

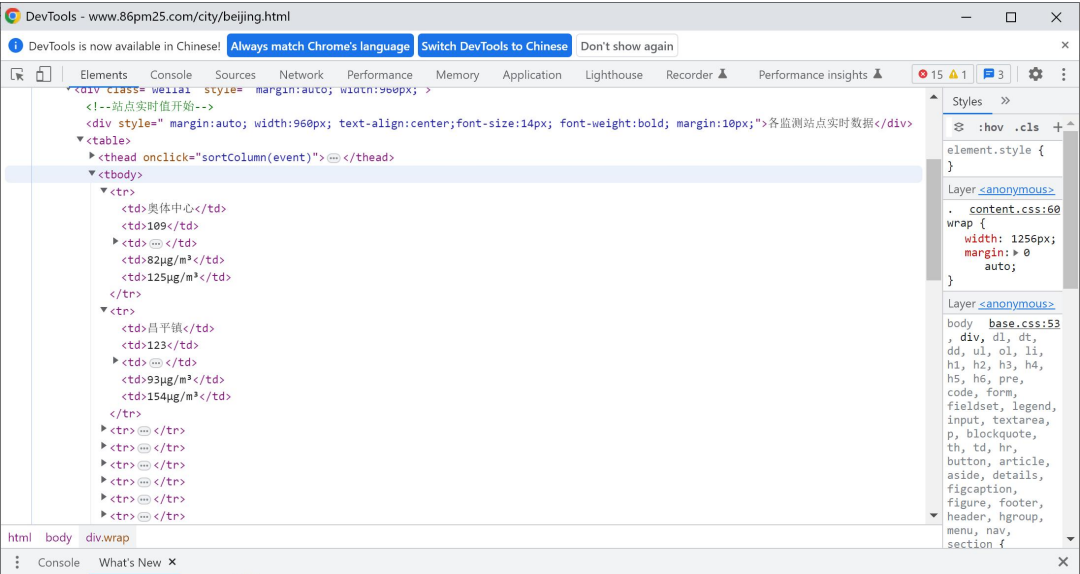
编程作业 1--正则表达式应用

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
丰台南岗	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 `wget` 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:

```

lhfh1@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="/" ></a></div>
          <div style="float:right;"><a href="/" ></a></div>

```

4、查看该文件，发现<div class="remark">标签和 tr 以及 td 标签里的数据是我们需要的：

```
var qualityStr = weatherinfovar.weatherinfo.quality; var aqiPercent = aqiPercent+(weatherinfovar.aqiLevel-idx);  
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-"+ 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 class="img"><img alt="AQI ruler showing current AQI value and corresponding color range." data-bbox="186 828 812 868"/>  
<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/wurandengjii/qing.gif'" /></td><td>82ug/m³</td><td>125ug/m³</td></tr>  
<tr><td>昌平镇</td><td>123</td><td><img src='../images/wurandengjii/qing.gif'" /></td><td>93ug/m³</td><td>154ug/m³</td></tr>  
<tr><td>大兴旧宫</td><td>114</td><td><img src='../images/wurandengjii/qing.gif'" /></td><td>86ug/m³</td><td>122ug/m³</td></tr>  
<tr><td>定兴</td><td>74</td><td><img src='../images/wurandengjii/liang.gif'" /></td><td>54ug/m³</td><td>96ug/m³</td></tr>  
<tr><td>东四</td><td>108</td><td><img src='../images/wurandengjii/qing.gif'" /></td><td>81ug/m³</td><td>128ug/m³</td></tr>  
<tr><td>房山燕山</td><td>129</td><td><img src='../images/wurandengjii/qing.gif'" /></td><td>98ug/m³</td><td>140ug/m³</td></tr>
```

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

```
lhfh1@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/m ³	125μg/m ³
昌平镇	123	93μg/m ³	154μg/m ³
大兴旧宫	114	86μg/m ³	122μg/m ³
定陵	74	54μg/m ³	96μg/m ³
东四	108	81μg/m ³	128μg/m ³
房山燕山	129	98μg/m ³	140μg/m ³
丰台小屯	110	83μg/m ³	131μg/m ³
丰台云岗	119	90μg/m ³	136μg/m ³
古城	130	99μg/m ³	148μg/m ³
官园	122	92μg/m ³	125μg/m ³
海淀万柳	122	92μg/m ³	144μg/m ³
怀柔新城	72	52μg/m ³	87μg/m ³
怀柔镇	78	57μg/m ³	95μg/m ³
门头沟三家店	135	103μg/m ³	166μg/m ³
密云新城	70	50μg/m ³	90μg/m ³
密云镇	68	49μg/m ³	81μg/m ³
农展馆	103	77μg/m ³	127μg/m ³
平谷新城	77	56μg/m ³	95μg/m ³
顺义新城	79	58μg/m ³	92μg/m ³
天坛	110	83μg/m ³	115μg/m ³
通州东关	108	81μg/m ³	127μg/m ³
万寿西宫	120	91μg/m ³	130μg/m ³
延庆石河营	79	58μg/m ³	99μg/m ³
延庆夏都	79	58μg/m ³	108μg/m ³

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.wri
```

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：

```
awk (~) - 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：

```
lhfh1@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：

```
l@h1cubuntu:~$ cat beijing.html | sed -e 's/[^<>]*>/ /g' -e 's/[年月]/-/g' | awk -f flow.awk | sed -e 's/ug.m*/ /g' -e 's/[更新：日]/ /g' -e 's/时/:00:00/g' | more
2023-03-19 19:00:00 奥体中心,82
2023-03-19 19:00:00 昌平镇,93
2023-03-19 19:00:00 大兴旧宫,86
2023-03-19 19:00:00 定陵,54
2023-03-19 19:00:00 东四,81
2023-03-19 19:00:00 房山燕山,98
2023-03-19 19:00:00 丰台小屯,83
2023-03-19 19:00:00 丰台云岗,98
2023-03-19 19:00:00 古城,99
2023-03-19 19:00:00 官园,92
2023-03-19 19:00:00 海淀万柳,92
2023-03-19 19:00:00 怀柔城,52
2023-03-19 19:00:00 怀柔镇,57
2023-03-19 19:00:00 门头沟三家店,103
2023-03-19 19:00:00 密云城,50
2023-03-19 19:00:00 密云镇,49
2023-03-19 19:00:00 农展馆,77
2023-03-19 19:00:00 平谷城,56
2023-03-19 19:00:00 顺义城,58
2023-03-19 19:00:00 天坛,85
2023-03-19 19:00:00 通州东关,81
2023-03-19 19:00:00 万寿西宫,91
2023-03-19 19:00:00 延庆石河营,58
2023-03-19 19:00:00 延庆夏都,58
```

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。