一、实验要求：

从网页上抓取北京各地区PM2.5的实时数据，使用sed和awk命令处理，按如下格式输出:

地点　时间　PM2.5浓度

二、实验步骤：

1. 用wget指令从http://www.soupm25.com/上抓取数据

```
wget http://www.soupm25.com/
```

2.查看抓取到的数据

```
cat index.html 或都用文本编辑器打开（方便查找）
```

3.使用sed命令初步筛选出时间和有用的表格数据,结果保存为index1.html

```
sed -n -e  '/北京空气质量指数更新时间/p' -e '/div class="content allsitelist"/,/<\/table>/ p' index.html > index1.html
#指令的第一部分
#sed -n -e '/北京空气质量指数更新时间/p'  作用是：将含有“北京空气质量指数更新时间”的行打印出来
#指令的第二部分
#sed -n -e ' '/div class="content allsitelist"/,/<\/table>/ p       作用是：打印div class="content allsitelist"到<\/table>之间的所有内容
```

4.删除所有的html标签和空行，结果保存为index3.html

```
sed 's/<[^<>]*>//g' index1.html >index2.html
#解释：
#将html标签替换为空
#正则表达式分析：
#匹配的表达式满足条件：  <>之间的为除了<和>之外的任意多个字符
sed '/^\s*$/d' index2.html >index3.html
#解释：
#将空行删除
#正则表达式分析：
#匹配的表达式满足条件：  以\s开头，同时以\s结尾
```

5.所有行置顶,结果保存为index4.html

```
sed 's/^\s*//g' index3.html > index4.html
#解释：
#将每行所需内容前的空白去掉，将所有行置顶
#正则表达式分析：
#匹配的表达式满足条件：  行以\s开头，后接若干\s
```

5.使用awk命令将index4.html处理成指定格式

- 创建index4.awk文件

  ```
  vim index4.awk
  ```

- 编写index4.awk文件

```
BEGIN{
```

```
    DATA;       #定义变量
    TIME;
    PLACE;
    PM25;
}
{
    RS = "\n";      #设置行分隔符为\n，不能是\r，因为\r划分出来的TIME变量会带上换行符
    if(NR==1){      #提取日期和时间
        DATA=$2;
        TIME=$3;
    }

    if(NR%4==2)     #利用NR标识各行，逐行处理
        PLACE = $1;
    else if(NR%4==3){
        PM25=$0;    #不能用$1,使用$1时不会带上单位
        printf("%10s \t %s %s \t %s\n",PLACE,DATA,TIME,PM25);
    }
}
```

- 处理index4.html文件得到最终的结果

```
awk -f index4.awk result.html
```

三、实验过程及遇到的问题

1.实验过程

这个实验我做了三次，途中操作系统坏了一次，最后赶在4月2号前完成，好在实验步骤都熟悉了，最后一次完成并没有花费多少时间。

2.Q && A

Q1: 如何将多条sed命令放在一条命令中执行

A: 方法一：使用 -e 命令,格式如下

```
sed -e '命令1' -e '命令2' 文件名
或
sed -e {'命令1'; '命令2'} 文件名
```

Note:

-n 命令应放在 -e 前面

方法二：使用管道

Note：

管道连接的命令顺序执行，前一条命令的输出可以作为下一条命令的输入

Q2: awk脚本文件中RS 和FS的作用：

A:

- RS是行分隔符
- FS是列分隔符，即将一行按FS分隔成多段

Q3: 在用awk处理得到的最终结果中，PM2.5的值总是会被归到新的一行

A:

- 这个问题我百思不得其解，开始怀疑是函数的问题，查了print 和printf的区别，print会自动换行，但换成printf也没有用
- 之后通过定义变量，调整printf语句，将所有的输出放在一块，也没有用
- 最后我发现换行是在TIME变量之后 ， 猜测TIME变量含有不可见字符"\r"
- 处理方案是：将RS定义成'\n',于是解决了问题

四、实验结果

站点	2019-04-01 21:00:00	PM2.5
琉璃河	2019-04-01 21:00:00	43 ug/m3
古城	2019-04-01 21:00:00	35 ug/m3
门头沟	2019-04-01 21:00:00	34 ug/m3
东高村	2019-04-01 21:00:00	32 ug/m3
北部新区	2019-04-01 21:00:00	29 ug/m3
昌平镇	2019-04-01 21:00:00	28 ug/m3
通州	2019-04-01 21:00:00	28 ug/m3
永定门内	2019-04-01 21:00:00	28 ug/m3
榆垡	2019-04-01 21:00:00	28 ug/m3
东四	2019-04-01 21:00:00	27 ug/m3
前门	2019-04-01 21:00:00	27 ug/m3
丰台花园	2019-04-01 21:00:00	27 ug/m3
西直门北	2019-04-01 21:00:00	26 ug/m3
延庆	2019-04-01 21:00:00	25 ug/m3
农展馆	2019-04-01 21:00:00	25 ug/m3
天坛	2019-04-01 21:00:00	25 ug/m3
东四环	2019-04-01 21:00:00	24 ug/m3
云岗	2019-04-01 21:00:00	24 ug/m3
奥体中心	2019-04-01 21:00:00	24 ug/m3
海淀区万柳	2019-04-01 21:00:00	24 ug/m3