斩断黑链

-机器学习实战

安全组

郭添森

2014/10

去哪儿? Qunar.Com 聪明你的旅行

关于我

- 2001/05 ~ 2011/05 艺龙网
  - 技术部/OPS：网络、系统、安全等
- 2011/05 ~ 现在 去哪儿网
  - 技术部/安全组：信息安全相关

什么是黑链

【北京┣电话记录彻底删除┣ 请加Q:97809598 敞血 ┗┣┣附近...

请加Q:97809598 敞血 ┗┣┣附近酒店, 北京┣电话记录彻底删除┣ 请加Q:97809598 敞血 ┗┣┣附近酒店预订, 北京┣电话记录彻底删除┣ 请加Q:97809598 敞血 ┗┣┣附近酒...

去哪儿网 - review.qunar.com/tag... 2014-07-27 - 快照 - 81%好评

北京》举报电话记录表→ 请加Q:97809598 狠遏丶「」附近酒店...

预定, 北京》举报电话记录表→ 请加Q:97809598 狠遏丶「」附近酒店查询 登录 不限 ¥200以下 ¥200-¥300 ¥300-¥500 ¥500以上 自定义 - 不限 经济型 二星级/其他 三星...

去哪儿网 - hotel.qunar.com/city... 2014-08-11 - 快照 - 57%好评

【北京▼时时彩稳赚人工后一怎么选号最准【总代Q78188883...

北京▼时时彩稳赚人工后一怎么选号最准【总代Q781888839大户必备】▿附近酒店, 北京▼时时彩稳赚人工后一怎么选号最准【总代Q781888839大户必备】▿附近酒店预订...

去哪儿网 - review.qunar.com/tag... 2014-05-26 - 快照 - 81%好评

【北京去哪儿机票怎么退票【400-822-9077】附近酒店】北京...

北京去哪儿机票怎么退票【400-822-9077】附近酒店详细介绍,北京去哪儿机票怎么退票【400-822-9077】附近酒店用户点评;为您预订北京去哪儿机票怎么退票【400-822-90...

去哪儿网 - review.qunar.com/tag... 2014-07-26 - 快照 - 81%好评

【北京去哪儿网人工退票服务热线是什么【400-822-9077】附...

北京去哪儿网人工退票服务热线是什么【400-822-9077】附近酒店, 北京去哪儿网人工退票服务热线是什么【400-822-9077】附近酒店预订, 北京去哪儿网人工退票服务热线是...

去哪儿网 - review.qunar.com/tag... 2014-07-22 - 快照 - 81%好评

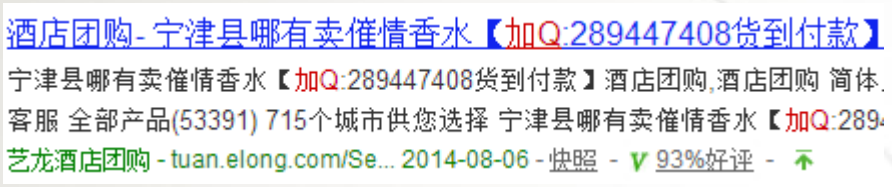
三哗仑哪里出售╬Q╝194562353╬oi1团购网站-三哗仑哪里出售╬Q╝...

共查询到0条"三哗仑哪里出售╬Q╝194562353╬oi1" 相关团品。您可以尝试下面的推荐产品正在团购(0) 往期团购(0) 1/1 3.9折 [酒店]三亚 | 三亚海翔缘...

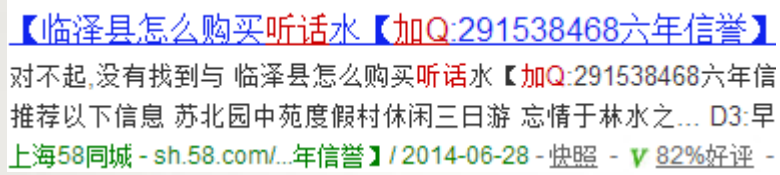
tuan.qunar.com/team/se... 2014-09-08 - V3 - 百度快照 - 评价

Qunar.Com

冰山一角

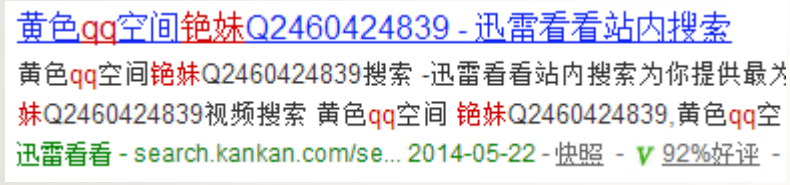


http://tuan.elong.com/SearchResult?q=宁津县哪有卖催情香水【加Q:289447408货到付款】



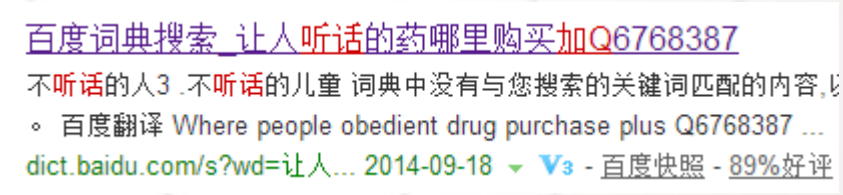
http://sh.58.com/huangpu/jdyd/jh_临泽县怎么购买听话水【加Q:291538468六年信誉】/

http://www.ebay.com/sch/sis.html?_nkw=临沂哪里有卖听话迷药【加Q:836812200顾客至上马/



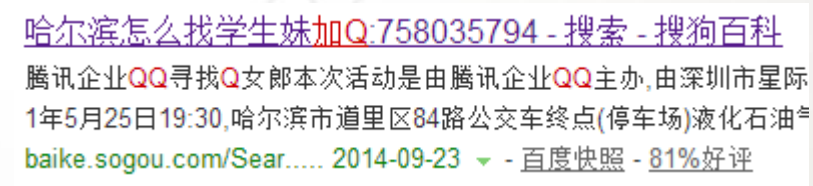
http://search.kankan.com/search.php?keyword=黄色qq空间【铯妹Q2460424839】&t=0

http://www.amazon.com/s/ref=nb_sb_noss?field-keywords=漳县催情香水女用 ┌【加Q:416804054八年信誉★☆货到付款



http://dict.baidu.com/s?wd=让人听话的药哪里购买加Q6768387

http://www.yy.com/index/s?wd=失忆听话水加Q6768387



http://baike.sogou.com/Search.e?sp=S哈尔滨怎么找学生妹加Q:758035794

Qunar.Com

黑链套餐	链接时间	本站特价	服务说明
黑链套餐[A]	若不被发现永久有效	30元[全球最低]	〖100个PR0~6的链接导入,质量完全随机,有好有坏,绝对超值,15/天内掉至60%以下包补一次〗
黑链套餐[B]	若不被发现永久有效	50元[全球最低]	〖200个PR0~6的链接导入,质量完全随机,有好有坏,绝对超值,15/天内掉至60%以下包补一次〗
黑链套餐[C]	若不被发现永久有效	100元[全球最低]	〖500个PR0~6的链接导入,质量完全随机,有好有坏,绝对超值,15/天内掉至60%以下包补一次〗
黑链套餐[D]	若不被发现永久有效	30/元[全球最低]	〖20个PR2~3的链接导入,绝对超值,15/天内掉至60%以下包补一次〗
黑链套餐[E]	若不被发现永久有效	30/元[全球最低]	〖12个PR4的链接导入,绝对超值,15/天内掉至60%以下包补一次〗
黑链套餐[F]	若不被发现永久有效	30/元[全球最低]	〖8个PR5的链接导入,绝对超值,15/天内掉至60%以下包补一次〗
特别公告	特别公告	特别公告	因本站价格底,有同行散布谣言恶意讲我们收钱不给货,在此申明本站所有交易首次为支付宝担保交易货到确认
本站业务可担保交易	暂无	暂无	〖唯一联系QQ：355311400 〗〖每天新货绝不重复,重一赔十,请勿拿我们价格和其他人比较〗

黑链类型

- 入侵型
  - 获取网站或服务器控制权后，发布或篡改数据
- 钓鱼型
  - 在url中
    - http://review.qunar.com/tag_beijing_city_去哪儿机票怎么退票【400-822-9077】_52.html
  - 在query string
    - http://tuan.qunar.com/team/search.php?kw=△www.xxoo.com●加淫钕Q：246４１５8460←上门●

利益链

- 在能控制的网页中嵌入钓鱼链接，发布黄色、反动、赌博、毒品、诈骗等信息，如

x

- 搜索引擎上钩，抓取&收录

  *.qunar.com的rank比较高，恶意信息排名会靠前。如果恶意信息在title或body等地方回显，危害更大。

- 黑产从业者
  - 以比较低的代价发布了违法信息，并有机会取得好的排名
- 用户
  - 自愿型：本身就有黄赌毒方面的需求，想依赖搜索引擎获取这些信息
  - 无辜型：被虚假信息欺骗进而带来财产损失，如假的客服电话
- 厂商
  - 流量浪费
  - 声誉损害
  - 合规风险

常规解法

- 在页面中不回显query string
  - 比较难符合业务需求
- 白名单
  - 运营成本vs业务需求
- 黑名单
  - 正则能跟上这些变形吗？
  - △www.xxoo.com●加淫钕Q：246４１５8460←上门●

“新”解法

- 机器学习十大经典算法之一：朴素贝叶斯
- 条件概率
  - P(A|B)表示事件B已经发生的前提下，事件A发生的概率，叫做事件B发生下事件A的条件概率。其基本求解公式为：P(A|B)=P(AB)/P(B)
- 贝叶斯定理
  - P(B|A)=P(A|B)P(B)/P(A)
- 正式定义
  - 假设F={F1,F2,…,Fn}为一个待分类项，每个F是一个特征属性
  - 有类别集合C={C1,C2,…,Cn}
  - 计算P(C1|F),P(C2|F),…,P(Cn|F)
  - 如果P(Cn|F)=max(P(C1|F),P(C2|F),…,P(Cn|F))，那么F∈Cn

分类示例

- 训练集

0 迷人景色 | 0 情人节 | 0 自来水 | 1 迷情药 | 1 迷魂水

- 训练结果

P(0)=3/5	P(1)=2/5
P(迷|0)=1/10	P(迷|1)=2/6=1/3
P(情|0)=1/10	P(情|1)=1/6
P(水|0)=1/10	P(水|1)=1/6

- 新数据分类

P(0|迷情水)=P(迷情水|0)P(0)/~~P(迷情水)~~=P(迷情水|0)P(0)=P(迷|0)P(情|0)P(水|0)P(0)=0.0006

P(1|迷情水)=P(迷情水|1)P(1)/~~P(迷情水)~~=P(迷情水|1)P(1)=P(迷|1)P(情|1)P(水|1)P(1)=0.0037

- 取自然对数，解决下溢出

Ln(a*b) = ln(a) + ln(b)

P(0|迷情水)= np.log(0.1)*3+np.log(3.0/5)= -7.4185809027481273

P(1|迷情水)= np.log(1.0/3)+np.log(1.0/6)*2+np.log(2.0/5)= -5.598421958998375

- 《机器学习实战》第四章有详细讲解和完整python示例代码

实战第一步：提取原始数据

- HiveQL + Python

```
FROM (
    FROM v4a
    MAP concat(uri,get)
    USING 'query_format_map.py'
    AS query
    WHERE
      method='GET'
      and d=$yestoday
    distribute by query
    ) map_output
  INSERT OVERWRITE DIRECTORY 'query_$yestoday'
  REDUCE map_output.query
  USING 'query_format_reduce.py'
  AS query2;
```

- 提取出来的数据类似为

莫泰上海赤峰轻纺市场
兴隆县没有副作用的安眠藥

实战第二步：人工分类

- 关键字、正则

淫妹|铯妹|骚女|一夜晴|交友|夜夜撸|粿聊|淫荡|激情|做爱|淫乱|淫欲|淫钕|黑丝|快播

大麻|性药|迷药|迷昏|迷情|迷幻|晕倒药|拍肩粉|苍蝇粉|魔鬼天使粉|乖乖粉|白粉|冰毒|毒药|说真话|迷魂|催情|安定片|安眠药|三唑侖|醚|听话

学历|老师|经理|窃听|离婚|跟踪|定位|婚外恋|驾驶证|克隆

- 分类好的训练数据类似为

0 上海赛车场

0 莫泰上海赤峰轻纺市场

1 兴隆县没有副作用的安眠藥

1 南方航空人工退票热线号码电话位置联系

实战第三步：生成唯一字列表

- Hadoop MapReduce

```
hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-2.2.0.jar \
 -input raw/${DATASTR} \
 -output out/${DATASTR}_vocab \
 -file map_createVocabList.py \
 -mapper map_createVocabList.py \
 -file reduce_createVocabList.py \
 -reducer reduce_createVocabList.py
```

- 一天的数据大概会产生一万个唯一字，数据类似为

帮 池 席 師 汤 遂 鸢 …

实战第四步：训练

- Hadoop Map/Reduce

```
hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-2.2.0.jar \
 -input raw/${DATASTR} \
 -output out/${DATASTR}_train \
 -numReduceTasks 1 \
 -file hadoop.vocab \
 -file map_train.py \
 -mapper "map_train.py hadoop.vocab" \
 -file reduce_train.py \
 -reducer "reduce_train.py hadoop.vocab"
```

- 训练生成的策略文件保存为json格式，类似为


```
{"pAb": 0.61635542683051781,
"VocabList": {"麂": 2931, "漂": 1669, "土": 663, ...},
"p0V": [-9.8788105648236879, -9.8788105648236879, -9.8788105648236879, ...],
"p1V": [-11.636133757271676, -11.636133757271676, -11.636133757271676, ...],
"cnt": 10341}
```

实战第五步：验证

- Hadoop Map/Reduce
- 验证结果

{"false_alarm_rate": "2.04%", "missing_report_rate": "0.75%", "bad": [16953.0, 2238800.0, 2255753.0], "good": [1375369.0, 28703.0, 1404072.0], "error": 0}

- 误报-把”好人”识别为”坏人”

人工分类错误，机器分类正确

result=1 expect=0 p0=-98.986217 p1=-98.259786 [拱墅区强力迷烟在哪里买得到]

人工分类正确，机器分类错误

result=1 expect=0 p0=-99.990711 p1=-94.922516 [山东省食品发酵工业研究设计院]

- 漏报-把”坏人”识别为”好人”

人工分类错误，机器分类正确

result=0 expect=1 p0=-100.003682 p1=-106.514324 [奥尔巴尼（纽约州）坦帕]

人工分类正确，机器分类错误

result=0 expect=1 p0=-100.004868 p1=-104.536461 [大武口区蝴蝶夫人催情口服液]

- 改进：重复第二、四、五步

实战第六步：拦截

- QAEGIS
  - 后端：决策中心
  - 前端：nginx模块
- 推送策略
  - 通过QAEGIS决策中心，把第四步生成的训练结果推送到nginx
- 实时拦截
  - Nginx模块，根据训练好的结果，用朴素贝叶斯算法对每个新的请求做分类，判断为恶意的即拦截
  - 每天拦截量约200多万

优缺点

- 优点
  - 朴素贝叶斯算法简单高效，对性能的影响几乎可以忽略不计
  - 保护全网所有应用，无需WEB应用配合做相关修改
- 缺点
  - 属监督学习，没教的不会，无法自我进化
  - 机器分类准确性严重依赖人工人类的准确性
  - 基于概率，不可避免总会有误报和漏报

Q&A

Qunar.Com