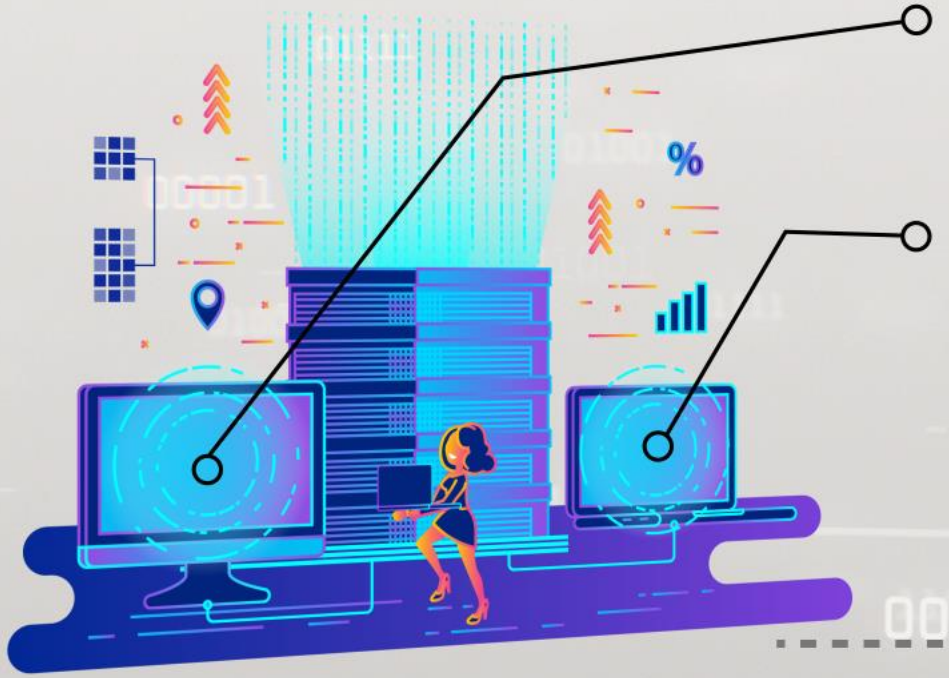
北京郵電大學
Beijing University of Posts and Telecommunications



正则表达式应用

蒋砚军 北京邮电大学计算机学院

四个例题

例1 累加作品发行量
例2 求一本书中出现频率最高的200个单词
例3 从网页中抽取数据绘制游客流量随时间变化的曲线
例4 根据捕获的网络流量观察TCP发送窗口变化

例3 绘制游客流量随时间变化的曲线

获取网页的命令：wget http://s.visitbeijing.com.cn/flow.php -o /dev/null; mv flow.php 180801.html

2018-08-01 北京景区舒适度指数播报(9点至16点每小时更新一次数据)

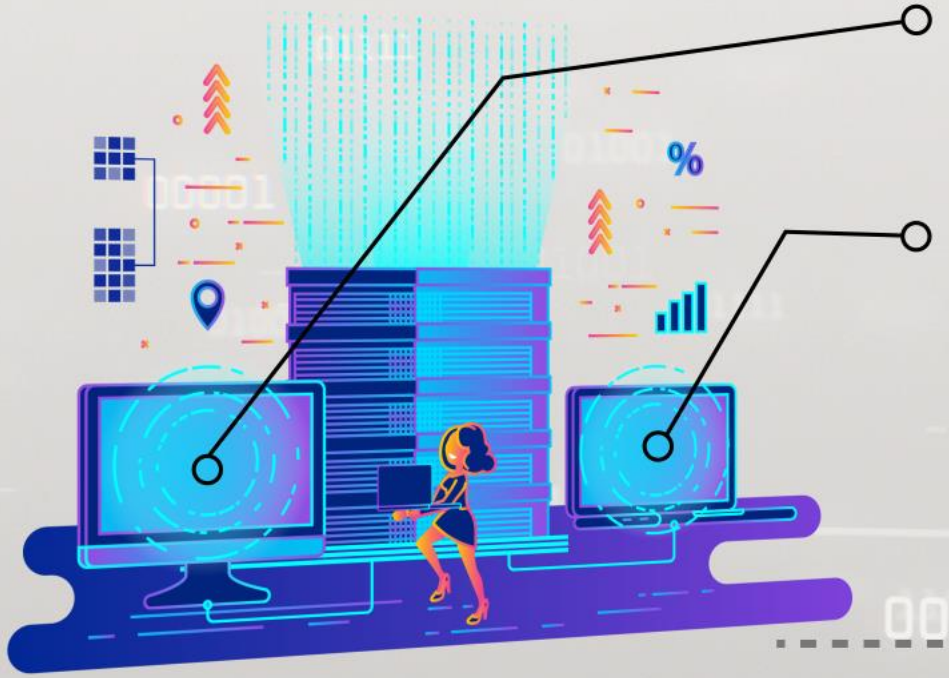
景区\时间	9：00		10:00		11:00		12:00		13:00		14:00		15:00		16:00		
	人数(万人)	舒适度	人数(万人)	舒适度	人数(万人)	舒适度	人数(万人)	舒适度	人数(万人)	舒适度	人数(万人)	舒适度	人数(万人)	舒适度	人数(万人)	舒适度	路况
故宫	约1.26	4	约2.01	4	约2.51	4	约2.45	4	约2.23	4	约1.82	4	约1.55	4	约1.29	4	查看
天坛公园	约1.06	5	约1.12	5	约1.09	5	约1.03	5	约0.96	5	约0.97	5	约1.00	5	约1.00	5	查看
恭王府	约0.07	5	约0.07	5	约0.08	5	约0.08	5	约0.08	5	约0.08	5	约0.09	5	约0.09	5	查看
国家体育场	暂无	暂无	暂无	暂无	暂无	暂无	暂无	暂无	暂无	暂无	暂无	暂无	暂无	暂无	暂无	暂无	查看
国家游泳中心	暂无	暂无	暂无	暂无	暂无	暂无	暂无	暂无	暂无	暂无	暂无	暂无	暂无	暂无	暂无	暂无	查看
奥林匹克森林公园	约4.72	5	约4.85	5	约5.13	4	约5.30	4	约5.49	4	约5.80	4	约5.97	4	约6.05	3	查看
颐和园	约1.26	5	约1.43	5	约1.56	5	约1.59	5	约1.61	5	约1.62	5	约1.61	5	约1.54	5	查看
八达岭	约0.93	4	约0.96	4	约0.78	4	约0.77	4	约0.72	4	约0.64	4	约0.60	5	约0.46	5	查看
十三陵	约0.89	5	约0.96	5	约1.04	5	约1.04	5	约1.05	5	约0.99	5	约0.95	5	约0.90	5	查看
慕田峪长城	约0.16	5	约0.21	5	约0.27	5	约0.28	5	约0.28	5	约0.25	5	约0.20	5	约0.16	5	查看
龙潭湖公园	约0.49	5	约0.50	5	约0.47	5	约0.43	5	约0.42	5	约0.44	5	约0.42	5	约0.41	5	查看
中山公园	约0.41	5	约0.41	5	约0.44	5	约0.35	5	约0.24	5	约0.23	5	约0.21	5	约0.19	5	查看

例3 绘制游客流量随时间变化的曲线

期望得到的输出结果（组织成.csv文件）：

```
故宫,2018-08-01 9:00,1.26
故宫,2018-08-01 10:00,2.01
故宫,2018-08-01 11:00,2.51
故宫,2018-08-01 12:00,2.45
故宫,2018-08-01 13:00,2.23
故宫,2018-08-01 14:00,1.82
故宫,2018-08-01 15:00,1.55
故宫,2018-08-01 16:00,1.29
天坛公园,2018-08-01 9:00,1.06
天坛公园,2018-08-01 10:00,1.12
天坛公园,2018-08-01 11:00,1.09
```

北京郵電大學
Beijing University of Posts and Telecommunications



谢 谢