Linux作业一

- 学号：2018211582
- 班级：2018211314
- 姓名：李志毅

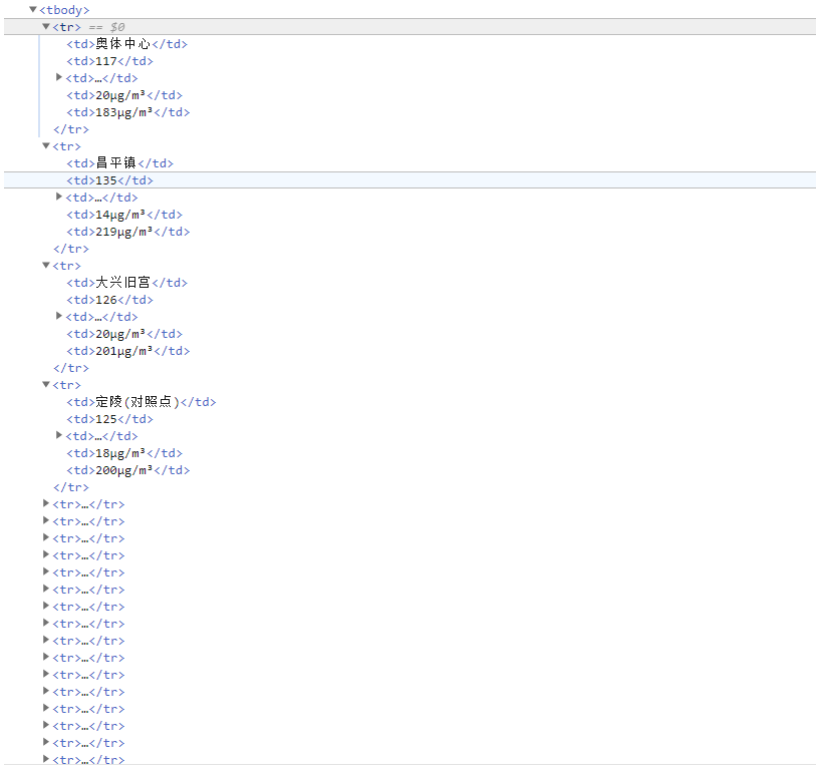
一、实验准备

搜索含有北京PM2.5的网站，找到网址为http://www.86pm25.com/city/beijing.html，网址含有的元素如下：

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	117	轻度污染	20μg/m³	183μg/m³
昌平镇	135	轻度污染	14μg/m³	219μg/m³
大兴旧宫	126	轻度污染	20μg/m³	201μg/m³
定陵(对照点)	125	轻度污染	18μg/m³	200μg/m³
东四	141	轻度污染	16μg/m³	232μg/m³
房山燕山	97	良	16μg/m³	144μg/m³
丰台小屯	135	轻度污染	25μg/m³	219μg/m³
丰台云岗	122	轻度污染	17μg/m³	194μg/m³
古城	125	轻度污染	19μg/m³	199μg/m³
官园	113	轻度污染	8μg/m³	176μg/m³
海淀万柳	124	轻度污染	21μg/m³	198μg/m³
怀柔新城	161	中度污染	17μg/m³	271μg/m³
门头沟三家店	144	轻度污染	21μg/m³	238μg/m³
密云新城	170	中度污染	18μg/m³	289μg/m³
密云镇	158	中度污染	18μg/m³	265μg/m³
农展馆	124	轻度污染	12μg/m³	197μg/m³
平谷新城	166	中度污染	20μg/m³	281μg/m³
顺义新城	167	中度污染	20μg/m³	283μg/m³
天坛	116	轻度污染	20μg/m³	182μg/m³
通州东关	163	中度污染	21μg/m³	275μg/m³
万寿西宫	112	轻度污染	19μg/m³	173μg/m³
延庆石河营	138	轻度污染	24μg/m³	225μg/m³
延庆夏都	147	轻度污染	18μg/m³	244μg/m³

使用F12观察网址HTML5结构，发现我们需要的数据存放在tr和td标签中，含有监测地点，PM2.5指数，在网站上部分标签中含有需要的时间数据



使用命令`wegt`将网页html5下载到服务器：

```
a582@Ubuntu:~$ wget http://www.86pm25.com/city/beijing.html
--2021-03-26 18:19:29--  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: 22055 (22K) [text/html]
Saving to: ‘beijing.html.1’

beijing.html.1                                      100%[=============================

2021-03-26 18:19:29 (1.21 MB/s) - ‘beijing.html.1’ saved [22055/22055]
```

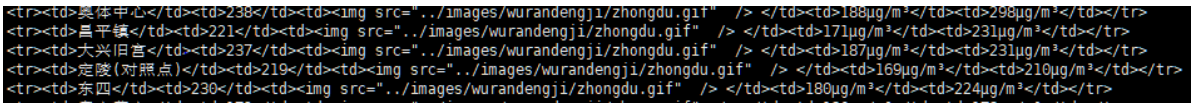
使用命令`cat`查看网页内容：

```
<div class="weilai" style=" margin:auto; width:960px;">
       <!--站点实时值开始-->

<div  style=" margin:auto; width:960px; text-align:center;font-size:14px; font-weight:bold; margin:10px;">各监测站点实时数据</div>
<table >        <thead onclick="sortColumn(event)"><tr style=" background-color:#EBEFF7"><th width="15%"  >监测站点</th><th width="10%" >AQI</th><t
%"  >PM10浓度</th> </tr></thead>
<tr><td>奥体中心</td><td>238</td><td><img src="../images/wurandengji/zhongdu.gif"  /> </td><td>188μg/m³</td><td>298μg/m³</td></tr>
<tr><td>昌平镇</td><td>221</td><td><img src="../images/wurandengji/zhongdu.gif"  /> </td><td>171μg/m³</td><td>231μg/m³</td></tr>
<tr><td>大兴旧宫</td><td>237</td><td><img src="../images/wurandengji/zhongdu.gif"  /> </td><td>187μg/m³</td><td>231μg/m³</td></tr>
<tr><td>定陵(对照点)</td><td>219</td><td><img src="../images/wurandengji/zhongdu.gif"  /> </td><td>169μg/m³</td><td>210μg/m³</td></tr>
<tr><td>东四</td><td>230</td><td><img src="../images/wurandengji/zhongdu.gif"  /> </td><td>180μg/m³</td><td>224μg/m³</td></tr>
<tr><td>房山燕山</td><td>172</td><td><img src="../images/wurandengji/zhong.gif"  /> </td><td>130μg/m³</td><td>173μg/m³</td></tr>
<tr><td>丰台小屯</td><td>209</td><td><img src="../images/wurandengji/zhongdu.gif"  /> </td><td>159μg/m³</td><td>249μg/m³</td></tr>
<tr><td>丰台云岗</td><td>170</td><td><img src="../images/wurandengji/zhong.gif"  /> </td><td>129μg/m³</td><td>230μg/m³</td></tr>
<tr><td>古城</td><td>199</td><td><img src="../images/wurandengji/zhong.gif"  /> </td><td>149μg/m³</td><td>227μg/m³</td></tr>
<tr><td>官园</td><td>216</td><td><img src="../images/wurandengji/zhongdu.gif"  /> </td><td>166μg/m³</td><td>237μg/m³</td></tr>
<tr><td>海淀万柳</td><td>217</td><td><img src="../images/wurandengji/zhongdu.gif"  /> </td><td>167μg/m³</td><td>248μg/m³</td></tr>
<tr><td>怀柔新城</td><td>195</td><td><img src="../images/wurandengji/zhong.gif"  /> </td><td>146μg/m³</td><td>177μg/m³</td></tr>
<tr><td>怀柔镇</td><td>203</td><td><img src="../images/wurandengji/zhongdu.gif"  /> </td><td>153μg/m³</td><td>172μg/m³</td></tr>
<tr><td>门头沟三家店</td><td>188</td><td><img src="../images/wurandengji/zhong.gif"  /> </td><td>141μg/m³</td><td>212μg/m³</td></tr>
<tr><td>密云新城</td><td>211</td><td><img src="../images/wurandengji/zhongdu.gif"  /> </td><td>161μg/m³</td><td>204μg/m³</td></tr>
<tr><td>密云镇</td><td>204</td><td><img src="../images/wurandengji/zhongdu.gif"  /> </td><td>154μg/m³</td><td>196μg/m³</td></tr>
<tr><td>农展馆</td><td>235</td><td><img src="../images/wurandengji/zhongdu.gif"  /> </td><td>185μg/m³</td><td>276μg/m³</td></tr>
<tr><td>平谷新城</td><td>130</td><td><img src="../images/wurandengji/qing.gif"  /> </td><td>99μg/m³</td><td>141μg/m³</td></tr>
<tr><td>顺义新城</td><td>210</td><td><img src="../images/wurandengji/zhongdu.gif"  /> </td><td>160μg/m³</td><td>226μg/m³</td></tr>
<tr><td>通州东关</td><td>223</td><td><img src="../images/wurandengji/zhongdu.gif"  /> </td><td>173μg/m³</td><td>225μg/m³</td></tr>
<tr><td>延庆石河营</td><td>150</td><td><img src="../images/wurandengji/qing.gif"  /> </td><td>115μg/m³</td><td>143μg/m³</td></tr>
<tr><td>延庆夏都</td><td>176</td><td><img src="../images/wurandengji/zhong.gif"  /> </td><td>133μg/m³</td><td>149μg/m³</td></tr></table>
       <!--站点实时值结束-->
       <!--站点实时值结束-->

       <!--广告开始-->
       <div  style=" margin:auto; width:960px; margin-top:10px; height:90px;">
       <script async src="https://pagead2.googlesyndication.com/pagead/js/adsbygoogle.js"></script>
<!-- 960x90 -->
```

二、实验步骤

1.替换标签

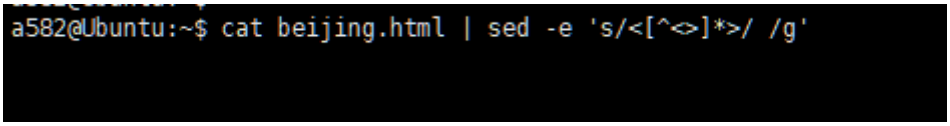
首先观察网站结构，查找我们需要的内容，监测地点和PM2.5的数据，在HTML5中包含在标签中，

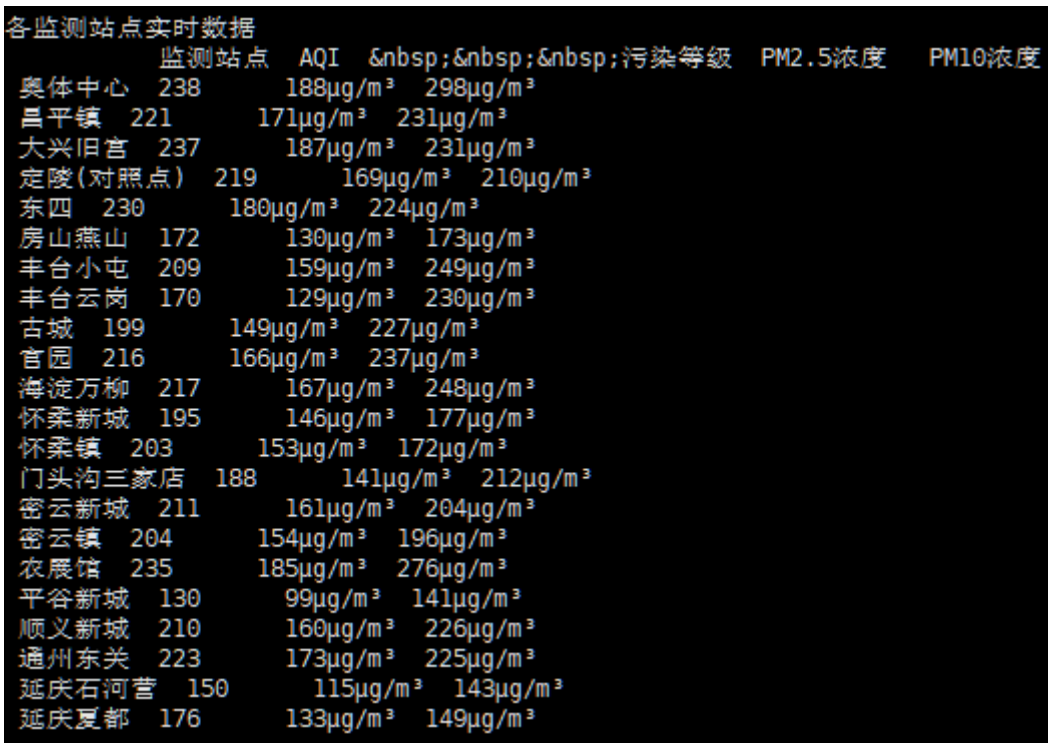


我们使用命令

```
sed -e 's/<[^<>]*>/ /g'
```

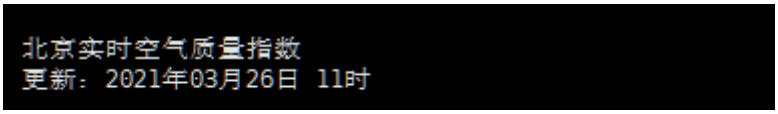
将以`<`开头，中间不是`<`或`>`，后接任意长度字符串并以`>`结尾的字符串，即html5中的标签，替换成空格，替换后截图如下：





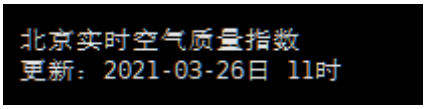
2.处理日期

之后处理标题日期，我们需要的日期在html5文件中出现在这一行：



首先我们将“年”和“月”替换成最后格式中的`"-"`

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



3.处理PM2.5数据并打印

之后我们要将含有时间的这一行和包含PM2.5数据的行筛选出来，我们可以编写`awk`文件`flow.awk`，将含有文本`"更新："`的一行中的第一个数据和第二个　数据分别赋值给变量date和time，将含有`m³`的一行中取第一个数据`监测地点`和第三个数据`PM2.5指数`，使用`printf`语句打印数据，`flow.awk`文件内容如下：

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

```
a582@Ubuntu:~$ cat beijing.html | sed -e 's/<[^<>]*>/ /g' -e 's/[年月]/-/g'| awk -f flow.awk | more
更新：2021-03-26日 11时,奥体中心,188μg/m³
更新：2021-03-26日 11时,昌平镇,171μg/m³
更新：2021-03-26日 11时,大兴旧宫,187μg/m³
更新：2021-03-26日 11时,定陵(对照点),169μg/m³
更新：2021-03-26日 11时,东四,180μg/m³
更新：2021-03-26日 11时,房山燕山,130μg/m³
更新：2021-03-26日 11时,丰台小屯,159μg/m³
更新：2021-03-26日 11时,丰台云岗,129μg/m³
更新：2021-03-26日 11时,古城,149μg/m³
更新：2021-03-26日 11时,官园,166μg/m³
更新：2021-03-26日 11时,海淀万柳,167μg/m³
更新：2021-03-26日 11时,怀柔新城,146μg/m³
更新：2021-03-26日 11时,怀柔镇,153μg/m³
更新：2021-03-26日 11时,门头沟三家店,141μg/m³
更新：2021-03-26日 11时,密云新城,161μg/m³
更新：2021-03-26日 11时,密云镇,154μg/m³
更新：2021-03-26日 11时,农展馆,185μg/m³
更新：2021-03-26日 11时,平谷新城,99μg/m³
更新：2021-03-26日 11时,顺义新城,160μg/m³
更新：2021-03-26日 11时,通州东关,173μg/m³
更新：2021-03-26日 11时,延庆石河营,115μg/m³
更新：2021-03-26日 11时,延庆夏都,133μg/m³
```

4.整理最终数据

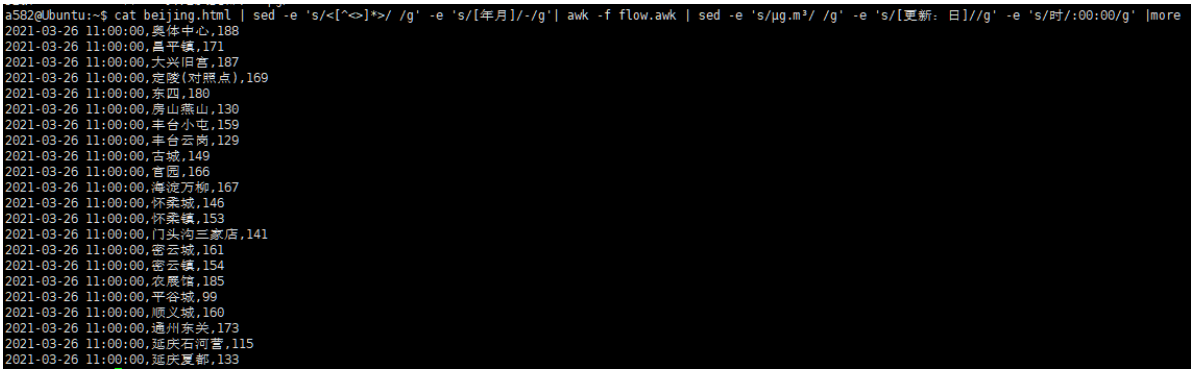
最后我们整理一下这个数据，将“更新：”替换成空，将“时”替换成“:00:00”，将`ug/m³`替换成空，将“日”替换成空，使用sed语句：

`sed -e 's/μg.m³/ /g' -e 's/[更新：日]//g' -e 's/时/:00:00/g' |more`

最终的完整的命令如下：

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

最终执行后的截图如下，包含监测时间、监测地点和监测地点的PM2.5指数，与实验要求一致



三、实验问题

在本次实验中我遇到了以下几个问题，并最终解决了这些问题

1.网址选择问题

包含北京PM2.5数据的网址有很多，不同的网站的HTML5结构不同，数据包含在不同的标签内，因此选择合适的网址可以进行更简便的正则匹配，最后在　综合考虑多方面因素后，我选择了本次实验的网址，他包含本次实验所需的所有数据，并且网站结构简单，没有复杂的标签，数据集中，适合于本次实验的正则 匹配实验。

2.sed指令使用不熟练

在刚开始实验的时候，sed指令只在课上学习过，并跟着MOOC上的案例实验过，并没有亲自动手在Linux系统上使用过，因此在最开始处理数据时，对于 sed用于替换的用法不熟悉，容易忘记单引号等。当需要复杂的替换语句时，更需要多次观看MOOC上的内容(例如第一个sed语句，对HTML5中的标签替换)，　来写出最简洁的正则匹配语句。

3.awk文件使用不熟练

在进行必要的一些替换后，需要编写awk脚本用户选择并打印出我们需要的数据的行，在最开始编写时，需要选择该行中标志性的特殊词，以此定位到某一行，例如需要时间这个数据，在H5中搜寻后发现，该行中包含“更新”两个特殊的语句，可以以此为切入点拿到数据，但拿到的该行第一列数据中除了时间外还包含有“更新”两个字，同时还有多余的“日”字，这些都在awk脚本执行后再执行sed语句处理，对于包含PM2.5数据的行，我们以m³为切入点，打印每一行，并在最后使用sed语句将m³剔除。编写awk文件时，找到所需行中的特殊字符或字符串很重要。

4.命令使用不够细节

在第一次编写完成并测试时，我发现我筛选出的数据都显示在一行中，多个PM2.5数据并没有换行，一开始我认为可能是元HTML5文件的tr标签之间并没有换行，我使用sed语句为每个后加入了换行符，发现最终的数据依旧没有换行，经过仔细的梳理，发现了原来是awk脚本中printf语句最后没有加入换行符`\n`，这是由于实验不够细致导致的，也为我以后进行实验敲响了警钟，要细致的处理数据，耐心的找寻出现的问题并解决。

四、实验总结

本次实验通过搜寻北京PM2.5Web网页并处理该网页上的HTML5数据，练习了正则匹配表达式的应用，本次实验让我正则表达式的运用能力得到了进一步的提升，对于sed语句和awk脚本文件的使用变得更加的熟练，对于在Linux系统上使用正则表达式处理数据有了更深的体会，从最开始使用指令都不熟悉到最后可以熟练根据文本内容分析出应该使用的语句，这次实验使我收获颇丰。