昆仑镜 - P2P舆情系统

数据匠团队

网站链接 APP下载

团队介绍

江少华: 项目&产品规划、NLP处理、后台、美工

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: 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: 网站搭建、前端设计

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:后台平台部分数据收集、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。

目录

项目背景

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

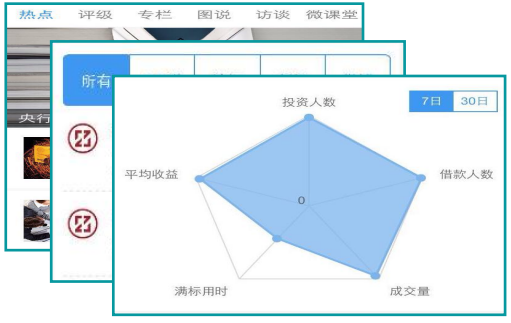
舆情产品

用户需求

- 资讯阅读
- 风险识别
- 了解行业现状
- 关注平台动态、口碑
- 专家、资深用户投资建议
- 交流

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
- 舆情分析不深入
- 平台、行业无洞察
- 无问题平台分析
- 数据单一

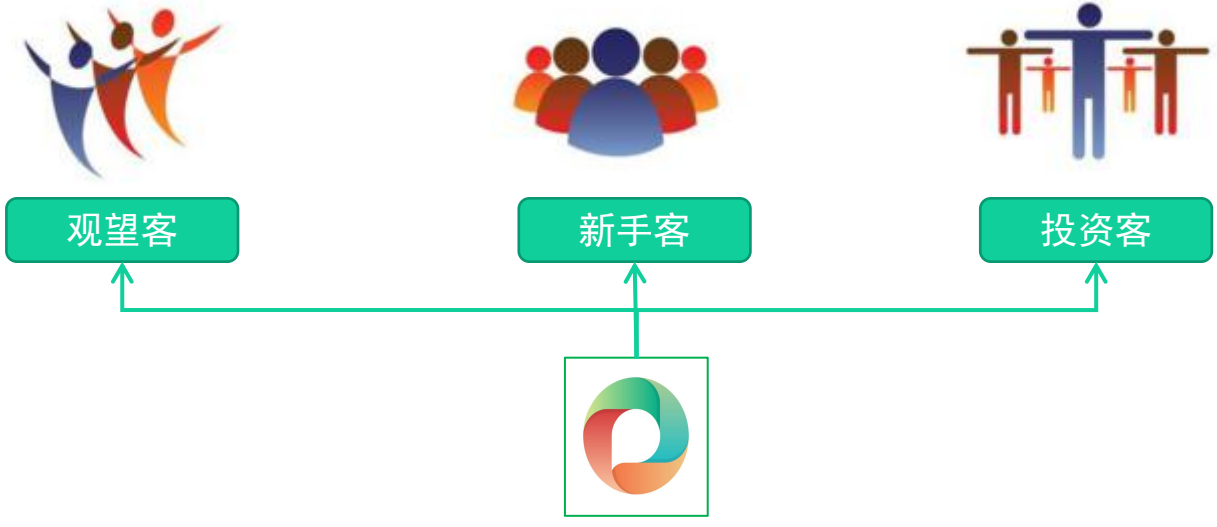
其　他　和讯网、网贷天眼、乐思舆情等

产品定位

- ✓ 了解P2P行业
- ✓ 找靠谱平台
- ✓ 找适合产品
- ✓ 投资顾问

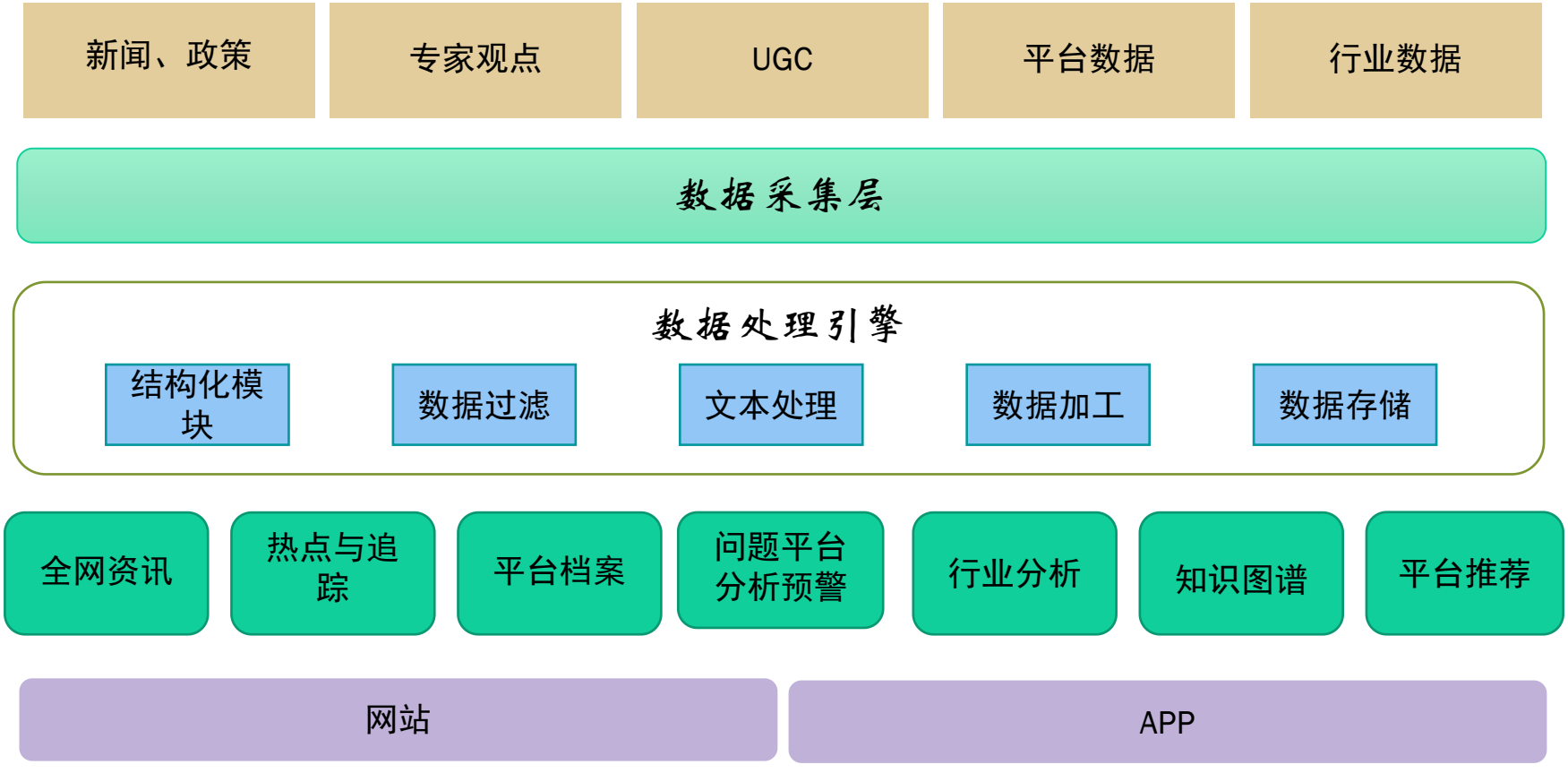
- ✓ 全方位了解关注平台
- ✓ 产品比对
- ✓ 相关资讯浏览

- ✓ 平台舆情平台
- ✓ 健康分
- ✓ 日常资讯阅读
- ✓ 关注新产品



昆仑镜，洞见一切

整体解决方案


新闻、政策
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
全网资讯
热点与追踪
平台档案
问题平台分析预警
行业分析
知识图谱
平台推荐
网站
APP

功能亮点

1. 舆情大盘 全面、深入的**行业舆情**分析监控，可定制化的**多维度平台数据可视化**，让投资者方寸之间，洞察网贷行情。

2. 全网资讯 汇聚新浪、网易、凤凰网、财新网、金融之家等**门户网站新闻**，网贷之家、融360等**论坛帖子**，微信、微博、知乎等**社交资讯**。

3. 热点话题 利用**机器学习与大数据技术**自动发现热点话题，对热点话题资讯进行专题统计分析，并进行后续追踪，让你全方面把握业界热点。

4. 平台档案 提供便捷的**平台导航**，运营数据、用户口碑评论、舆情关键字等**多维度数据**，并据此综合评估，让投资者更快更准地了解平台。

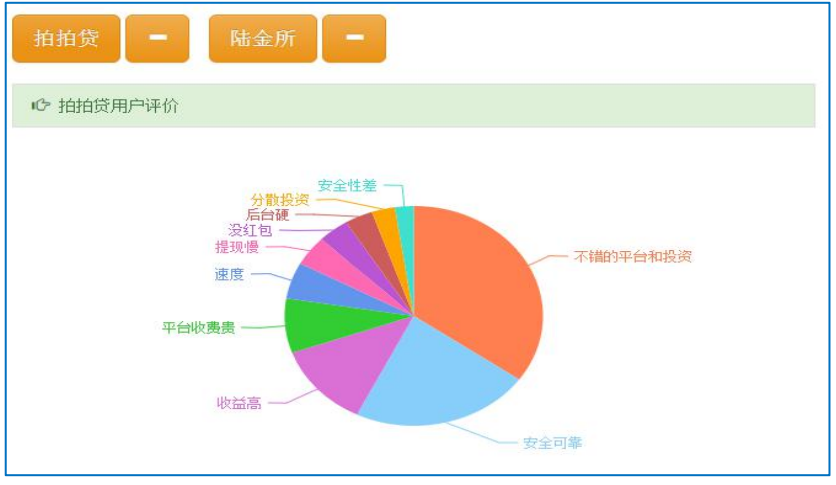
功能亮点

5. 舆情防雷 对历史问题平台**多维度分析**，对平台发生问题前的舆情走向、运营情况、高层变动、产品变动等信息进行量化、建模，并对当前活跃平台计算**健康分**。

6. 投资顾问 为新手推荐指标最佳的平台，同时提供一套**风险问卷**，测试投资者的风险承受能力、投资风格，从而推荐合适的平台。

7. 网站+APP **网站端**提供全面、专业的系统功能，**APP端**提供轻巧、关键的系统功能，全方位满足投资者理财需求。

舆情大盘



定制化的关注平台版块

- ✓ 成交量年图
- ✓ 利率年图
- ✓ 资金流图
- ✓ 舆情关键字
- ✓ 官网日均UV、PV

行业洞察

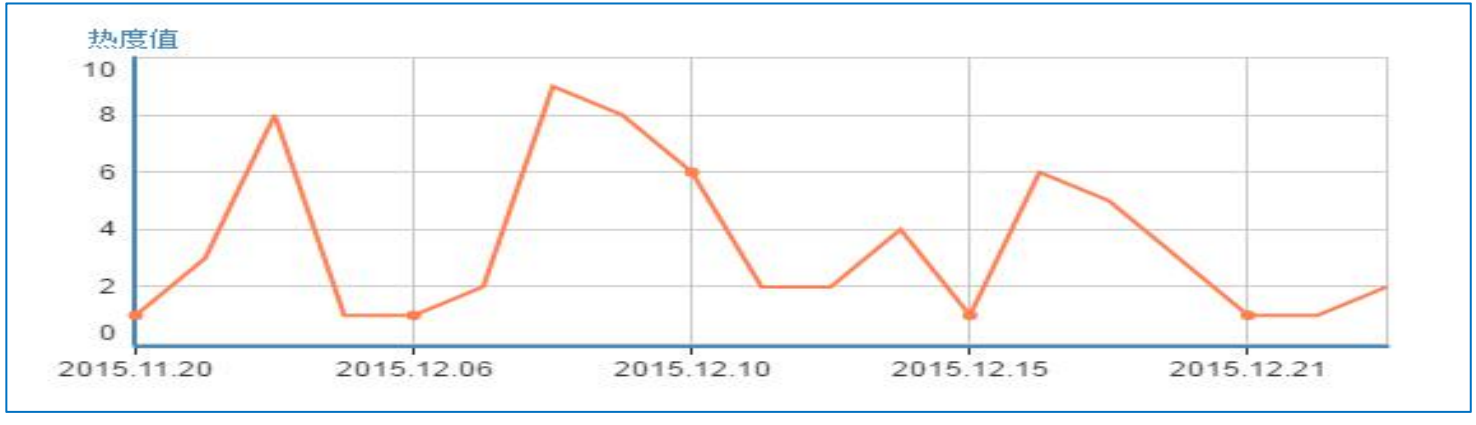
- ✓ 舆情动态图
- ✓ 平台分布图
- ✓ 投资人数变化图
- ✓ 借款金额汇总图
- ✓ 综合利率变化图

全网资讯

- ✓ 热点话题入口
- ✓ 今日/本周/本月热点词汇
- ✓ 新闻分类展示
- ✓ 资讯涉及平台展示
- ✓ 资讯标签展示（平台高层变动、融资、提现困难、平台跑路）

热点话题

- ✓ 热度变化趋势分析
- ✓ 话题关键词
- ✓ 话题追踪
- ✓ 话题资讯汇总

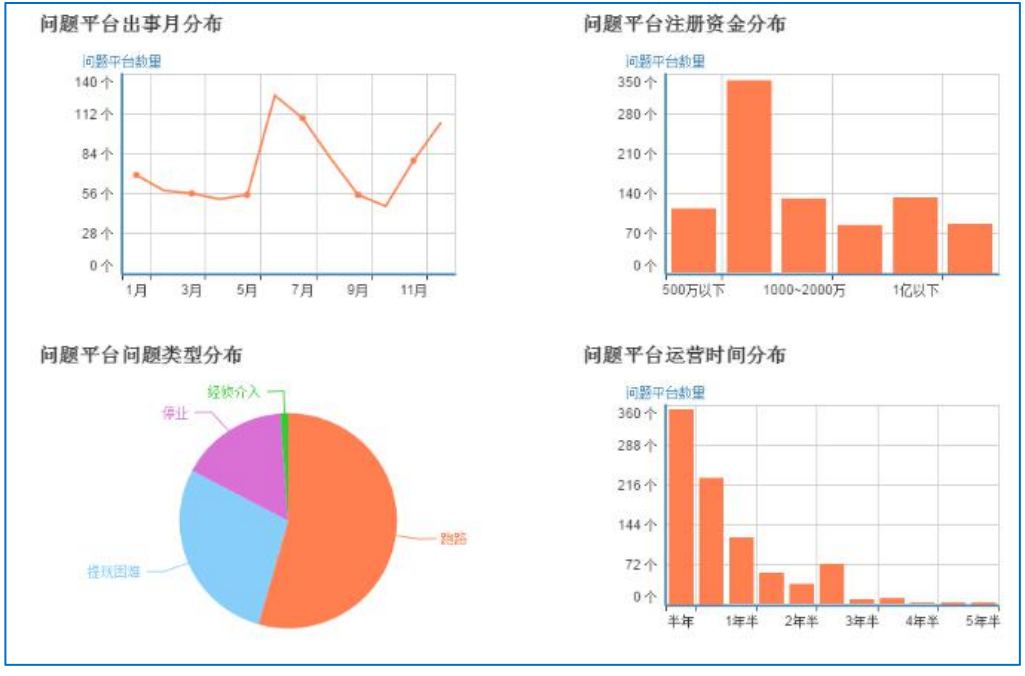


平台档案

- ✓ 可自由排序的排行榜
- ✓ 平台详情
  - 注册信息、平台背景、管理团队
  - 运营情况
- ✓ 用户评论
- ✓ 用户情绪、口碑及可视化
- ✓ 平台舆情关键字
- ✓ 平台近期新闻

积木盒子

舆情防雷



- ✓ 问题平台分析
  - 运营时长
  - 注册资金
  - 月份分布
  - 地域分布
- ✓ 雷区-问题平台列表
- ✓ 平台健康分

投资顾问

- ✓ 排行榜推荐
  - 综合排名 Top5
  - 运营时长 Top5
  - 平均收益 Top5
  - 注册资金 Top5
- ✓ 风险问卷
  - 风险承受能力分析
  - 投资风格分析

测试结果

- 风险态度：稳健型

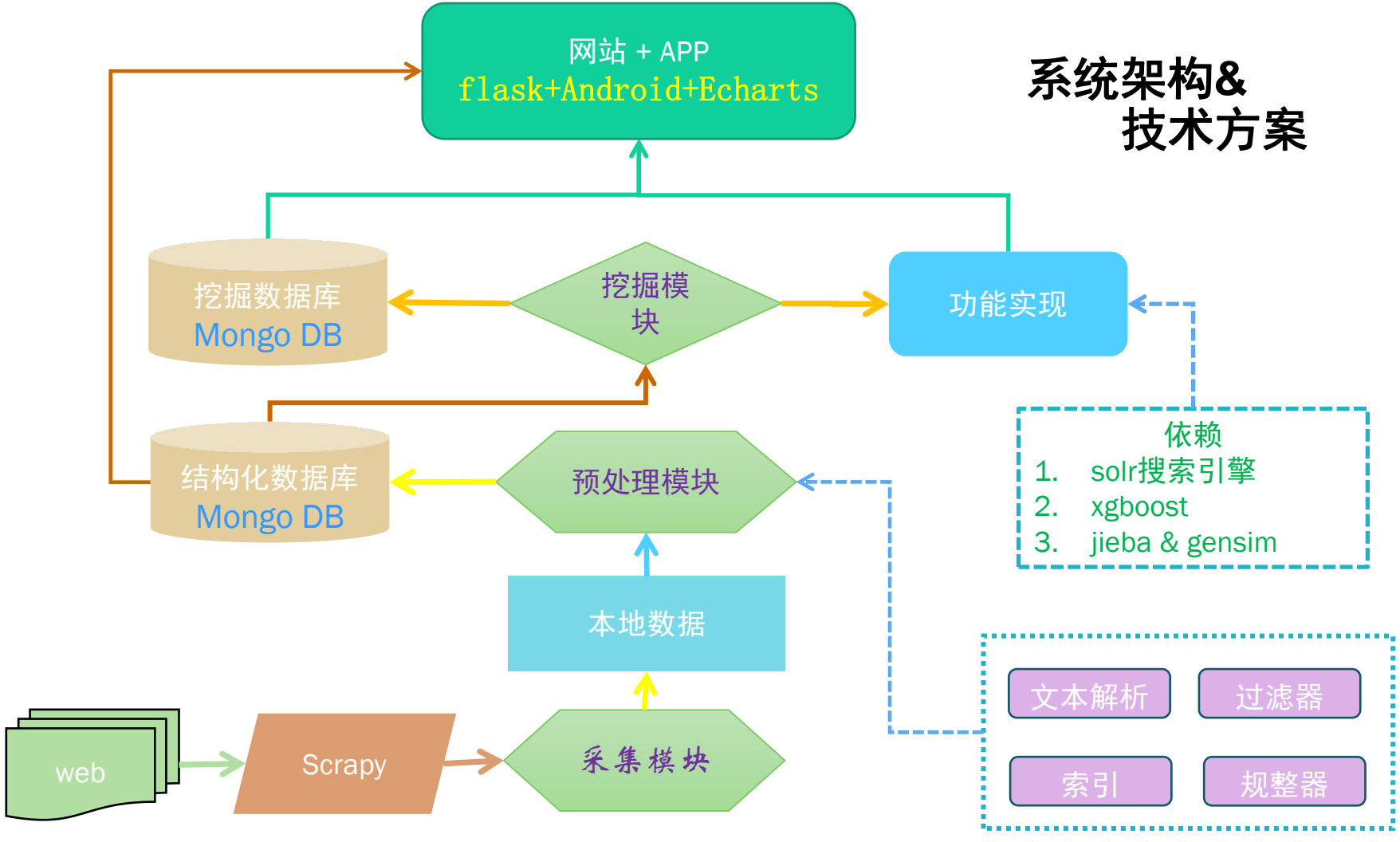
稳健型投资者既担忧风险也渴望收益，希望在较低风险下获取稳健的收益。理财时要对投资本金的安全性给予适当关注，主要是投资者的资产可能保持一个稳步上升的态势。

- 风险承受能力：100 分
- 获利期待：中等报酬
- 推荐平台：陆金所 宜人贷 人人贷 微贷网

技术亮点

1. 运用大量NLP算法处理文本，包括N-Gram、Word2Vec、Doc2Vec、LDA等算法、针对行业的词典、情感分析、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. 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
4. 系统模块间耦合性低、可扩展性强
5. 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
6. 优秀的数据可视化
7. 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
   - ✓ 用户论坛评论、微博
   - ✓ 知乎问答
   - ✓ 国家政策等
2. 可收集的平台经营数据、行业交易数据、官网UV、PV等指标数据
3. 平台百科等知识库数据

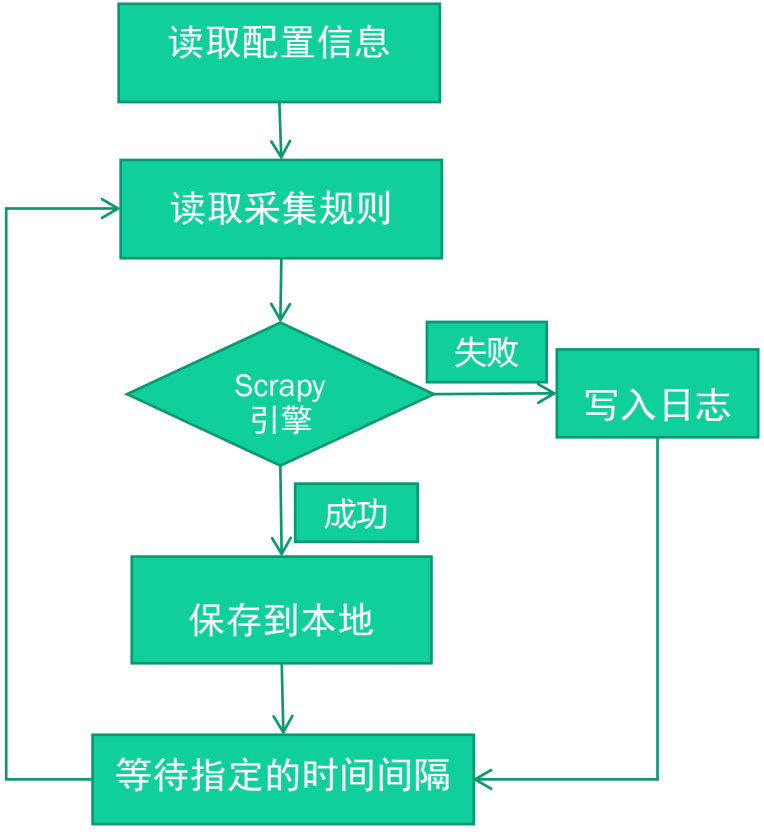
需满足：

多样性好、数据量大

可结构化

实时性好、针对性好

数据采集

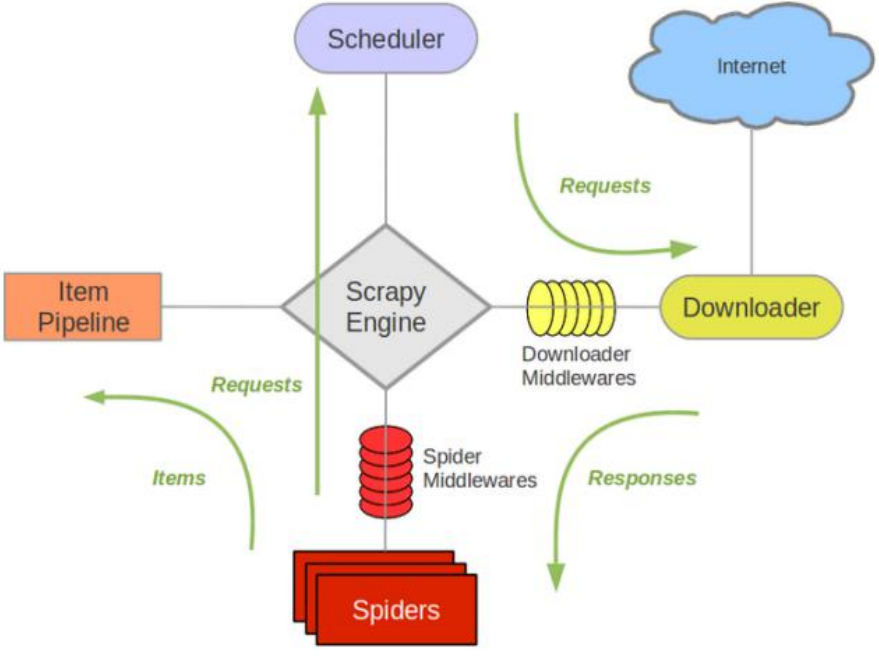


爬虫流程图

共爬取60G文本数据

- ✓ 新闻类
  - 新浪财经 （71196篇）
  - 网易财经 （221747篇）
  - 凤凰网、财新网 （47907篇）
  - 金融之家、和讯网 （2400篇）
  - 微信公众号 （31054篇）
  - 网贷之家（4400篇）
  - 中申网、金评媒（8400篇）
- ✓ 用户评论、观点类
  - 网贷之家论坛 （20万帖子评论）
  - 微博 （15万微博评论）
  - 知乎 （68万问答）
- ✓ 平台 （全部收录）
- ✓ 国家政策 （相关的470篇）

Scrapy



Step1 定制新闻列表爬虫

Step2 将爬虫提交给Scheduler，生成Request进行网页请求

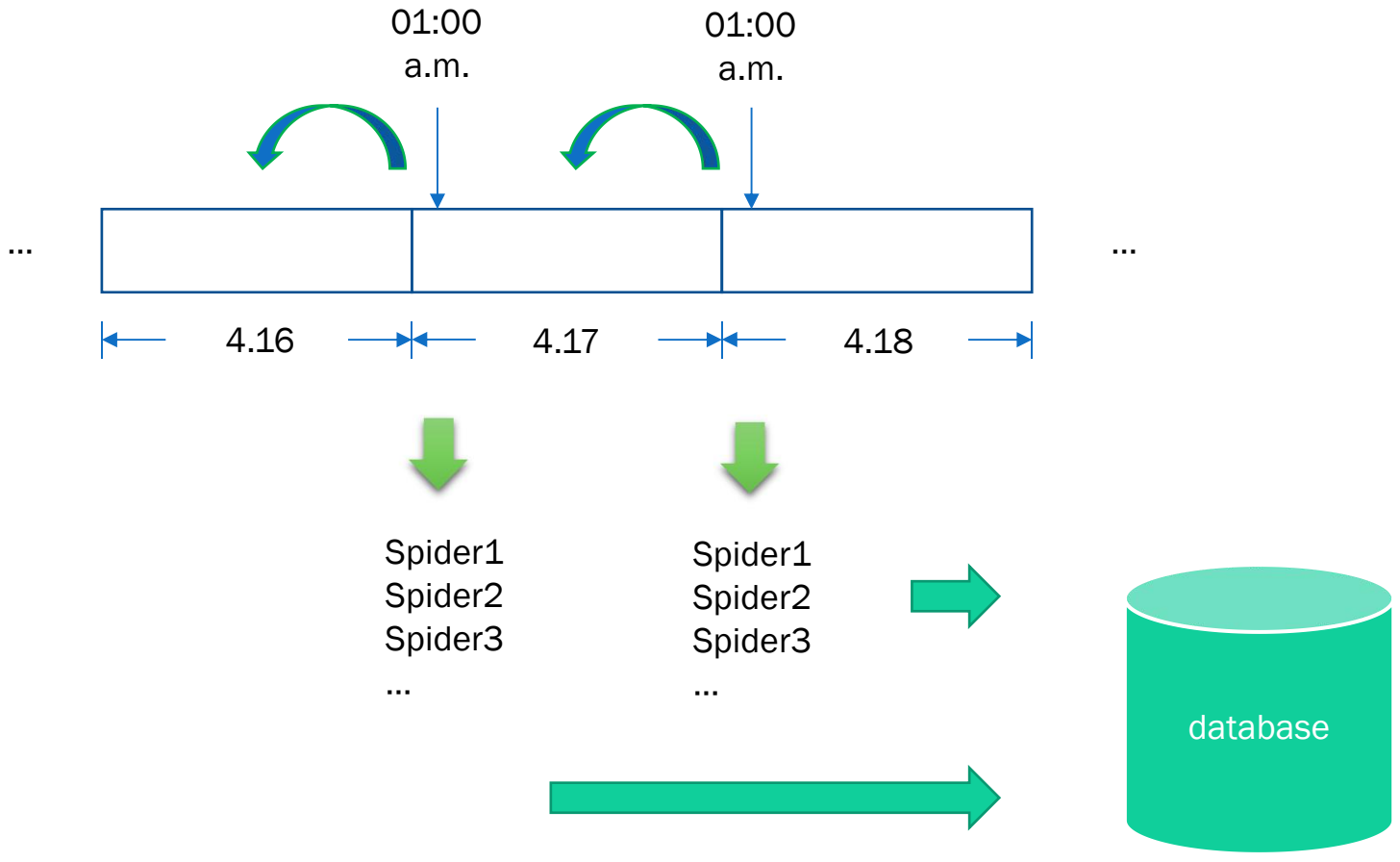
Step3 使用xpath解析Responses，将获取的url列表存储于中间文件

Step4 通过中间文件定制新闻详细内容爬虫

Step5 将爬虫提交给Scheduler，生成Request进行网页请求

Step6 使用xpath解析Responses，获取Title，Author，item_pub_time和content等信息，存储本地文件

增量式爬虫


01:00
a.m.
01:00
a.m.
...
...
4.16
4.17
4.18
Spider1
Spider2
Spider3
...
Spider1
Spider2
Spider3
...
database

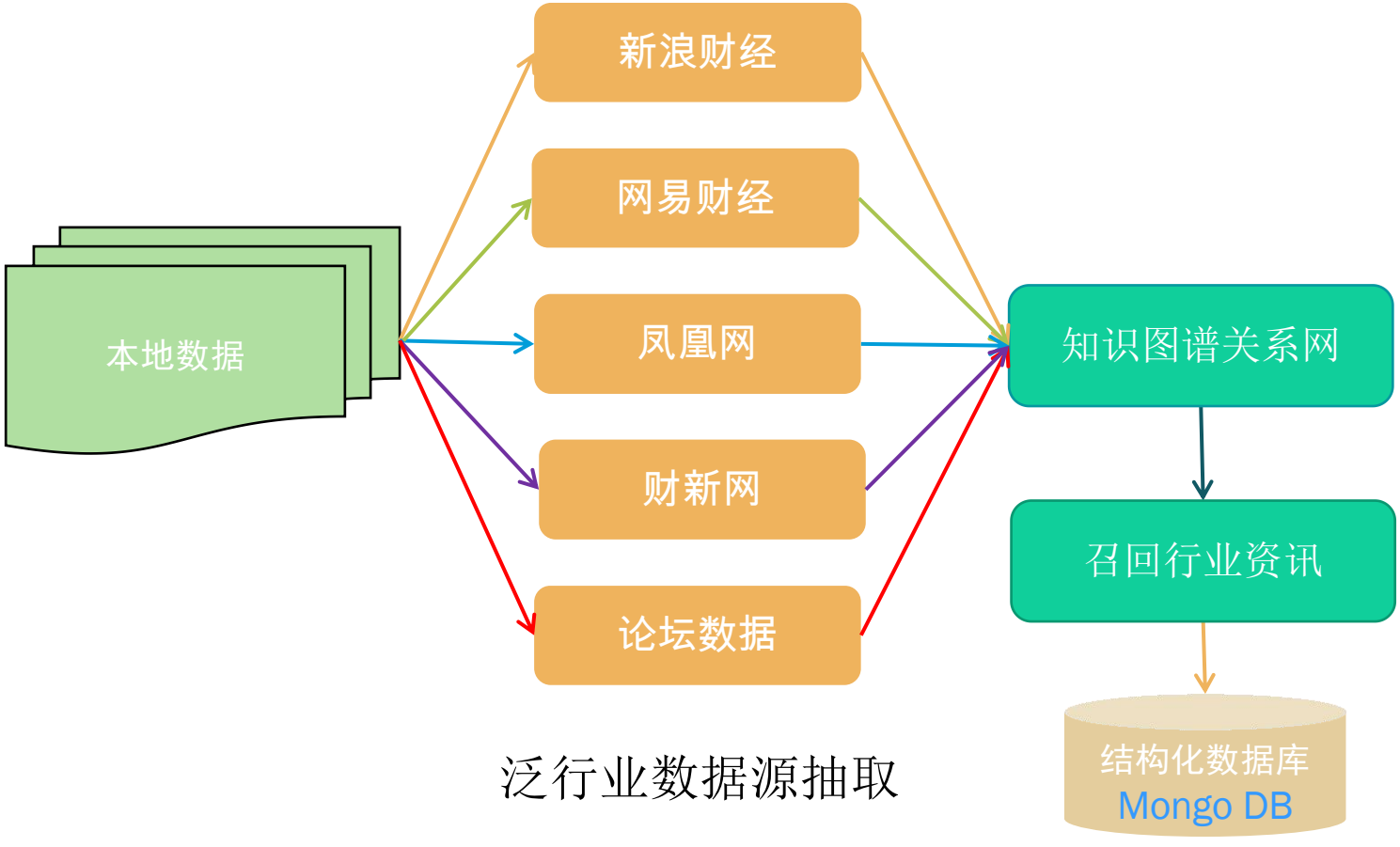
一、数据采集模块

二、预处理模块

三、数据挖掘模块

四、网站、APP与可视化

行业资讯召回



泛行业数据源抽取

过滤器

新闻语料库训练

UGC质量分计算公式

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

利用 $Q > 0.01$过滤，约保留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

- ✓ 文本处理
  - 杂乱文字、符号过滤
  - 各网站分段格式统一
  - 统一日期格式
- ✓ 资讯分类
- ✓ 添加资讯标签
- ✓ 添加索引、建倒排表
- ✓ 资讯统一编号

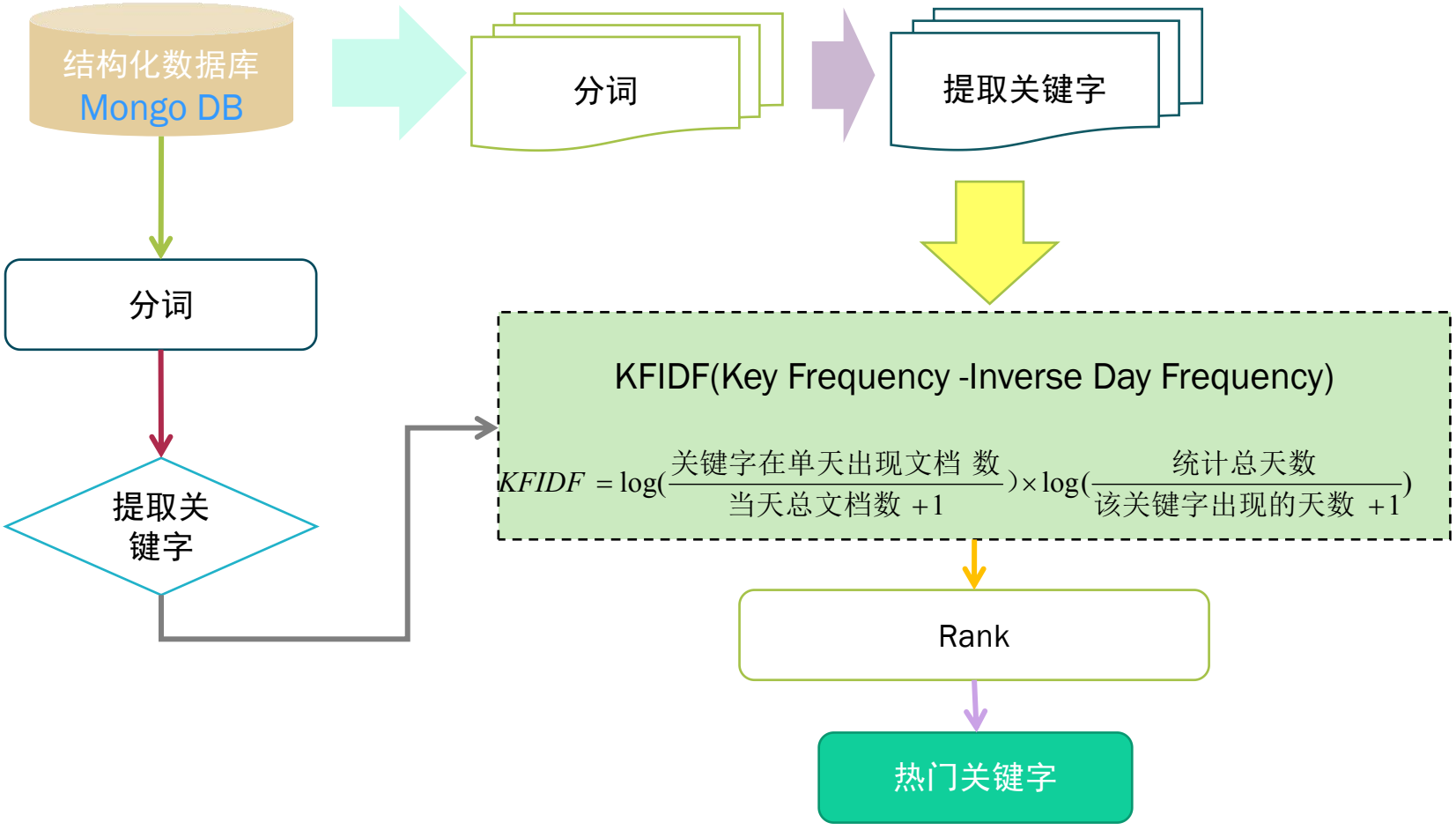
一、数据采集模块

二、预处理模块

三、数据挖掘模块

四、网站、APP与可视化

热门关键字

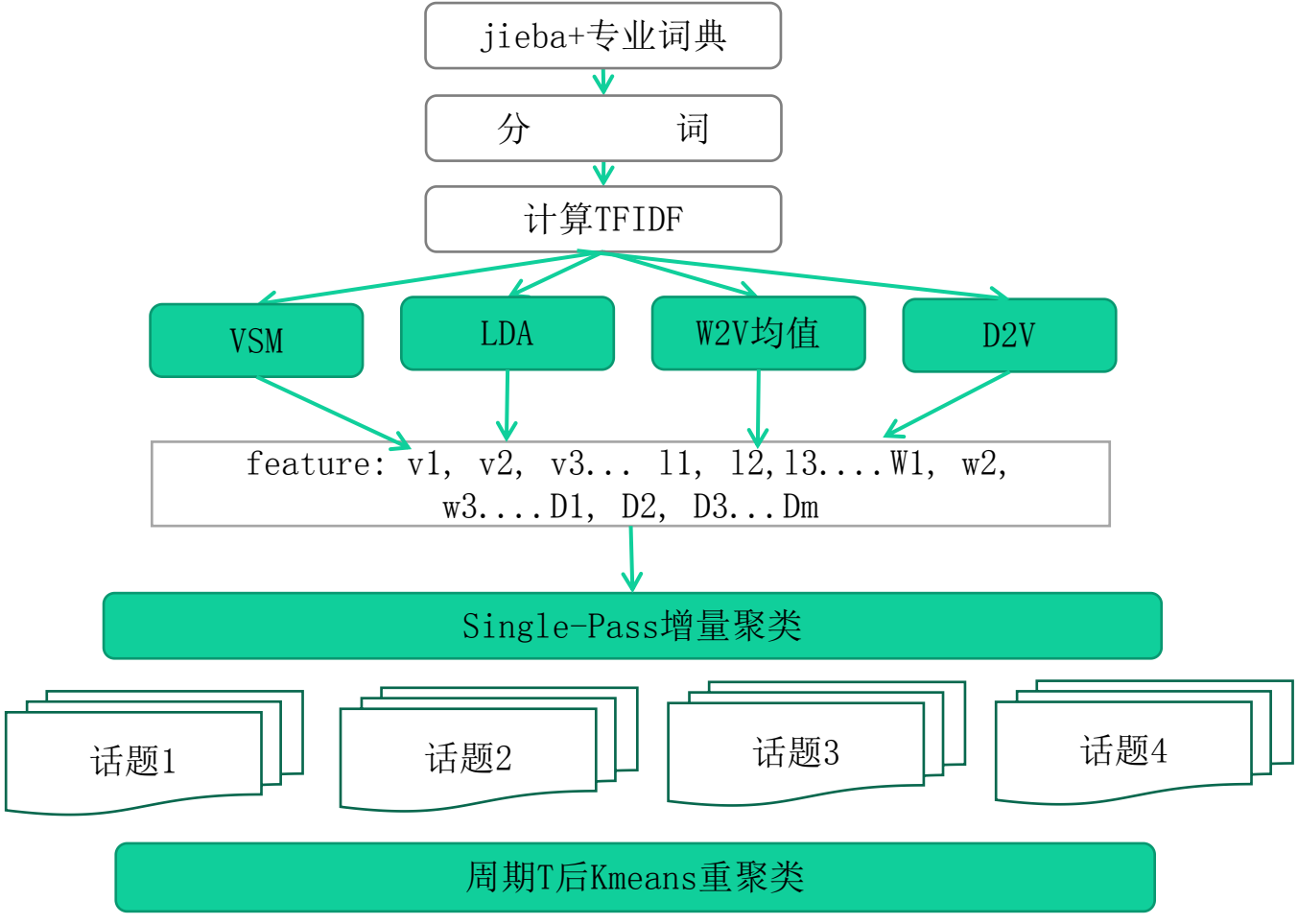

结构化数据库
Mongo DB
分词
提取关键字
分词
提取关
键字
KFIDF(Key Frequency -Inverse Day Frequency)
$KFIDF = \log(\frac{关键字在单天出现文档数}{当天总文档数+1}) \times \log(\frac{统计总天数}{该关键字出现的天数+1})$
Rank
热门关键字

热门话题发现

预处理

特征提取

聚类模块

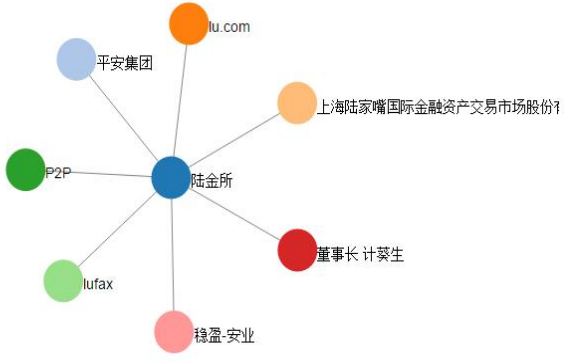


舆情防雷-健康分

- ✓ 利用机器学习训练分类器，得到平台出问题的概率。
- ✓ 使用特征
  - 平台基础信息特征
  - 舆情特征
  - 平台管理层特征
  - 股东信息特征、动态特征
  - 平台产品特征
  - 网贷之家、网贷天眼、融360等第三方平台评分

搜　索

Solr　　+　　知识图谱



资讯搜索、平台搜索、资讯搜索

知识图谱

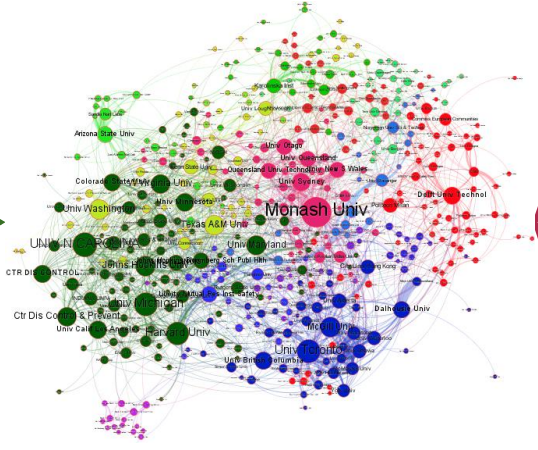
实体识别

实体消歧

关系抽取

知识推理

知识表示



7类实体
50万个节点
300万对关系

改善NLP模型效果

优化搜索

全方位平台信息

用户自追踪

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

- 开源库：异步通信Afinal/Gson + Jsoup
- 图表：MPAndroidChart + 图片缓存UImageLoader
- 第三方平台：友盟 + 百度推送

谢谢