



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](http://www.86pm25.com/city/beijing.html)，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：

```
88 <tr><td>奥体中心</td><td>500</td><td></td><td>187 μg/m³</td>
89 <tr><td>昌平镇</td><td>500</td><td></td><td>159 μg/m³</td><
90 <tr><td>大兴旧宫</td><td>500</td><td></td><td>193 μg/m³</td>
91 <tr><td>定陵</td><td>500</td><td></td><td>179 μg/m³</td><td>
92 <tr><td>东四</td><td>500</td><td></td><td>171 μg/m³</td><td>
93 <tr><td>房山燕山</td><td>373</td><td></td><td>111 μg/m³</td><td>
94 <tr><td>丰台小屯</td><td>500</td><td></td><td>151 μg/m³</td>
95 <tr><td>丰台云岗</td><td>468</td><td></td><td>114 μg/m³</td>
96 <tr><td>古城</td><td>471</td><td></td><td>108 μg/m³</td><td>
97 <tr><td>官园</td><td>500</td><td></td><td>177 μg/m³</td><td>
98 <tr><td>海淀万柳</td><td>500</td><td></td><td>149 μg/m³</td>
99 <tr><td>怀柔新城</td><td>407</td><td></td><td>153 μg/m³</td>
100 <tr><td>怀柔镇</td><td>380</td><td></td><td>105 μg/m³</td><td>
101 <tr><td>门头沟三家店</td><td>457</td><td></td><td>114 μg/m³</td>
102 <tr><td>密云新城</td><td>491</td><td></td><td>117 μg/m³</td>
103 <tr><td>密云镇</td><td>424</td><td></td><td>123 μg/m³</td><td>
104 <tr><td>农展馆</td><td>500</td><td></td><td>182 μg/m³</td><td>
105 <tr><td>平谷新城</td><td>500</td><td></td><td>146 μg/m³</td>
106 <tr><td>顺义新城</td><td>500</td><td></td><td>117 μg/m³</td>
107 <tr><td>天坛</td><td>500</td><td></td><td>182 μg/m³</td><td>
108 <tr><td>通州东关</td><td>500</td><td></td><td>173 μg/m³</td>
109 <tr><td>万寿西宫</td><td>500</td><td></td><td>171 μg/m³</td>
110 <tr><td>延庆石河营</td><td>500</td><td></td><td>140 μg/m³</td>
111 <tr><td>延庆夏都</td><td>500</td><td></td><td>128 μg/m³</td>
```

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.org/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(AQI)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'
```

```
<tr><td>奥体中心</td><td>500</td><td></td><td>226µg/m³</td><td>1145µg/m³</td></tr>
<tr><td>昌平镇</td><td>500</td><td></td><td>229µg/m³</td><td>1194µg/m³</td></tr>
<tr><td>大兴旧宫</td><td>500</td><td></td><td>213µg/m³</td><td>1126µg/m³</td></tr>
<tr><td>定陵</td><td>500</td><td></td><td>257µg/m³</td><td>1208µg/m³</td></tr>
<tr><td>东四</td><td>500</td><td></td></tr>
```

图 3: 标签替换前

```
奥体中心 500      226µg/m³  1145µg/m³
昌平镇 500      229µg/m³  1194µg/m³
大兴旧宫 500      213µg/m³  1126µg/m³
定陵 500      257µg/m³  1208µg/m³
东四 500      201µg/m³  1159µg/m³
房山燕山 500      196µg/m³  996µg/m³
丰台小屯 500      149µg/m³  850µg/m³
丰台云岗 500      211µg/m³  1179µg/m³
古城 500      195µg/m³  1203µg/m³
官园 500      222µg/m³  1111µg/m³
海淀万柳 500      177µg/m³  1004µg/m³
怀柔新城 500      226µg/m³  755µg/m³
怀柔镇 500      141µg/m³  742µg/m³
门头沟三家店 500      205µg/m³  1170µg/m³
密云新城 500      137µg/m³  866µg/m³
密云镇 500      170µg/m³  778µg/m³
```

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'
```

```
北京实时空气质量指数
更新: 2023-03-22日 10时
```

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“m³”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：

```
c1616@Ubuntu-bupt: ~  
/更新:/{date=$1;time=$2;}  
/m³/{  
    printf("%s %s,%s,%s\n",date,time,$1,$3);  
}
```

图 6: awk 文件内容

```
c1616@Ubuntu-bupt:~$ cat beijing.html | sed -e 's/<[^<>]*>/ /g' -e 's/[年月]/-/g' | awk -f flow.awk | more  
更新: 2023-03-22日 10时,奥体中心,226µg/m³  
更新: 2023-03-22日 10时,昌平镇,229µg/m³  
更新: 2023-03-22日 10时,大兴旧宫,213µg/m³  
更新: 2023-03-22日 10时,定陵,257µg/m³  
更新: 2023-03-22日 10时,东四,201µg/m³  
更新: 2023-03-22日 10时,房山燕山,196µg/m³  
更新: 2023-03-22日 10时,丰台小屯,149µg/m³  
更新: 2023-03-22日 10时,丰台云岗,211µg/m³  
更新: 2023-03-22日 10时,古城,195µg/m³  
更新: 2023-03-22日 10时,官园,222µg/m³  
更新: 2023-03-22日 10时,海淀万柳,177µg/m³  
更新: 2023-03-22日 10时,怀柔新城,226µg/m³  
更新: 2023-03-22日 10时,怀柔镇,141µg/m³  
更新: 2023-03-22日 10时,门头沟三家店,205µg/m³  
更新: 2023-03-22日 10时,密云新城,137µg/m³  
更新: 2023-03-22日 10时,密云镇,170µg/m³  
更新: 2023-03-22日 10时,农展馆,224µg/m³  
更新: 2023-03-22日 10时,平谷新城,218µg/m³  
更新: 2023-03-22日 10时,顺义新城,150µg/m³  
更新: 2023-03-22日 10时,天坛,183µg/m³  
更新: 2023-03-22日 10时,通州东关,215µg/m³  
更新: 2023-03-22日 10时,万寿西宫,205µg/m³  
更新: 2023-03-22日 10时,延庆石河营,192µg/m³  
--More--
```

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
```



```
c1616@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。