P2P舆情2.0产品说明书

李图209

说明

P2P舆情监控系统2.0在上一版本的基础上，主要做了以下改进

扩大了数据来源的范围：以贷罗盘和网贷天眼为载体，补充了平台曝光、投诉贴及平台交易数据等

深化了评级流程：结合抓取到的平台新闻、问题曝光、平台交易数据及用户情感分析结果，确定风险指标，对平台进行评级和聚类

增加了新的模块和功能：事件提取功能，问题平台曝光模块、平台监测预警功能

细化了数据获取与清洗步骤：如删除重复新闻，补充新闻时间戳等

改进了demo展示：完善了web端架构，丰富了可视化展示的图表，如预警平台词云展示等

目录

我们的舆情监控系统采用python和html相结合的架构，通过python抓取网页新闻并分析，分析结果由基于html的web端展现。链接如下：http://1.209litu.applinzi.com/test.html

1 数据抓取及预处理

2 文本、情感分析

3 平台评级及预警

4 可视化展现

数据来源

数据时间跨度为2014年至今，内容包括“行业新闻”12000篇、“平台新闻”10000篇、“论坛发帖”20000篇、“平台事件”“平台投诉”10000条

行业新闻

以新浪新闻为载体，收集所有与“P2P”、“互联网金融”有关的新闻

平台新闻

以新浪新闻为载体，根据网贷之家的平台数据，收集包括“拍拍贷”、“陆金所”等多个P2P平台的有关新闻

论坛发帖

以网贷之家论坛、百度贴吧为载体，收集所有与P2P的有关的讨论帖

平台数据

以贷罗盘、网贷之家为载体，收集网站上4所有注册平台近3个月的成交量、投资金额、利率、借款人数等9项数据

平台曝光

以网贷天眼为载体，收集所有平台曝光贴和平台投诉贴

数据获取

Urllib2获取网页数据、服务器JSON数据

- 一级网页直接提取
- 二级或者多级网页分步提取
- 新闻网页特殊提取

正则表达式 BeautifulSoup框架 提取需要的数据

- 指标项较少时可直接写正则表达式，简单快捷
- 类别种类较多考虑BeautifulSoup框架

数据初步处理及保存

- 正则表达式容易提取噪声信息，需要提前进行替换、删除等简单处理
- 保存以txt和csv格式为主

数据清洗

缺失值处理

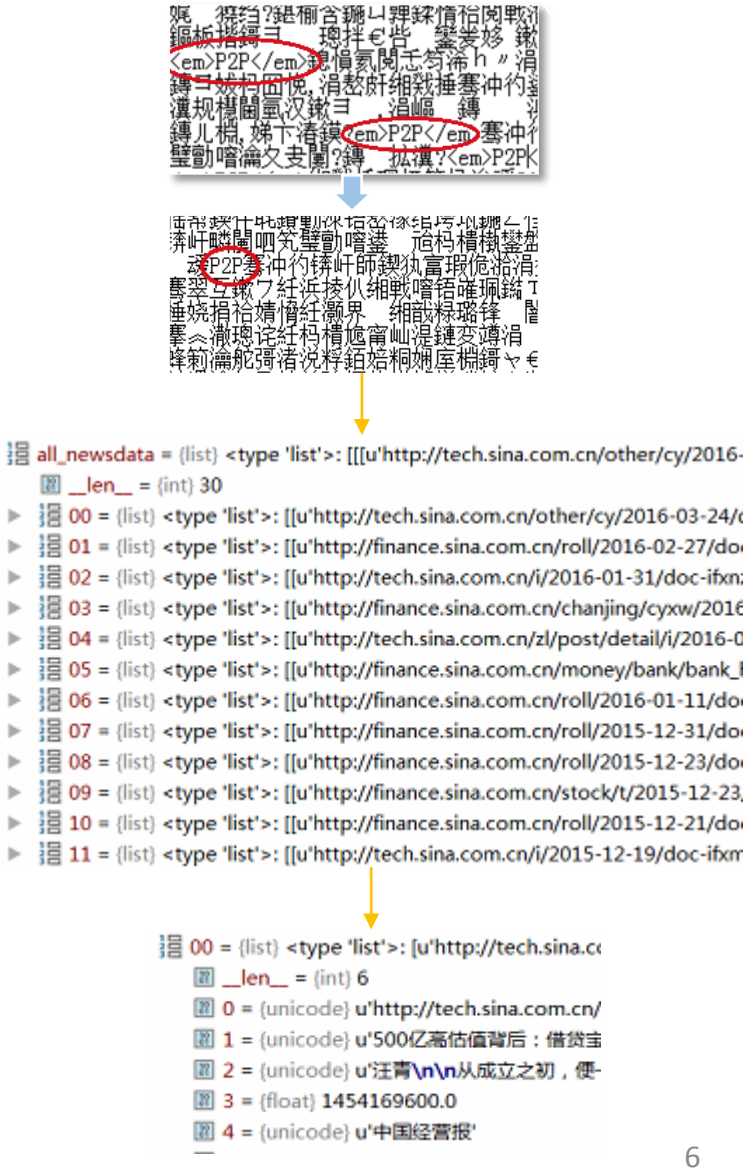
针对各程序之间既定的接口格式，在前期数据处理中使用全局变量NULL填充缺失数据。

结构化处理

针对网上数据的特点，如HTML，JSON，XML以及众多非结构化数据，根据最初定义的数据结构，分别设计转化方法将数据转化为统一的结构化数据。

数据维度规约

对来自各数据源的数据，对其进行维度规约，剔除重复冗余数据，以及进行数据文本分析等工作，剔除掉无关的数据内容。



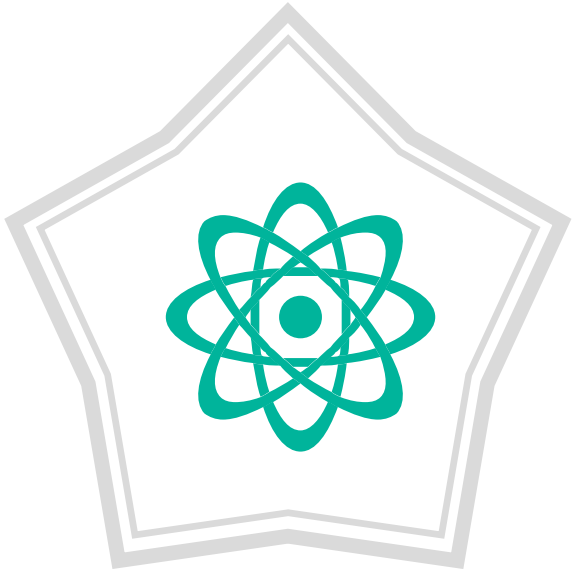
文本分析--政策、媒体及平台新闻分类

特征词提取

利用TFIDF算法提取
行业新闻文本中的关键词

新闻文本分类

基于朴素贝叶斯方法，将
新闻划分为政策规定新闻
或普通媒体报道新闻



新闻分类展示

分为政策规定、媒体以及
平台三块分别展现相应类
别下的新闻报道

文本分析--新词学习

输入文本：
- 3月17日百度搜索上所有P2P新闻
- 网贷之家上所有P2P平台名称

基于Trie树结构实现高词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图（DAG)

运用动态规划查找最大概率路径, 找出基于词频的最大切分组合

对于现有词典中的未登录词，使用HMM模型和Vertibi算法进行识别

识别出的部分新词

关键词	词频	词性
P2P	987	n
网贷	239	n
315	38	n
e租宝	14	nt
CEO	9	n
Online	8	a
Finance	8	n
360	8	nt
微信	7	n
曝出	6	v
年化	6	n
小编	5	n
12315	5	n
开鑫贷	5	nt
线上	5	a

部分平台名称

平台名称	词频	词性
陆金所	1	nt
红岭创投	1	nt
链家理财	1	nt
鑫合汇	1	nt
微贷网	1	nt
你我贷	1	nt
搜易贷	1	nt
团贷网	1	nt
投米网	1	nt
爱钱进	1	nt
翼龙贷	1	nt
融金宝	1	nt
易贷网	1	nt
聚有财	1	nt
拍拍贷	1	nt

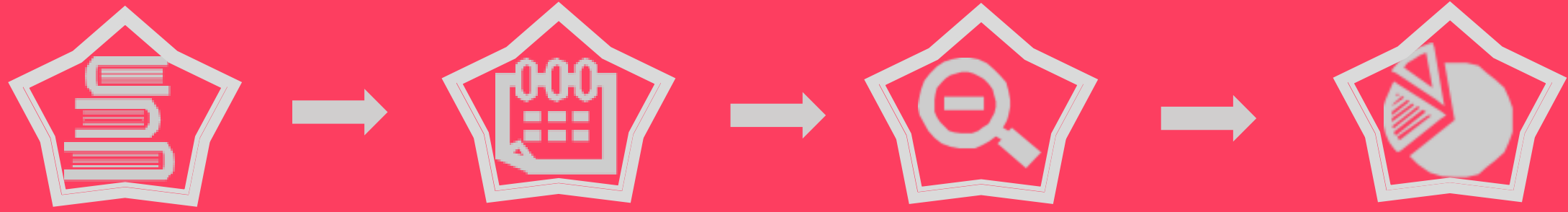
文本分析--新闻聚类

每周行业新闻的聚类结果

时间	新闻标题	热度	新闻总数
2016/3/27	第三届互联网金融全球峰会北大论坛将于4月举办	59	804
2016/3/20	万亿“三农”市场将成互联网金融新蓝海	2	181
2016/3/13	赖小民代表：互联网金融监管 “管机构”不如“管业务”	37	216
2016/3/6	新华社：互联网金融迎来规范发展年	52	166
2016/2/28	P2P行业被误解 专家：真正的P2P不会出问题	29	121
2016/2/20	汽车金融迎来互联网新势力：后来者下沉渠道至县	1	71
2016/2/11	近百家上市公司涉足P2P 揭秘含金量最高平台	11	71
2016/2/4	P2P平台傍信托一元起售产品 或不合规	1	165
2016/1/28	地产信托行业流行新玩法 拥抱互联网金融	2	323
2016/1/21	P2P借贷平台“融资城”被立案侦查	9	348
2016/1/13	1月12日全国P2P网贷平台20排行榜	9	217
2016/1/6	拆解7个平台4个“跑路”的P2P投资“雷点”	59	181
2015/12/30	2015互联网金融监管破局：短板与护城河	17	290
2015/12/23	P2P“监管时代”倒计时	49	166

文本分析--热点分析

通过文本聚类获得每天、每周、每月各个平台和总体行业的热点新闻



新词学习

以3月17日200条P2P行业新闻作为训练集，通过HMM算法识别新词并存储至字典

时间分类

将行业或平台新闻按照1天、7天、30天分类，统计各个时间段的新闻量

关键词提取

运用“Jieba分词”提取新闻中有代表性若干的关键词及其权重

新闻聚类

根据向量空间模型，余弦夹角最小的两个新闻将被看做同一话题，话题中新闻数量最多者为热点

文本分析--事件提取

数据清洗与分段：删除文中的空格及特殊符号，对新闻正文进行断句处理

词性标注：将句子分词并标注词性（以现代汉语语料库为参考），删除不符合主谓宾结构的句子并设定不同权重，如标题的权重大于正文中的句子

模式提取：以上一部分的句子集合作为数据集，“jieba分词”提取的关键词作为一项频繁项集，运用FP-tree算法找到关键词的在文章中出现次数最多的组合

模式匹配：在原文中搜索对应的关键词组合，调整组合顺序并输出主谓宾结构的句子

由前一部分的新闻聚类我们发现，通过分词工具得到的关键词只是一系列离散的词汇，无法体现出新闻的主题，而有些新闻标题的结构往往不够规范，经常会出现多个不同标题的新闻涉及同一事件的情况。因此，我们补充了事件提取的功能

情感分析--用户口碑分析

口碑分析

数据来源：网贷之家论坛等相关评论

主要内容：分析行业、各平台的各时间粒度的口碑值，提取用户对行业及各平台印象的特征

1 情感特征库

利用带标签评论集，基于朴素贝叶斯算法构建正面与负面情感特征库

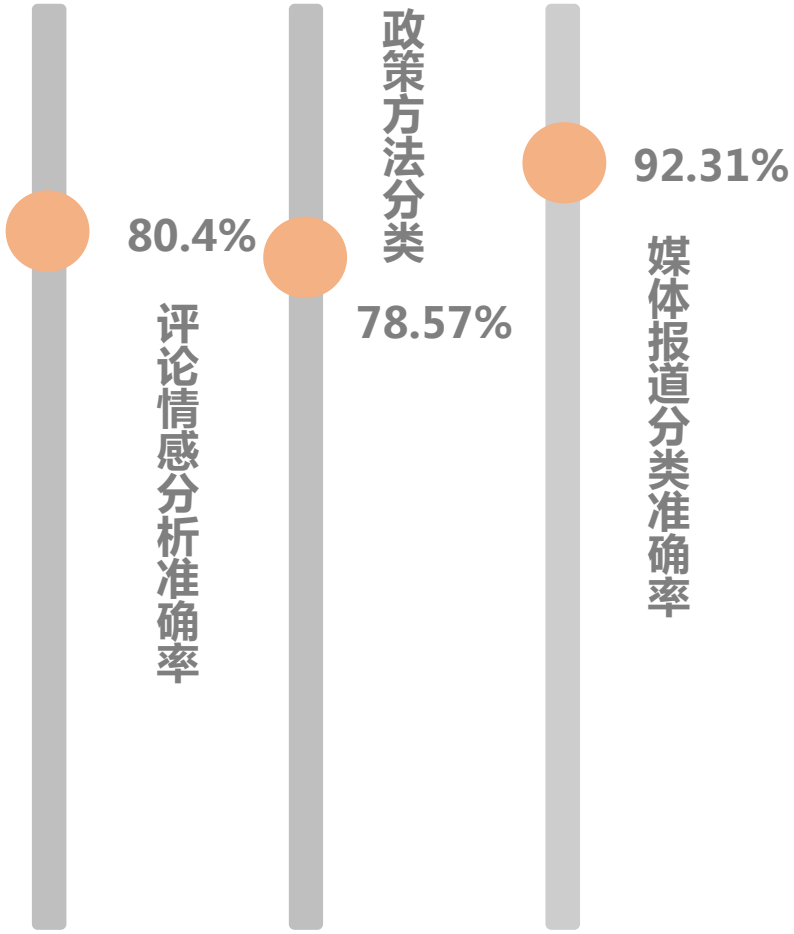
2 口碑计算

基于朴素贝叶斯算法，对评论进行情感分类。在此基础上对行业所有评论以及各平台评论按照给定的时间粒度分类，分析行业及平台的口碑值

3 用户印象

基于情感模式的分析，提取用户点评中对于行业及平台印象的特征

情感分析--用户口碑分析



口碑实例分析

选取184条用户评论，算法结果：准确率=80.4%，召回率=85.1%

行业用户口碑：56.5%用户持正面态度

行业用户印象：利率低（278），体验不错（175），资金安全（134）

新闻分类实例分析

政策方法新闻分类准确率：78.57%

行业媒体报道分类准确率：92.31%

平台评级与分类

根据网贷之家提供的100个平台数据，运用主成分分析和聚类分析对各大平台进行评级

1.指标选取
- 数据标准化处理，检验变量间的相关性
- 通过碎石图确定主成分个数

2.相关性检验及主成分个数设定

3.确定主成分及因子旋转
- 获取每个主成分中各个指标的权重值
- 因子斜交旋转使得各个指标尽可能包含在某个主成分中

4.计算平台各主成分的得分及总评分
- 计算各平台得分，根据各个总得分按照正态分布五等分级法将所有平台分成AA, A, B, C，D 五个等级
- 对每个等级的平台进行描述分析

5.根据每个主成分对平台进行聚类分层
- 对于每个主成分，根据包含变量的特征，赋予标签
- 通过权重设定，将所有平台聚类分为不同层次，建立筛选机制

- **各个平台的交易数据及固有属性**：月成交量，平均利率 ，平均借款期限，投资人数 ，借款标数,，注册资本， 满标用时 ,借款人数，人均投资金额 ，人均借款金额,，资金杠杆，运营时间，信息透明度等
- **各个平台的评论数据**：问题事件次数，好评率

平台评级结果展示

主成分展示

成分标签	人气类标签	风险类标签	规模类标签
1	成交量	平均利率	成交量
2	投资人数	运营时间	注册资本
3	借款人数	平台透明度	借款人数
4	资金杠杆	好评率	满标用时
5	平均借款期限	问题事件次数	

部分平台评级

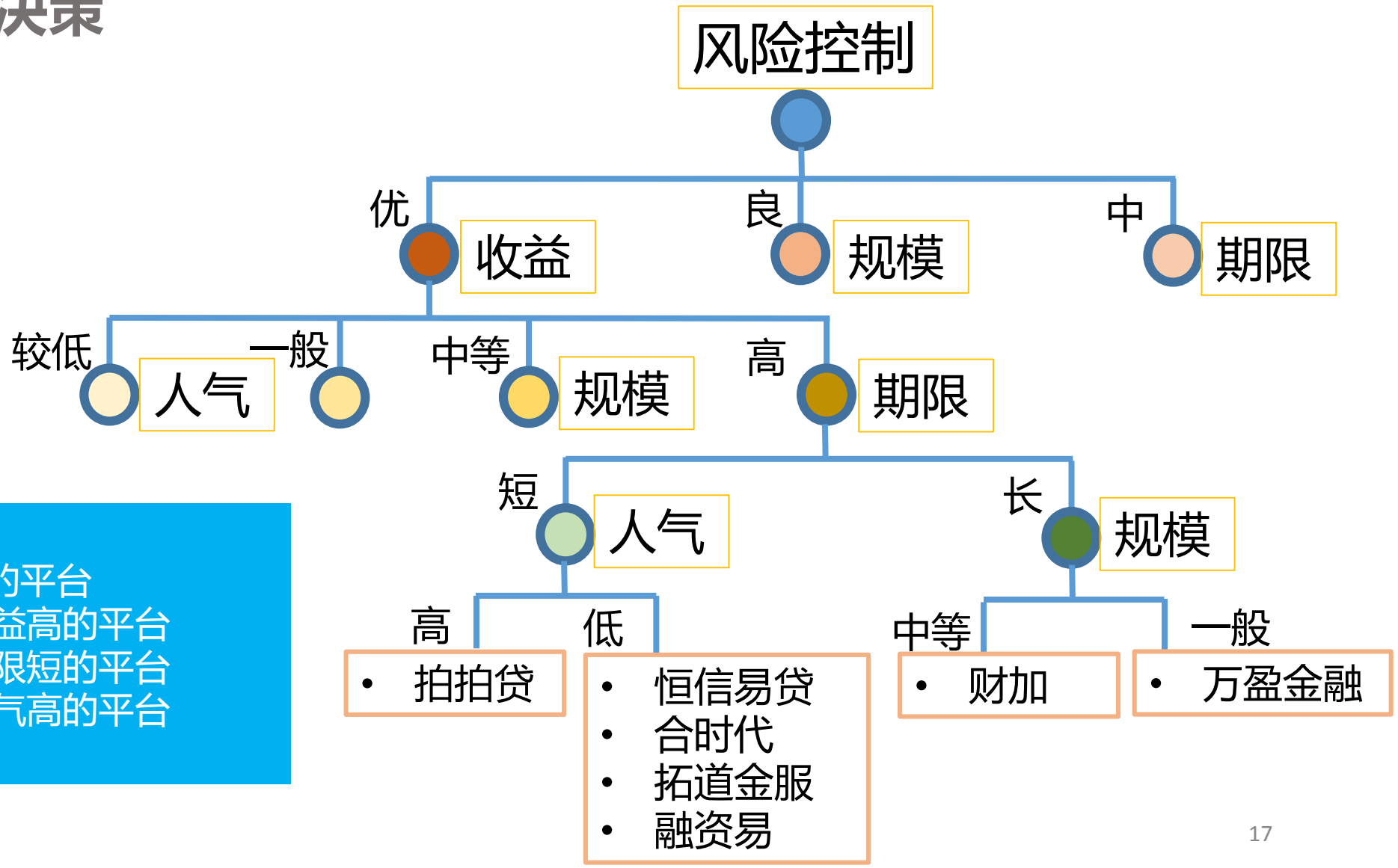
排名	名称	平台人气	风险控制	平台规模	总得分	评级
1	拍拍贷	65.14	105.76	69.08	78.78	AA
2	爱钱进	86.31	65.29	61.18	74.03	AA
3	陆金所	67.87	62.1	96.11	72.44	AA
4	红岭创投	76.98	57.74	69.5	69.25	AA
5	你我贷	81.74	57.89	57.1	68.69	AA
6	宜人贷	85.37	53.76	45.66	66.47	AA

平台聚类结果展示

排名	名称	平台人气	风险控制	平台规模	平台收益	收益期限	评级
1	拍拍贷	高	优	较大	高（13-20%）	短（0-3个月）	AA
2	爱钱进	高	优	较大	高（13-20%）	很长（大于24个月）	AA
3	陆金所	高	优	极大	较低（6-9%）	很长（大于24个月）	AA
4	红岭创投	高	优	较大	较低（6-9%）	短（0-3个月）	AA
5	你我贷	高	优	较大	中等（11-13%）	很长（大于24个月）	AA
6	宜人贷	高	优	一般	中等（11-13%）	很长（大于24个月）	AA
7	聚宝匯	一般	中	极大	一般（9-11%）	长（9-24个月）	A
8	有利网	高	优	一般	一般（9-11%）	长（9-24个月）	A

- 平台人气分为高、一般、低三个等级；风险控制分为优、良、中三个等级；平台规模分为极大、较大、中等、一般四个等级；平台收益分为高、中等、一般、较低四个等级；收益期限分为很长、长，较短，短四个等级进行区分
- 评级AA表示特别推荐级，各个指标排名相对靠前，对于新手或者保守安全型客户为主
  评级A表示一般推荐级，可针对有经验的客户进行合理选择
  评级B,C,D级在人气、风控和规模上都有一定劣势，一般不推荐，需要详细了解后再进行决策

平台评级—模拟决策



模拟决策步骤
1. 首先选择风险控制为优的平台
2. 在 1 的基础上，考虑收益高的平台
3. 在 2 的基础上，考虑期限短的平台
4. 在 3 的基础上，考虑人气高的平台
5. 确定目标平台

平台预警分析

平台预警

从各平台大量的交易信息和用户评论信息中分析出平台是否有跑路或出现其他问题的风险，让用户做出更合理的投资、撤资等决策

1 负面消息监控

基于负面内容的监控机制，可直接获取各平台当前的负面新闻消息内容以及各论坛社区投诉事件问题的程度

2 基本指标监控

从3000余家平台近12个月的成交量、投资金额、利率、借款人数等9项数据中，选取平均借款期限、人均借款金额和利率等三项数据进行实时监控，利用离群点分析方法进行异常预警

平台预警--负面新闻实时监控

在新闻事件追踪的基础上，对各平台发生的新闻事件进行情感分析，提取情感关键词，通过关键词的情感特征进行分类，追踪各平台实时的负面新闻事件，从而使得用户了解各平台最新负面新闻报道

新闻情感分类举例

时间	新闻事件	正负面
2016-3-9	央行下属中国互联网金融协会3月25日上海挂牌	非负面
2016-3-23	伪国资系P2P安心金融官网打不开 客服称停业	负面
2016-4-8	警钟为你敲响，“中晋系”为何招摇过市无人监管	负面
2016-4-11	易乾财富被指庞氏骗局 多项信息成谜	负面
2016-4-12	PPmoney陈宝国赴美考察 布局全球资产配置	非负面

平台预警--平台问题事件监控

对爬取自网贷天眼的平台曝光贴，平台投诉贴和平台事件贴进行监控分析。用基于规则以及语义分析的方法，提取文本中曝光的关键内容，并将问题划分为三类，程度上从高到低依次为一级、二级、三级

一级问题：已被证实的平台跑路停业型
- 发布停业公告
- 连续多日联系不上，法人失联
- 确认跑路

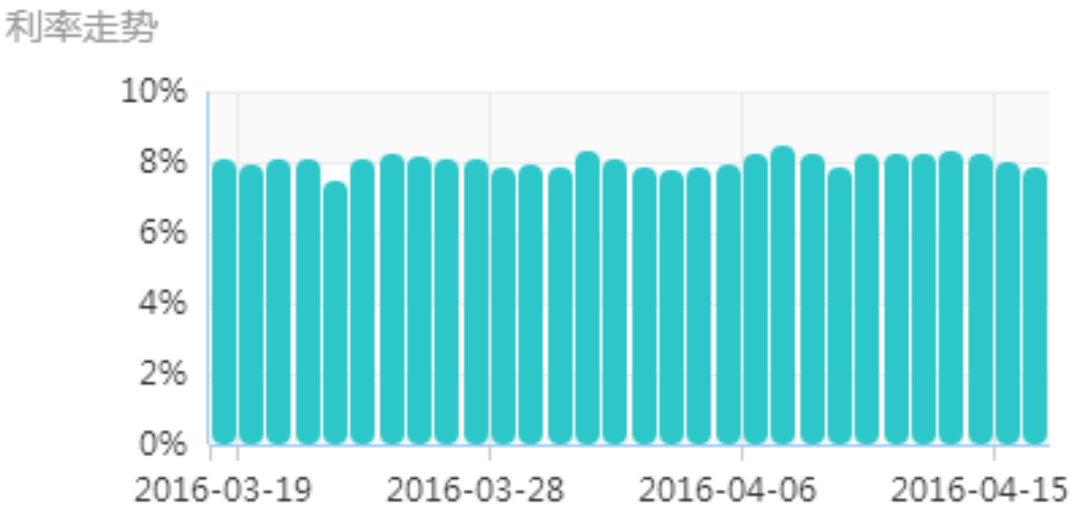
二级问题：用户严重投诉型
- 提现不到账、提现困难
- 电话无人接
- 网站突然打不开、登录不了

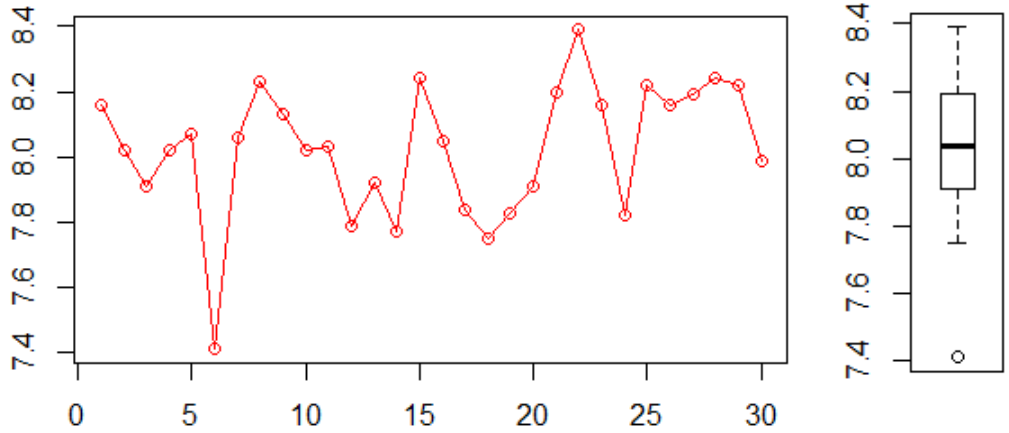
三级问题：用户个人抱怨型
- 注册等活动承诺未兑现
- 不诚信，服务态度差

时间	平台	事件问题	预警程度
2016/4/12	远虑财富	连续多日联系不上	一级预警
2016/4/12	信诺金融	网站长期无法访问	一级预警
2016/4/12	中融国润资本	网站长期无法访问	一级预警
2016/3/16	新联在线	提现不到账	二级预警
2016/3/16	红岭创投	自融	二级预警
2016/2/6	宝点网	提现不到账	二级预警
2016/2/6	壹佰金融	电话无人接	二级预警
2016/2/23	凤凰金融	有挂羊头的嫌疑	二级预警
2016/4/4	网利宝	注册不给红包、体验差	三级预警
2016/3/2	诺诺镑客	财神爷爷推广活动不诚信	三级预警
2016/3/18	鑫合汇	耍赖奖励不兑现	三级预警

平台预警--基本指标监控

实时监控3000余家平台每日的平均借款期限、人均借款金额和利率等三项数据，以过去一个月的数据作为基准，利用离群点分析方法进行异常预警





- 以陆金所过去1个月利率为研究对象，通过散点图和箱线图得到陆金所利率范围在7.75%-8.4%之间，下四分位数为7.91%，上四分位8.2%，中位数8.04%
- 以超过上下边界的利率将作为异常值进行预警，状态控制分为绿色（安全正常），黄色（提醒处于边界周围），红色（预警超过边界）

平台预警

- 根据问题平台的研究分析，发现当一个平台容易发生停业跑路状况时，平均利率设置较高，借款金额较大，借款期限较短，因此，希望通过此类预警指标结合用户评论投诉事件，对各个平台进行风险预警
- 以陆金所为例，以2016/3/18至2016/4/15的数据为基准，平均利率预警界限是7.75-8.4%，平均借款金额界限是1.6-7.6万元，平均借款期限界限是27.7-35.3个月

时间	平均利率	平均借款金额	平均借款期限	负面新闻数量	问题投诉	结论
2016/4/10	7.82	4.17	29.19	0	无	--
2016/4/11	8.22	5.6	32.89	0	无	--
2016/4/12	8.16	6.67	32.03	0	无	--
2016/4/13	8.19	6.45	32.2	0	无	--
2016/4/14	8.24	6.76	32.95	0	无	--
2016/4/15	8.22	7.07	33.16	0	无	--
2016/4/16	7.99	5.17	30.07	0	无	正常
2016/4/17	7.78	3.6	26.56	0	无	借款期限有微弱减少

可视化展示--前期作品

通过搜索新闻文本中的对应关键词，将舆情分为政策新闻、行业新闻、评论三类，将近期舆情展示在web端首页

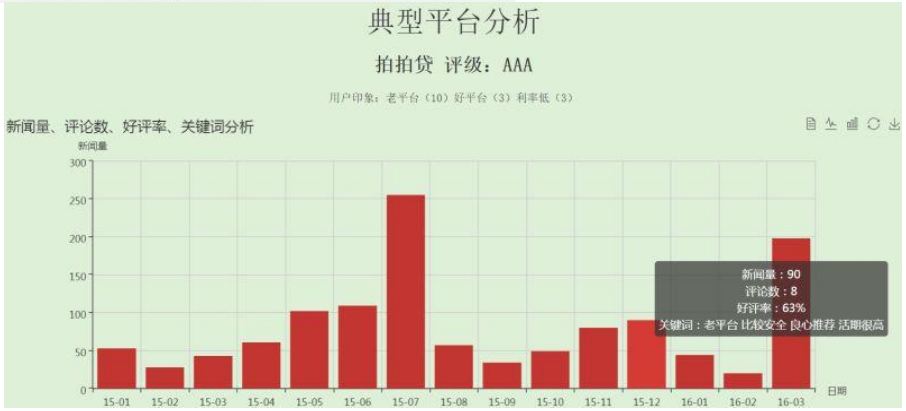
近期舆情分类展示

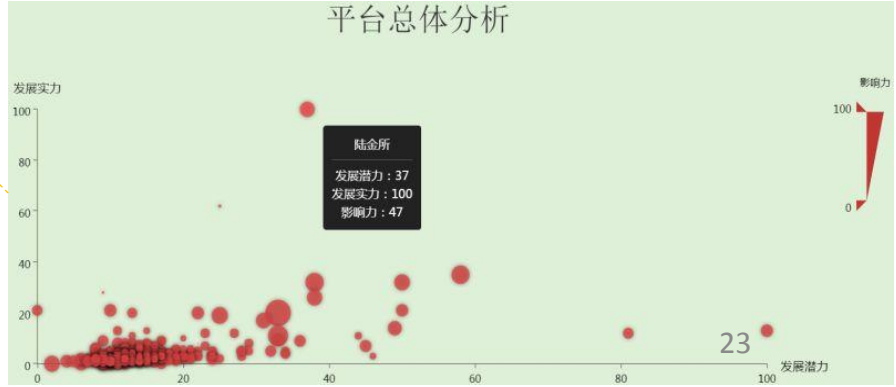
行业分析

平台分析

包括行业新闻量趋势，行业热点新闻、行业关键词

包括平台热点新闻趋势、用户情感分析、平台分级等





可视化展示--补充模块

对一个月以来出现跑路问题的平台进行词云展示

平台曝光

平台预警

平台聚类

实时监控平台交易数据并进行异常值预警

对100个平台进行分析评级与聚类

平台总体分析

平台综合评级

排名	名称	人气	风控	规模	收益	平均收益期限	综合得分	综合评级
1	拍拍贷	65.14/高	105.76/优	69.08/较大	高	短	78.78	AA
2	爱钱进	86.31/高	65.29/优	61.18/较大	高	很长	74.03	AA
3	陆金所	67.87/高	62.1/优	96.11/极大	较低	很长	72.44	AA
4	红岭创投	76.98/高	57.74/优	69.5/较大	较低	短	69.25	AA
5	你我贷	81.74/高	57.89/优	57.1/较大	中等	很长	68.69	AA
6	宜人贷	85.37/高	53.76/优	45.66/一般	中等	很长	66.47	AA
7	聚宝匯	54.02/一般	45.76/中	101.07/极大	一般	长	62.06	A
8	有利网	77.09/高	52.24/优	44.91/一般	一般	长	62.02	A
9	积木盒子	55.36/一般	67.14/优	62.5/较大	较低	较短	60.67	A
10	人人贷	66.99/高	57.13/优	52.04/中等	一般	很长	60.52	A
11	诺诺镑客	59.4/一般	59.91/优	57.89/较大	中等	很长	59.22	A
12	麻袋理财	64.3/高	54.18/优	51.26/中等	中等	很长	58.18	A
13	微贷网	56.15/一般	60.11/优	58/较大	一般	短	57.81	A
14	翼龙贷	64.38/高	48.43/良	54/中等	一般	长	57.03	A

2 平台交易数据监控分析

日期	平均利率	平均借款金额	平均借款期限	负面新闻数量	问题投诉	结论
2016/4/10	7.82	4.17	29.19	0	无	--
2016/4/11	8.22	5.6	32.89	0	无	--
2016/4/12	8.16	6.67	32.03	0	无	--
2016/4/13	8.19	6.45	32.2	0	无	--
2016/4/14	8.24	6.76	32.95	0	无	--
2016/4/15	8.22	7.07	33.16	0	无	--
2016/4/16	7.99	5.17	30.07	0	无	正常
2016/4/17	7.78	3.6	26.56	0	无	借款期限有微弱减少

谢谢！