R七种武器之网络爬虫RCurl 第3周

法律声明

【声明】本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

课程详情访问炼数成金培训网站

http://edu.dataguru.cn

关注炼数成金企业微信

- 提供全面的数据价值资讯，涵盖商业智能与数据分析、大数据、企业信息化、数字化技术等，各种高性价比课程信息，赶紧掏出您的手机关注吧！

本周内容

- 字符串处理基础
- 正则表达式初识
- 提取新浪财经股票数据
- 腾讯财经小试牛刀

字符串处理基础

- 赋值
- 长度、个数：nchar，length
- 替换：chartr（原始字符，替换字符，字符串）
- 连接：paste 参数sep，collapse
- 切割：strsplit
- 比较：>、<、==、!=
- 并集、交集、补集：union，intersect，setdiff
- 截取：substr，substring
- 匹配：match，pmatch，charmatch

正则表达式初识

- 对字符串操作的一种逻辑公式
- 主要应用对象：文本
- 作用：
  - 逻辑过滤
  - 精准抓取
- 正则表达式的特点是：
  - 1. 灵活性、逻辑性和功能性非常的强；
  - 2. 可以迅速地用极简单的方式达到字符串的复杂控制；
  - 3. 对于刚接触的人来说，比较晦涩难懂。

正则表达式初识

- \：转义字符
- .：除了换行以外的任意字符
- ^：一行字符串的起始
- $：一行字符串的结束
- *：零个或者多个之前的字符
- +：一个或者多个之前的字符
- ?： 零个或者一个之前的字符
- 保留字符都需要转义字符 \ 来转义表示

正则表达式初识

- 方括号[]，代表可以匹配其中任何一个字符。而^在[]中代表“非”，-代表“之间”
  - [qjk]：q，j，k中任意一个字符
  - [^qjk]：非q，j，k的任意其它字符
  - [a-z]：a至z中任意一个小写字符
  - [^a-z]： 非任意一个a至z小写字符的其它字符（可以是大写字符）
  - [a-zA-Z]：任意一个英文字母
  - [a-z]+： 一个或者多个小写英文字母
- |：或者
- 小括号()与花括号｛ ｝配合“|”使用

正则表达式初识

- 常用的特殊转义字符含义
- \n：换行符
- \t：tab
- \w：任意字母（包括下划线）或者数字 即 [a-zA-Z0-9_]
- \W：\w的反义 即[^a-zA-Z0-9_]
- \d：任意一个数字 即[0-9]
- \D：\d的反义 即[^0-9]
- \s：任意一个空格，比如space, tab, newline 等
- \S：\s的反义，任意一个非空格

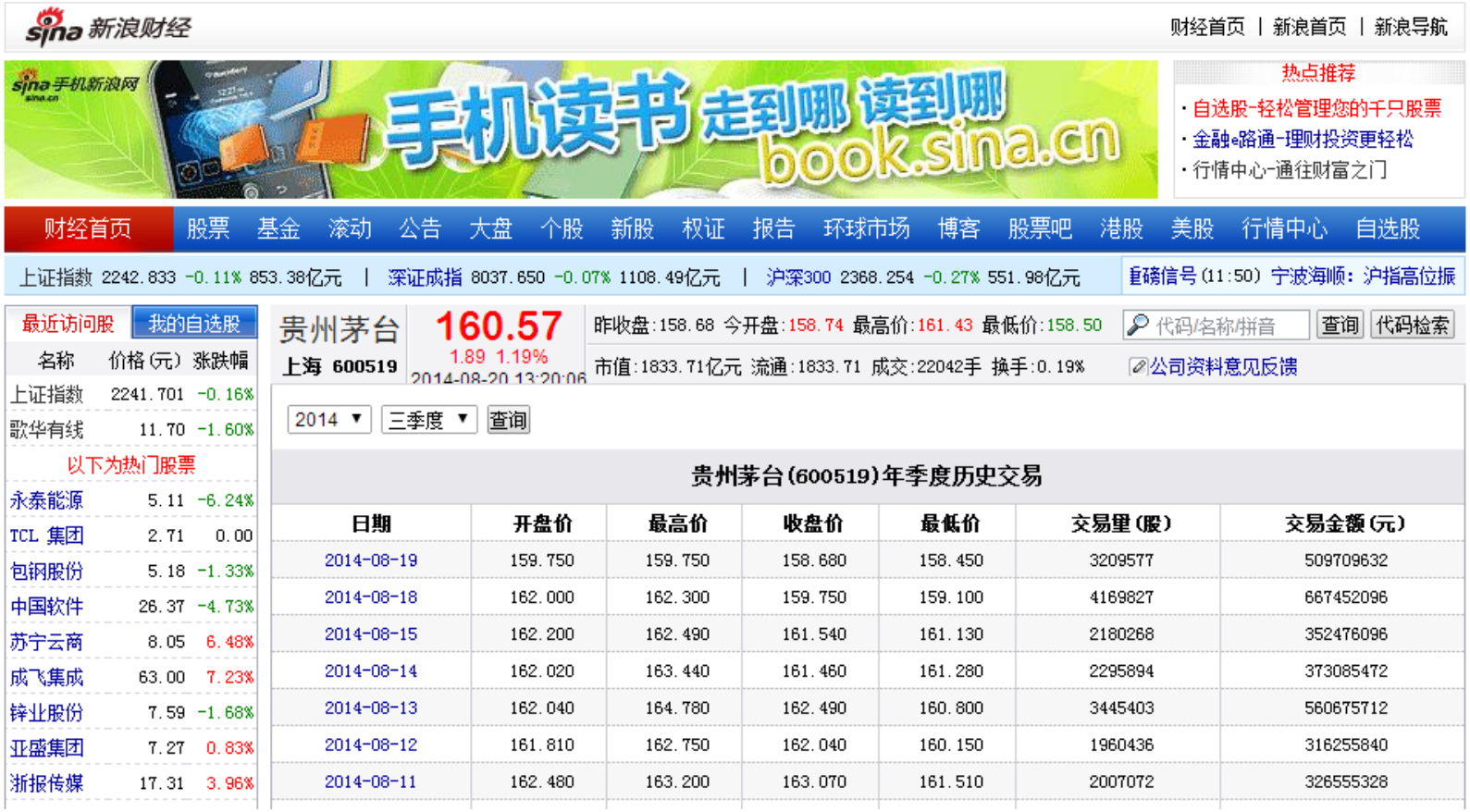
正则常用函数

- grepl：返回一个逻辑值
- grep：返回匹配的id

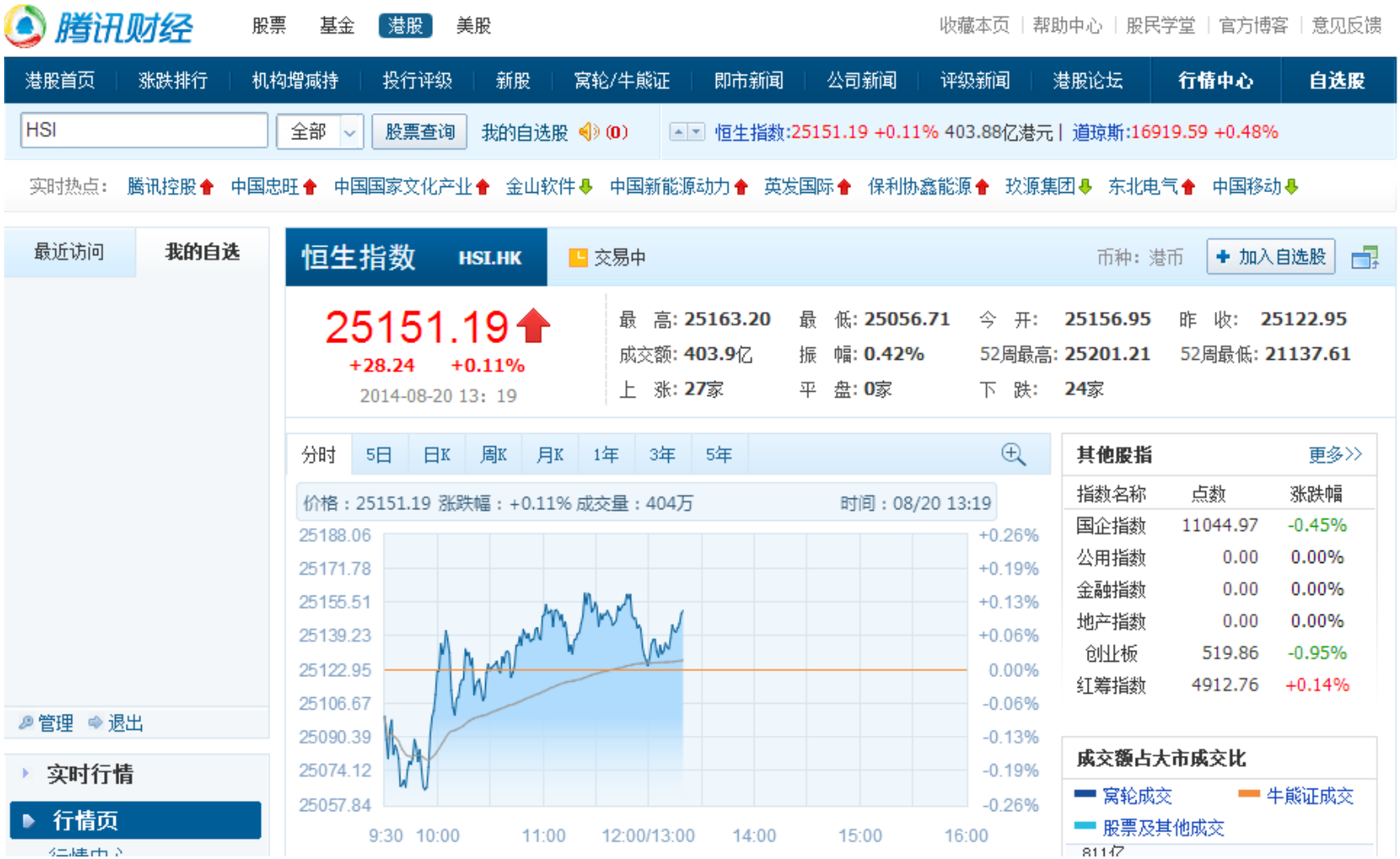
- 正则替换：sub和gsub

- regexpr：返回一个数字，1表示匹配，-1表示不匹配，以及两个属性，匹配的长度和是否使用useBytes
- regexec：返回一个list，字符串中第一个匹配及其长度以及是否使用useBytes
- gregexpr：返回一个list， 每一个匹配及其长度以及是否使用useBytes

抓取新浪财经的数据


新浪财经
财经首页 | 新浪首页 | 新浪导航
手机读书 走到哪 读到哪
book.sina.cn
热点推荐
· 自选股-轻松管理您的千只股票
· 金融e路通-理财投资更轻松
· 行情中心-通往财富之门
财经首页 股票 基金 滚动 公告 大盘 个股 新股 权证 报告 环球市场 博客 股票吧 港股 美股 行情中心 自选股
上证指数 2242.833 -0.11% 853.38亿元 | 深证成指 8037.650 -0.07% 1108.49亿元 | 沪深300 2368.254 -0.27% 551.98亿元
重磅信号(11:50) 宁波海顺：沪指高位振
最近访问股 我的自选股
名称 价格(元) 张跌幅
上证指数 2241.701 -0.16%
歌华有线 11.70 -1.60%
以下为热门股票
永泰能源 5.11 -6.24%
TCL 集团 2.71 0.00
包钢股份 5.18 -1.33%
中国软件 26.37 -4.73%
苏宁云商 8.05 6.48%
成飞集成 63.00 7.23%
锌业股份 7.59 -1.68%
亚盛集团 7.27 0.83%
浙报传媒 17.31 3.96%
贵州茅台
上海 600519
160.57
1.89 1.19%
2014-08-20 13:20:06
昨收盘:158.68 今开盘:158.74 最高价:161.43 最低价:158.50
市值:1833.71亿元 流通:1833.71 成交:22042手 换手:0.19%
代码/名称/拼音 查询 代码检索
公司资料意见反馈
2014 三季度 查询
贵州茅台(600519)年季度历史交易
日期 开盘价 最高价 收盘价 最低价 交易里(股) 交易金额(元)
2014-08-19 159.750 159.750 158.680 158.450 3209577 509709632
2014-08-18 162.000 162.300 159.750 159.100 4169827 667452096
2014-08-15 162.200 162.490 161.540 161.130 2180268 352476096
2014-08-14 162.020 163.440 161.460 161.280 2295894 373085472
2014-08-13 162.040 164.780 162.490 160.800 3445403 560675712
2014-08-12 161.810 162.750 162.040 160.150 1960436 316255840
2014-08-11 162.480 163.200 163.070 161.510 2007072 326555328

腾讯财经牛刀小试


腾讯财经
股票 基金 港股 美股
收藏本页 | 帮助中心 | 股民学堂 | 官方博客 | 意见反馈
港股首页 | 涨跌排行 | 机构增减持 | 投行评级 | 新股 | 窝轮/牛熊证 | 即市新闻 | 公司新闻 | 评级新闻 | 港股论坛 | 行情中心 | 自选股
HSI
全部
股票查询
我的自选股 (0)
恒生指数:25151.19 +0.11% 403.88亿港元 | 道琼斯:16919.59 +0.48%
实时热点: 腾讯控股 中国忠旺 中国国家文化产业 金山软件 中国新能源动力 英发国际 保利协鑫能源 玖源集团 东北电气 中国移动
最近访问
我的自选
管理 退出
实时行情
行情页
恒生指数 HSI.HK
交易中
币种：港币
加入自选股
25151.19
+28.24 +0.11%
2014-08-20 13：19
最 高: 25163.20
最 低: 25056.71
今 开: 25156.95
昨 收: 25122.95
成交额: 403.9亿
振 幅: 0.42%
52周最高: 25201.21
52周最低: 21137.61
上 涨: 27家
平 盘: 0家
下 跌: 24家
分时 5日 日K 周K 月K 1年 3年 5年
价格：25151.19 涨跌幅：+0.11% 成交量：404万
时间：08/20 13:19
25188.06 25171.78 25155.51 25139.23 25122.95 25106.67 25090.39 25074.12 25057.84
+0.26% +0.19% +0.13% +0.06% 0.00% -0.06% -0.13% -0.19% -0.26%
9:30 10:00 11:00 12:00/13:00 14:00 15:00 16:00
其他股指
更多>>
指数名称 点数 涨跌幅
国企指数 11044.97 -0.45%
公用指数 0.00 0.00%
金融指数 0.00 0.00%
地产指数 0.00 0.00%
创业板 519.86 -0.95%
红筹指数 4912.76 +0.14%
成交额占大市成交比
窝轮成交
牛熊证成交
股票及其他成交

炼数成金逆向收费式网络课程

- Dataguru（炼数成金）是专业数据分析网站，提供教育，媒体，内容，社区，出版，数据分析业务等服务。我们的课程采用新兴的互联网教育形式，独创地发展了逆向收费式网络培训课程模式。既继承传统教育重学习氛围，重竞争压力的特点，同时又发挥互联网的威力打破时空限制，把天南地北志同道合的朋友组织在一起交流学习，使到原先孤立的学习个体组合成有组织的探索力量。并且把原先动辄成千上万的学习成本，直线下降至百元范围，造福大众。我们的目标是：低成本传播高价值知识，构架中国第一的网上知识流转阵地。
- 关于逆向收费式网络的详情，请看我们的培训网站 http://edu.dataguru.cn

Thanks

FAQ时间