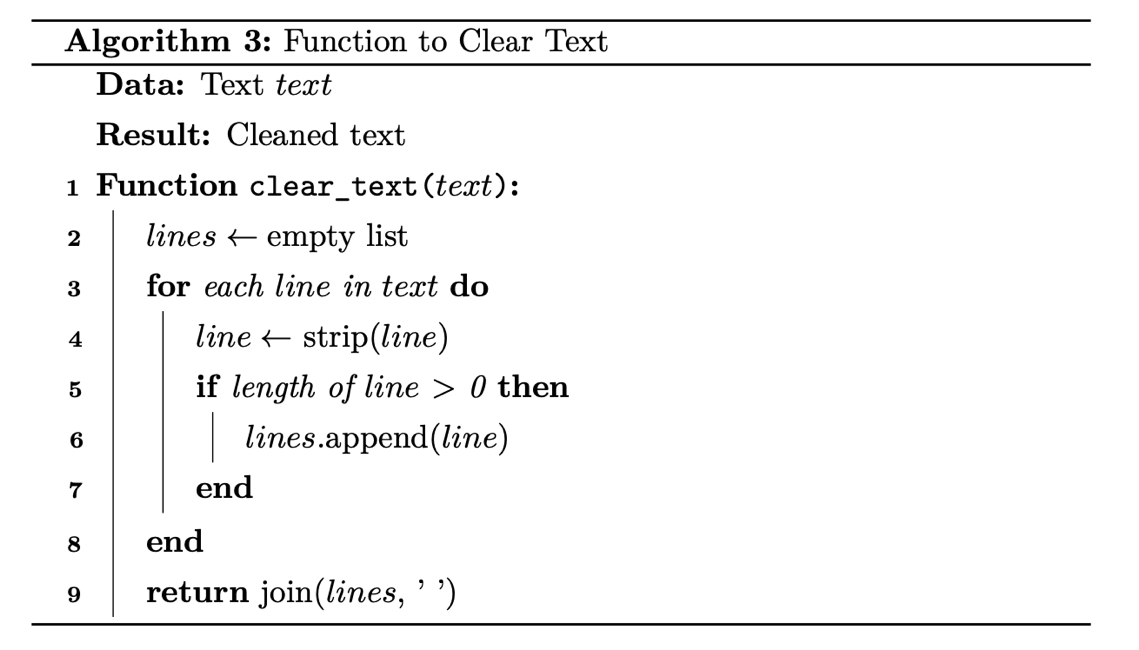
函数re\_findall作用:

此函数同样接受两个参数：p，是已编译的正则表达式模式，text，是要查找匹配项的输入文本。

它使用正则表达式模式 p 的 findall 方法，在输入文本中查找所有非重叠的匹配项。

它返回一个包含所有在输入文本中找到的匹配项的列表

函数clear\_text伪代码



函数clear\_text作用:

此函数接受一个参数 text，即要处理的输入文本。

它使用换行符 ('\n') 将输入文本拆分成多行，然后去除每行的前导和尾随空格，将非空行添加到名为 lines 的列表中。

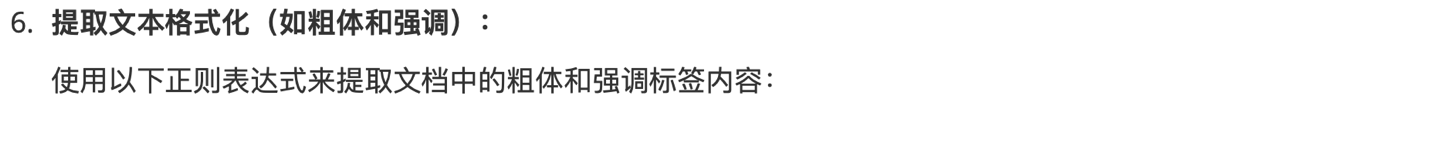
最后，它将清理后的行连接成一个用空格分隔的单一字符串，并返回该清理后的文本。

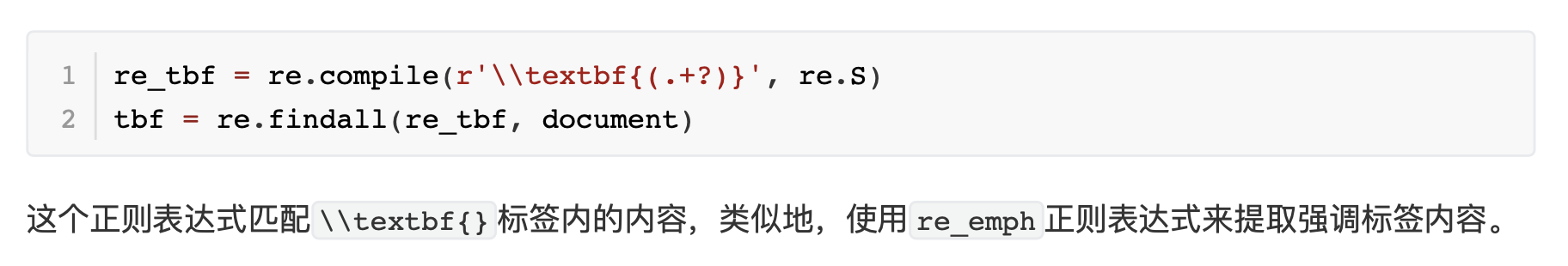
提取latex文档的各个标签











提取latex中的表格的部分:







1. 实验总结

本次实验的主要目的是从一个LaTeX文档中提取内容，将其转换为HTML格式，最终将HTML转换为PDF格式。以下本次试验的总结:

1.正则表达式提取内容：

使用正则表达式从LaTeX文档中提取标题、摘要、章节、子章节、itemize环境、tabular环境以及文本格式化（如粗体和强调）等各种元素。

内容清理和结构化：

2.清理提取的内容，去除多余的空格和换行符。

3.使用HTML标签对提取的内容进行结构化，如使用 <h1> 标签表示标题，<p> 标签表示段落，<ul> 标签表示无序列表，<table> 标签表示表格，以及 <strong> 和 <em> 标签表示粗体和强调文本。

HTML生成：

4.生成HTML文本，将提取的内容以合适的HTML标记嵌入其中，以保持文档的结构和格式。

HTML转PDF：

5.最终使用 html2pdf 库将生成的HTML文本转换为PDF文件。