编程作业 1--正则表达式应用

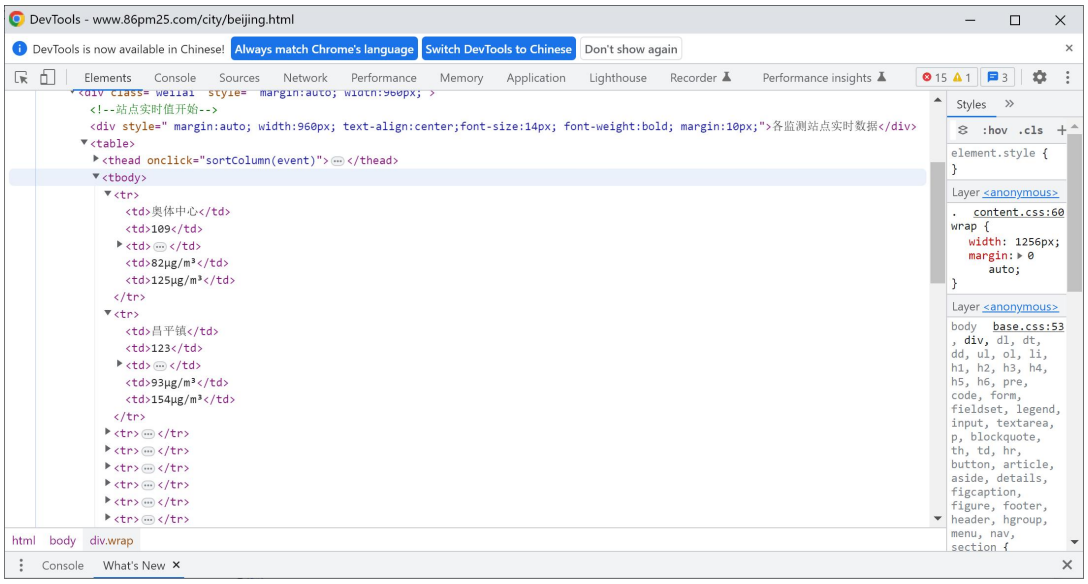
一、实验步骤

1、从因特网上搜索相关 Web 网页，找到含有北京 PM2.5 的网站如下：
http://www.86pm25.com/city/beijing.html，查看网页如下：

各监测站点实时数据

监测站点	AQI	污染等级	PM2.5浓度	PM10浓度
奥体中心	109	轻度污染	82μg/m³	125μg/m³
昌平镇	123	轻度污染	93μg/m³	154μg/m³
大兴旧宫	114	轻度污染	86μg/m³	122μg/m³
定陵	74	良	54μg/m³	96μg/m³
东四	108	轻度污染	81μg/m³	128μg/m³
房山燕山	129	轻度污染	98μg/m³	140μg/m³
丰台小屯	110	轻度污染	83μg/m³	131μg/m³
丰台云岗	119	轻度污染	90μg/m³	136μg/m³
古城	130	轻度污染	99μg/m³	148μg/m³
官园	122	轻度污染	92μg/m³	125μg/m³
海淀万柳	122	轻度污染	92μg/m³	144μg/m³
怀柔新城	72	良	52μg/m³	87μg/m³
怀柔镇	78	良	57μg/m³	95μg/m³
门头沟三家店	135	轻度污染	103μg/m³	166μg/m³
密云新城	70	良	50μg/m³	90μg/m³
密云镇	68	良	49μg/m³	81μg/m³
农展馆	103	轻度污染	77μg/m³	127μg/m³
平谷新城	77	良	56μg/m³	95μg/m³
顺义新城	79	良	58μg/m³	92μg/m³
天坛	110	轻度污染	83μg/m³	115μg/m³
通州东关	108	轻度污染	81μg/m³	127μg/m³
万寿西宫	120	轻度污染	91μg/m³	130μg/m³
延庆石河营	79	良	58μg/m³	99μg/m³
延庆夏都	79	良	58μg/m³	108μg/m³

从控制台可以看到我们需要的地点和 PM2.5 数据在 tr 和 td 标签中，



2、使用命令 wegt 下载网页：

```
lhfhl@ubuntu:~$ wget http://www.86pm25.com/city/beijing.html
--2023-03-19 05:09:01--  http://www.86pm25.com/city/beijing.html
Resolving www.86pm25.com (www.86pm25.com)... 120.27.42.216
Connecting to www.86pm25.com (www.86pm25.com)|120.27.42.216|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 22209 (22K) [text/html]
Saving to: 'beijing.html'

beijing.html                        100%[===================================================================================>]  21.69K  --.-KB/s    in 0.03s

2023-03-19 05:09:02 (867 KB/s) - 'beijing.html' saved [22209/22209]

lhfhl@ubuntu:~$
```

3、使用 cat 命令查看下载的文件：

```
lhfhl@ubuntu:~$ 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() {
        var url = location.href;

        if ((url.indexOf('www.86pm25.com') != -1) && navigator.userAgent.match(/(iPhone|iPod|Android|ios|iPad)/i)) {
            var url = location.href;
            url = url.replace("www.", "m.");
            location.href = url;
        }
    })();</script>
<body>
        <div class="wrap">

                <div class="layout">

                        <div class="content">
                        <div class="aqi-site" style="height:100px; border:0; padding-top:10px;">
                        <div style=" float:left;"><a href="/" ><img src="../images/logo.jpg" border=0 /></a></div>
                        <div style=" float:right;"><a href="/" ><img src="../images/dingbu.jpg" border=0 /></a></div>
```

4、查看该文件，发现 标签和 tr 以及 td 标签里的数据是我们需要的：

```
        var idx = weatherinfovar.weatherinfo.idx; var aqiLevel = weatherinfovar.weatherinfo.aqiLevel;
    }
    else { var qualityStr = "轻度污染"; var aqiPercent = "25.75"; var idx = "103"; var aqiLevel = "3"; }
 </script>
<script type="text/javascript">    document.write("<div class='aqi aqi-lv" + aqiLevel + "'>");</script>
<div  style="text-align:center">
<h3 style="font-size:14px; font-weight:bold">北京实时空气质量指数</h3></div>
<div class="remark">更新：2023年03月19日 19时</div>
<div  style="background:url(http://www.86pm25.com/images/aqi-ruler-140616.png) no-repeat;margin:0 auto 8px;width:330px;height:33px;background-position:-1px 0">
<div class="panel"><script type="text/javascript">                    document.write("<b class='cur' style='left:" + aqiPercent + "%;'>" + idx + "</b>");</script></div>
```

```
<tr><td>奥体中心</td><td>109</td><td><img src="../images/wurandengji/qing.gif"  /> </td><td>82μg/m³</td><td>125μg/m³</td></tr>
<tr><td>昌平镇</td><td>123</td><td><img src="../images/wurandengji/qing.gif"  /> </td><td>93μg/m³</td><td>154μg/m³</td></tr>
<tr><td>大兴旧宫</td><td>114</td><td><img src="../images/wurandengji/qing.gif"  /> </td><td>86μg/m³</td><td>122μg/m³</td></tr>
<tr><td>定陵</td><td>74</td><td><img src="../images/wurandengji/liang.gif"  /> </td><td>54μg/m³</td><td>96μg/m³</td></tr>
<tr><td>东四</td><td>108</td><td><img src="../images/wurandengji/qing.gif"  /> </td><td>81μg/m³</td><td>128μg/m³</td></tr>
<tr><td>房山燕山</td><td>129</td><td><img src="../images/wurandengji/qing.gif"  /> </td><td>98μg/m³</td><td>140μg/m³</td></tr>
```

5、使用命令 cat beijing.html | sed -e 's/<[^<>]*>/ /g'将标签替换成空格

```
lhfhl@ubuntu:~$ cat beijing.html | sed -e 's/<[^<>]*>/ /g'
```

```
各监测站点实时数据
            监测站点  AQI     污染等级  PM2.5浓度  PM10浓度
奥体中心  109      82μg/m³  125μg/m³
昌平镇  123      93μg/m³  154μg/m³
大兴旧宫  114      86μg/m³  122μg/m³
定陵  74      54μg/m³  96μg/m³
东四  108      81μg/m³  128μg/m³
房山燕山  129      98μg/m³  140μg/m³
丰台小屯  110      83μg/m³  131μg/m³
丰台云岗  119      90μg/m³  136μg/m³
古城  130      99μg/m³  148μg/m³
官园  122      92μg/m³  125μg/m³
海淀万柳  122      92μg/m³  144μg/m³
怀柔新城  72      52μg/m³  87μg/m³
怀柔镇  78      57μg/m³  95μg/m³
门头沟三家店  135      103μg/m³  166μg/m³
密云新城  70      50μg/m³  90μg/m³
密云镇  68      49μg/m³  81μg/m³
农展馆  103      77μg/m³  127μg/m³
平谷新城  77      56μg/m³  95μg/m³
顺义新城  79      58μg/m³  92μg/m³
天坛  110      83μg/m³  115μg/m³
通州东关  108      81μg/m³  127μg/m³
万寿西宫  120      91μg/m³  130μg/m³
延庆石河营  79      58μg/m³  99μg/m³
延庆夏都  79      58μg/m³  108μg/m³
```

6、上述处理后的日期数据如下：

```
北京实时空气质量指数
更新： 2023年03月19日  19时

                        document.write(" " + idx + " ");
```

使用如下命令：

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

将“年”和“月”替换成所需格式的“-”，如下：

```
北京实时空气质量指数
更新： 2023-03-19日  19时

                        document.wr
```

7、为了筛选出时间和地点以及 PM2.5 的数据，编写 flow.awk 文件如下，其功能是将含有“更新：”的一行中的第一个和第二个数据分别赋值给 date 和 time，将含有 m³ 的一行中第一个数据地点和第三个数据 PM2.5，使用 printf 语句打印：

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

使用如下命令：

```
cat beijing.html | sed -e 's/<[^<>]*>/ /g' -e 's/[年月]/-/g'| awk -f flow.awk | more
```

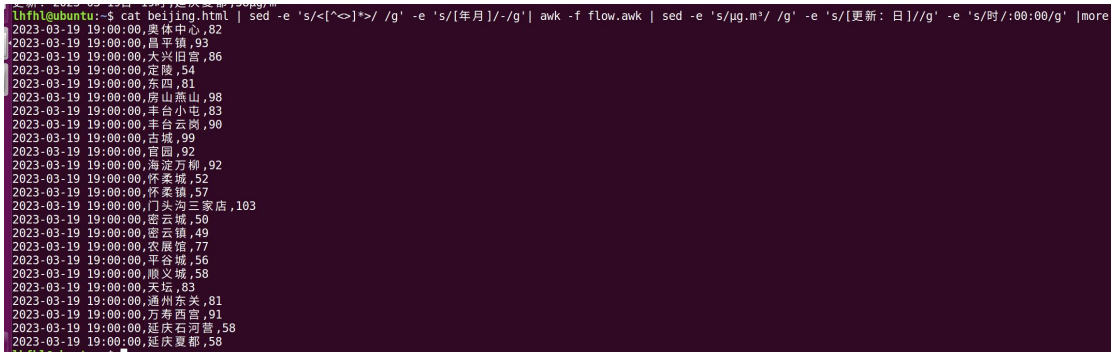
可得如下数据：

```
lhfhl@ubuntu:~$ cat beijing.html | sed -e 's/<[^<>]*>/ /g' -e 's/[年月]/-/g'| awk -f flow.awk | more
更新： 2023-03-19日  19时,奥体中心,82μg/m³
更新： 2023-03-19日  19时,昌平镇,93μg/m³
更新： 2023-03-19日  19时,大兴旧宫,86μg/m³
更新： 2023-03-19日  19时,定陵,54μg/m³
更新： 2023-03-19日  19时,东四,81μg/m³
更新： 2023-03-19日  19时,房山燕山,98μg/m³
更新： 2023-03-19日  19时,丰台小屯,83μg/m³
更新： 2023-03-19日  19时,丰台云岗,90μg/m³
更新： 2023-03-19日  19时,古城,99μg/m³
更新： 2023-03-19日  19时,官园,92μg/m³
更新： 2023-03-19日  19时,海淀万柳,92μg/m³
更新： 2023-03-19日  19时,怀柔新城,52μg/m³
更新： 2023-03-19日  19时,怀柔镇,57μg/m³
更新： 2023-03-19日  19时,门头沟三家店,103μg/m³
更新： 2023-03-19日  19时,密云新城,50μg/m³
更新： 2023-03-19日  19时,密云镇,49μg/m³
更新： 2023-03-19日  19时,农展馆,77μg/m³
更新： 2023-03-19日  19时,平谷新城,56μg/m³
更新： 2023-03-19日  19时,顺义新城,58μg/m³
更新： 2023-03-19日  19时,天坛,83μg/m³
更新： 2023-03-19日  19时,通州东关,81μg/m³
更新： 2023-03-19日  19时,万寿西宫,91μg/m³
更新： 2023-03-19日  19时,延庆石河营,58μg/m³
更新： 2023-03-19日  19时,延庆夏都,58μg/m³
```

8、为了满足格式要求，需要将上图中“更新：”替换成空，将“时”替换成“:00:00”，将“ug/m³”替换成空，将“日”替换成空，则，最终的命令如下：

```
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
```

执行该命令后，可得满足条件的数据，如下图：



二、实验总结

本次实验中的多种命令使用起来并不熟悉，用的时候感觉很难，后来通过查找资料和观看视频逐渐熟练了起来，经历了这个过程，我对这些命令有了更深的体会，同时通过本次实验也让我对正则表达式有了更好的掌握，收获了很多。