Linux 开发环境及应用上机作业一：正则表达式应用 实验报告

毛子恒
2019211397
北京邮电大学 计算机学院

日期：2022 年 3 月 22 日

1 实验内容

从因特网上搜索 Web 页，用 `wget` 获取网页，处理网页 HTML 文本数据，从中提取出当前时间点北京各监测站的 PM2.5 浓度，输出 CSV 格式数据。

2 实验步骤

获取数据 使用命令 `wget http://www.86pm25.com/city/beijing.html -o /dev/null` 取得数据，结果保存在 `beijing.html` 中。

去除 HTML 标签 使用命令 `sed -e 's/<[^<>]*>/ /g'`，去除其中的 HTML 标签，其中的正则表达式匹配以尖括号包含的、其中不含尖括号的字符串。

获取时间和处理数据 去除标签之后，得到的数据表格如图 1。

另外可以找到时间信息如“更新：2022 年 03 月 19 日 11 时”，日期在第一个区域，小时在第二个区域。

因此，采用“更新”关键词获取到时间，之后对于每一个包含 `μg/m³` 的行，气象站地点在第一个区域，PM2.5 的值在第三个区域，调整格式并打印，最后编写的 awk 文件如下：

```
/更新/{ date = $1; time = $2;}
/μg\/m³/{printf("%s %s:00:00,%s,%s\n", date, time, $1, $3);}
```

使用命令 `awk -f work.awk`，得到的数据表格如图 2。

调整格式 去掉“更新”二字，调整时间格式为 `yyyy-mm-dd hh:00:00`，去掉 `μg/m³`，使用命令 `sed -e 's/更新://g' -e 's/\([0-9]*\) 年\([0-9]*\) 月\([0-9]*\) 日/\1-\2-\3/g' -e 's/时//g' -e 's/μg\/m³//g'`。

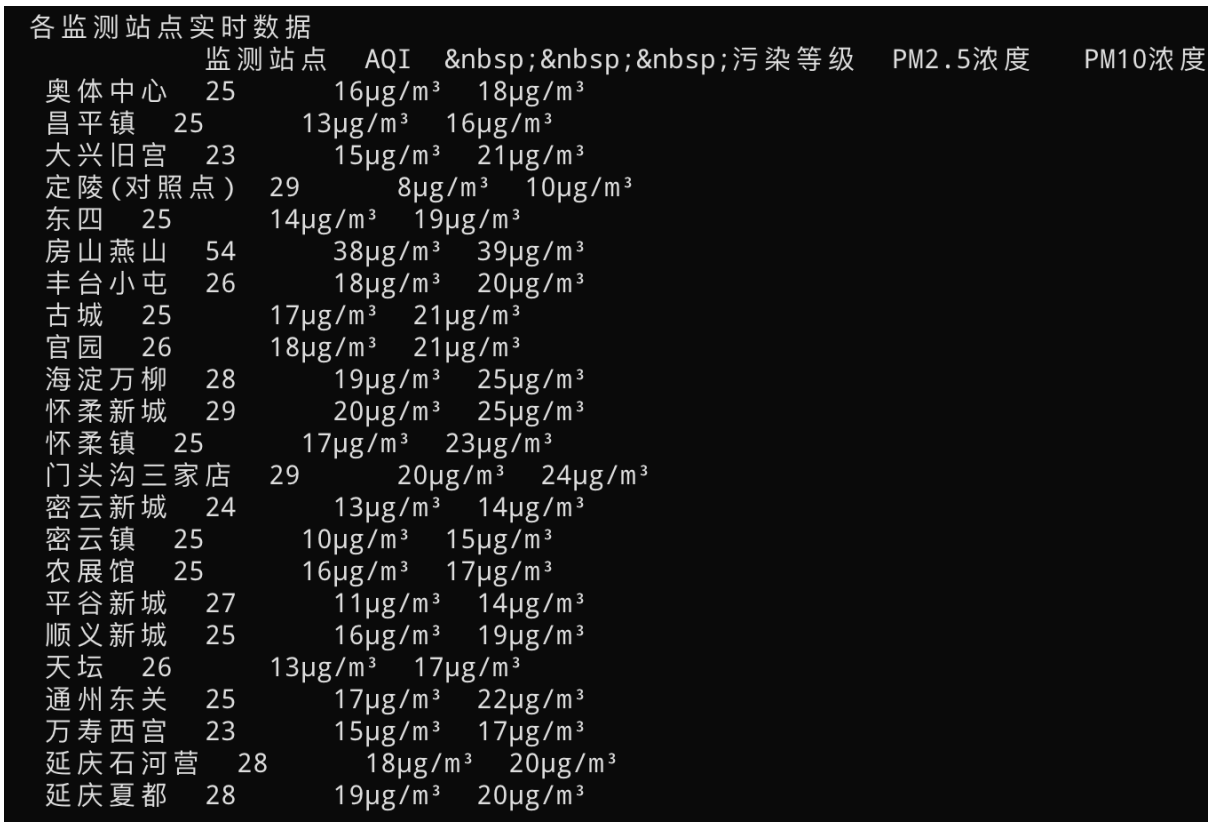

图 **1**

图 **2**

输出到文件 使用命令 `tee result.csv >/dev/null` 输出到 `result.csv`，最终结果如图 **3**。

所有命令连起来如图 **4**。

```
2022-03-19 11:00:00,奥体中心,16
2022-03-19 11:00:00,昌平镇,13
2022-03-19 11:00:00,大兴旧宫,15
2022-03-19 11:00:00,定陵(对照点),8
2022-03-19 11:00:00,东四,14
2022-03-19 11:00:00,房山燕山,38
2022-03-19 11:00:00,丰台小屯,18
2022-03-19 11:00:00,古城,17
2022-03-19 11:00:00,官园,18
2022-03-19 11:00:00,海淀万柳,19
2022-03-19 11:00:00,怀柔新城,20
2022-03-19 11:00:00,怀柔镇,17
2022-03-19 11:00:00,门头沟三家店,20
2022-03-19 11:00:00,密云新城,13
2022-03-19 11:00:00,密云镇,10
2022-03-19 11:00:00,农展馆,16
2022-03-19 11:00:00,平谷新城,11
2022-03-19 11:00:00,顺义新城,16
2022-03-19 11:00:00,天坛,13
2022-03-19 11:00:00,通州东关,17
2022-03-19 11:00:00,万寿西宫,15
2022-03-19 11:00:00,延庆石河营,18
2022-03-19 11:00:00,延庆夏都,19
```

图 3

```
b397@Ubuntu-bupt:~/work1$ wget http://www.86pm25.com/city/beijing.html -o /dev/n
ull
b397@Ubuntu-bupt:~/work1$ cat beijing.html | sed -e 's/<[^<>]*>/ /g' | awk -f wo
rk.awk | sed -e 's/更新： //g' -e 's/\([0-9]*\)年\([0-9]*\)月\([0-9]*\)日/\1-\2-\
3/g' -e 's/时//g' -e 's/μg\/m³//g' | tee result.csv >/dev/null
```

图 4

3 实验总结

本次实验中我熟悉了 sed,awk,tee 等命令的功能和正则表达式的基本语法，并且成功运用它们处理了简单的文本文件。