1. **实验内容**

从因特网上搜索相关Web网页，处理网页html数据，从中提取出当前时间点北京各监测站的PM2.5浓度，输出格式如下。要求：写出各个处理步骤，并给出解释。

2020-03-09 13:00:00,海淀区万柳,73

2020-03-09 13:00:00,昌平镇,67

2020-03-09 13:00:00,奥体中心,66

2020-03-09 13:00:00,海淀区万柳,73

2020-03-09 13:00:00,昌平镇,73

2020-03-09 13:00:00,奥体中心,75

1. **实验步骤**

选用网页：http://www.86pm25.com/city/beijing.html



监测站点及PM2.5浓度值：



更新时间：



|  |  |
| --- | --- |
| 命令 | 含义 |
| cat data.html | 原始网页 |
| sed -e 's/<tr/\n<tr/g' | 观察到奥体中心所在行与其他行格式不一致，在<tr前添加一个换行符 |
| sed -e 's/<[^<>]\*>/ /g' | 去除形如<[^<>]\*>的所有html超文本标记 |
| sed -e 's/更新：/更新 /g' | 添加一个空格，将更新与日期分隔在不同的列 |
| sed -e 's/[年月]/-/g' | 将所有的”年”、”月”替换为”-”，以符合输出格式的要求 |
| sed -e 's/日//g' | 将所有的”日”去除，以符合输出格式的要求 |
| sed -e 's/时/:00:00/g' | 原文本中时间只精确至小时，因而将”时”替换为” :00:00”， 以符合输出格式的要求 |
| sed -e 's/μg\/m³/ μg\/m³/g' | 在”μg/m³”前添加一个空格，将PM2.5浓度值与单位分隔在不同的列 |
| awk '/更新/{date=$2; time=$3}; /— μg\/m³/{printf("%s %s,%s,%s\n"), date, time, $1, $3;}' | 搜索文本中的”更新”字样，即更新时间所在行，此时日期在第二列，小时数在第三列，将其作为变量保存；  搜索文本中的” — μg/m³”字样，即监测站点及PM2.5浓度值所在行。按格式输出之前存储的日期和小时，以及位于第一列和第三列的监测站点名称和PM2.5浓度值。 |

1. **实验结果**

将以上命令进行简单的融合：

cat data.html | sed -e 's/<tr/\n<tr/g' -e 's/<[^<>]\*>/ /g' -e 's/更新：/更新 /g' -e 's/[年月]/-/g' -e 's/日//g' -e 's/时/:00:00/g' -e 's/μg\/m³/ μg\/m³/g' | awk '/更新/{date=$2; time=$3}; /— μg\/m³/{printf("%s %s,%s,%s\n"), date, time, $1, $3;}'

