昆仑镜 - P2P舆情系统

数据匠团队

用一颗匠心打造数据产品

Demo http://120.24.245.17:8086/home

团队介绍

江少华：队长，负责项目规划、NLP处理、后台功能实现

中国人民大学硕士，目前就职于蚂蚁金服。参加过IJCAI Competition，阿里的移动推荐、穿衣搭配比赛，职位预测竞赛竞赛等，分别获第4、亚军、第5、季军。

吕　超：负责网络爬虫编写、结构化数据

北京大学硕士，目前实习于微软亚太研究院。参加TREC2014微博检索，TREC2015微博时间线，均为冠军。

闵大为：负责网站搭建、前端

浙江大学硕士，主要研究交通仿真、交通拥堵指数计算等交通问题。参加过阿里的公交线路客流预测比赛（142名），获得过国家奖学金、校优秀毕业生。

孙志远：负责后台平台部分功能实现

中国科学院硕士，主要研究数据挖掘、大数据处理与架构。曾参与亿级PV平台数据收集存储、计算架构设计与搭建，完成千万用户产品恶意信息过滤系算法设计。

市场调研

网贷之家
WWW.WDZJ.COM

融360
RONG360.COM

- 界面杂乱
- 非垂直化
- 无UGC数据
- 功能简单
- 数据单一

其　他　　和讯网、P2P观察网、网贷天眼等

项目背景与定位

存在问题

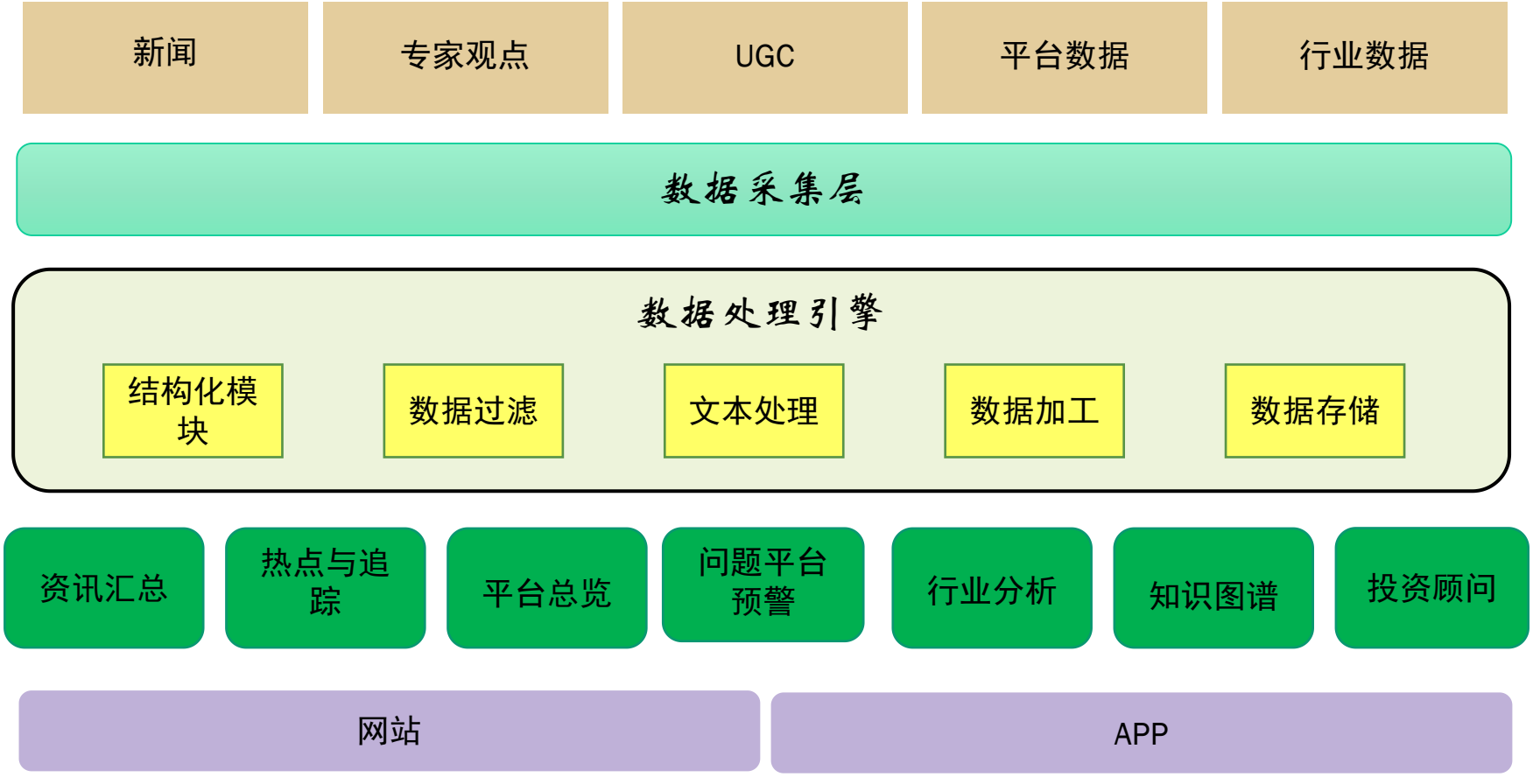
- 平台跑路、歇业
- 投资风险大
- 平台信息不透明
- 用户不了解平台
- 平台不了解用户的想法
- 政府不了解市场的真实情况

解决方案

用户需求

- 资讯阅读
- 风险识别
- 资金安全
- 专家、资深用户投资建议
- 其他

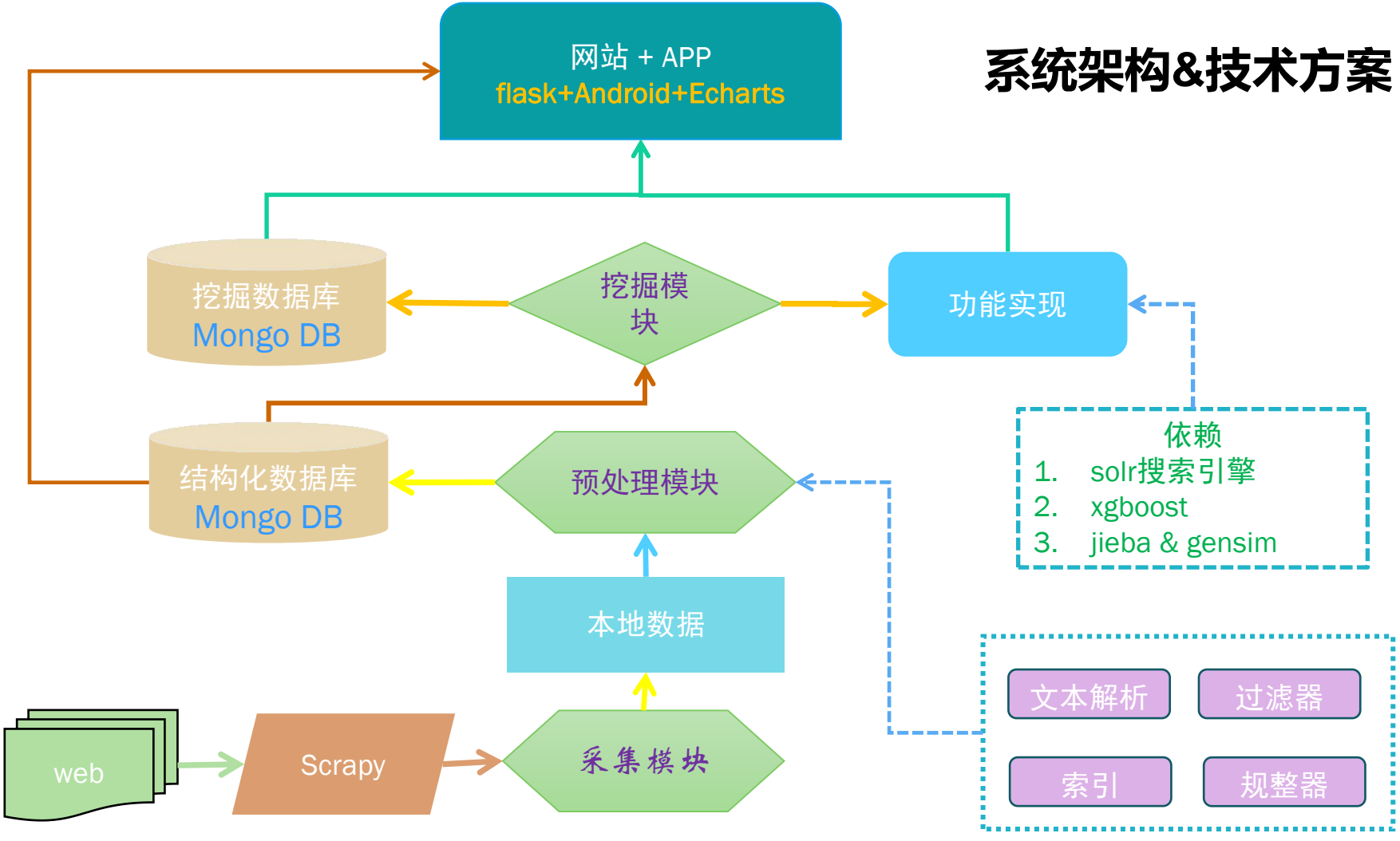
整体解决方案


新闻
专家观点
UGC
平台数据
行业数据
数据采集层
数据处理引擎
结构化模块
数据过滤
文本处理
数据加工
数据存储
资讯汇总
热点与追踪
平台总览
问题平台预警
行业分析
知识图谱
投资顾问
网站
APP

项目爆点

1. 运用**大量NLP算法**做文本数据处理，包括N-Gram语言模型、**Word2Vec、Doc2Vec、LDA**等算法。生成针对P2P行业的分词词典、情感分析算法、**UGC质量分算法、热点话题**检测算法。
2. 系统功能完备，在赛题要求的基础上增加**投资顾问、知识图谱搜索**模块。
3. 在核心痛点-**资讯、平台产品、问题平台剖析**这3个需求上投入大量精力，深度开发。
4. 数据源丰富，包括难以采集的**知乎、微博、微信**数据，增量更新，满足**实时性**
5. 系统模块间耦合性低、可扩展性强
6. 系统运用**Scrapy、Mongo DB**、轻量级web框架**flask、solr**等高性能框架进行开发，开发成本低，性能高
7. 优秀的数据可视化
8. 提供线上**网站&APP**供用户访问

系统架构&技术方案


网站 + APP
flask+Android+Echarts
挖掘数据库
Mongo DB
挖掘模块
功能实现
结构化数据库
Mongo DB
预处理模块
依赖
1. solr搜索引擎
2. xgboost
3. jieba & gensim
本地数据
文本解析
过滤器
索引
规整器
web
Scrapy
采集模块

一、数据采集模块

二、预处理模块

三、数据挖掘模块

四、网站、APP与可视化

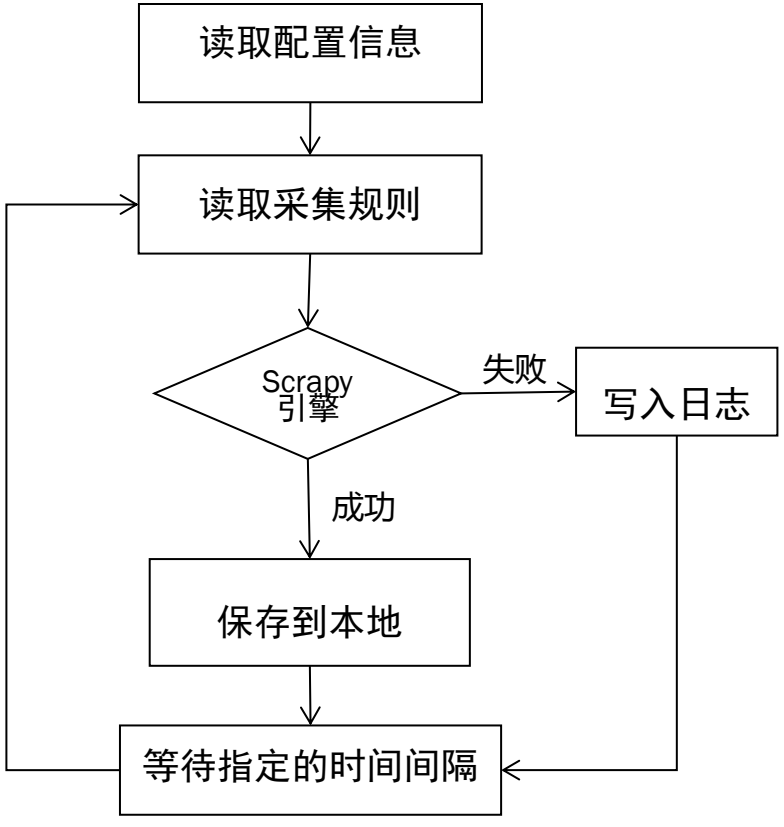
数据源

1. P2P相关主题的文本数据：
   - 媒体新闻
   - 微信公众号文章
   - 专家观点
   - 用户评论
   - 知乎问答
   - 国家政策等

需满足：
多样性好、数据量大
可结构化
实时性好、针对性好

1. 可收集的平台经营数据、行业交易数据
2. 平台百科等知识库数据

数据采集



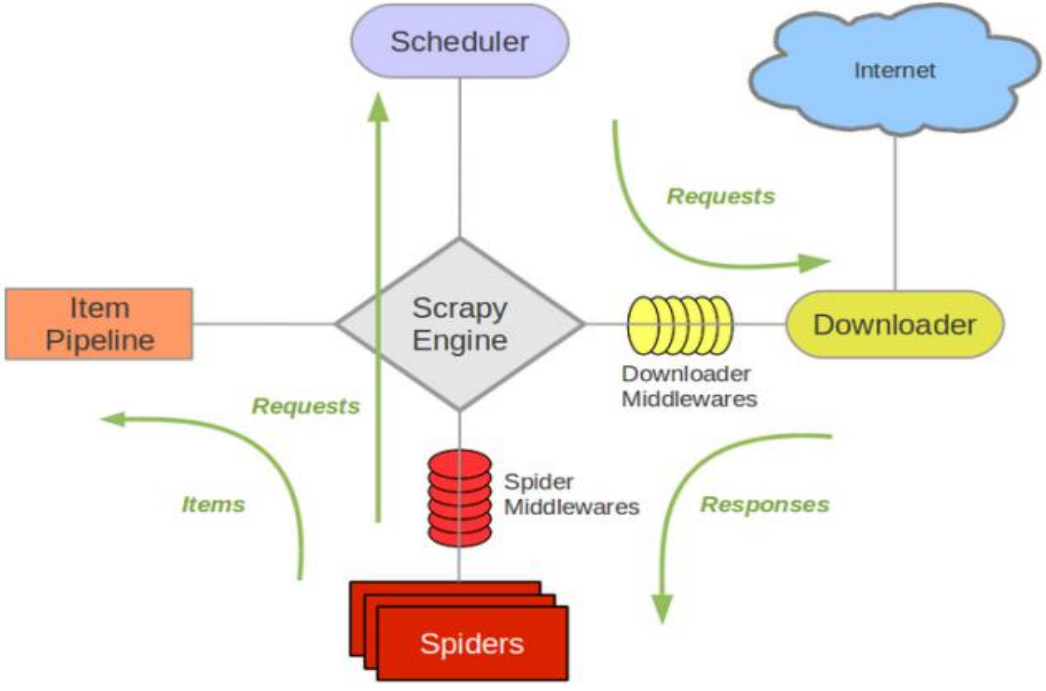
a) 新闻类
- 金融之家
- 和讯网
- P2P观察网
- 微信公众号
- 网贷之家-行业(3300篇)
- 网贷之家-平台(1100篇)

b) 用户评论、观点类
- 天眼论坛，整个版块抓取
- 用户分享
- 融360 P2P论坛
- 关于P2P的微博、评论
- 网贷人论坛
- 知乎，相对专业的评论

c) P2P平台&公司
- 网贷天眼网贷平台汇总 (4189个)
- P2P观察网平台信息
- 融360，P2P平台评级

d) 国家政策
- 和讯网政策版块
- 网贷之家政策版块(340篇)

Scrapy


Scheduler
Internet
Requests
Item Pipeline
Scrapy Engine
Downloader
Downloader Middlewares
Requests
Items
Spider Middlewares
Responses
Spiders

一、数据采集模块

二、预处理模块

三、数据挖掘模块

四、网站、APP与可视化

文本解析器

xpath + beautifulsoup

过滤器

UGC质量分计算公式

$$Q = 0.7 * ngram\_w + 0.1 * len\_w + 0.1 * sen\_sim\_w + 0.1 * bad\_word\_w$$

保留 $Q > 0.01$，约10%

• badbadbad	3.12739740596e-05
• 抢个沙发	6.25279481192e-05
• 乐贷通，官方投资群298771722，全国统一电话4007117177	0.0007
• 少见多怪平台做促销活动超短期的福利而已	0.0020002
• 36%的收益，你敢投吗？今早上班无聊走窜了一些平台的APP，偶然发现一个平台顿时吓死宝宝了	0.008
• 实地考察平台，在投资之前也可以去实地考察一下，这样比较有安全感并且也可以更多的了解公司的情况	0.014047

数据归整器

- 杂乱文字、符号过滤
- 分类
- 添加标签
- 添加索引
- 统一编号

一、数据采集模块

二、预处理模块

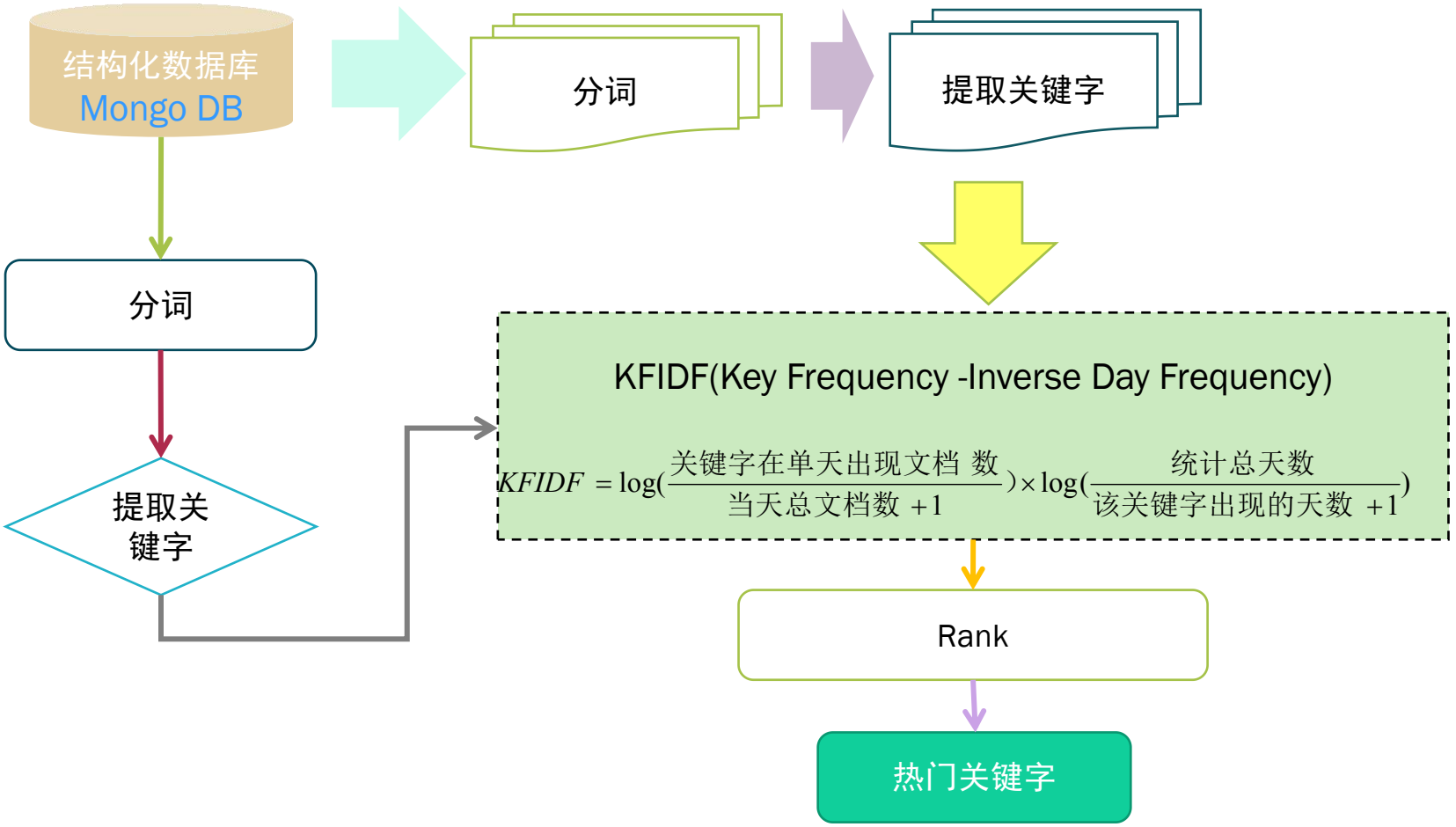
三、数据挖掘模块

四、网站、APP与可视化

资讯汇总

- 今日（本周、本月）关键字版块
- 热门话题版块
- 资讯（新闻、政策、专家观点、UGC）

关键字

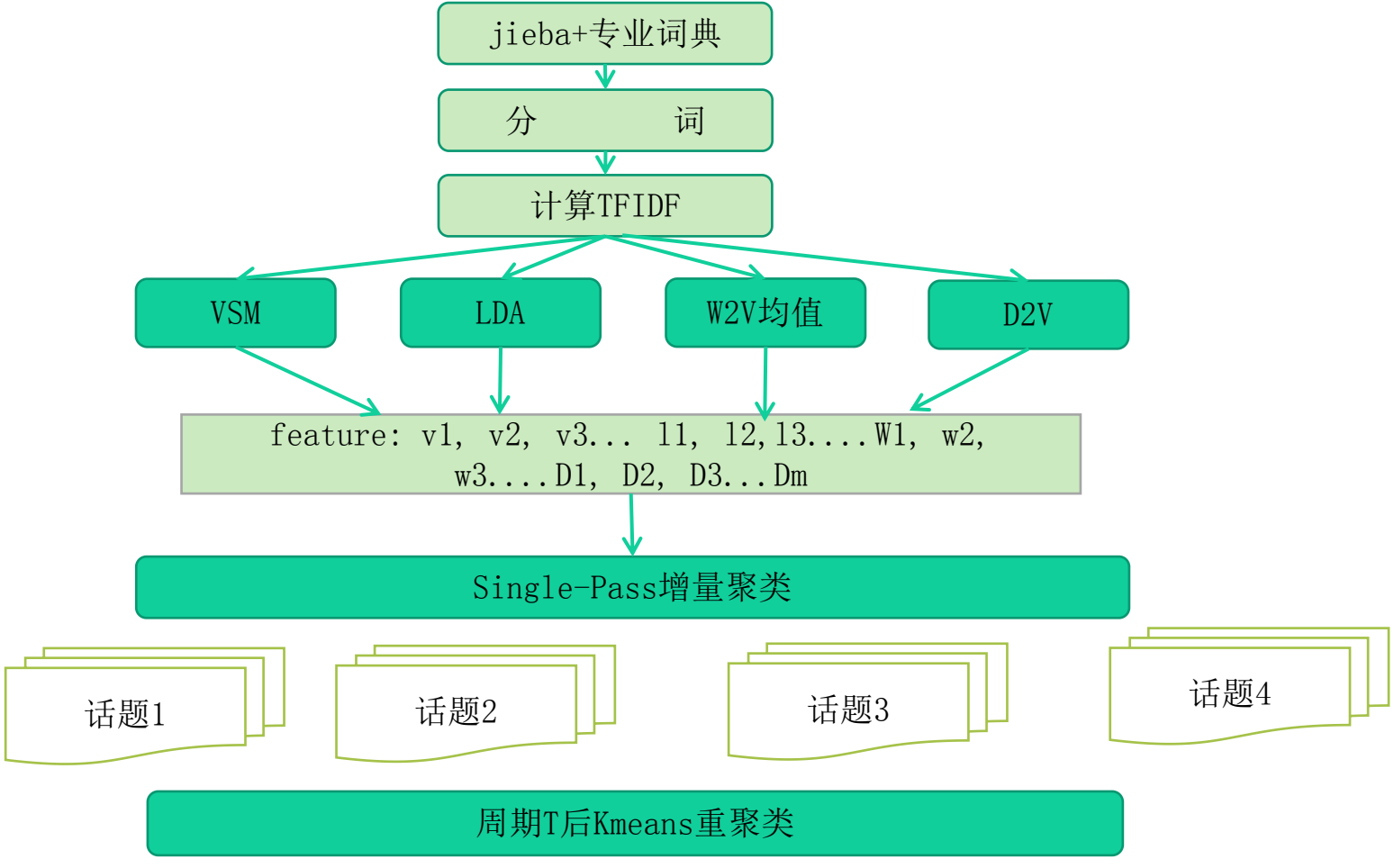

结构化数据库
Mongo DB
分词
提取关键字
分词
提取关
键字
KFIDF(Key Frequency -Inverse Day Frequency)
$KFIDF = \log(\frac{\text{关键字在单天出现文档 数}}{\text{当天总文档数 }+1}) \times \log(\frac{\text{统计总天数}}{\text{该关键字出现的天数 }+1})$
Rank
热门关键字

热门话题发现

预处理

特征提取

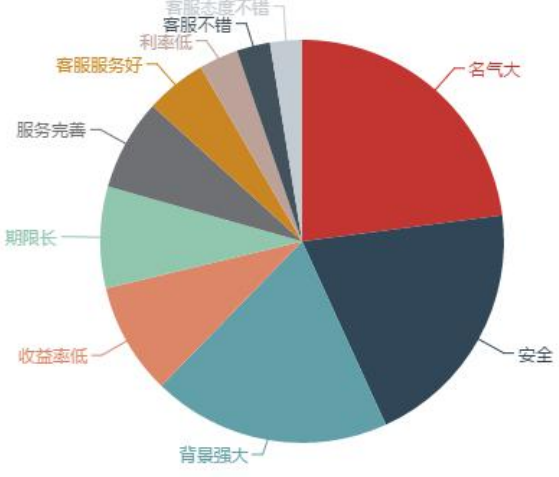
聚类模块



平台详情

基础信息 + 用户评价 + 近期新闻

评级	A
人气指数	100
平均收益	6.82%
平台背景	上市公司背景
上线时间	2012-03-02
平均利率	7.44%
成交量	5815.97万元
平均借款期限	22.94月
累计待还金额	3663421.73万元



iPhoneSE亟待认可 陆金所、光大分利宝、点融网实力认证 New
聚焦博鳌 陆金所、光大分利宝、拍拍贷响应监管 New
互金协会未排斥网贷平台 陆金所等为常务理事单位 New
兴业前行长李仁杰出任陆金所董事长 计葵生任CEO New
互金协会一年会费上亿元!一图看懂陆金所、宜信等10余家理事单位... New
2016安全稳健理财平台财猫网VS陆金所 New
陆金所当选中国互联网金融协会常务理事 New
iPhoneSE热度大减 陆金所、惠投无忧、点融网获更多关注 New
"百万任我行"登陆陆金所 平安人寿、陆金所联手打造互联网保险 New

问题平台

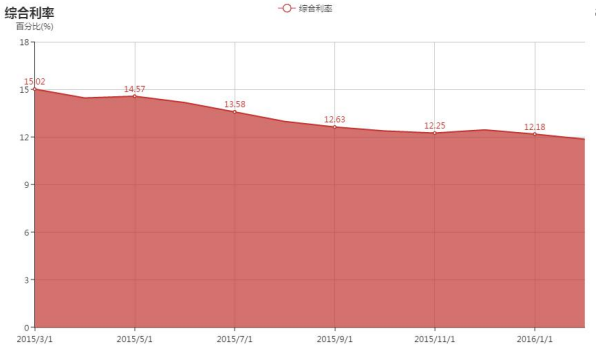
历史问题平台剖析

- 排行榜
- 深度分析
- 可视化展示

未来问题平台预测

- 二分类预测问题

行业总览



综合利率走向

+

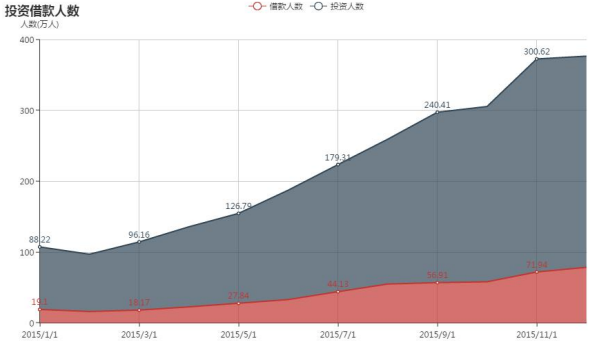
行业分布

+

行业规模多维度

+

行业热度

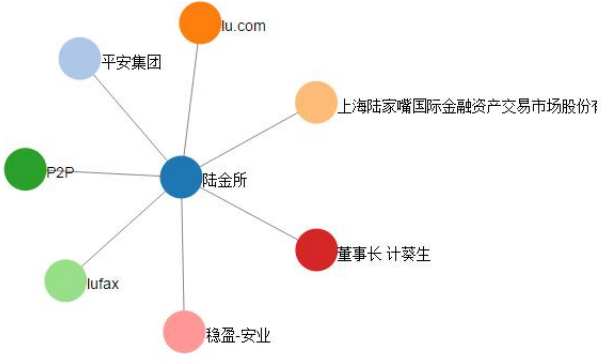


Echarts 可视化

搜索

Solr + 知识图谱


Apache
Solr


lu.com
平安集团
上海陆家嘴国际金融资产交易市场股份
P2P
陆金所
董事长 计葵生
lufax
稳盈-安业


More images
Barack Obama
Recent posts on Google+
图像
文本
社交主页
28 MILLION AMERICANS

投资顾问

V1.0 问卷测试+推荐

例如： 根据您以往投资的经验，当平台的资金被分配到高风险的股票或是其他不确定收益的项目中时，您通常：

A. 非常焦虑
B. 有一些焦虑
C. 完全放心

通过测试获知用户风险承受能力

V2.0 基于用户行为、兴趣的个性化推荐

通过用户在网站上的点击、浏览资讯与产品、评论搜索等信息判断用户兴趣点、投资点，再结合UGC数据中的一些专业投资建议，给出推荐产品。

用户导向、个性化、专业

一、数据采集模块

二、预处理模块

三、数据挖掘模块

四、网站、APP与可视化

用户界面

网站开发

- 前端：HTML5 + CSS + JavaScript+JSON
- 后台：Python轻量级Web应用框架Flask
- 服务器：阿里云ECS

APP开发

- 开发android系统的APP

项目规划与进展

已开发

待跟进

项目一期 进度：100%

周期：2月28日~3月5日

- 市场调研
- 组建团队、确定分工
- 框架设计、模块划分、确定各模块技术方案
- 构建简单的本地demo、新闻数据爬取

项目二期 进度: 80%

周期：3月6日~3月31日

- 新闻数据、专家观点、UGC数据、平台数据、行业数据的爬取
- 后台功能开发
- 搭建完备、可访问的网站

项目三期 进度:未开始

周期：4月1日~4月18日

- 增加产品数据、微信公众号文章抓取
- 网站爬虫增加定时增量爬取逻辑
- 优化图谱搜索模块，扩展知识库
- APP开发
- 系统优化
- 调研是否需要spark平台
- 调研是否接入第三方平台数据

谢谢