DataFunSummit

自然语言处理

在线峰会

产业创新与实践 论坛

2021.07.10（周六） 09：00~17：30

2021

DataFunSummit

基于NLP的产业链构建与应用

嘉宾 深证信息 毛瑞彬

目录

CONTENTS

01
理论与背景

Subject

产业链建设驱动力、理论、行业现状和建设目标

驱动力-价值发现和风险识别

01 助力监管科技
上市审核和持续监管

02 服务金融市场
标的筛选、借贷管理、风控管理

03 区域经济发展
产业升级、招商和企业扶持

产业链理论

■ 劳动分工

✓人类起源于群居者之间的分工

✓亚当斯密为代表的古典主流经济学家从微观层面分析劳动分工和专业化对提高企业劳动生产率的作用。

✓新古典理论和新古典学派经济学代表人物马歇尔将分工扩展到企业之间，强调企业之间分工协作的重要性，从而开启了对产业链的研究。

■ 产业链和价值链

✓Porter（1985）《竞争优势》《竞争战略》最早提出价值链理论，把企业各部分分为基本活动和支持活动，强调企业的竞争优势。

✓Krugman（1995）分析过企业将内部各个价值环节在不同地理空间进行配置的能力问题，这开启了价值链理论中价值链治理模式与产业空间转移之间关系的重要研究领域。

■ 国家产业链安全

行业现状

现有的产品主要聚焦二级市场的投顾、投研和风控，少数面向一级市场。

行业分类

面向B端的数据产品

行业上下游

面向C端投顾

热点概念

面向C端投顾

要素描述

仅限于行业和公司

建设目标

应用自然语言处理和软件工程思想，基于国民经济行业分类、投入产出表等，从产业、行业到企业，构建一个多层次多维度产业链知识图谱，为监管、投融资和招商等应用领域提供服务。

产业

了解产业链完整视图，产业链的发展，上中下游重点细分行业，核心监管机构等

行业

了解上下游和替代行业发展情况，以及本行业规模、竞争格局、发展历史与趋势、政策法规等

公司

通过可比公司重点财务数据，结合产业链细分行业的价格、销量、融资等信息，了解企业的发展前景。

02
架构与流程

Subject

依据目标，设计产业链本体，建设路径，
系统架构和自动构建流程

建设路径

搜集资料

互联网海量文本、研究报告、第三方报告等资料收集。

人工审核

审核产业链重点要素的数据数量及质量，核心难点依靠互联网众包平台辅助审核。

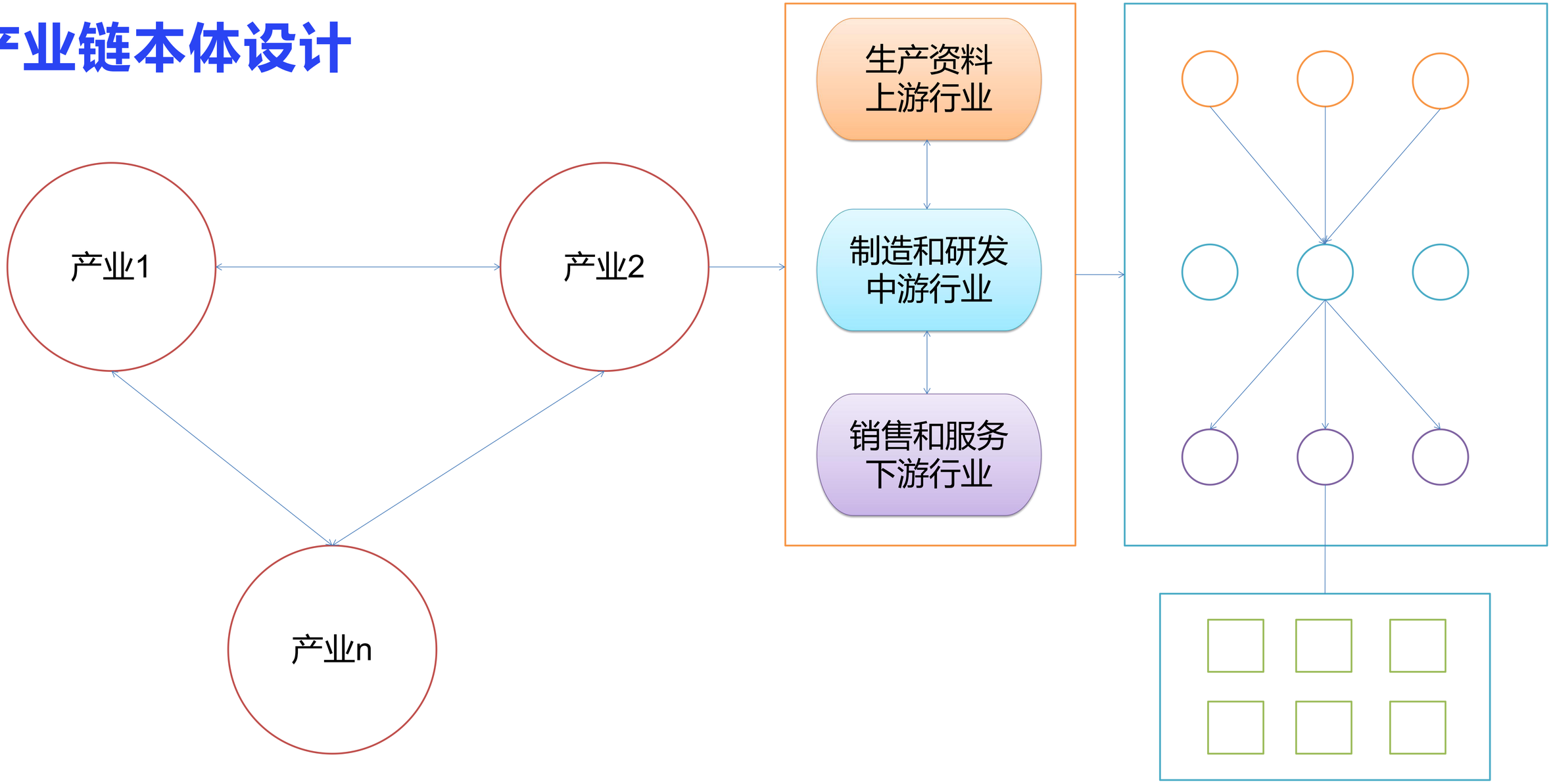
建立框架

定义产业链本体、样例公司和样例数据

自动构建

PDF文档解析，行业数据抽取、行业及上下游识别、要素结构化。

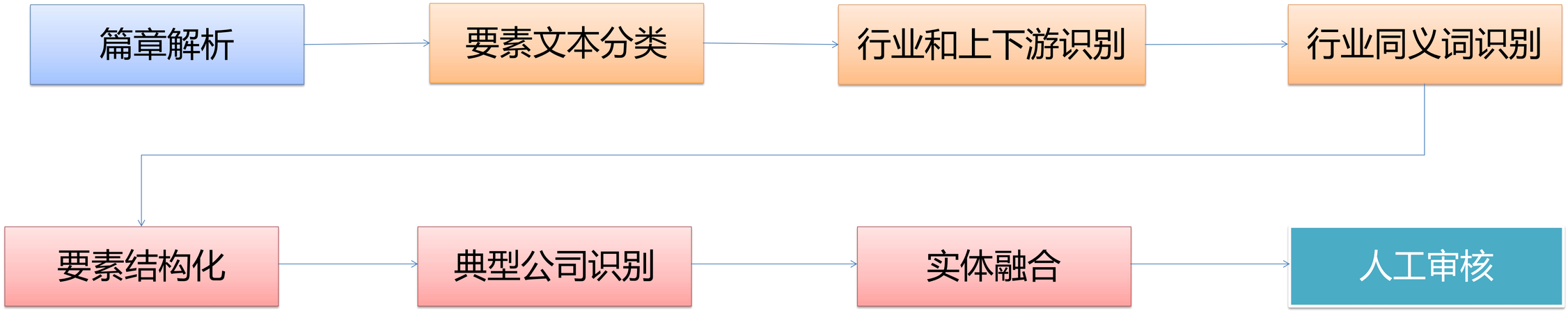
产业链本体设计



系统架构


服务层
监管
投融资
产业规划
知识层
产业
行业
企业
要素
本体库
计算层
机器学习
批处理
图数据库
数据源
研报/公告/新闻
国民经济行业分类
GICS
全国投入产出表

自动构建流程



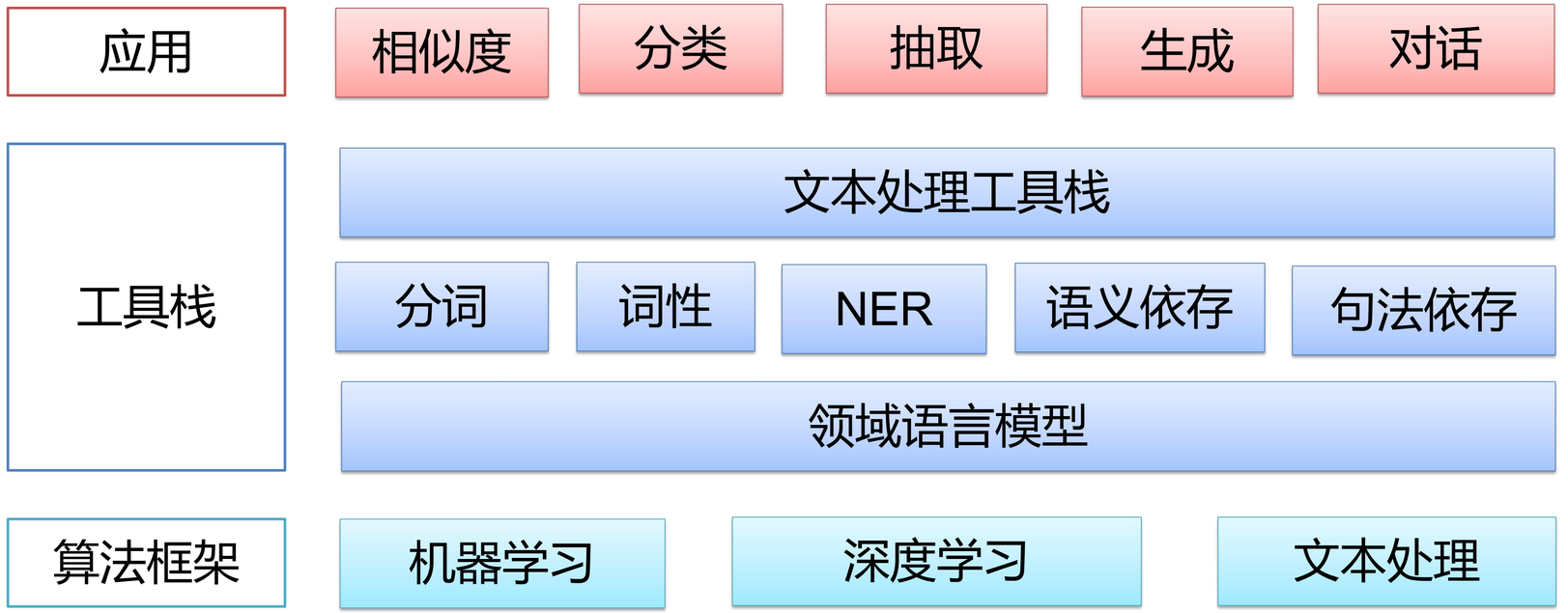
分类　相似度　NER　关系抽取

03
关键方法

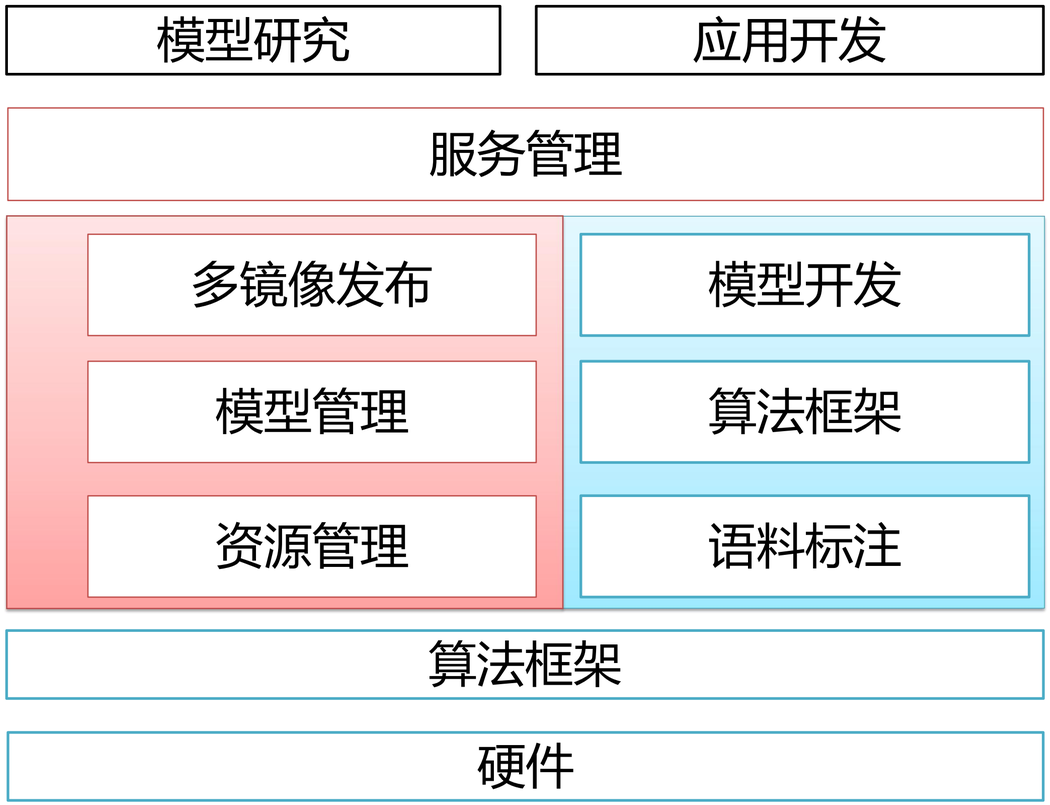
Subject

包括基础设施和自动构建环节中的关键技术，基础设施为模型任务提供平台和工具

基础设施-算法架构


应用
相似度
分类
抽取
生成
对话
工具栈
文本处理工具栈
分词
词性
NER
语义依存
句法依存
领域语言模型
算法框架
机器学习
深度学习
文本处理

基础设施-机器学习平台

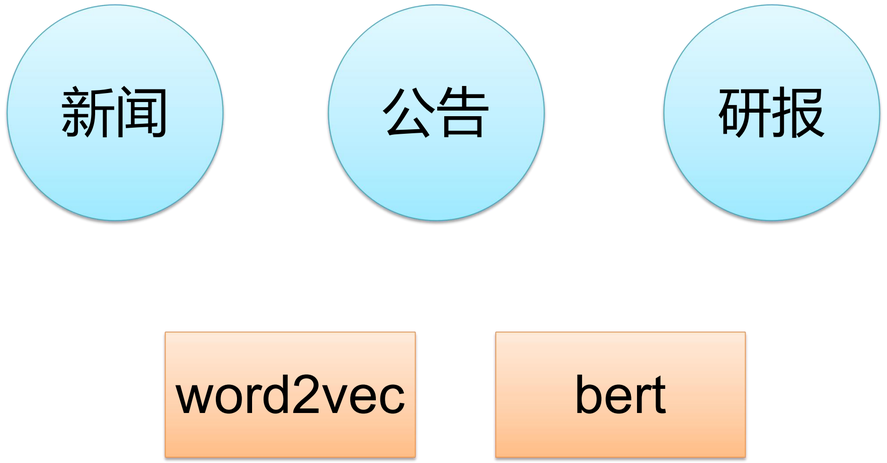


✓基于NLP构建产业链的核心是文本处理能力和模型能力

✓面向领域的NLP工具（分词、NER、句法和语义）和语言模型是快速建模的支撑

✓算法流程管理和自动化辅助是快速应用的基础

基础设施-语言模型



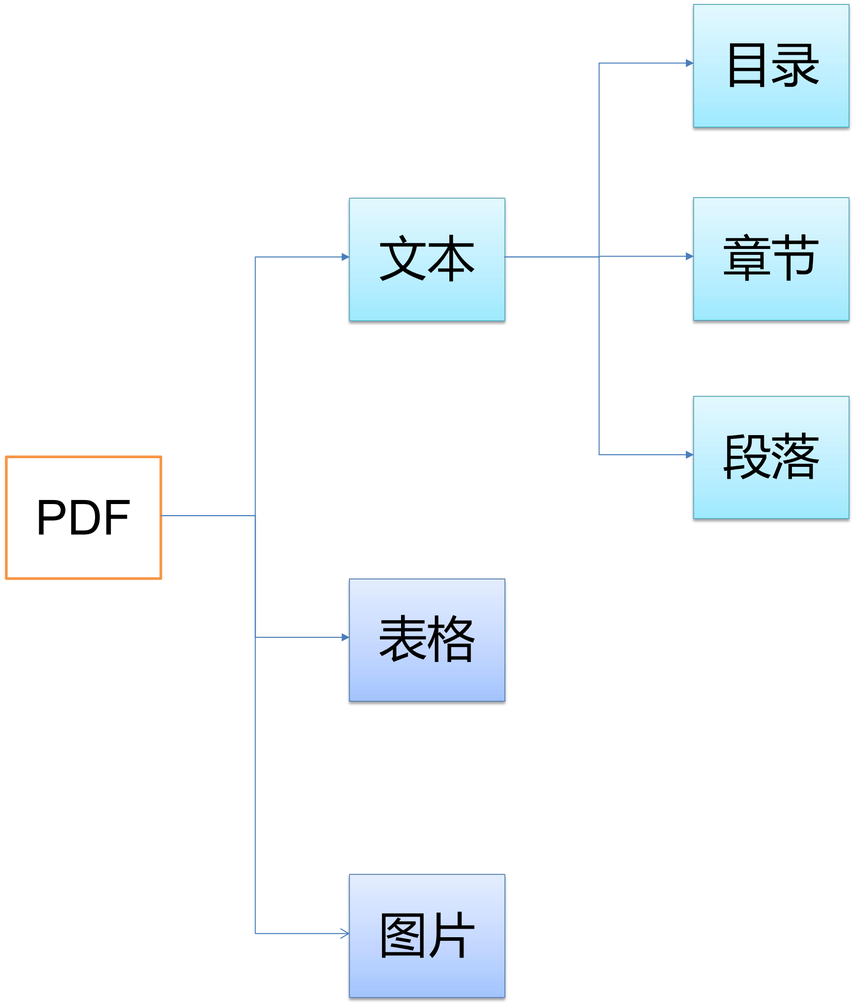
- ✓证券领域三种文本类型，新闻、公告和研报
- ✓面向不同任务和应用场景，同时支持word2vec和bert
- ✓基于语言模型可高效实现相似度、分类、实体和关系识别等下游应用

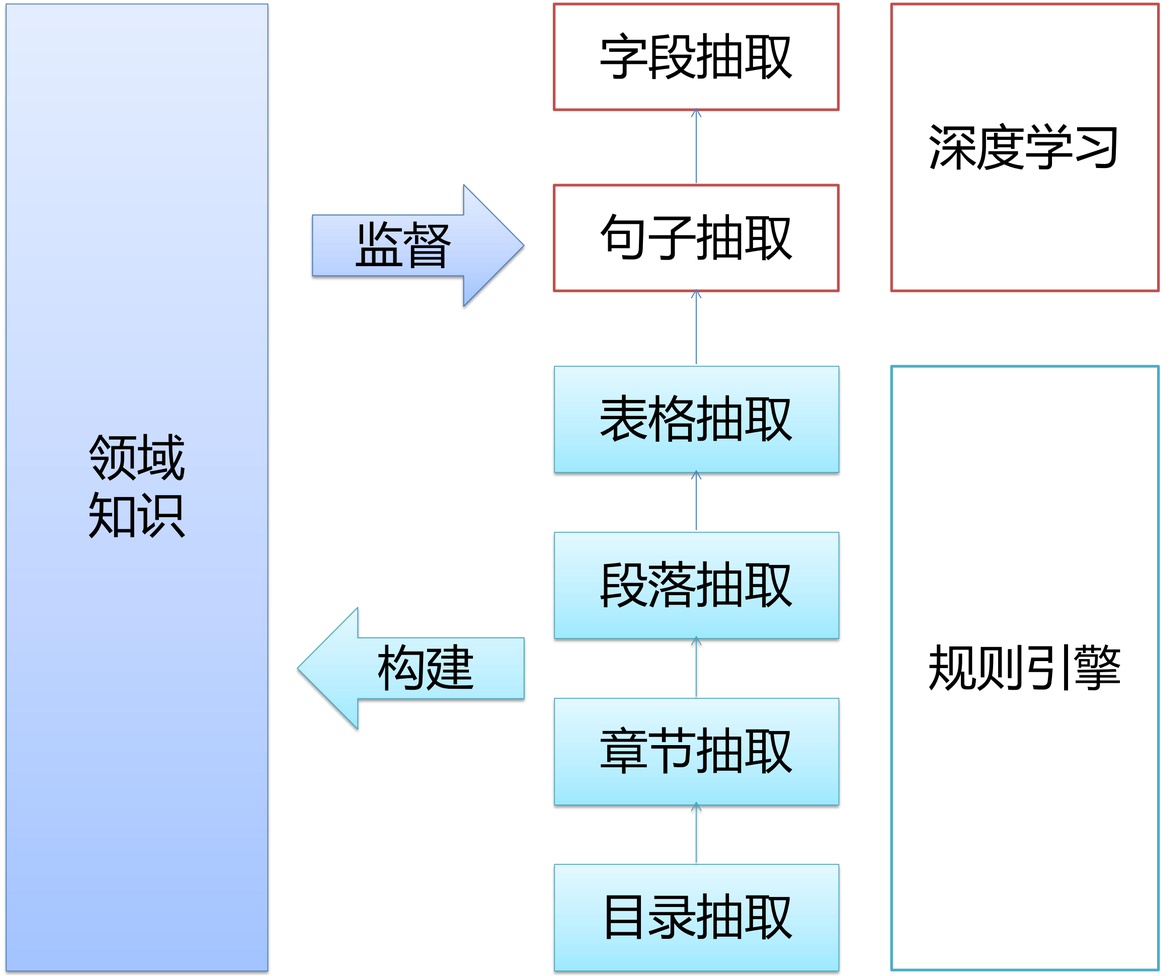
基础设施-词法和句法分析

分词方法	召回率	准确率	F值
领域词法	0.969	0.955	0.962
中科院	0.910	0.852	0.880
hanlp	0.894	0.860	0.876
jieba	0.832	0.832	0.832

- 面向证券领域的句法和语义工具
  - ✓支持证券领域典型的复杂句式
  - ✓句法的输入是词法工具的输出
  - ✓操作粒度全流程保持一致性
  - ✓实体单独作为成分
  - ✓通过化简降低处理难度

自动构建-篇章解析


PDF
文本
表格
图片
目录
章节
段落

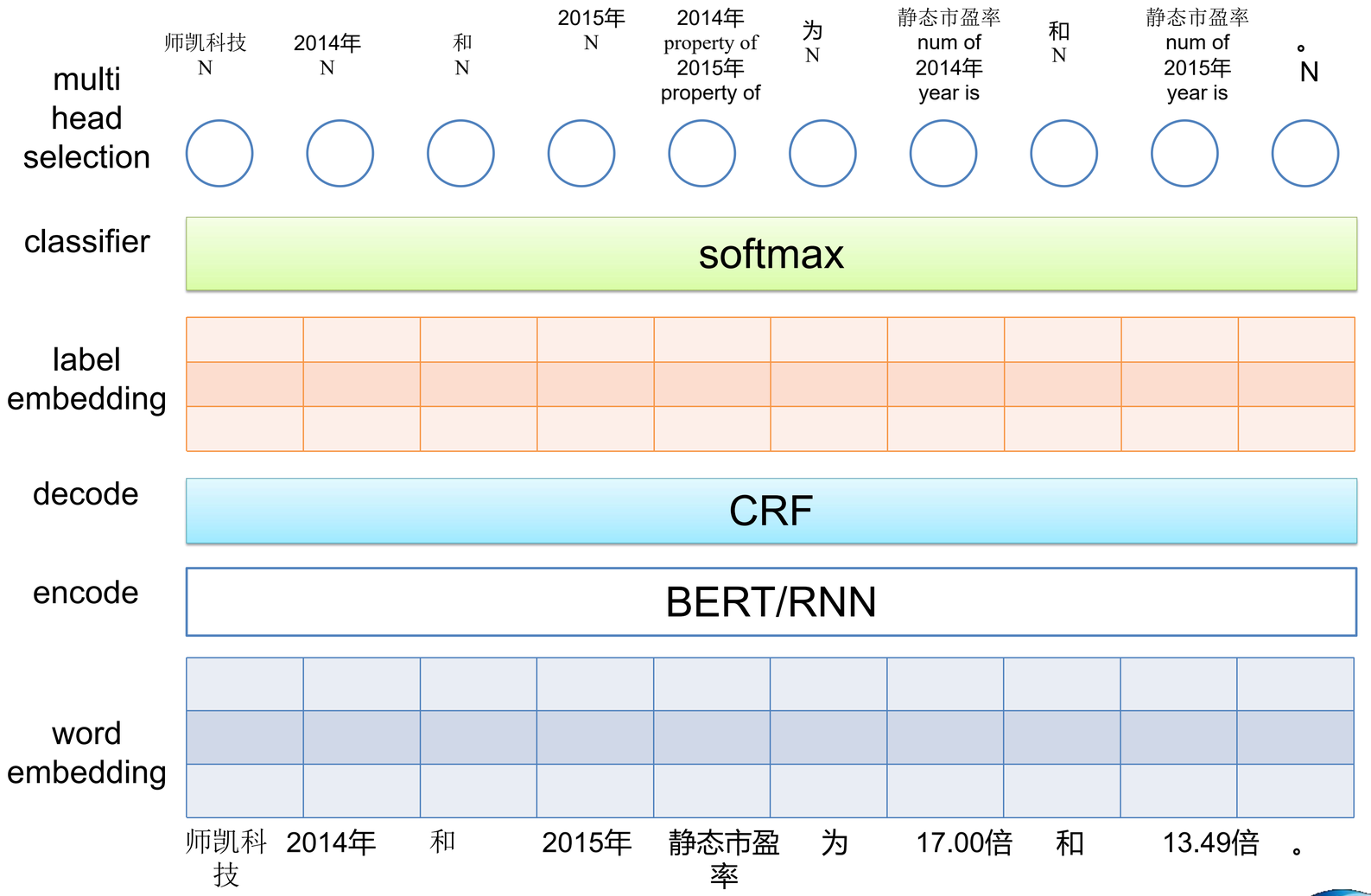

领域
知识
监督
构建
字段抽取
句子抽取
表格抽取
段落抽取
章节抽取
目录抽取
深度学习
规则引擎

自动构建-行业要素分类

类别	特征
行业规模	行业名称、行业规模指标词、数字
竞争格局	公司/地域/产品、百分比
政策法规	官方机构、书名号、行业名称
发展历史与趋势	时间段、行业名称
节点定义	行业名称、产品名称、定义指示词
行业壁垒	行业名称、技术名词、指示词
…	…

段落 → NER/句法 → 特征构建 → 分类器 → 分类结果

自动构建-行业和上下游识别



- ✓基于领域语言模型
- ✓实体和关系联合学习
- ✓识别行业名称
- ✓识别多个行业之间的上下游关系
- ✓联合学习能够解决pipline模型误差传播的问题。

自动构建-行业和上下游识别

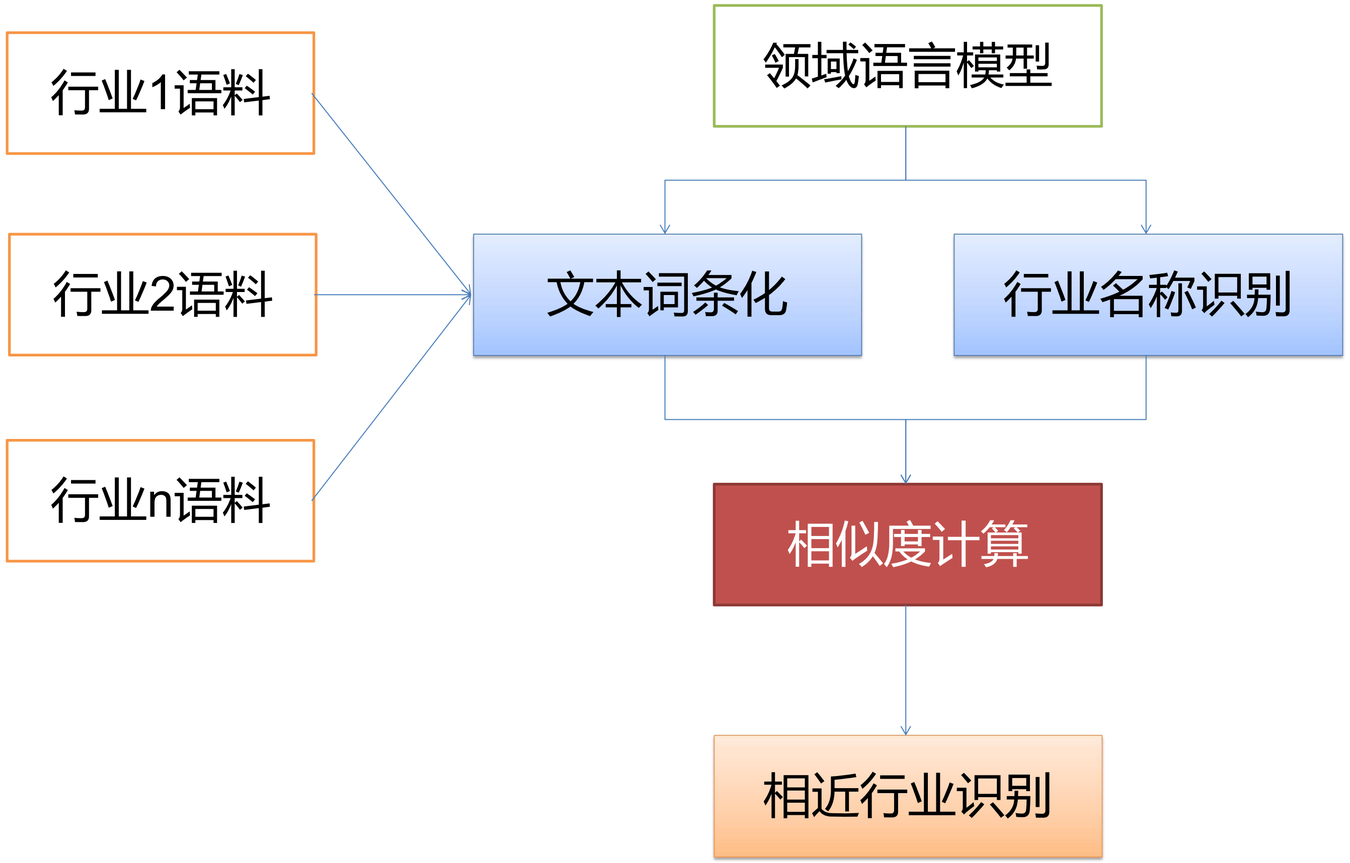
行业	行业	关系	句子
聚酯化纤	纺织、服装、汽车	上游	我国聚酯化纤产业经过数年调整,产业结构优化,下游以纺织、服装、汽车等为主的民用和工业终端消费需求保持稳定增长,出口企稳回升,聚酯纤维应用替代范围不断加大,聚酯产业链景气回升。
MIM (金属注射成形)	汽车、医疗器械	上游	此外,MIM在汽车、医疗器械领域具有巨大的潜在市场空间,根据我们测算的数据,假设未来MIM在汽车零部件、医疗器械领域有1%的渗透率,MIM在国内汽车零部件领域市场空间将超过320亿元,在全球医疗器械领域市场空间将接近40亿美元。
油墨、涂料、塑料	有机颜料	下游	基于有机颜料具有鲜艳的色光、较高的着色强度以及色谱齐全等特点,是生产多种产品不可缺少的着色剂,其应用用途十分广泛,世界有机颜料的消费主要集中在油墨、涂料和塑料领域,产品具有相当的刚需属性,近年全球有机颜料整体消费增速在3~5%区间,未来有机颜料需求仍能够保持稳速增长。
家电、工控、汽车、 通讯、电力、能源、 安防、航空航天	继电器	下游	继电器作为最主要的基础元件之一，是整机电路控制系统中必要的、核心的电控基础元件，广泛应用于家电、工控、汽车、通讯、电力、能源、安防、航空航天等领域，主要作用是实现"自动、远程"控制。

自动构建-要素结构化

从市场规模来看，近年来全球体外诊断市场增长稳定，2013年全球体外诊断市场规模达到了554亿美元，2015年全球体外诊断市场规模约为634.27亿美元，预计2015年到2018年，将以7%的年度复合增长率平稳增长，到2018年预计可以达到777.01亿美元。

地域范围	行业	时间	属性	金额
全球	体外诊断	2013年	市场规模	554亿美元
全球	体外诊断	2015年	市场规模	634.27亿美元
全球	体外诊断	2018年	市场规模	777.01亿美元

自动构建-行业同义词识别



●除了上下游以外，还有一些相同相近行业，也需要进行识别

✓通过领域文本词条化对行业名称进行扩展

✓利用语言模型和相似度计算进行识别

行业	相近行业
防腐涂料	卷材涂料、隔热涂料、聚氨脂涂料
沙格列汀	利格列汀、瑞格列汀、恒格列净
手机游戏	移动游戏、街机游戏、体感游戏

自动构建-典型公司识别

产品信息

行业网站

公告和研报

电商网站

公司信息

公司简介

公司网站

新闻

工商信息

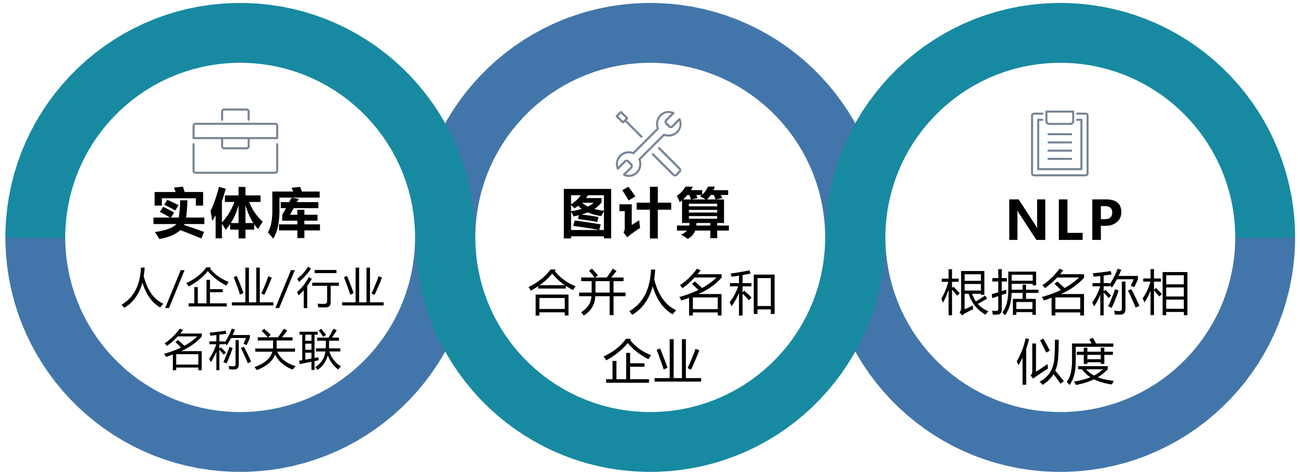
软著

商标

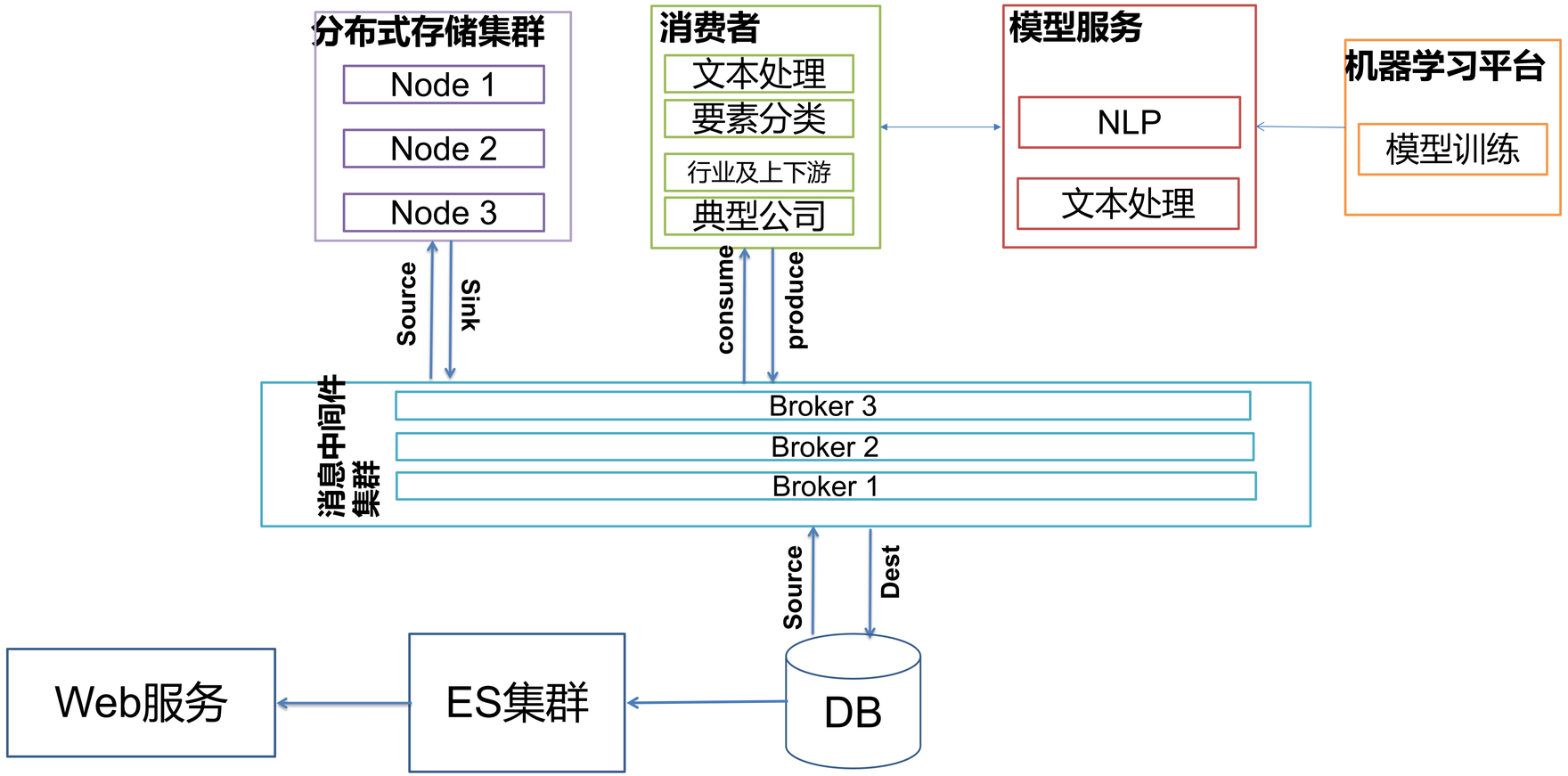
主营业务

自动构建-实体融合

利用实体库、图计算和自然语言处理等方法对实体进行融合



自动构建-文本批处理



- 基于流式处理，把产业链构建过程变成管道处理模式
  - ✓集成文本处理工具
  - ✓集成NLP工具
  - ✓集成数据源
  - ✓集成消费者组件
  - ✓容器化部署
  - ✓实时自动化构建

产业链系统

全球主要证券市场深度数据

燧石星火平台独家项目

全国1.4 亿工商主体多维信息

三大数据基础

产业链
细分赛道

重构资本市场数据使用方式

自建了一套包括 7000 余细分行业的分类体系，基础框架参考 GICS 标准

全国企业定位至数千个细分行业，支持单公司多行业分类，更精准查找可比公司

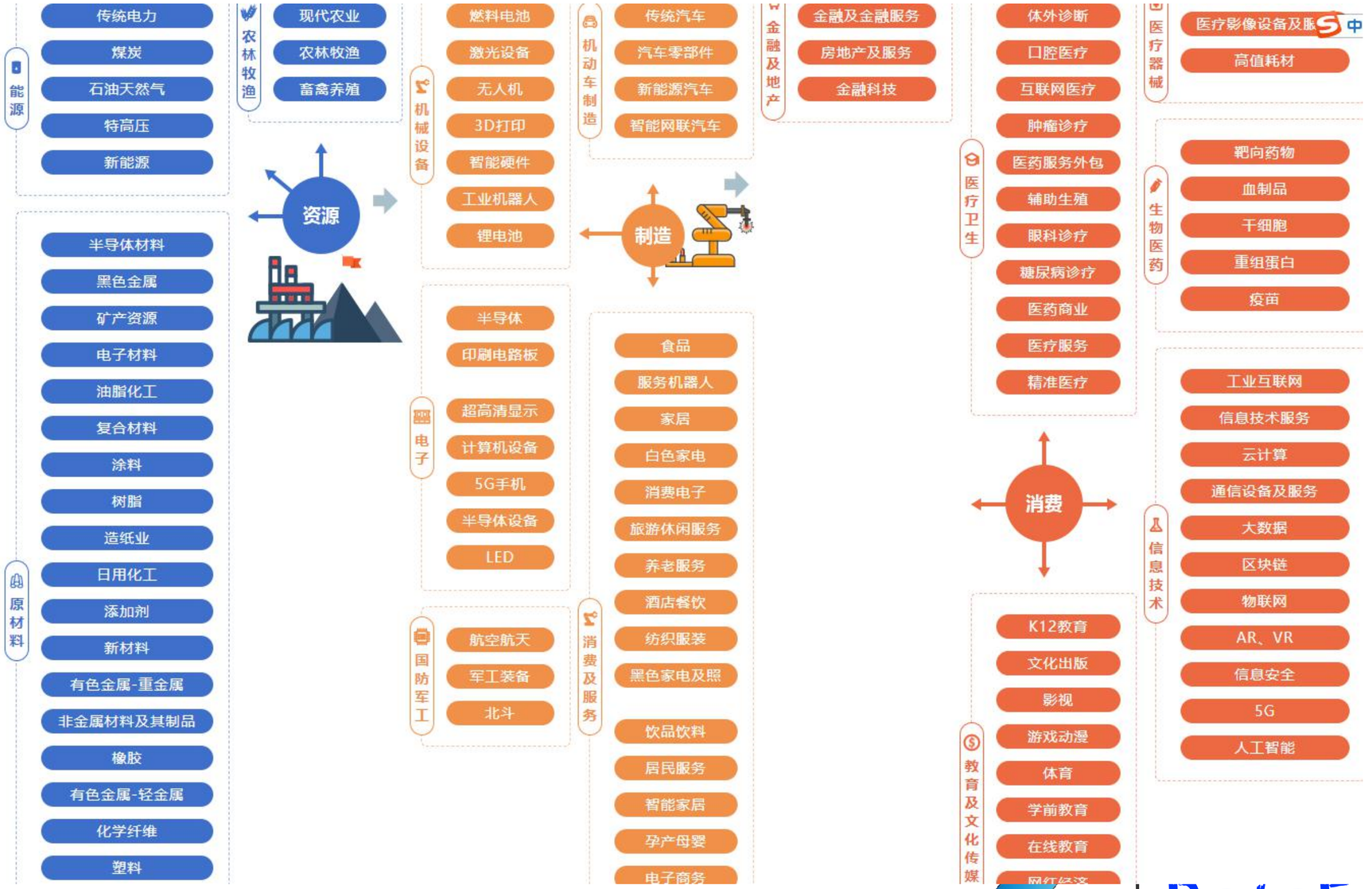
基于行业研报及其他资料，智能透视市场规模、竞争格局等行业数据

04
示例和应用

Subject

包括基础设施和自动构建环节中的关键技术，基础设施为模型任务提供平台和工具

产业链



示例-产业层面的上下游行业

示例-行业上下游

节点产业链　节点简介　典型公司　智能行业透视　财务指标　产品指标

节点产业链

节点简介

锂电池关键四大主要材料包括正极、负极、电解液、隔膜等，其革新技术的产业化落地关系着锂电产业的前进速度。其中，正极材料是锂电池最为关键的原材料，直接决定了电池的安全性能和电池能否大型化，同时也是锂电池成本占比最高的材料，约占锂电池电芯材料成本的40%左右。

正极材料作为锂电池最为关键，成本占比最高的原材料，对于锂电池生产发展具有重要影响。高工产研锂电研究所(GGII)数据显示，2016-2019年中国锂电正极材料出货量持续增长，增速不断加快。2019年中国锂电正极材料出货量40.4万吨，同比增长32.5%，加快2.5个百分点。其中，三元正极材料出货量19.2万吨，同比增幅40.7%;磷酸铁锂材料出货量8.8万吨，同比增长29.3%;钴酸锂材料出货量6.62万吨;锰酸锂材料出货量5.7万吨。

从竞争格局角度分析，2019年正极材料出货量排名前三的企业分别为厦门钨业、天津巴莫和德方纳米，市场占比分别为9.92%、6.44%和6.31%。厦门钨业受益于钴酸锂以及三元材料的双重增长，2019年其正极材料出货量同比增长超过50%;天津巴莫受益于产能释放以及终端客户需求量提升，其三元正极材料出货量增速明显，总出货市场排名上升至第二位。

从正极材料细分类型占比趋势来看，2019年，三元正极材料和钴酸锂材料出货量占比较2018年有所下降，占正极材料出货量比重分别为47.62%和16.42%;磷酸铁锂材料和锰酸锂材料出货量占比较2018年有所增加，2019年占比分别为21.83%和14.14%。

在国家一系列政策的大力支持下，新能源汽车得到了大力发展，进而带动动力电池及正极材料等行业的快速发展。高工锂电调研数据显示，2017年，中国正极材料总产值达417.1亿元，同比增长95.1%，2018年我国锂电池正极材料产值为531.5亿元，产值增速有所放缓;初步核算，2019年我国锂电池正极材料产值达到737亿元，千亿市场规模未来可期。

示例-行业要素

节点产业链 | 节点简介 | 典型公司 | 智能行业透视 | 财务指标 | 产品指标

典型公司

- 北京当升材料科技股份有限公司 (300073)
- 湖南杉杉能源科技股份有限公司 (835930)
- 宁波杉杉股份有限公司 (600884)
- 厦门钨业股份有限公司 (600549)
- 横店集团东磁股份有限公司 (002056)
- 湘潭电化科技股份有限公司 (002125)
- 贵州安达科技能源股份有限公司 (830809)
- 宁波容百新能源科技股份有限公司 (688005)
- 湖南中科电气股份有限公司 (300035)
- 天津巴莫科技有限责任公司
- 江门市科恒实业股份有限公司 (300340)
- 贵州振华新材料股份有限公司 (K0522)
- 科达制造股份有限公司 (600499)
- 国轩高科股份有限公司 (002074)
- 欧赛新能源科技股份有限公司 (836058)
- 烟台卓能电池材料股份有限公司 (834314)
- 湖南长远锂科股份有限公司 (K0265)
- 江西正拓新能源科技股份有限公司 (831980)
- 绵阳富临精工股份有限公司 (300432)
- 湖南瑞翔新材料股份有限公司
- 浙江华友钴业股份有限公司 (603799)
- 湖南杉杉能源科技股份有限公司 (835930)
- 中信国安信息产业股份有限公司 (000839)
- 格林美股份有限公司 (002340)
- 深圳中华自行车（集团）股份有限公司 (200017)
- 贝特瑞新材料集团股份有限公司 (835185)
- 上海璞泰来新能源科技股份有限公司 (603659)
- 深圳市德方纳米科技股份有限公司 (300769)
- 北大先行科技产业有限公司
- 江西紫宸科技有限公司

点击下拉查看更多

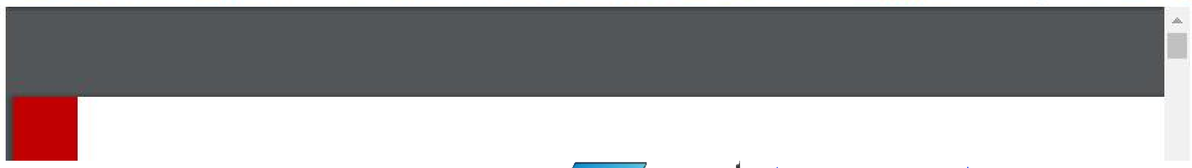
智能行业透视

全部 | 行业规模结构化 | 行业规模 | 竞争格局 | 发展历史与趋势 | 政策法规 | 产业链 | 行业细分 | 节点定义 | 行业壁垒 | 行业驱动力

文本行业	区域	时间	数据类型	数据值	披露机构	披露文件
正极材料		2021年1月	产量	6.1万吨		
正极材料	国内	2021年1月	月产量	6万吨	ICC鑫椤...	

新能源汽车数据月报（2021年1月）：产业链迎来开门红，景气度迈上新台阶

全屏阅读



应用-IPO审核

大宗商品原材料采购价格比对：

自动抽取发行人招股书披露的大宗商品采购价格，并与行业价格数据进行比对，对存在重大差异的情况进行提示。

应用-IPO审核

上市公司分产品毛利率比对：

- 通过将上市公司收入项目对齐至自有行业分类体系，实现分产品项目数据比对；
- 毛利率为最具综合性的财务指标，分产品毛利率比对有助于识别经营/财务异常。

应用-股权投资

以工业机器人为例进行企业筛选：

- 按照企业规模、经营状况进行排序，优先显示路演、融资和高新技术企业；
- 根据企业所在行业与行业的典型公司进行多维度对比，从而筛选投资标的。

武汉海默机器人有限公司	行业典型公司 路演 融资 高新技术企业 未上市挂牌			协作机器人
法定代表人：肖海峰	注册时间：2016-09-27	注册资本：4822.3626万元人民币		
深圳市大寰机器人科技有限公司	行业典型公司 路演 融资 未上市挂牌			末端执行器
法定代表人：孙杰	注册时间：2016-12-23	注册资本：122.2449万元人民币		
青岛宝佳自动化设备有限公司	行业典型公司 路演 高新技术企业 未上市挂牌			码垛机器人 工业互联网软件
法定代表人：高明作	注册时间：2009-03-18	注册资本：12098万		
上海达野智能科技有限公司	行业典型公司 路演 未上市挂牌			并联机器人（Delta）
法定代表人：刘增杰	注册时间：2017-06-27	注册资本：1000万元人民币		
勃肯特（天津）机器人技术有限公司	行业典型公司 路演 未上市挂牌			并联机器人（Delta）
法定代表人：王岳超	注册时间：2017-11-02	注册资本：1104.1667万元人民币		

应用-股权投资

以POCT和工业机器人为例筛选行业：

- 通过对比当前国内行业规模和全球行业规模筛选；
- 也可以通过对比当前行业规模和未来行业规模的空间进行筛选
- 还可以对比多个行业的行业规模差异进行筛选。

智能行业透视

全部 | 行业规模结构化 | 行业规模 | 竞争格局 | 发展历

文本行业	区域	时间	数据类型	数据值	披露机构	披露文件
工业机器人	我国	2023年	业规模	156亿美元		
工业机器人	全球	2023年	市场规模	335亿美元		
工业机器人	我国	2023年	年销量	26万台	高工机器...	

智能行业透视

全部 | 行业规模结构化 | 行业规模 | 竞争格局 | 发展历史

文本行业	区域	时间	数据类型	数据值	披露机构	披露文件
POCT	全球	2026年	市场规模	240亿美元		
POCT	我国	2026年	市场规模	15亿美元		
POCT	我国	2025年	市场规模	225亿元	TriMark	
POCT	中国	2025年	市场规模	225亿元		

DataFunSummit

THANKS!

今天的分享就到这里...

Ending