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

02 架构与流程

03 关键方法

04 应用及示例

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。



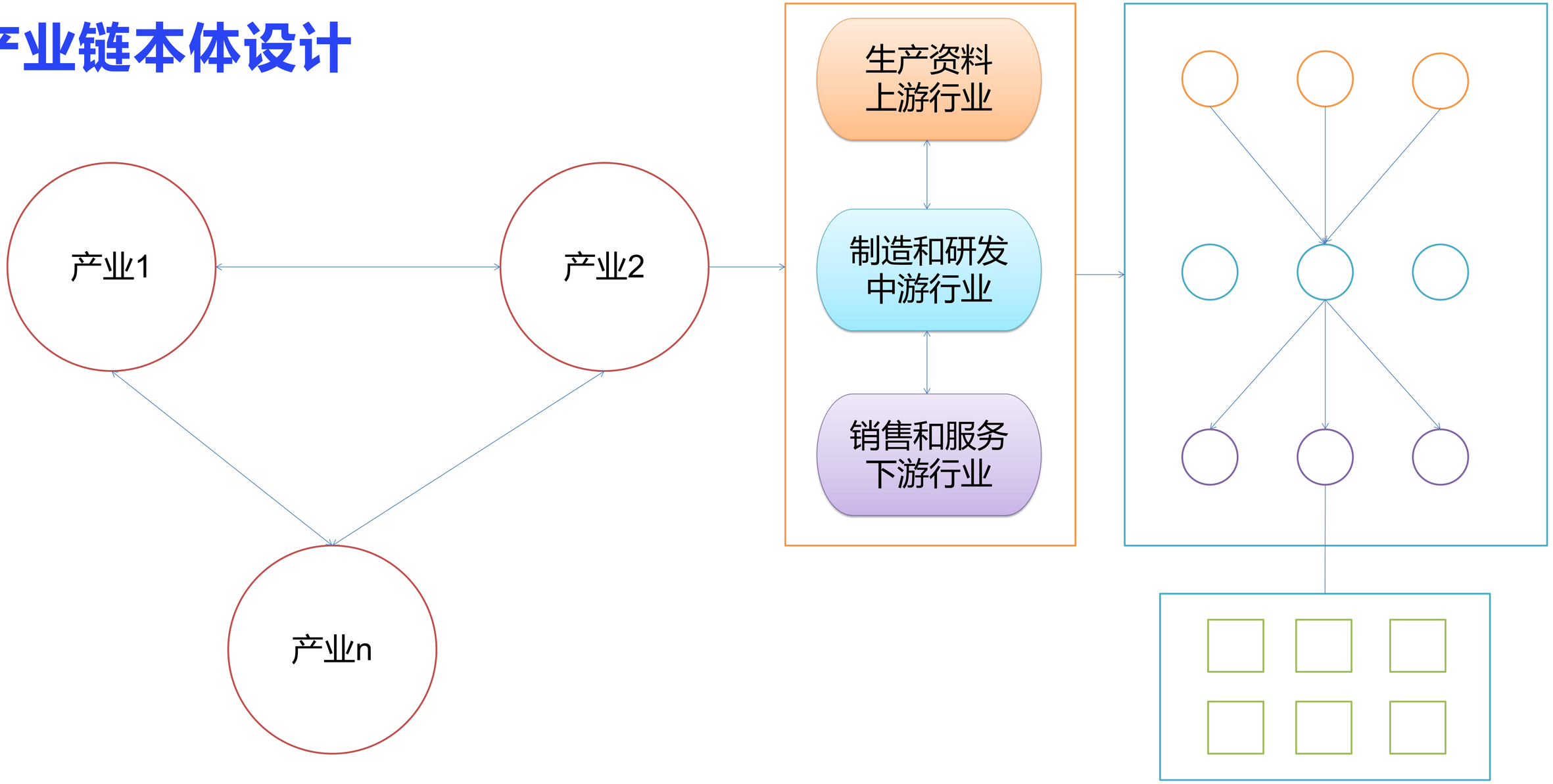
建立框架

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。

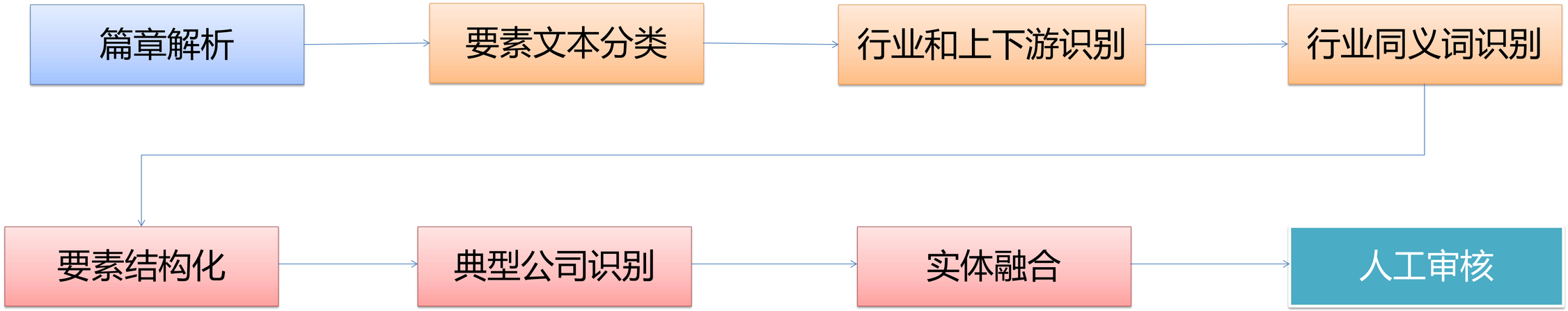
产业链本体设计



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



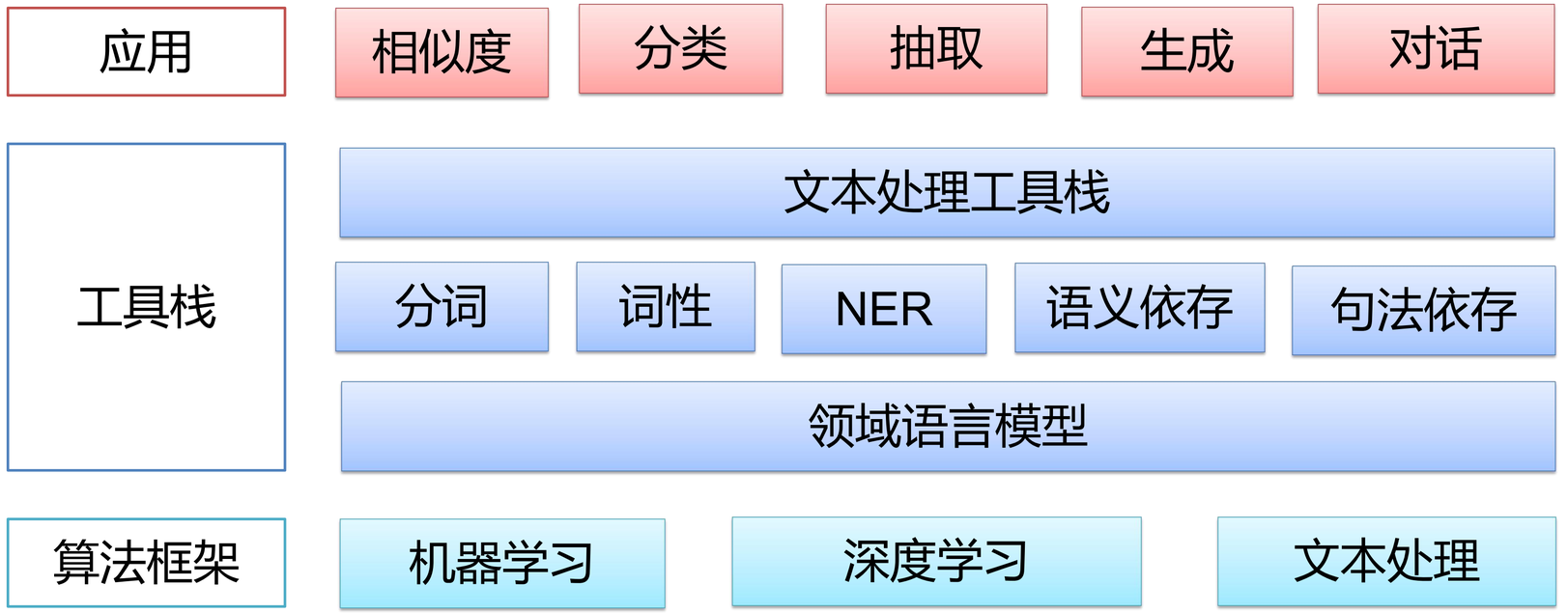
分类　相似度　NER　关系抽取

03
关键方法

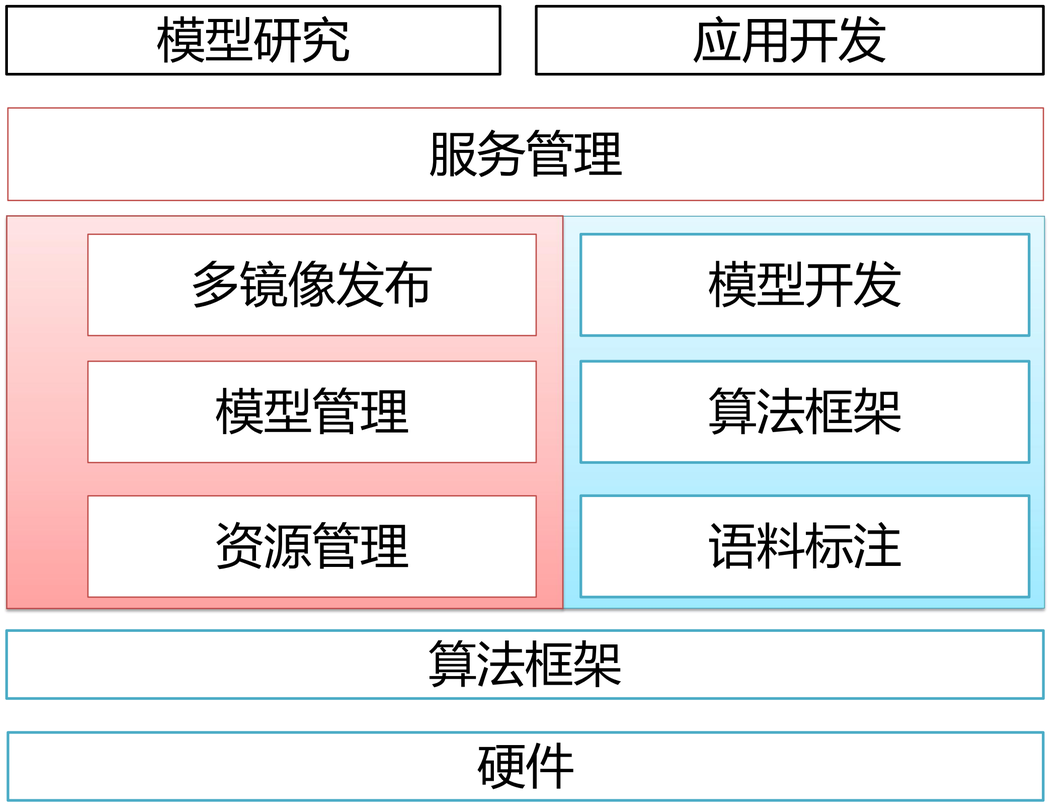
Subject

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

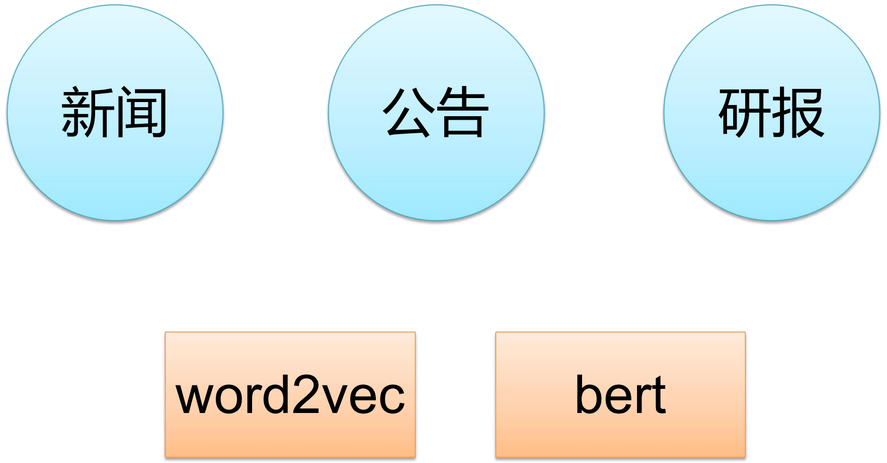


✓基于NLP构建产业链的核心是文本处理能力和模型能力

✓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



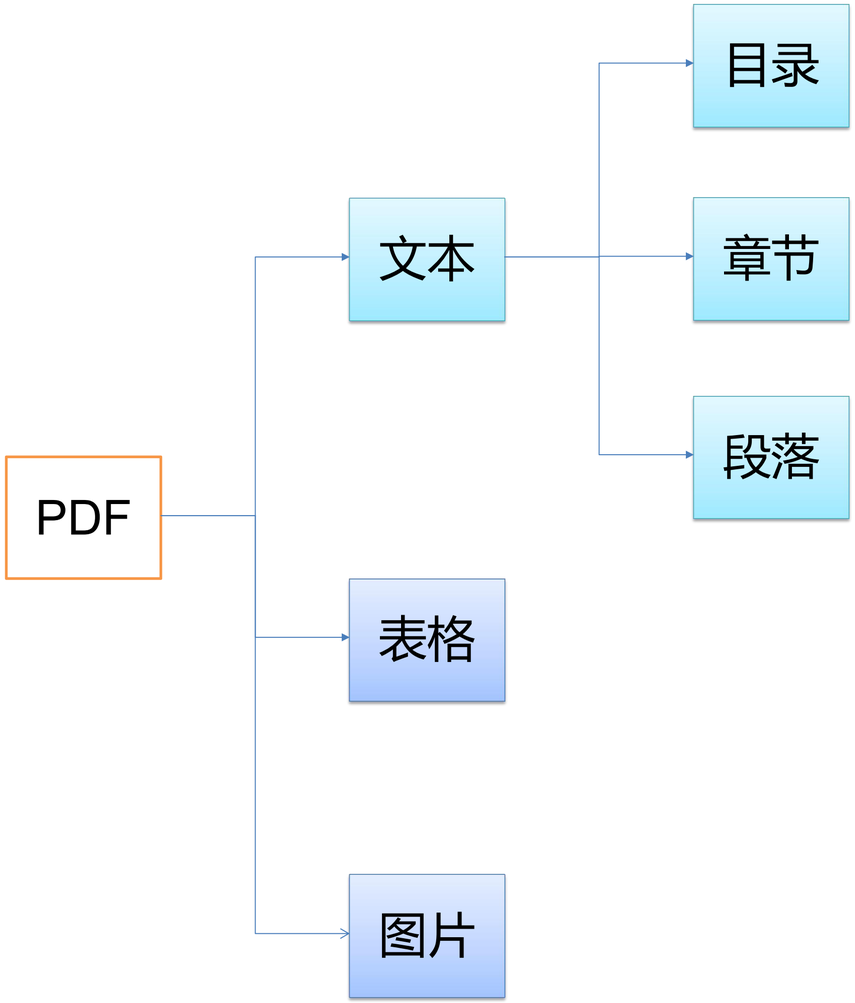
- ✓证券领域三种文本类型，新闻、公告和研报
- ✓面向不同任务和应用场景，同时支持word2vec和bert
- ✓基于语言模型可高效实现相似度、分类、实体和关系识别等下游应用

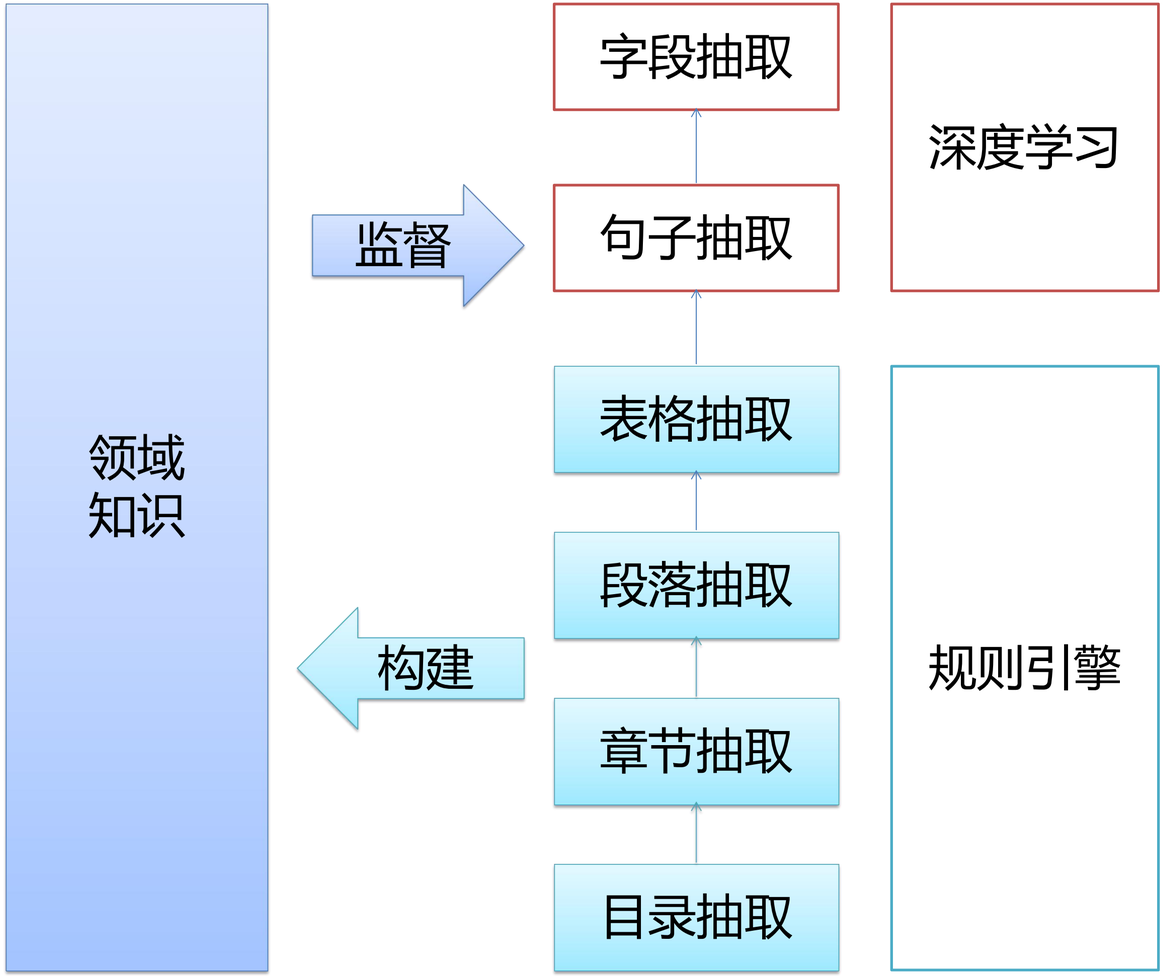
基础设施-词法和句法分析

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

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：https://d.book118.com/628031045024006032