

第五届泰迪杯数据挖掘挑战赛竞赛

——《通用论坛正文提取》赛题解读

互动派科技-苏淦

题目背景：

1. 数据爆炸

- 每半年产生的数据，即以往数据总量之和

2. 互联网数据价值巨大

- 舆情分析（事件、活动）
- 社会调查（企业、品牌、产品、人群）
- 决策支撑（营销、生产、渠道）

3. 面临挑战

- 海量数据挖掘需要海量的抓取、存储、计算能力
- **但是：** 真正有效**信息分散**在半结构化的HTML页面中、**提取成本高**



网易首页 > 新闻中心 > 新闻 > 正文

重要信息

什么事情让王健林、董明珠和刘强东感到耻辱？

2017-02-16 23:25:28 来源: 新京报即时新闻(北京)

⚠ 举报

分享到：



📄 2149

(原标题：什么事情让王健林、董明珠和刘强东感到耻辱?)

即便没有奶茶妹妹光环加持，京东CEO刘强东本人也堪称一台巨大的“话题制造机”。最近，因为公开“晒三观”，爱说话的东哥又火了一把。

春节期间，这位大佬在自己的京东集团内部走了一次基层。面对随行的央视的镜头，刘强东没有聊到“你幸福吗”的经典话题，反而大谈起了“耻辱”。

什么是耻辱？什么是企业家视之为耻辱的事情？公司秘闻（ID：high3c）梳理了

交友

广告、噪声



JD.COM 京东



一个高中理科生的求助帖

楼主: i玖夜 时间: 2017-02-16 00:02:00 点击: 188 回复: 23

- 守护
- 脱水
- 打赏
- 看楼主
- 设置

现在高一刚刚分了文理科 数理化特别差 但是我觉得只要肯努力就一定会可以学好 所以选择了理科 可是现在才学了两天就发现自己完全学不懂化学 而物理也仅仅是能够听懂老师再讲什么 想要理转文可是教务处年级组长不让转科怎么办? 希望有经验的学长学姐能帮助一下我 谢谢了

楼主发言: 6次 发图: 0张 | 更多

来自 天涯社区客户端 | 举报 | 分享 | 楼主

天涯赏金牛人

春节榜单

未知学院

股市论谈

原创写手



打碗面汤



洋补牢



90歌手小千



Cindy小宝



老猫批八字

楼主: i玖夜 时间: 2017-02-16 00:02:48

现在的自己真的特别迷茫 都怪自己太任性了 但是还是希望得到别人的建议和帮助

来自 天涯社区客户端 | 举报 | 1楼 | 打赏 | 回复 | 评论 | 赞

作者: 还看今朝扬 时间: 2017-02-16 07:48:31

问题&目标：

Q: 对于任意提供的BBS类型的网页(或url)，获取其HTML源码内容，设计一个能够**智能地**提取该页面主贴、所有回帖的算法，自动提取该页面的**所有主贴和回帖**的下列核心字段：

- 用户名
- 标题（回帖若无可为空）
- 帖子内容
- 帖子的发布日期

问题&目标：

规则：

1. 不依赖于特定网站、或网页模板
2. 简化场景，只考虑对静态HTML的渲染后的结果进行提取
 - 降低对Ajax、js动态生成页面的处理难度



作品回顾 & 企业收获：

1. 集思广益 — “好犀利”

- 效果好的参赛作品，虽然不依赖“特定网站”模板，本质还是离不开“规则库”的构建
- 企业从参赛作品里面，看到了很多构建“规则库”的新花样，或者对这个方法论的数学证明。

2. 抛砖引玉 — “求简历”

- 做数据采集类的赛题比较少见，提取数据这个话题看似简单，但是要做到“通用”却很难，实际考虑场景更加复杂
- 大数据行业对海量的互联网数据的高效抓取和提取需求很大，越来越多人开始重视互联网数据的综合运用，这方面的人才需求也较大（快到碗里来~~）

3. 企业代表 — “棒棒哒”

- 荣幸地能作为企业代表之一参与到“泰迪杯”全国数据挖掘竞赛里面

Thank You !