

算法项目实践 案例简介

人工智能中心

Friday, February 22, 2019

目录

机器学习技术

自然语言处理技术

未来规划

立项流程

1

机器学习技术方向

1、验证码识别

目标

对公司爬虫在爬取网页（极验，易盾、百度等）时遇到的验证码，使用机器学习算法进行识别。

背景：关于爬虫

- 爬虫的本质：模拟用户打开浏览器网页，然后获取网页中想要的信息。所以，只要用户可以访问的数据，爬虫都能获取到。
- 困难：对于用户登录才能访问的网站，爬虫就必须模拟用户的登录操作。所以，爬虫先要识别各式各样、日益复杂的验证码。



字母数字类



滑动式



点击式



1、验证码识别

解决方案



案例一：识别字母数字类验证码

6718

CJHKSS

j q z c f u

j p x d

- 对于目前市面上 **95%** 的数字字母类验证码，服务的平均识别精度在 **93%** 以上。
- 目前上线的识别服务，gpu单节点qps在 **300** 左右。

—— qps (Queries Per Second) ，即每秒查询率，可衡量服务器性能

1、验证码识别

案例二：识别Geetest 验证码



- 服务对于目前所有类型的Geetest验证码都有很高的识别精度，平均准确率大于 **95%**。
- 特别是对于语序点选验证，已累计服务上亿次。
- 目前上线的识别服务，gpu单节点qps在 **20** 左右。

目标

在财务报表数据中，找出疑似异常样本的数据，从而为相关人员提供合理的稽查线索。

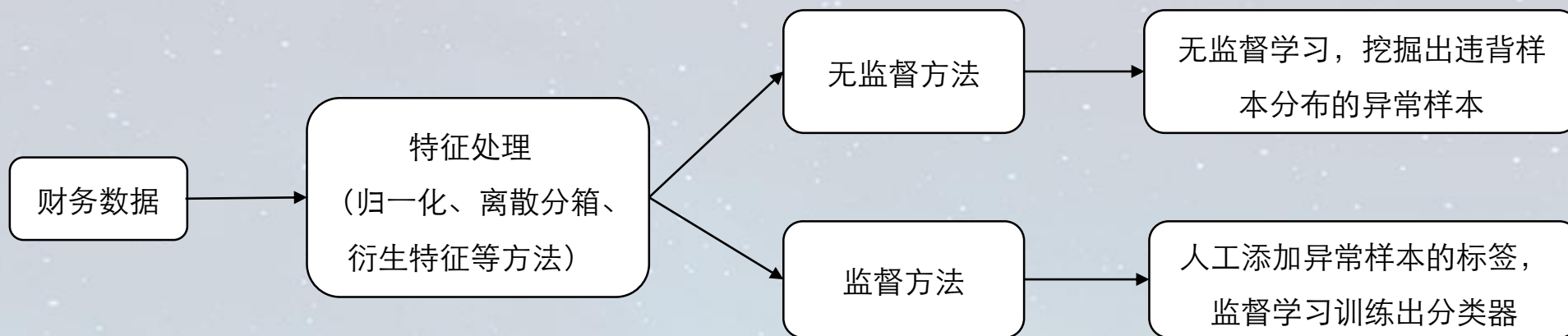
背景：财务舞弊

- **财务舞弊的手段：**主要是对财务报表信息进行有意的错报或忽略。
 - 包括虚构收入、变更利润、隐藏负债、资产重组、利用资产减值等
- **识别财务舞弊的传统方法：**由注册会计师对财务数据进行人工分析，计算相关指标，并观察识别。
 - 观察的指标包括毛利率、应收账款周转率、存货周转率、资产负债率、流动比率、净资产收益率等
- **传统方法的局限：**
 - 1) 依赖于注册会计师的个人经验与技能，缺乏系统性。
 - 2) 财务舞弊的手段众多，某些隐蔽性的异常很难被发现。

所以，机器学习技术可以在这个领域发挥很大的作用。

2、财务异常挖掘

解决方案



案例一：证监会财务异常分析

- 目前异常识别的准确率达到 80% 以上。

案例二：深交所财务异常分析

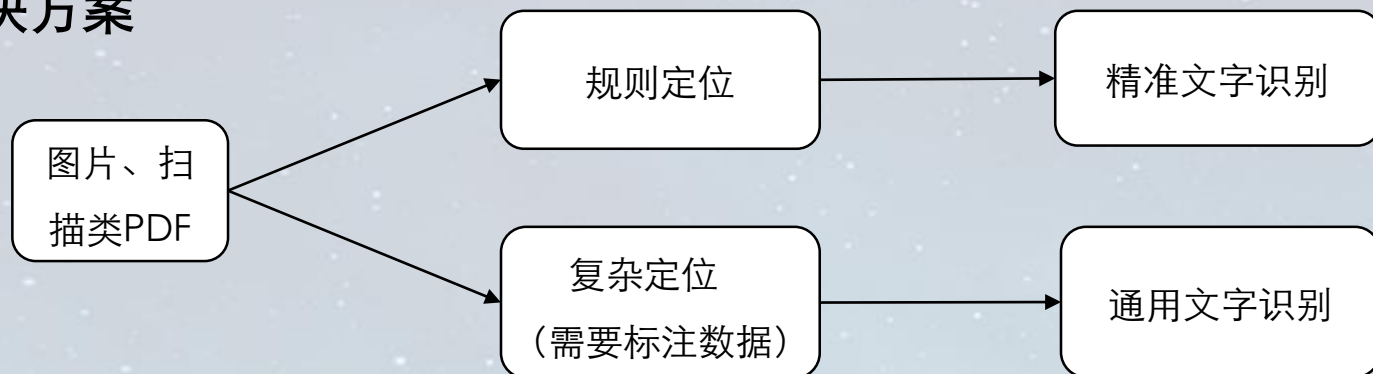
- 在深交所最终给出的未公开样本中，异常识别的准确率达到 80% 以上。

3、OCR（图像文字识别）

目标

对网页截图和扫描类PDF文件进行快速文字识别。此服务可以用于某些网站的爬虫。

解决方案



案例：大众点评的网页SVG图片

- 服务支持 **6400** 个字符（包括汉字、字母、数字、标点符号）的精准识别。
- 支持超长序列（最长 **33** 个字符）的OCR识别。
- 在大众点评svg图片的文字识别上达到了 **95%** 以上的准确率。
- 动态显存占用率为纯模型600mb显存，GPU单节点平均QPS **20**。



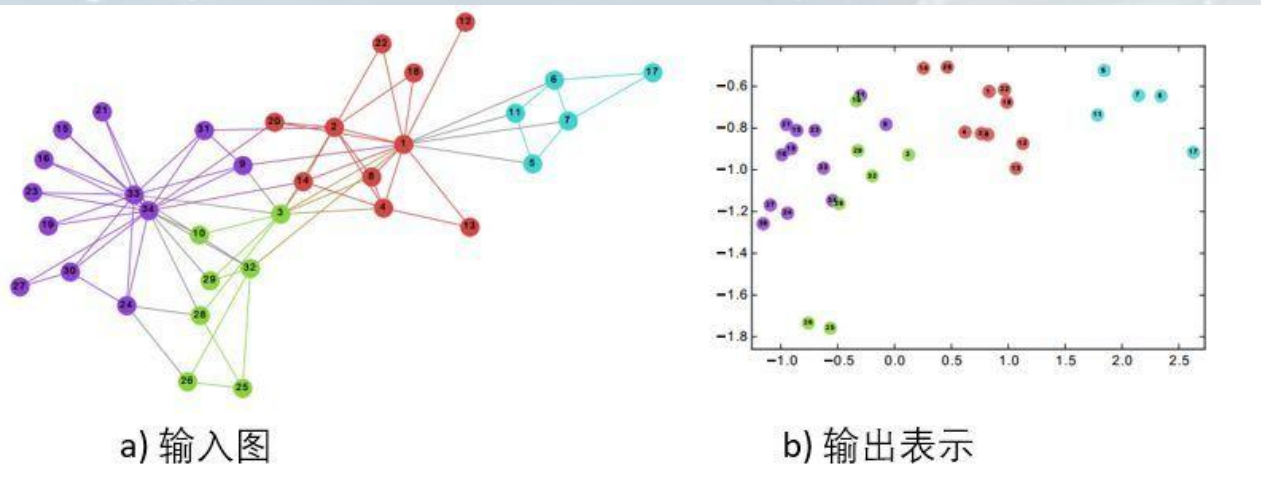
4、图挖掘（网络表征学习）

目标

通过对图结构的连接关系的学习，得到图中每个节点的向量表征（即网络表征学习）。

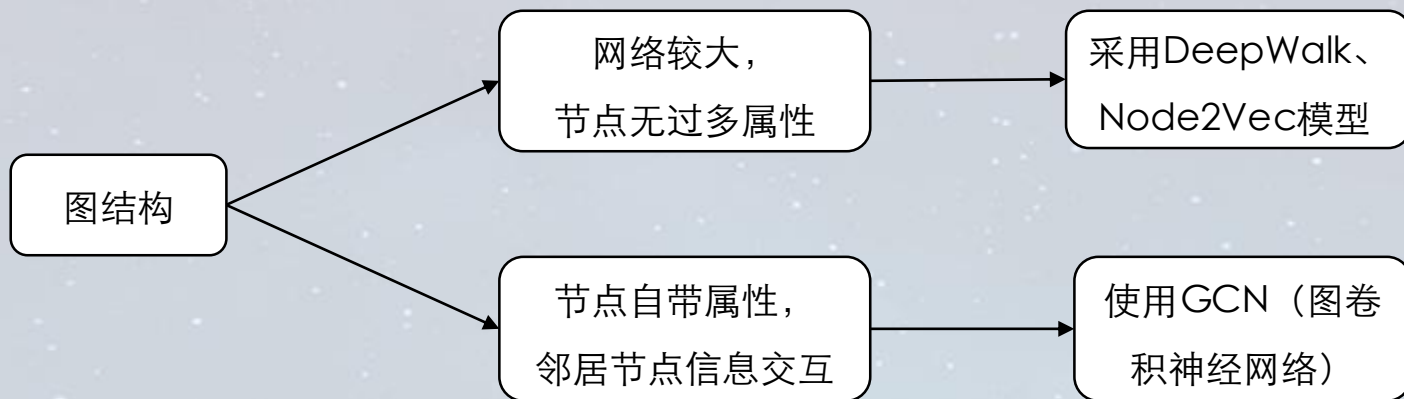
背景：网络表征学习

- **图结构**：由节点和连接节点的边组成的结构。
 - 例如在社交网络的图结构中，图的节点表示单个用户，图的边表示用户之间的互联关系
- **网络表征学习**：针对图结构，用向量的形式表示网络的节点属性的机器学习方法。
 - 常用的方法有DeepWalk、Node2Vec等
- **网络表征学习的本质**：将高维图结构用更简单的低维向量表示，并且保留网络结构和节点属性的特征信息（如右图）。



4、图挖掘（网络表征学习）

解决方案



案例：基于网络节点连接的阿拉丁评分标签预测

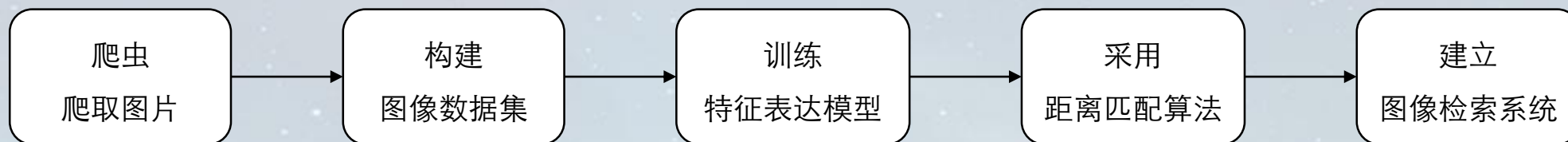
- 计算了企业投资网络的 200万+ 节点的表征。
- 使用阿拉丁评分标签（约 13万 黑样本）检测模型的分类效果，得到 0.7 的F1。
- 模型局限性：模型需要整体加载到内存中，耗用空间很大（约30G），所以实验只能在CPU环境下进行。

5、图像检索

目标

提供 CBIR（基于内容的图像检索, Content Based Image Retrieval）服务，实现精准的以图搜图功能。此服务可以广泛应用于商品检索系统和多媒体知识图谱系统。

解决方案



案例：战神网信 —— 图像检索

- 满足 203 类装备实体检索，其中42类实体检索精准率大于 90%。
- CPU单节点QPS 30。

2

自然语言处理技术

目录

CONTENTS

01

词法分析

02

实体

03

关系

04

事件

05

文本

06

KBQA

01

词法分析

词法分析是NLP的基础技术，是大部分NLP任务的第一步，包括分词，词性标注。

词法
分析

```
graph LR; A((词法分析)) --> B((分词)); A --> C((词性标注))
```

分词

分词是中文处理的特有任务，因为中文词语之间没有间隔，需要将字序列变换成为词序列。

词性
标注

词性标注为分词结果中的每个单词标注一个正确的词性，即确定每个词是名词、动词、形容词等。

12月12日上午，“2018信用建设暨（第四届）信用高峰论坛”在北京召开，本届论坛由北京市经济和信息化局、中国人民银行营业管理部、中关村科技园区管理委员会作为支持单位，中关村发展集团、中关村企业信用促进会、北京市中小企业公共服务平台联合主办，以“信用优化营商环境助力创新与高质量发展”为主题，聚焦新时代优化营商环境新要求，探讨信用改善环境新方向。

分词

词性标注

12月 12日 上午 , “ 2018 信用 建设 暨 (第四 届) 信用
高峰 论坛 ” 在 北京 召开 , 本届 论坛 由 北京市 经济 和
信息化局 、 中国 人民 银行 营业 管理部 、 中关村 科技园区 管理
委员会 作为 支持 单位 , 中关村 发展 集团 、 中关村 企业 信用
促进会 、 北京市 中小 企业 公共 服务 平台 联合 主办 , 以 “
信用 优化 营商 环境 助力 创新 与 高 质量 发展 ” 为 主题 ,
聚焦 新 时代 优化 营商 环境 新 要求 , 探讨 信用 改善 环境
新 方向 。

词性标注

介词 代词 动词
区别词 名词 形容词
数词 时间词 标点
连词 量词

02

实体

问题描述

实体识别是指给定一个文本，识别出文中出现的命名实体，如人名、地名、公司名等，具体的类别与任务有关。

实体识别



NER 基础解析

使用开庭公告、人民法院公告、招标、中标、执行、失信被执行、舆情等9个数据源训练公司、人名、组织名的机构名3类实体，实现了97.2%的F1值。

案例展示



12月12日上午， “ 2018 信用 建设 暨 （ 第四 届 ） 信用 高峰 论坛 ” 在 北京 召开 ， 本届 论坛 由 北京市经济和信息化局 、 中国人民银行营业管理部 、 中关村科技园区管理委员会 作为 支持 单位 ， 中关村发展集团 、 中关村企业信用促进会 、 北京市中小企业公共服务平台联合主办 ， 以 “ 信用 优化 营商 环境 助力 创新 与 高 质量 发展 ” 为 主题 ， 聚焦 新 时代 优化 营商 环境 新 要求 ， 探讨 信用 改善 环境 新 方向 。

实体类识图示

公司名 地名 时间 组织名

实体链接

问题描述

实体链接就是一种将文本中实体指称项链接到知识库里对应实体的技术，能够有效解决命名实体歧义这一问题。实体链接技术需要庞大的知识库，包含各实体的各种属性。



解决方案

- Step 1 生成备选实体。备选实体主要依赖字符串匹配、实体别名字典和搜索引擎技术产生。
- Step 2 备选实体排序。排序标准通常包括：实体普及程度，文本前后文，候选实体之间主题相关性。具体排序方法包括有监督方法和无监督方法。常用的有监督方法包括 learn to rank，概率模型，基于图的方法。无监督方法指向量空间模型，将实体指称项和候选实体都用向量表示，并计算向量相似性，将相似度最高的作为目标实体。

实体链接

公司、组织、人名三个实体的f1值可以达到97.2%
链接错误率为2.8%，数据缺失率为25.13%

对于文章中提到的公司实体或者其简称，能够返回完整的公司名，并且将公司名和公司实体库关联起来。

案例

疑似关联方

最终模型的f1值为0.944.利用规则之后
二度关联方的正确率在95%以上。

项目实现利用企查查的check_id数据，分别生成50万正负样本，最终能够确定不同公司的同名人士是否为同一人。

03

关系

如对于 公司名(CMP) 这一实体，通过预先定义及关系抽取，最终可以获得以下结果

<CMP_1, CMP_2, Invest_0> 投资关系(-->): CMP_1 投资 CMP_2

<CMP_1, CMP_2, Partner> 合作伙伴: CMP_1和CMP_2是合作伙伴

关系抽取

关系抽取是信息抽取的重要子任务，关系抽取主要负责在实体识别基础上，抽取实体间的语义关系。

方案一

基于规则的关系抽取。

人工编写大量的规则从文本中抽取实体之间的关系。

方案二

基于半监督的关系抽取。

通过预先准备的高质量关系数据生成抽取关系的规则，再通过规则从文本库中得到更多关系数据以生成新规则，周而复始，直至满足某种终止条件。

方案三

监督方法。

将关系抽取转化为分类问题，需要高质量的标注数据。

战神实体关系抽取 F1值 92.4%

从文本中找出实体之间的关联关系。

涉及实体：

机构名(ORG) 公司名(CMP) 地名(LOC) 人名(PER)
时间(TIM) 职位(PST) 产品(PRD) 政策实体(POL)

关联关系：

<ORG, CMP, Invest> 投资关系：ORG 投资 CMP

<ORG, CMP, Member> 成员关系：CMP是ORG的一员
等14项。



半结构化信息抽取——描述

半结构化数据，主要是各种文书（招标投标文书，法律文书，企业公告等等），这类数据存在结构化文本和非结构化文本混杂的情况。半结构化文本解析指从上述文本中解析出结构化数据，包括实体，关系，事件要素。

半结构化信息抽取——方案

半结构化数据抽取通常包含两个步骤：1、数据转换，2、文本解析。

- 数据转换：
 - ①pdf解析，结合pdf2json,pdfplumber等工具解析pdf文件，存在速度较慢的问题
 - ②html解析，结合beautifulsoup解析html格式数据
- 文本解析
 - ①表格解析，通过规则将html和pdf中的表格还原成计算机可处理的关系表。
 - ②实体识别
 - ③基于规则或者模型确定实体的角色。
 - 基于规则时需要产品给出规则；
 - 基于模型主要依赖各种深度学习模型，需要产品给出标注数据，对于自然文本通常推荐使用基于模型的方法。

通股质押，具体情况如下：

一、股份质押的具体情况

1、股份质押的基本情况

出质人：谢晓东先生将持有本公司限售流通股 2,160.00 万股（占公司总股本为 20.97%）质押给中泰证券股份有限公司，用于办理股票质押式回购业务。初始交易日为 2018 年 4 月 18 日，购回交易日为 2019 年 4 月 17 日。

2、股份累计质押的情况

截至 2018 年 4 月 19 日，谢晓东先生持有本公司股份总数为 3,224.66 万股，占公司总股本比例为 31.31%。本次质押后累计质押股份为 2,160.00 万股，占其所持有公司股份总数的 66.98%，占公司总股本的比例为 20.97%。

二、

谢晓东先生持有本公司股份总数为 3,224.66 万股，占公司总股本比例为 31.31%。本次质押后累计质押股份为 2,160.00 万股，占其所持有公司股份总数的 66.98%，占公司总股本的比例为 20.97%。

从文本中解析

上市公司公告
解析
(质押事件)

增持人	增持方式	增持日期	增持数量 (股)	增持金额(元)	均价(元/股)	占公司总股本比例
陈清州	竞价交易	2018年2月5日	800,000	8,461,595.00	10.5770	0.0441%
	竞价交易	2018年2月9日	449,900	5,024,923.00	11.1690	0.0248%
	竞价交易	2018年2月13日	1,558,900	18,460,446.16	11.8420	0.0859%
	竞价交易	2018年2月14日	1,445,500	17,019,882.88	11.7744	0.0796%
	竞价交易	2018年3月1日	1,333,600	17,666,868.10	13.2475	0.0735%
	竞价交易	2018年4月25日	2,388,700	25,871,598.33	10.8308	0.1316%
	竞价交易	2018年4月26日	1,316,910	14,124,487.04	10.7255	0.0726%
合计			9,293,510	106,629,800.51	11.4736	0.5120%
许诺	竞价交易	2018年4月25日	1,027,300	11,008,931.00	10.7164	0.0566%
	合计		1,027,300	11,008,931.00	10.7164	0.0566%

从表格中解析

3、增持计划实施前后的持股变化情况

增持人	本次增持前持有股份		本次增持后持有股份	
	持股数量(股)	持股比例	持股数量(股)	持股比例
陈清州	933,401,296	51.42%	942,694,806	51.93%
许诺	0	0	1,027,300	0.06%
合计	933,401,296	51.42%	943,722,106	51.99%

最终解析结果准确率在90%左右。

04

事件

事件抽取——问题描述

事件抽取是把含有事件信息的非结构化文本以结构化的形式呈现出来。事件抽取任务可由下面两个步骤组成：

- 事件检测：主要是根据上下文识别出触发词以及代表的事件类型和子类型
- 事件元素识别：事件元素是指事件的参与者。事件元素识别是指根据事件所属的事件模版，抽取相应的元素，并为其标上正确的元素标签。

事件抽取——解决方案

1. 基于模式匹配：结合语法树或正则表达式的事件抽取系统。基于模式匹配的方法在特定领域中性能较好，知识表示简洁，便于理解和后续应用，但对于语言、领域和文档形式都有不同程度的依赖，覆盖度和可移植性较差。此外需要产品配合给出规则。
2. 基于统计模型的方法：这种方法需要产品提供质量较高的标注数据。
由于事件检测包含了多个步骤，因此事件检测方法可分为两类：
 - 管道抽取，依次将事件抽取多个子任务变为分类任务
 - 联合推理，同时进行事件检测和事件要素抽取，避免误差传播导致的性能下降。在具体的分类任务中，传统的机器学习方法和深度学习算法都有所应用。

上市公司公告增减持事件抽取

需要从pdf文件中解析出

公司名、股权变动人（股东名称）、股权变动方式（减持方式）、权益变动日期（减持期间）、股权变动数（减持股数（股））、股权变动比例（减持比例）、股份性质、原持股数等信息。



持股变动实施完毕公告.pdf



使用3000条标注的增减持事件语料，使用NER和QA模型，解决该项目中事件抽取的问题，实现字段平均94%的F1值。

广东威华股份有限公司

公司名

关于持股 5%以上股东减持公司股份的公告

本公司及董事会全体成员保证信息披露内容的真实、准确和完整，没有虚假记载、误导性陈述或重大遗漏。

2018年4月16日，广东威华股份有限公司（以下简称“公司”）收到公司持股5%以上股东李建华先生的《关于减持公司股份的告知函》，李建华先生于2018年4月16日通过深圳证券交易所大宗交易系统减持公司股份2,449,000股，占公司总股本的0.4575%。具体情况如下：

一、股东减持情况

（一）股东减持股份情况

股东名称	减持方式	减持期间	减持均价 (元/股)	减持股数 (股)	减持比例
李建华	大宗交易	2018.4.16	13.51	2,449,000	0.4575%

（二）股东本次减持前后持股情况

股东名称	股份性质	本次减持前持有股份		本次减持后持有股份	
		股数(股)	占总股本比例	股数(股)	占总股本比例
李建华	无限售条件股	2,258,400	8.64%	43,809,400	8.1834%

先持股比例

二、其他相关说明

原持股数

原持股比例

现持股数

（一）本次减持行为未违反《证券法》、《深圳证券交易所中小企业板上市

文本摘要抽取

通过自动分析给定的一篇文档或多篇文档，提炼、总结其中的要点信息，最终输出一篇长度较短、可读性良好的摘要。该摘要中的句子可直接出自原文，也可为重新撰写所得。



国信信用探测文本摘要

原新闻：

5日，备受关注的电影《黄金时代》在北京举行了电影发布会，导演许鞍华和编剧李樯及汤唯、冯绍峰等众星悉数亮相。据悉，电影确定将于10月1日公映。本片讲述了“民国四大才女”之一的萧红短暂而传奇的一生，通过她与萧军、汪恩甲、端木蕻良、洛宾基四人的情感纠葛，与鲁迅、丁玲等人一起再现上世纪30年代的独特风貌。电影原名《穿过爱情的漫长旅程》，后更名《黄金时代》，这源自萧红写给萧军信中的一句话：“这不正是我的黄金时代吗？”

新闻摘要: [查看文档](#)

50%

40%

30%

20%

15日，备受关注的电影《黄金时代》在北京举行了电影发布会，导演许鞍华和编剧李樯及汤唯、冯绍峰等众星悉数亮相。据悉，电影确定将于10月1日公映。电影原名《穿过爱情的漫长旅程》，后更名《黄金时代》，这源自萧红写给萧军信中的一句话：“这不正是我的黄金时代吗？”

新闻摘要:

50%

40%

30%

20%

15日，备受关注的电影《黄金时代》在北京举行了电影发布会，导演许鞍华和编剧李樯及汤唯、冯绍峰等众星悉数亮相。据悉，电影确定将于10月1日公映。

05

文本分类

PART ONE

01

文本分类

给定一个文本输入，通常是自然文本，输出可枚举的类别标签，标签可以是互斥或者相互独立



PART TWO

02

文本相似性

给定两个文本，预测他们的相似程度



PART THREE

03

文本聚类

给定多个文本，根据他们的相似度，将相似的聚在一起，形成一个类别，从而将所有文本分成多个类别。



PART FOUR

04

情感识别

给定一篇文章，预测文章的正负面情感；高级应用的时候，预测对不同目标实体的情感



问题描述

给定一个文本输入，通常是自然文本，输出可枚举的类别标签，标签可以是互斥或者相互独立。

新闻

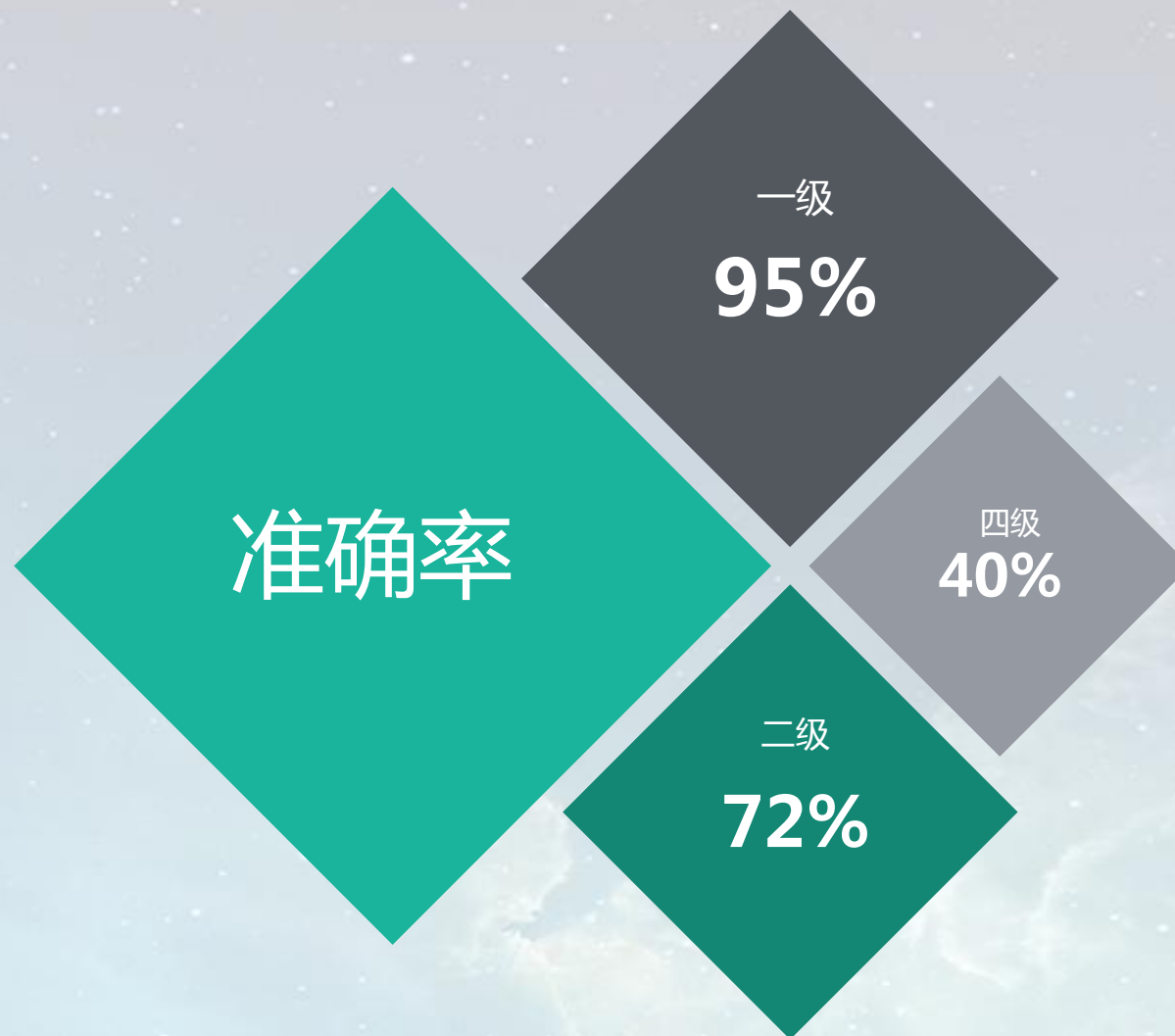
5日，备受关注的电影《黄金时代》在北京举5日，备受关注的电影《黄金时代》在北京举行了电影发布会，导演许鞍华和编剧李樯及汤唯、冯绍峰等众星悉数亮相。据悉，电影确定将于10月1日公映。本片讲述了“民国四大才女”之一的萧红短暂而传奇的一生，通过她与萧军、汪恩甲、端木蕻良、洛宾基四人的情感纠葛，与鲁迅、丁玲等人一起再现上世纪30年代的独特风貌。电影原名《穿过爱情的漫长旅程》，后更名《黄金时代》，这源自萧红写给萧军信中的一句话：“这不正是我的黄金时代吗？”

新闻分类:

体育 教育 财经 社会 娱乐 军事 国内 科技 互联网 房产 国际 女人 汽车 游戏

企业行业分类（一级、二级、四级）

根据企业的名称、经营范围 的描述文本，按照国家行业分类规范文档，划定当前企业所隶属的一、二、四级行业类别。



给定两个文本，判定它们的相似程度。

整体方法

利用现有模型框架，直接完成对两文本相似程度的判断。常用模型有BERT、BM25、Match-Tensor。



分步方法

- 1) 将文本转化成向量。向量化的方法包括TFIDF、simHash、doc2vec以及深度学习方法。
- 2) 计算2个向量之间的距离（距离越小，相似度越高）。距离度量方法包括海明距离、欧氏距离、余弦距离等。

红盾舆情

采取分步方法：
TFIDF+simHash和海明距离的方法，对红盾新闻中的重复文本进行去除。

案例

基于编辑距离的去重算法，
设置75%相似度阈值，对短
新闻进行去重过滤。

CCTV项目



文本聚类

给定多个文本，根据他们的相似度，将相似的聚在一起，形成一个类别，从而将所有文本分成多个类别。

01



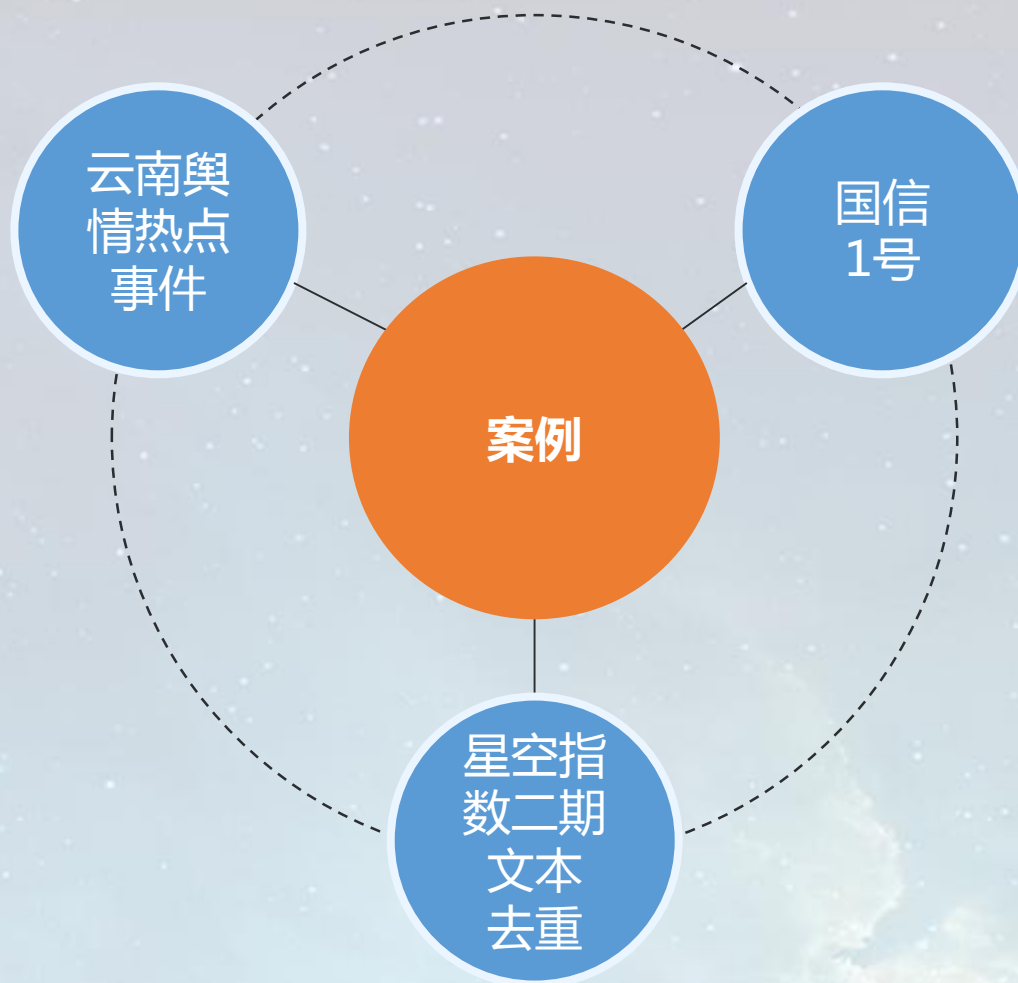
利用文本相似性算法得到向量或距离定义，然后利用经典的数据挖掘的聚类算法，比如K-means，DBSCAN

02



利用主题聚类LDA完成文本聚类任务

统计top-K重复新闻后聚类toP-N新闻主题，产生N个热度主题。



提供语义文本相似度、聚类算法的演示；提供事件发现和多文本分析

采用simhash算法计算文本hashcode，具有相同hashcode的文本即为重复文本。

给定一篇文章，预测文章的正负面情感；
高级应用的时候，能够预测对不同目标实体的情感。

方案一

使用通用情感词典，
做基于词典的情感
识别，基于词典标
注的模型识别

方案二

使用领域情感词+通
用情感词

方案三

使用标注数据，做
有监督的模型训练

方案四

使用领域标注数据，
通用模型，做迁移
学习

15日，备受关注的电影《黄金时代》在北京举行了电影发布会，导演许鞍华和编剧李樯及汤唯、冯绍峰等众星悉数亮相。据悉，电影确定将于10月1日公映。本片讲述了“民国四大才女”之一的萧红短暂而传奇的一生，通过她与萧军、汪恩甲、端木蕻良、洛宾基四人的情感纠葛，与鲁迅、丁玲等人一起再现上世纪30年代的独特风貌。电影原名《穿过爱情的漫长旅程》，后更名《黄金时代》，这源自萧红写给萧军信中的一句话：“这不正是我的黄金时代吗？”

情感分析:

负面指数: 0.0106



正面指数: 0.9894

通用

汽车

厨具

餐饮

新闻

微博

■ 正面

■ 负面

负面指数:

0~0.5之间判断为负面,

0.5~1之间判断为正面。

可调高级接口获取负面指数, 自定义判断标准。

01

证监会私募画像 (方案1)

采用情感词典对文本进行情感词匹配, 计算情感得分, 对文本进行情感正负倾向进行打分、分类。

02

国信情感

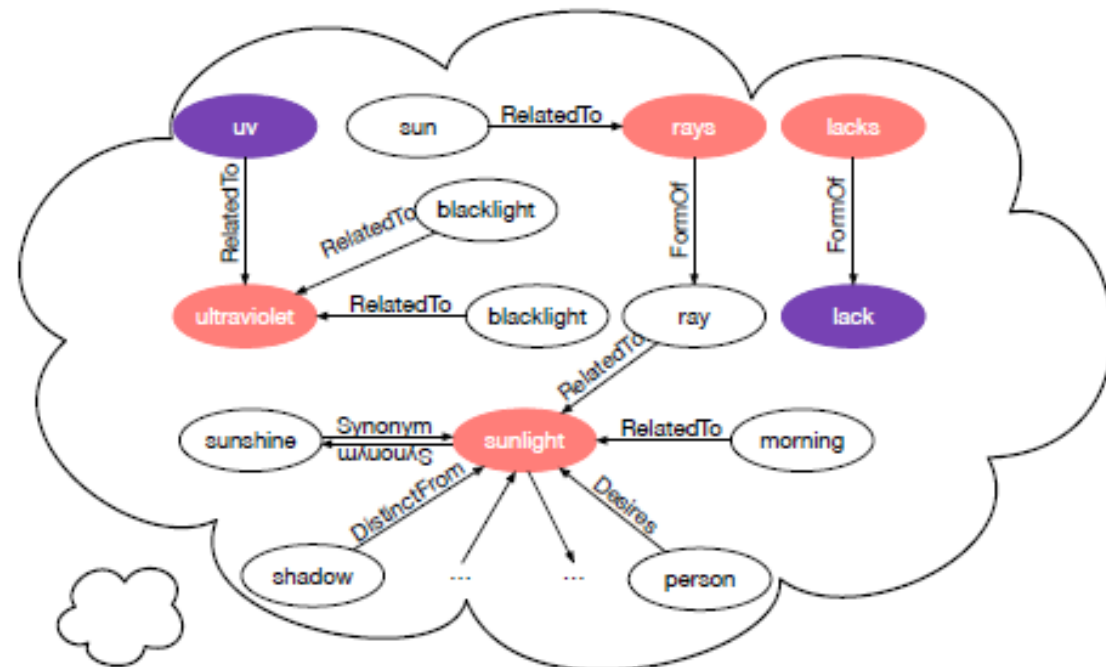
全国信用信息共享平台(二期)项目的子模块, 使用33000多个经过整理的正负向情感词结合电荷传播模型, 解决新闻舆情中的情感识别问题。

06

知识图谱

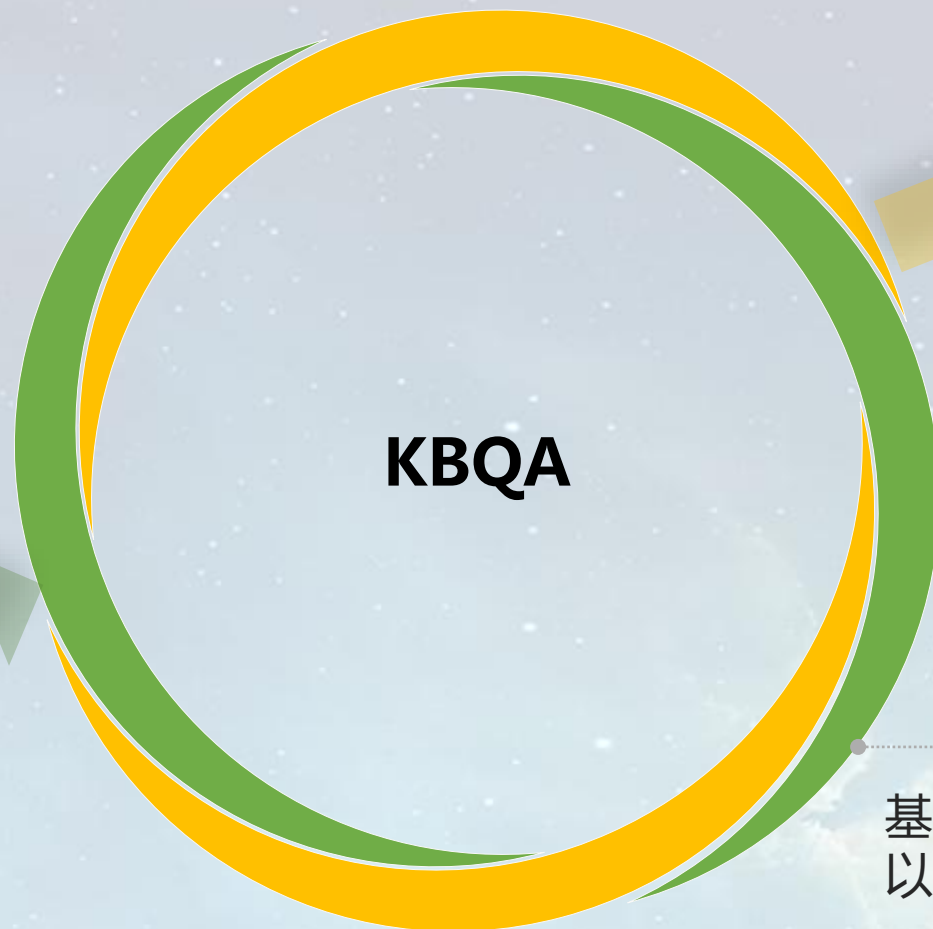
知识图谱

通过三元组（<entity, relation, entity>）
来作为知识的表现形式。



Moonlight lacks the ultraviolet rays of sunlight. → I don't think that's a lack of uv.

Moonlight lacks the ultraviolet rays of sunlight. → I'm not sure what you're saying.



描述

给定一个知识图谱和一个自然语言问题，基于图谱回答问题。

方案

从问题中解析出关键要素，然后从知识图谱中查询得到答案。

战神网信-军事搜索

基于知识图谱进行1度关系以内的知识问答。

3

未来规划

1、社会信用体系建设

参考政策

国务院关于印发社会信用体系建设规划纲要（2014—2020年）的通知

http://www.gov.cn/zhengce/content/2014-06/27/content_8913.htm

具体业务

- 围绕舆情数据
- 识别行为数据预测信用问题：
 - 1) 事件的类型（触发词）
 - 2) 涉事的主体
 - 3) 事件的抽取
- 特定领域的舆情算法集
- 及时性：实时处理，实时监控，实时推送

2、打击非法集资

具体业务

- 围绕特定金融或涉金融企业，P2P企业，跟踪其在更广泛舆情上的表现
- 识别非法集资的倾向
 - 1) 事前：尽可能在出问题、出状况之前发现并提前预警
 - 2) 事中：安抚、疏导、维稳
 - 3) 事后：监管打击后的效果验收
- 特定领域的舆情算法集

3、资本市场监管

具体业务

- 财务造假，数据挖掘
- 信息披露，多源数据（公告、舆情、其他）抽取的相互印证
- 内幕交易

4、中小微企业信用贷款（基础解析）

具体业务

- 围绕裁判文书、招投标、行政处罚、舆情等基础数据
- 构建企业关系图谱
 - 1) 实体库
 - 2) 事件库
 - 3) 关系识别
 - 4) 相关的消歧（链接）

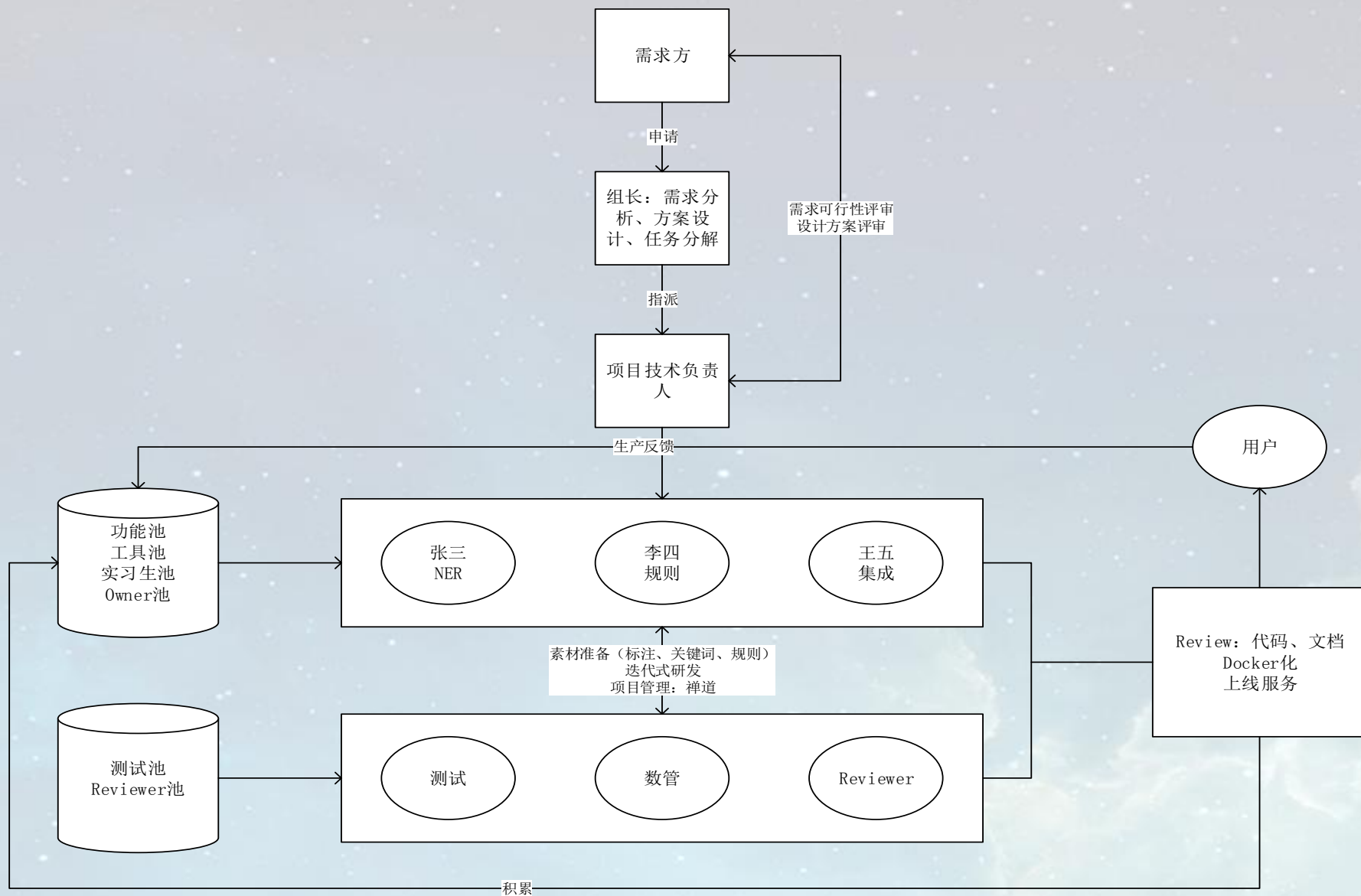
5、舆情指数ETF (yy)

具体业务

- 将市场上被讨论最多的若干有代表性股票的舆情数据都搜集起来做情感识别（做空做多）。
 - 包括社交网络数据，论坛数据，主流媒体数据等
- 指数本身就是一个投资组合，通过舆情来决定做多或做空哪些股票（需相关业务专家支持）。
 - 纯技术驱动的技术产品，技术结果直接影响着盈亏

4 立项流程

立项流程图



注意事项:

- 以邮件立项为起点，发送给曾途（CEO）、吴桐、赵磊、刘世林、李倩、胡兵、邢杰、张祎迪，向CEO说明该项目的商业或技术价值，CEO（或赵磊代理）审批之后，才会进入后续立项流程
- 张祎迪将项目信息同步到禅道，项目管理协调各方进行任务的指定和指派，任务完成之后要及时结项，否则下次立项降低优先级

THANK YOU FOR WATCHING

Friday, February 22, 2019

www.bbdservice.com