上机作业 1：正则表达式应用

088 于海鑫
2017211240
name1e5s@bupt.edu.cn

版本：8
更新：*March 23, 2020*

1 实验要求

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，输出如下 CSV 格式数据。

```
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
```

撰写实验报告。

2 实验过程

2.1 获取数据

在 `bing` 上搜索”北京各监测站的 PM2.5 浓度“并筛选，最终我们选定了最为简单的 `http://www.86pm25.com/city/beijing.html` 作为筛选的目标。在这个网页中，我们要进行处理的 HTML 片段如下：

```
1        <!--站点实时值开始-->
2
3  <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>
4  <table >     <thead onclick="sortColumn(event)"><tr style="␣
     background-color:#EBEFF7"><th width="15%"  >监测站点</th><th width
     ="10%" >AQI</th><th width="20%"  >   污染等级</th><
     th width="20%"  >PM2.5浓度</th> <th width="20%"  >PM10浓度</th> </
     tr></thead>
5  <tr><td>奥体中心</td><td>25</td><td><img src="../images/wurandengji/
     you.gif"  /> </td><td>8μg/m³</td><td>21μg/m³</td></tr>
6  <tr><td>昌平镇</td><td>33</td><td><img src="../images/wurandengji/you
     .gif"  /> </td><td>15μg/m³</td><td>33μg/m³</td></tr>
7  <tr><td>定陵</td><td>26</td><td><img src="../images/wurandengji/you.
     gif"  /> </td><td>12μg/m³</td><td>24μg/m³</td></tr>
8  <tr><td>东四</td><td>27</td><td><img src="../images/wurandengji/you.
     gif"  /> </td><td>14μg/m³</td><td>25μg/m³</td></tr>
9  <tr><td>古城</td><td>25</td><td><img src="../images/wurandengji/you.
     gif"  /> </td><td>16μg/m³</td><td>21μg/m³</td></tr>
10 <tr><td>官园</td><td>30</td><td><img src="../images/wurandengji/you.
     gif"  /> </td><td>11μg/m³</td><td>20μg/m³</td></tr>
11 <tr><td>海淀区万柳</td><td>27</td><td><img src="../images/wurandengji
     /you.gif"  /> </td><td>15μg/m³</td><td>—μg/m³</td></tr>
12 <tr><td>怀柔镇</td><td>29</td><td><img src="../images/wurandengji/you
     .gif"  /> </td><td>6μg/m³</td><td>—μg/m³</td></tr>
13 <tr><td>农展馆</td><td>27</td><td><img src="../images/wurandengji/you
     .gif"  /> </td><td>16μg/m³</td><td>22μg/m³</td></tr>
14 <tr><td>顺义新城</td><td>29</td><td><img src="../images/wurandengji/
     you.gif"  /> </td><td>2μg/m³</td><td>12μg/m³</td></tr>
15 <tr><td>天坛</td><td>43</td><td><img src="../images/wurandengji/you.
     gif"  /> </td><td>9μg/m³</td><td>43μg/m³</td></tr>
16 <tr><td>万寿西宫</td><td>23</td><td><img src="../images/wurandengji/
     you.gif"  /> </td><td>5μg/m³</td><td>22μg/m³</td></tr></table>
17    <!--站点实时值结束-->
```

2.2 数据筛选

2.2.1 提取表格

我们首先要做的就是把这一部分表格从整个 `HTML` 文本中筛选出来，经过观察可以发现该网页的表格十分有规律，**有且只有**表格部分的开头为 `<tr><td>` ，以此为切入点，我们使用 `grep` 配合正则表达式抓取表格片段。

```
bash$ grep "^<tr><td>"
```

2.3 删除 TAG

HTML 作为富文本文档，在转换为我们需要的格式前必须要做的就是把多余的 TAG 部分给清除掉，这里我们使用 `sed` 来解决问题，代码如下：

```
bash$ sed -E 's/(<[^>]+>)+/~/g' | sed 's/~/ /g'
```

这一段代码的作用时先把全部的 TAG 转换为单个的波浪号字符，之后再把该字符转换为空格。此时的输出如下：

```
name1e5s@sumeru:~$ wget -q -O -  http://www.86pm25.com/city/beijing.
   html | grep "^<tr><td>" | sed -E 's/(<[^>]+>)+/~/g' | sed 's/~/ /g
'
 奥体中心 28   13μg/m³ 21μg/m³
 昌平镇 28   7μg/m³ 10μg/m³
 定陵 29   7μg/m³ 12μg/m³
 东四 30   17μg/m³ 28μg/m³
 古城 29   13μg/m³ 17μg/m³
 官园 33   11μg/m³ 18μg/m³
 海淀区万柳 30   10μg/m³ 10μg/m³
 农展馆 34   17μg/m³ 34μg/m³
 顺义新城 32   22μg/m³ 28μg/m³
 天坛 37   17μg/m³ 37μg/m³
 万寿西宫 24   13μg/m³ 24μg/m³
```

2.4 删除单位

`μg/m³` 作为不需要的单位，使用 `sed` 剔除掉。

```
bash$ sed 's%μg/m³%%g'
```

此时的输出如下：

```
name1e5s@sumeru:~$ wget -q -O -  http://www.86pm25.com/city/beijing.html | grep "^<tr><td>" | sed -E 's/(<[^>]+>)+/~/g' | sed 's/~/ /g' | sed 's%μg/m³%%g'
 奥体中心 28   13 21
 昌平镇 28   7 10
 定陵 29   7 12
 东四 30   17 28
 古城 29   13 17
 官园 33   11 18
 海淀区万柳 30   10 10
 农展馆 34   17 34
 顺义新城 32   22 28
 天坛 37   17 37
 万寿西宫 24   13 24
```

2.5 打印结果

在获取到必要的信息之后，我们可以使用 `awk` 输出结果。

```
bash$ awk '{print datenow "," $1 "," $3}' datenow="`date "+%Y-%m-%d %H:%M:%S"`"
```

此时的输出如下：

```
name1e5s@sumeru:~$ wget -q -O -  http://www.86pm25.com/city/beijing.html | grep "^<tr><td>" | sed -E 's/(<[^>]+>)+/~/g' | sed 's/~/ /g' | sed 's%μg/m³%%g'| awk '{print datenow "," $1 "," $3}' datenow="`date "+%Y-%m-%d %H:%M:%S"`"
2020-03-23 14:35:43,奥体中心,13
2020-03-23 14:35:43,昌平镇,7
2020-03-23 14:35:43,定陵,7
2020-03-23 14:35:43,东四,17
2020-03-23 14:35:43,古城,13
2020-03-23 14:35:43,官园,11
2020-03-23 14:35:43,海淀区万柳,10
2020-03-23 14:35:43,农展馆,17
2020-03-23 14:35:43,顺义新城,22
```

```
2020-03-23 14:35:43,天坛,17
2020-03-23 14:35:43,万寿西宫,13
```

符合要求。