Python大作业：文本格式转换和简单统计

一、任务简述

本任务利用正则表达式解析给定的《The Merchant of Venice》HTML网页文件，并将文件内容按Markdown格式存储至文件中。

主要考察学生以下几个方面：

- 程序设计能力及Python编程模式的理解
- 运算符、表达式、内置函数及序列结构的运用能力
- Python分支结构、循环结构及函数设计的掌握情况
- 运用正则表达式处理字符串的能力
- Python读写文本文件

二、背景：HTML和Markdown

超级文本标记语言（英文缩写：HTML）是为“网页创建和其它可在网页浏览器中看到的信息”设计的一种标记语言。HTML是一种规范，一种标准，它通过标记符号来标记要显示的网页中的各个部分。网页文件本身是一种文本文件，通过在文本文件中添加标记符，可以告诉浏览器如何显示其中的内容（如：文字如何处理，画面如何安排，图片如何显示等）。一般右键点击网页打开菜单，选择查看源代码，即可查看页面对应HTML代码。Markdown是一种轻量级标记语言，简单易学，用途广泛，易于与HTML进行转换。

HTML文本示例

```
<p>
Act 1, Scene 1: <a href="./merchant/merchant.1.1.html">Venice. A street.
</a><br>
Act 1, Scene 2: <a href="./merchant/merchant.1.2.html">Belmont. A room in
PORTIA'S house.</a><br>
Act 1, Scene 3: <a href="./merchant/merchant.1.3.html">Venice. A public
place.</a><br>
</p>
```

对应网页浏览器显示如下

Act 1, Scene 1: Venice. A street.
Act 1, Scene 2: Belmont. A room in PORTIA'S house.
Act 1, Scene 3: Venice. A public place.

HTML 标记也称HTML 标签(HTML tag)，详细标签含义可查阅HTML 参考手册或其他网络资料，这里只做简单介绍：

- HTML 标记是由尖括号包围的关键词，比如 `<html>`。
- HTML 标记通常是成对出现的，比如 `<p>` 和 `</p>`。

- 标记对中的第一个标记是开始标记，第二个标记是结束标记。
- HTML 标记英文字母不区分大小写，即<a> 和 <A>等同。

三、程序说明

程序的输入、输出及流程

程序首先询问要解析的《The Merchant of Venice》主网页文件（即名为Merchant of Venice_ List of Scenes.html的文件）位置，然后根据输入的文件位置获取文件内容，并开始解析文件内容，主要是利用正则表达式提取网页内容。如果提取的内容为链接，则要进一步读取和解析链接所指向的网页内容。如果提取内容为文本，则要将它转换成Markdown的标记格式，并依序存入同一个文本文件中，具体格式请参考下面**文本输出格式**。程序在解析过程中，还要统计所处理的所有网页文件中的各个HTML标签的出现次数。程序最终输出两个结果：一个是Markdown文件，另一个是在屏幕上分行显示出现最多的三个标记及其出现次数。注意：对于成对的标签，如 仅记做 出现一次。

输入数据说明

附件包含data和document两个子目录，data放置需处理的HTML文件数据，document放置输出格式示例文件。需处理的HTML文件是一个呈二级结构的《The Merchant of Venice》的网页文件，其一级网页即为目录，文件名为Merchant of Venice_ List of Scenes.html，即需要输入到程序中的文件名，该网页记载了各场景剧本文件的相对路径，可用浏览器打开预览，也可使用记事本方式打开查看源码。附件的merchant目录中存放着各场景的剧本网页，这些文件**禁止**直接输入文件名访问，需从Merchant of Venice_ List of Scenes.html网页中的<a>标记中获取场景剧本网页的路径及文件名。

输出格式说明

程序从主网页文件开始提取剧本并存入 名为The Merchant of Venice.md 的Markdown文件中：

1. 剧本名单独一行，且前后各空一行，即被空行包围，并在名称前加一个 # ，并以空格分隔。
2. 幕号 ACT 单独一行，且前后各空一行，即被空行包裹，作为二级标题，即在名称前加 ## 即可，同样使用空格同文字分隔，仅在每一幕的开头添加。可从 Merchant of Venice_ List of Scenes.html 文件中提取。
3. 场名 SCENE 单独一行，且前后各空一行，即被空行包裹，作为三级标题，即在名称前加 ### ，同样使用空格同文字分隔。场名可从子级剧本网页文件中提取，通常在 <h3> 标记内，如 <h3>SCENE I. Venice. A street.</H3> 。
4. 人物名单独一行，作为 ** 包裹，即 **NAME** 。使用 <a> 标记内，如 <A NAME=speech1>ANTONIO ，此时 *NAME=speech1* 代表设置标记的 *NAME* 属性为 *speech1* 。
5. 台词根据提取文本分行，一个标记内的一段话即为一行，直接输出即可。使用 <a> 标记内，如 <A NAME=1>In sooth, I know not why I am so sad:</A> ，此时 *NAME=1* 代表设置标记的 *NAME* 属性为 *1* ，代表台词序号。 注：人物名与台词均使用<a>标记，区别在于 *NAME* 属性设置值不同，详细情况可将剧本幕的网页通过记事本的方式打开查看。
6. 舞台说明又叫舞台提示，单独一行，且前后各空一行。舞台提示为斜体，使用 标记，如 Enter ANTONIO, SALARINO, and SALANIO。

输出文件示例

附件 document 目录下的文件 Example for The Merchant of Venice.md 展示了两场 *ACT* 剧本的输出格式，即附件 data\merchant目录下的merchant.1.1.html和merchant.1.2.html解析转换后的结

果。文件`Example for Markdown to PDF.pdf`是`Example for The Merchant of Venice.md`生成的pdf文件，这里不作为要求，仅作展示。

文件`Example for The Merchant of Venice.md`的部分内容节录展示如下：

```
# The Merchant of Venice

## ACT 1

### SCENE 1: Venice. A street.

*Enter ANTONIO, SALARINO, and SALANIO*

**ANTONIO**

In sooth, I know not why I am so sad:
It wearies me; you say it wearies you;
But how I caught it, found it, or came by it,
What stuff 'tis made of, whereof it is born,
I am to learn;
And such a want-wit sadness makes of me,
That I have much ado to know myself.

......

*Enter BASSANIO, LORENZO, and GRATIANO*

### SCENE 2: Belmont. A room in PORTIA'S house.

......
```

四、程序功能函数建议

根据解析流程，可将程序划分为不同函数。在把函数连接成一个大程序之前，请仔细测试每个函数。

1. *get_list_scene(file_path: str) -> list*: 读取路径名称对应的HTML目录文件，并解析出各幕`SCENE`网页文件的路径,并以 *list* 类型作为函数返回。
2. *get_scene_script(file_path: str) -> str*: 读取路径名称对应的HTML文件，解析出该幕`SCENE`中的剧本，并将按格式存储的剧本以 *str* 格式作为函数返回。
3. *write_script(file_name: str, content: str)*: 传入file_name，并以追加写入的方式打开，并将 *str* 写入.
4. *get_list_tags(file_path: str) -> set*: 读取路径名称对应的HTML文件，并解析出内含的所有标记名,注意不区分大小写,并以 *set* 类型存储作为函数返回。
5. *get_tagnum(file_path: str, tag_name: str) -> int*: 读取路径名称对应的HTML文件，并统计传入值 *str* 代表的标记数量,并将该值作为返回。

以上仅列出部分功能函数，供同学们参考，请同学们根据需要，自行修改或添加更多功能函数。