Linux 开发环境及应用实验报告

实验一：正则表达式应用

付容天

学号 2020211616

班级 2020211310

计算机学院（国家示范性软件学院）

2023 年 3 月 22 日

1 实验准备

在这部分实验中，我首先找到目标网站清新空气-北京空气质量监控，查看网页源代码，可以发现我们需要的数据存放在 tr 和 td 标签中，如下图所示：

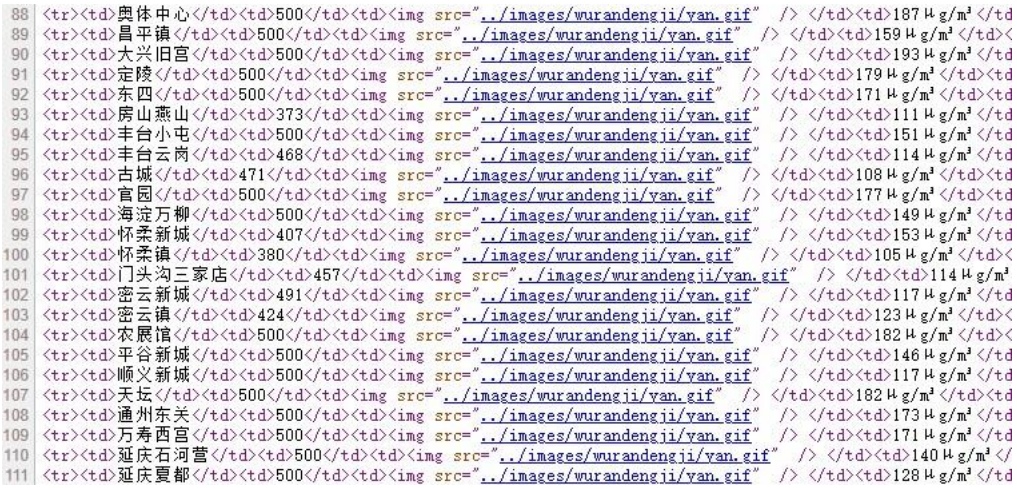

图 1：网站 HTML 源代码分析

我们可以使用下面的代码将网页的 HTML 源码下载到服务器，并通过 cat 命令查看到如图 2 所示的网页内容：

```
wget http://www.86pm25.com/city/beijing.html
```

```
cl616@Ubuntu-bupt:~$ ls
beijing.html
cl616@Ubuntu-bupt:~$ cat beijing.html
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.or
g/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
 <title>北京PM2.5实时查询和北京空气质量指数(AQI)--PM2.5查询</title>
<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<meta name="description" content="北京PM2.5实时数据查询及北京空气质量污染指数(AQ
I)查询" />
<meta name="keywords" content="北京PM2.5,北京空气质量指数(AQI),北京空气污染指数"
 />
<link href="../css/main.css" type="text/css" rel="stylesheet" />
<link href="../css/mainaqi.css" type="text/css" rel="stylesheet" />

<script language="JavaScript" src="../js/jquery.min.js" type="text/javascript"><
/script>
<script language="JavaScript" src="../js/highcharts.js" type="text/javascript"><
/script>
</head>
<script type="text/javascript">    (function() {
```

图 2：beijing.html 文件内容

2 实验步骤

2.1 标签替换

为了得到标签 tr 和 td 中的所需信息，我们首先需要将标签进行替换，使用下面的命令将“<”开头、中间不是“<”或“>”，并且后接任意长度字符串然后以“>”结尾的字符串（即 HTML 中的标签）替换成空格，替换前后对比如图 3 和图 4 所示。

```
cat beijing.html | sed ‘s/<[^<>]*>/ /g’
```

图 3：标签替换前

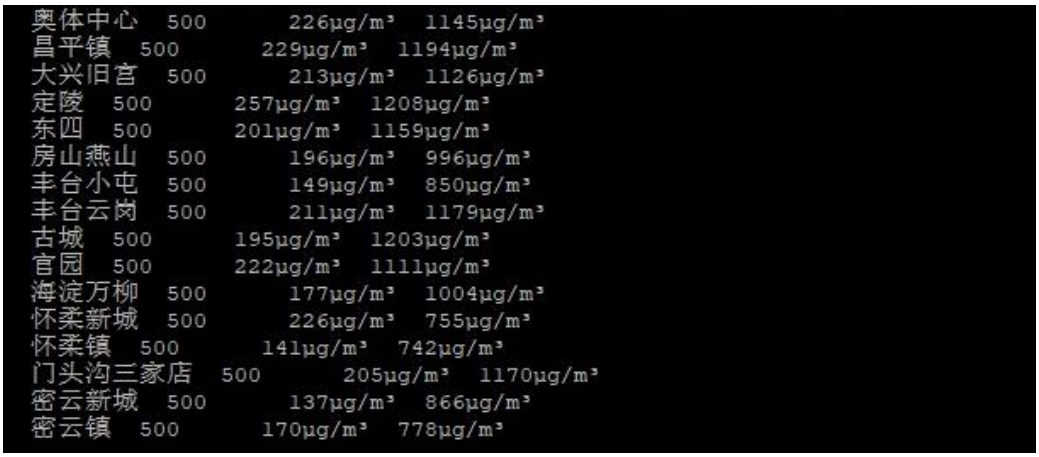

图 4：标签替换后

2.2 日期处理

然后处理标题日期，使用如下所示的指令，提取出日期如下图 5 所示：

```
cat beijing.html | sed -e ’s/<[^<>]*>/ /g’ -e ‘s/[年月]/-/g’
```

图 5：日期处理结果

2.3 处理 PM2.5 数据并打印

现在我们需要将含有时间和包含 PM2.5 的数据筛选出来并打印到屏幕上，我们可以编写 awk 文件，将含有文本“更新”的一行中的第一个数据和第二个数据分别赋给变量 date 和 time，将含有“m^3”的一行取第一个数据“监测地点”和第三个数据“PM2.5 指数”，使用 printf 语句打印数据。awk 文件内容与执行结果分别如图 6 和图 7 所示：

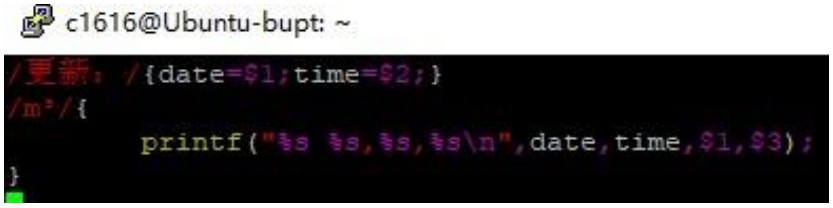

图 6：awk 文件内容

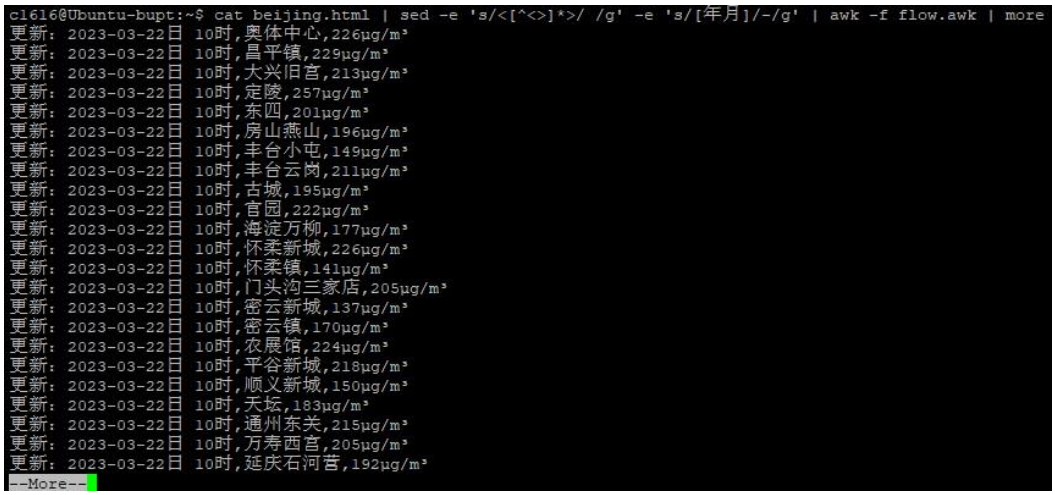

图 7：执行结果

2.4 整理数据格式

最后我根据实验要求，将“更新：”替换成空，将“时”替换成“00:00”，将“μg.m³”替换成空，将“日”替换成空，完整命令如下所示，最终得到结果如图 8 所示（与实验要求一致）：

```
cat beijing.html | sed -e 's/<[^<>]*>/ /g' -e 's/[年月]/-/g'| awk -f flow.awk | sed -e 's/μg.m³/ /g' -e 's/[更新：日]//g' -e 's/时/:00:00/g' |more
```

```
cl6l6@Ubuntu-bupt:~$ cat beijing.html |
> sed -e 's/<[^<>]*>/ /g' -e 's/[年月]/-/g' |
> awk -f flow.awk |
>  sed -e 's/μg.m³/ /g' -e 's/[更新：日]//g' -e 's/时/:00:00/g' | more
2023-03-22 10:00:00,奥体中心,226
2023-03-22 10:00:00,昌平镇,229
2023-03-22 10:00:00,大兴旧宫,213
2023-03-22 10:00:00,定陵,257
2023-03-22 10:00:00,东四,201
2023-03-22 10:00:00,房山燕山,196
2023-03-22 10:00:00,丰台小屯,149
2023-03-22 10:00:00,丰台云岗,211
2023-03-22 10:00:00,古城,195
2023-03-22 10:00:00,官园,222
2023-03-22 10:00:00,海淀万柳,177
2023-03-22 10:00:00,怀柔城,226
2023-03-22 10:00:00,怀柔镇,141
2023-03-22 10:00:00,门头沟三家店,205
2023-03-22 10:00:00,密云城,137
2023-03-22 10:00:00,密云镇,170
2023-03-22 10:00:00,农展馆,224
2023-03-22 10:00:00,平谷城,218
2023-03-22 10:00:00,顺义城,150
2023-03-22 10:00:00,天坛,183
2023-03-22 10:00:00,通州东关,215
2023-03-22 10:00:00,万寿西宫,205
2023-03-22 10:00:00,延庆石河营,192
2023-03-22 10:00:00,延庆夏都,176
```

图 8: 最终实验结果

3 实验问题与实验总结

在本次实验中我也遇到了一些问题，现记录如下：

（1） 网址选择问题：许多网站上都有空气质量的数据，但是通常具有很复杂的 HTML 结构。经过寻找的对比，我选择了上面的网站；

（2） 正则表达式不熟练：一开始在应用正则表达式结合 sed 指令对数据进行处理时，总是需要回看帮助手册，并且总是有格式问题。经过一段时间的练习，解决了这个问题；

（3） awk 理解与应用不熟练：awk 脚本可以用来执行用户预先设计好的一些指令，一开始我对 awk 脚本的机制理解不够透彻，导致在一些细节上出问题。后来我复习了理论知识，结合一些网上的资料，解决了这个问题。

总的来说，在本次实验中，我通过收集和分析北京 PM2.5 网页的 HTML 数据，练习了正则匹配表达式相关知识，还练习了 sed 语句与 awk 脚本文件的使用，对于在 Linux 系统上使用正则表达式处理数据有了更深入的理解，本次实验我收获满满。