20201203 作业

题目说明：

文件 www.gov.cn.html 是中国政府网网站首页，存储为 utf-8 文本文件，可用各种文本编辑器打开浏览，也可以在浏览器中打开后观察其网页效果。利用所学的 Python 字符串函数，处理网页代码。本次作业完成，不得使用 import 引入其他扩展库。部分代码截图如下：

```
<li class="mainlevel"><a target="_blank" href="/xinwen/index.htm">新闻</a>
  <ul class="sub_nav_033" style="display: none;">
  <li><a href="/xinwen/yaowen.htm" target="_blank">要闻</a></li>
  <li><a href="/xinwen/zhuanti/index.htm" target="_blank">专题</a></li>
  <li><a href="/xinwen/lianbo/index.htm" target="_blank">政务联播</a></li>
  <li><a href="/xinwen/fabu/index.htm" target="_blank">新闻发布</a></li>
  <li><a href="/xinwen/renmian/index.htm" target="_blank">人事</a></li>
  <li><a href="/xinwen/tuku/index.htm" target="_blank">图片</a></li>
  <li><a href="/xinwen/gundong.htm" target="_blank">滚动</a></li>
  </ul>
</li>
```

作业1：输出该文件中所有中文字符到文件“学号_hz.txt”。仅统计 Unicode 编码界于 4E00-9FA5（十六进制）的中文字符；在连续中文字符后，如果遇到非界于 4E00-9FA5 的字符，则在其后增加一个\n（即换行）。如上图顶部的“新闻”之后是<，则在“新闻”之后输出一个\n；

作业2：网页中的图片都是以<img 开始，到最临近的>结束，如：<img src="/govweb/xhtml/2016gov/images/public/201712_video1.png">。网页代码对网页元素的大小写不敏感，如<img 或<Img 都是合法的表达，本题仅考虑<img 情况。输出所有图

片的数据，即<img 开始到最近邻的>结束，每个 img 占一行，输出文件名为：学号_img.txt；

作业3：编写一个名为 findPair(strSrc, First, Last)，其 strSrc 表示将被处理的字符串，First 参数开始字符串，Last 是结束字符串，查找字符串中开始为 First 和结束为 Last 的位置，返回值为开始字符串的位置编号和结束字符串的位置编号。如：输入字符串 strSrc 为：abc<link rel="canonical" href="https://www.gov.cn/index.htm"/>XYZ，First 参数为<link，Last 参数为>，则返回值为(3,61)。如果 First 不存在与 strSrc 中，则返回值为(-1,-1)，如 Last 不存在与 srtSrc 中，则返回值为(X, -1)其中 X 为 First 第一次出现的位置，如果 First 存在，且在 First 之后存在 Last，则返回出现位置；

作业4：利用已经定义的 findPair()函数，按一下步骤清除网页中指定内容：

1. 清除网页中以<!—开始以-->结束的内容；
2. 清除网页中以<style 开始以</style>结束的内容；
3. 清除网页中以<script 开始以</script>结束的内容；
4. 清除网页中以<开始以>结束的内容。

第 4 项清除必须在最后，前三项顺序可以根据需要调整。将清除后的网页内容存入文件学号_html.txt 之中。

计分规则：

- 1 个正确，得 40 分；

- 2 个正确，得 70 分
- 3 个正确，得 90 分
- 4 个正确，得 100 分