

**Linux开发环境及应用**



**正则表达式应用**

**姓 名 鄭毓恒**

**学 院 计算机学院**

**专 业 计算机科学与技术**

**班 级 2020211302**

**学 号 2020211262**

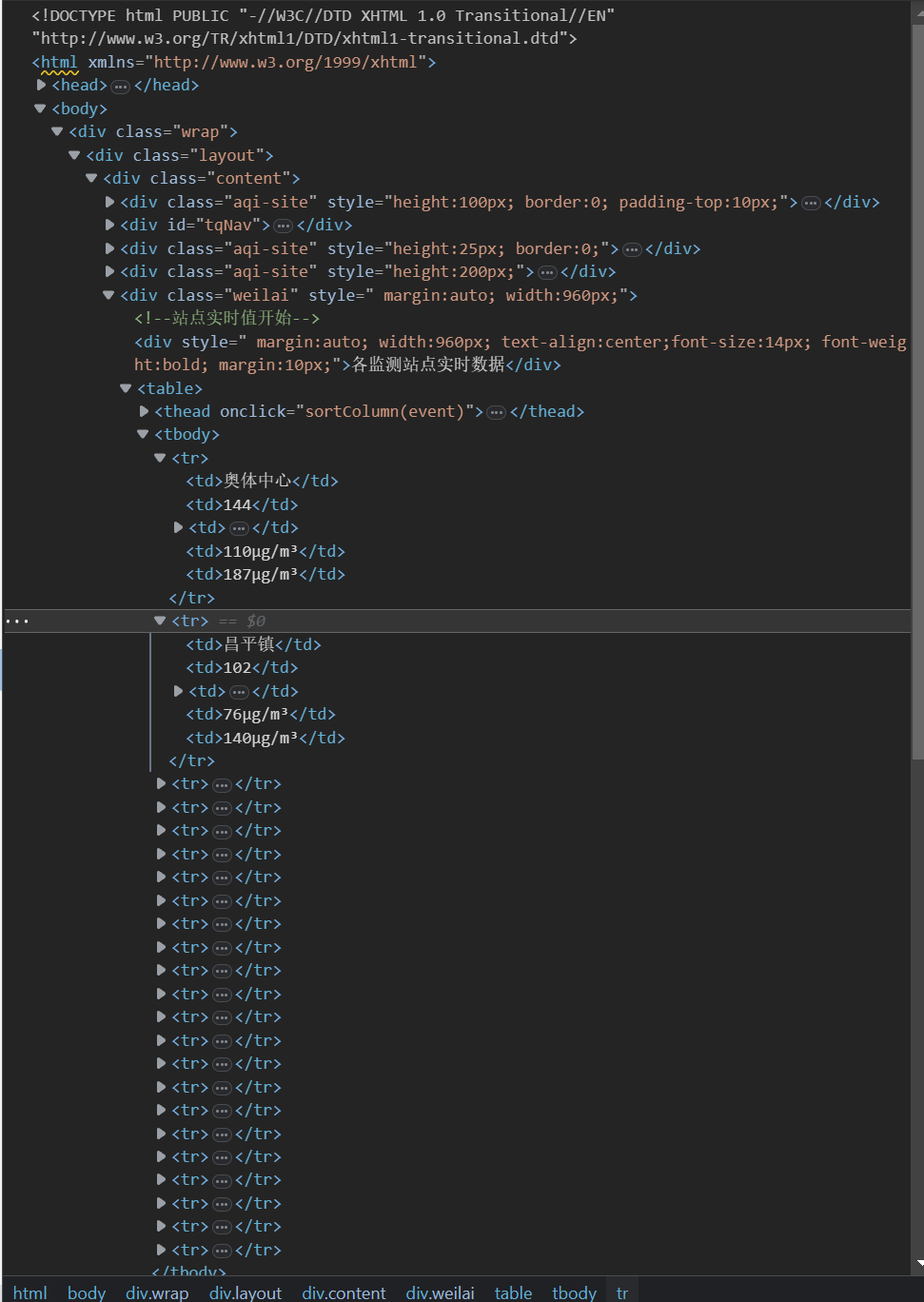
**任课教师 张雪松**

**2023年 3 月**

**搜索合适网站**

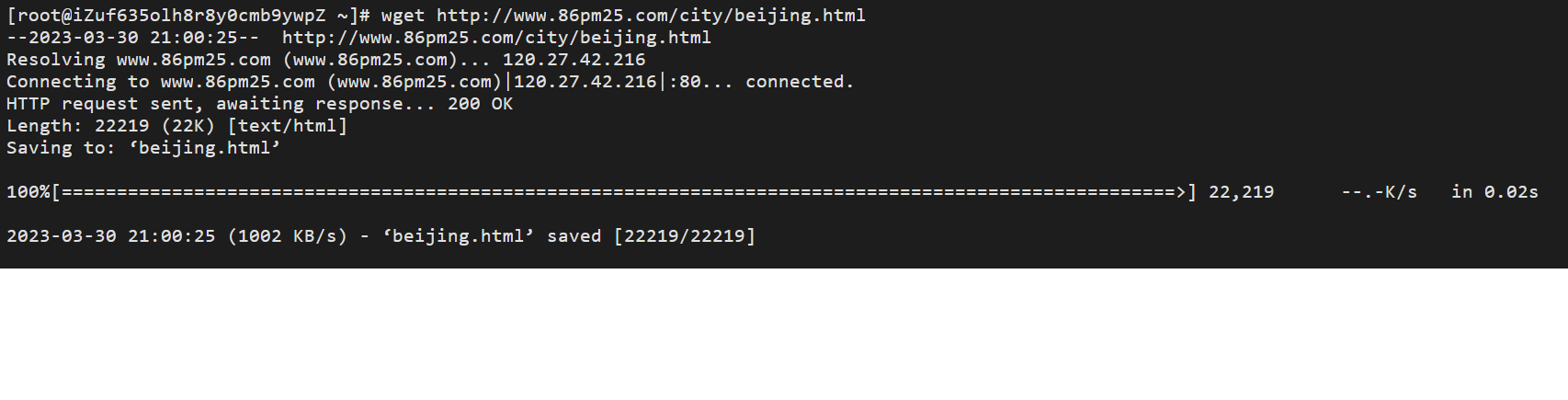


在网站<http://www.86pm25.com/city/beijing.html>，可以搜索到北京各监测站的PM2.5浓度。

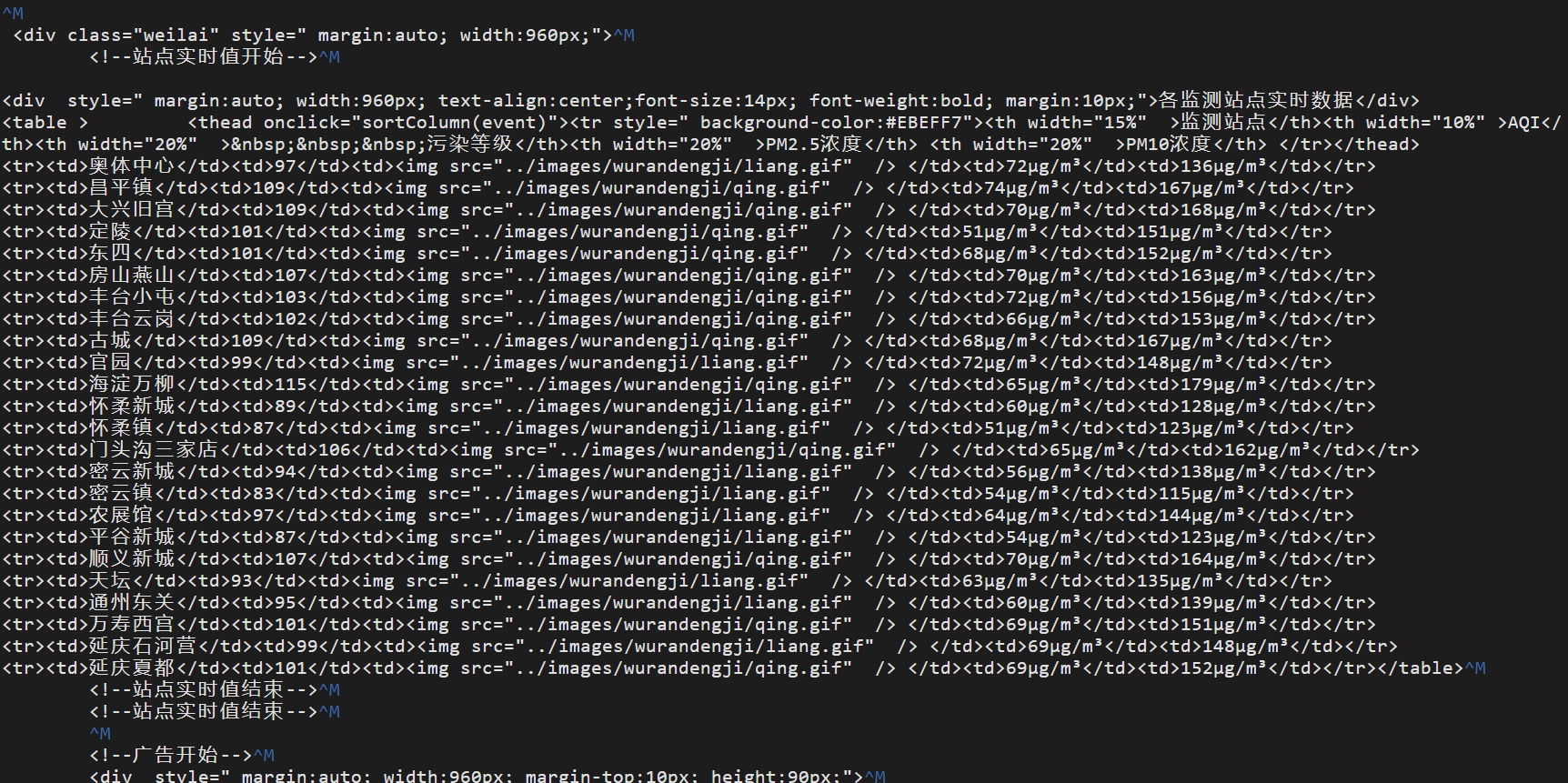


检查网站的HTML结构，可以找到监测站地点和PM2.5指数的位置。

**下载作业所需文件**

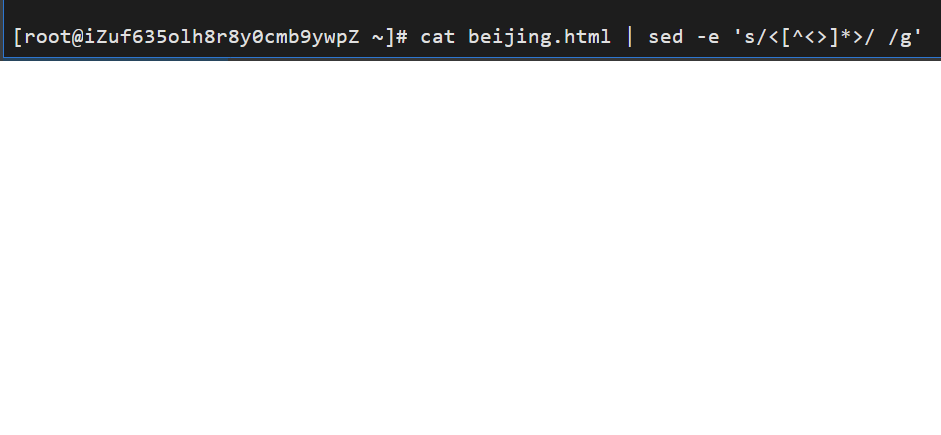


在Linux环境下，使用wget指令，下载网页的HTML。

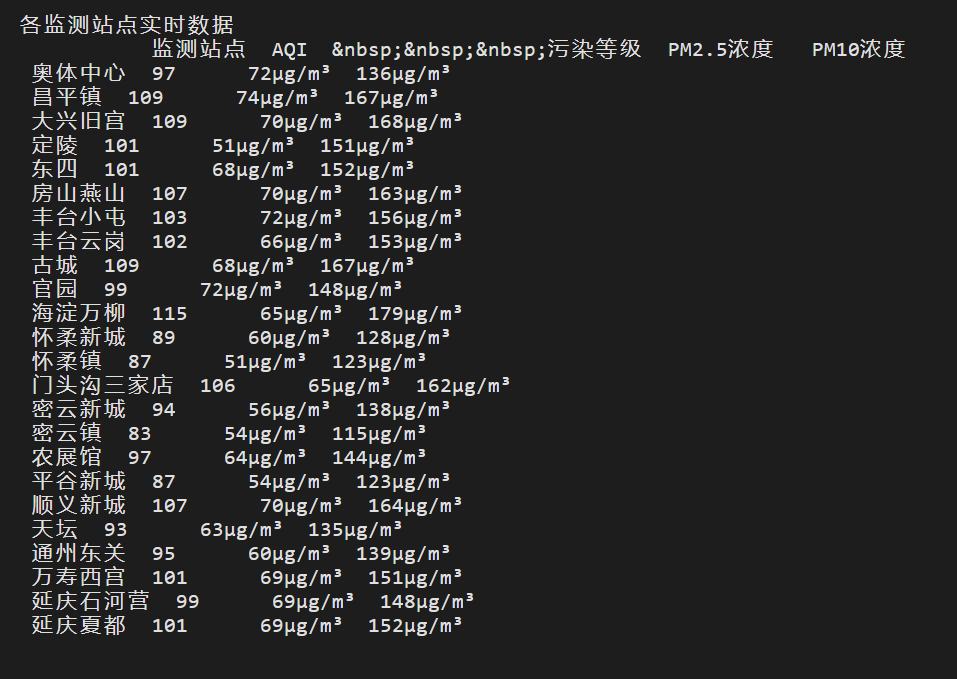


打开下载好的html文件，内容如上。

**去除HTML标签**

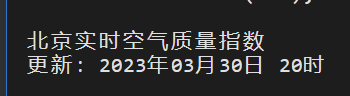


因为我们需要的监测站点名称和PM2.5浓度都在标签<tr><td></td></tr>中，可使用以上命令将HTML文件中的标签去除。

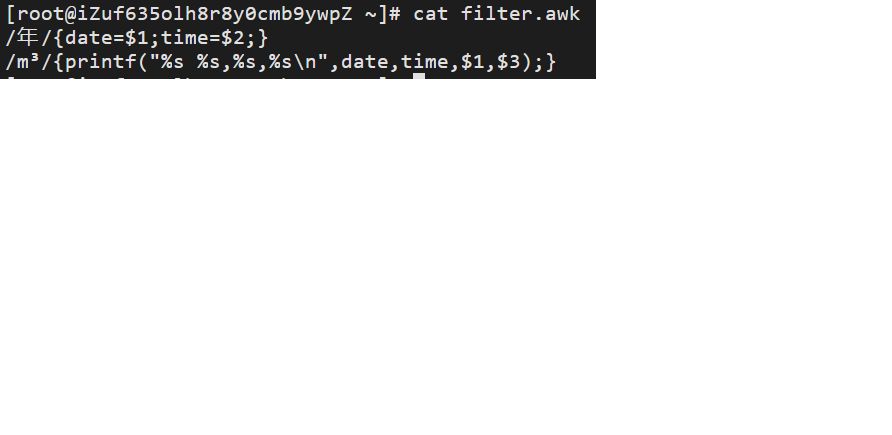


去除完标签后的输出结果如上图。

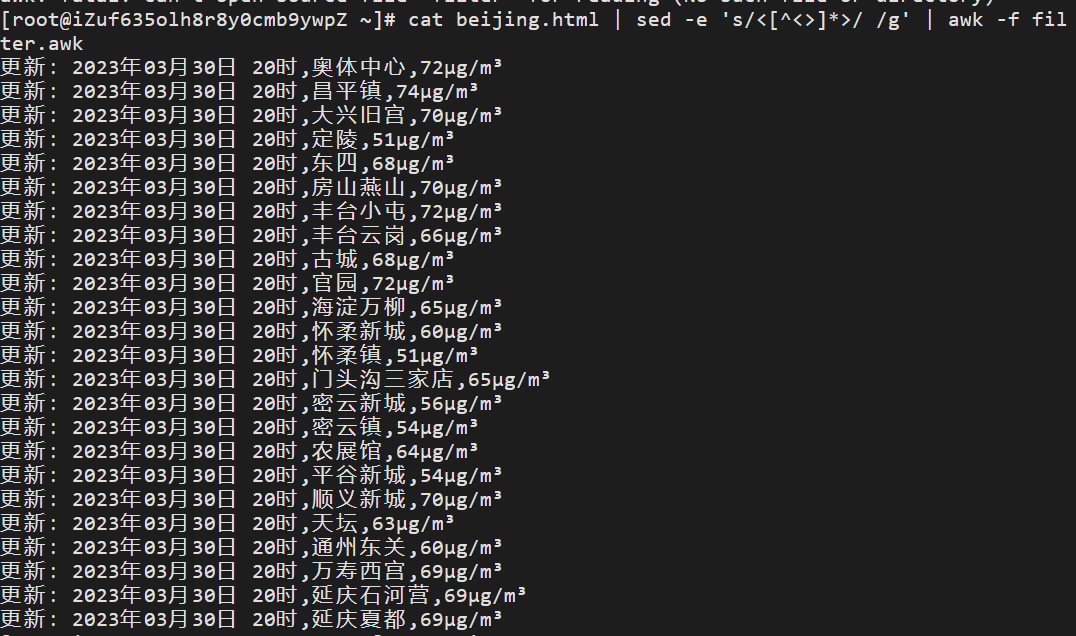
**使用awk处理格式**



根据作业要求，每一行都需要数据的日期和时间。经过查看文件，只有含有日期时间的这一行有“年”字。



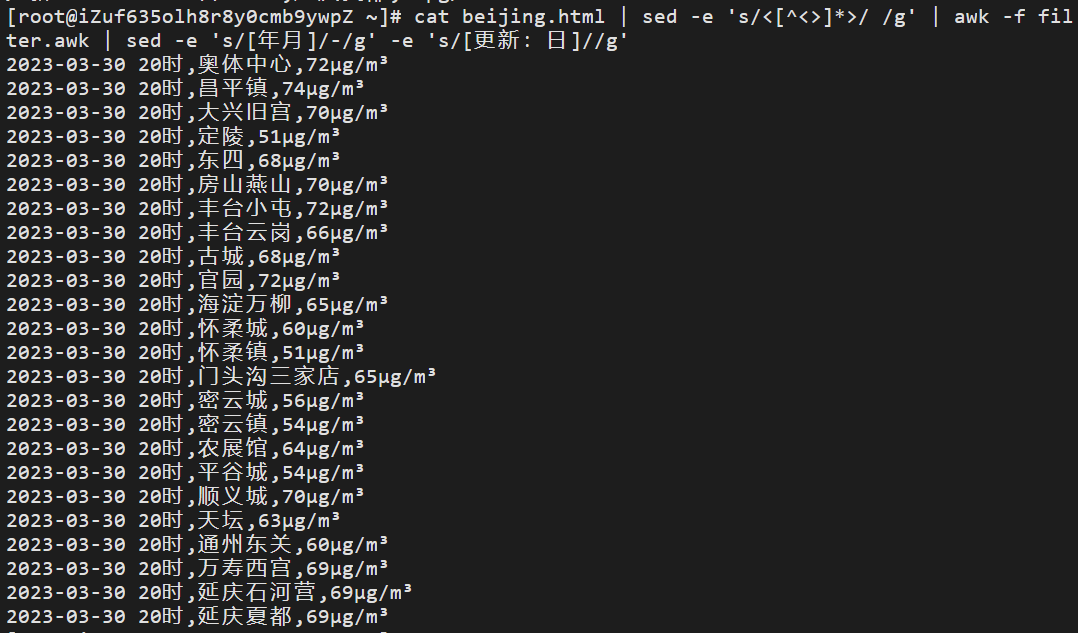
新增awk文件filter.awk，内容如上。这将会把含有文本“年”的一行的第一域和第二个域的内容分布赋值给变量date和time，并取含有m³的一行的第一域内容监测站地点和第三域内容PM2.5浓度，按printf内的格式打印。



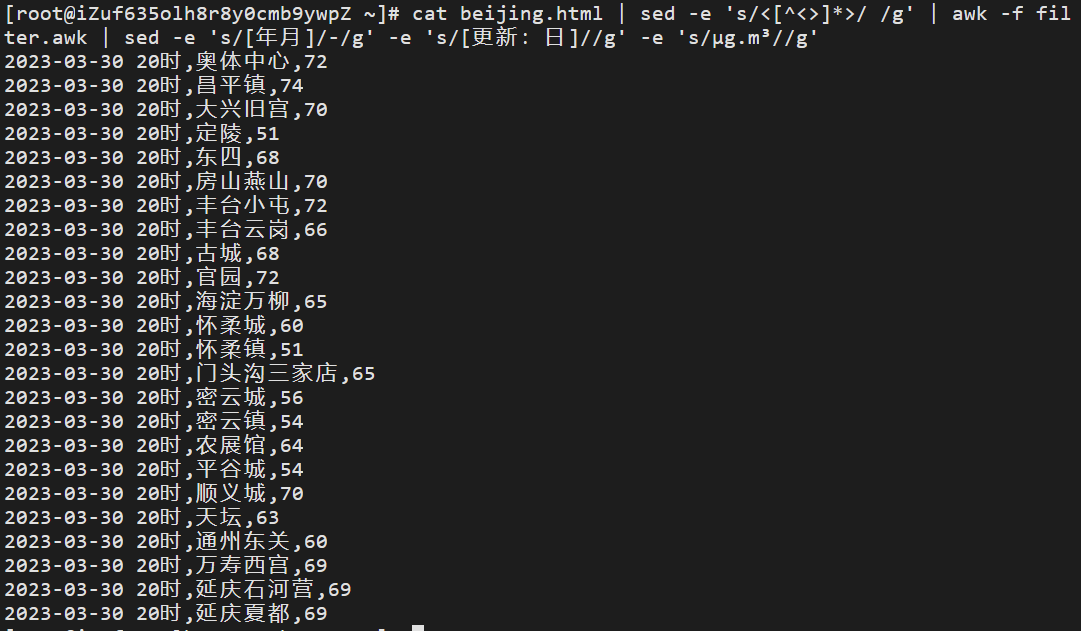
执行结果如上。

**修改格式**

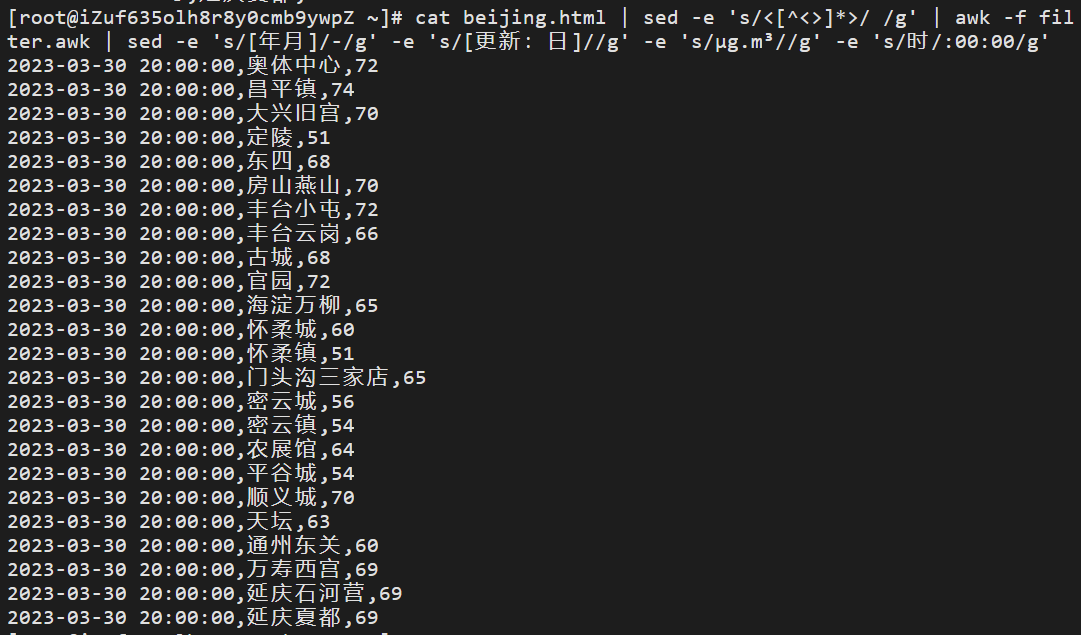
经过awk文件处理后的数据还不符合作业要求，比如“更新：”和PM2.5浓度的单位需要消除。



输入以上指令，将“更新：”、“年”、“月”替换为空。



输入以上指令，将PM2.5浓度的单位去除。



最后，将“时”替换为“:00:00”。得到符合格式要求的输出结果。