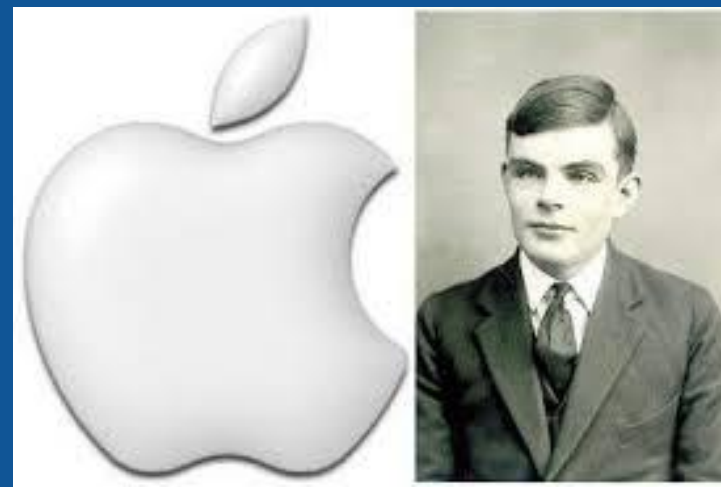
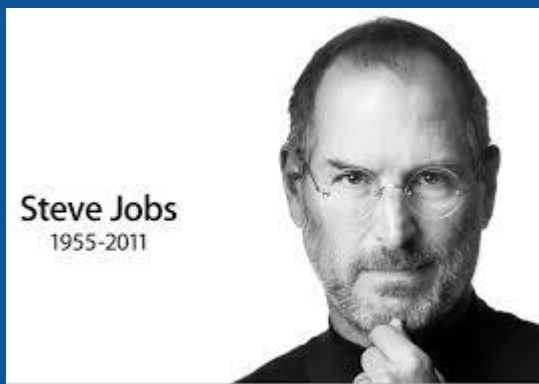


知识图谱

技术分享

一句话说明什么是知识图谱

“苹果公司的创始人是乔布斯”



知识图谱就是实体及其链接

目录

C O N T E N T S

01

图谱简介

02

图谱实践

03

命名实体抽取

04

关系抽取

05

图谱问答

06

总结扩展



1

the part one

知识图谱简介



1、知识图谱简介

1.1、发展脉络

1960 语义网络
1989 WEB
1998 语义网
2006 链接数据
2012 知识图谱
2017 事理图谱

1.2、技术体系

“理论”：命名实体抽取、关系抽取、知识融合、嵌入模型。

“实践”：图谱构建、图谱存储、图谱应用、NLP增强

知识图谱

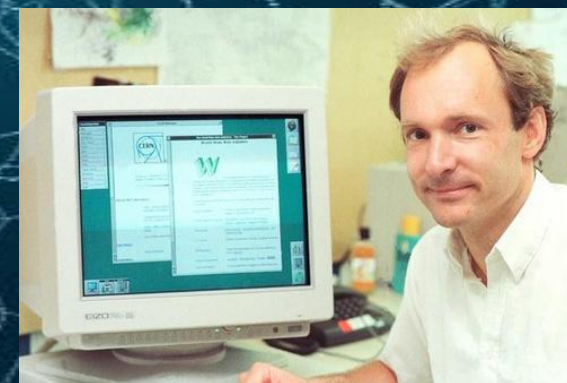
1.3、开源项目



1.4、应用方向

推荐系统
安全领域
金融风控
智能对话

1.1、发展脉络——超文本与WEB



1969年
ARPANET

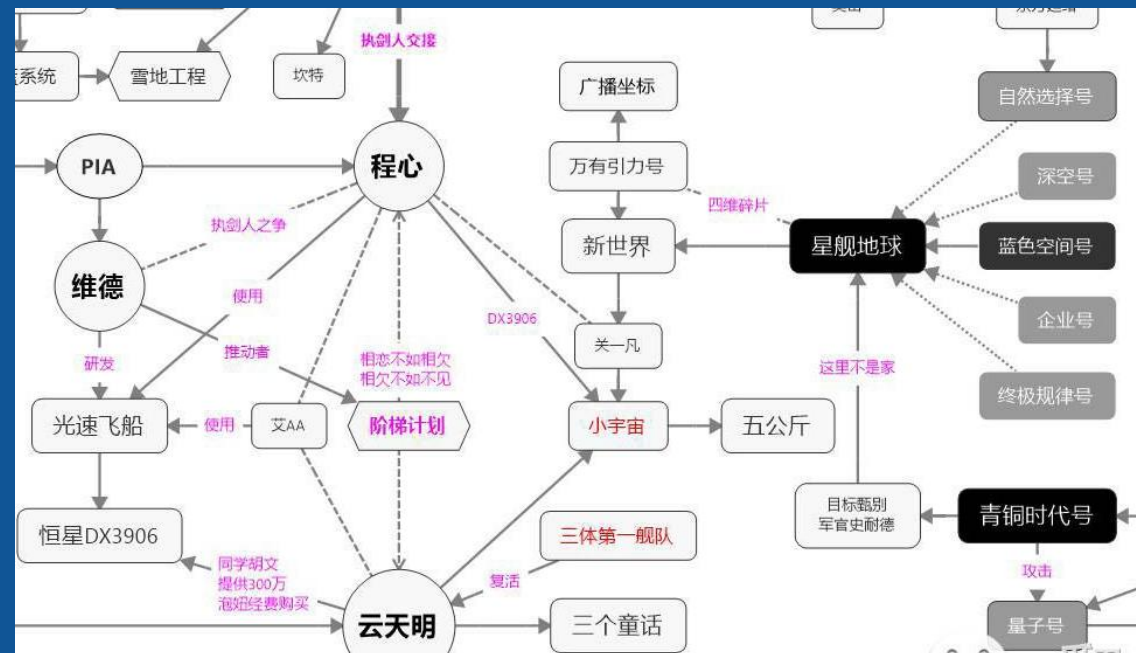


1973、1974年
TCP/IP
Internet



1989年
WWW, 万维网

1.1、发展脉络



链接文档

是一个概念

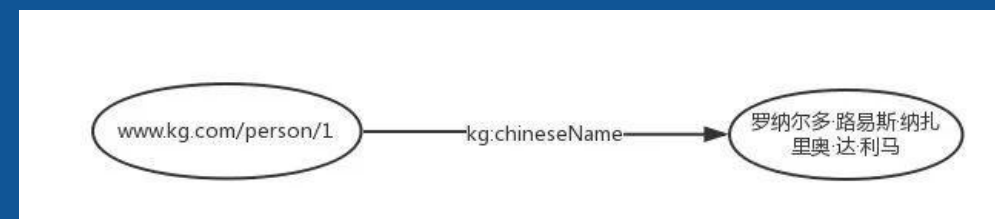
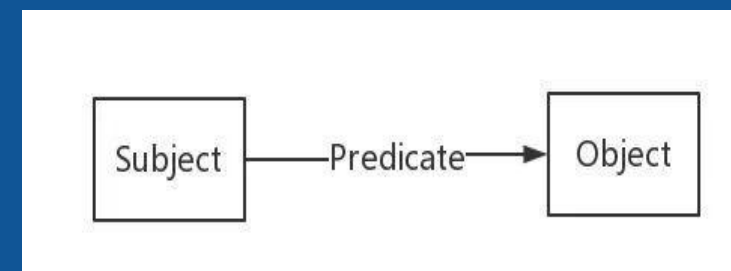
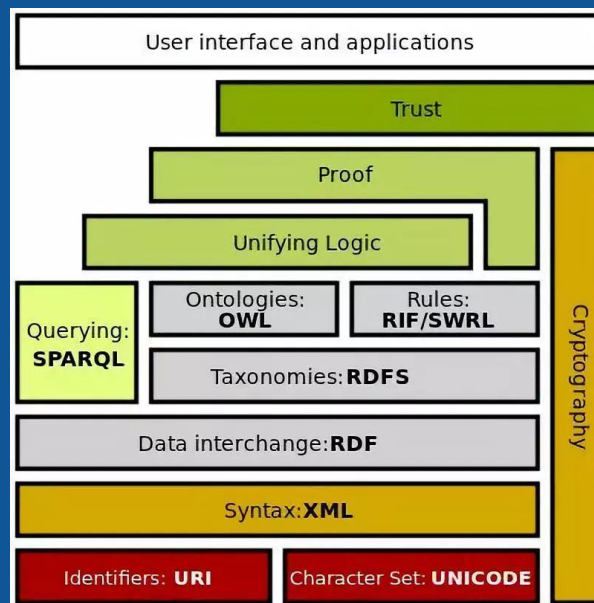
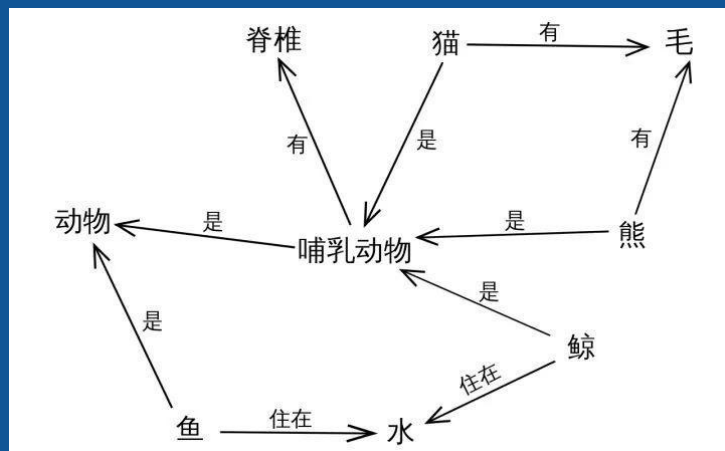
也是一个词典的名字



链接数据

也是一个词典的名字

1.1、发展脉络



语义网络

Semantic Network

优点：1、容易理解和展示。2、相关概念容易聚类。

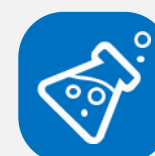
缺点：1、节点和边的值没有标准2、多源数据融合比较困难，3、无法区分概念节点和对象节点



语义网

Semantic Web

RDF, RDFS, OWL



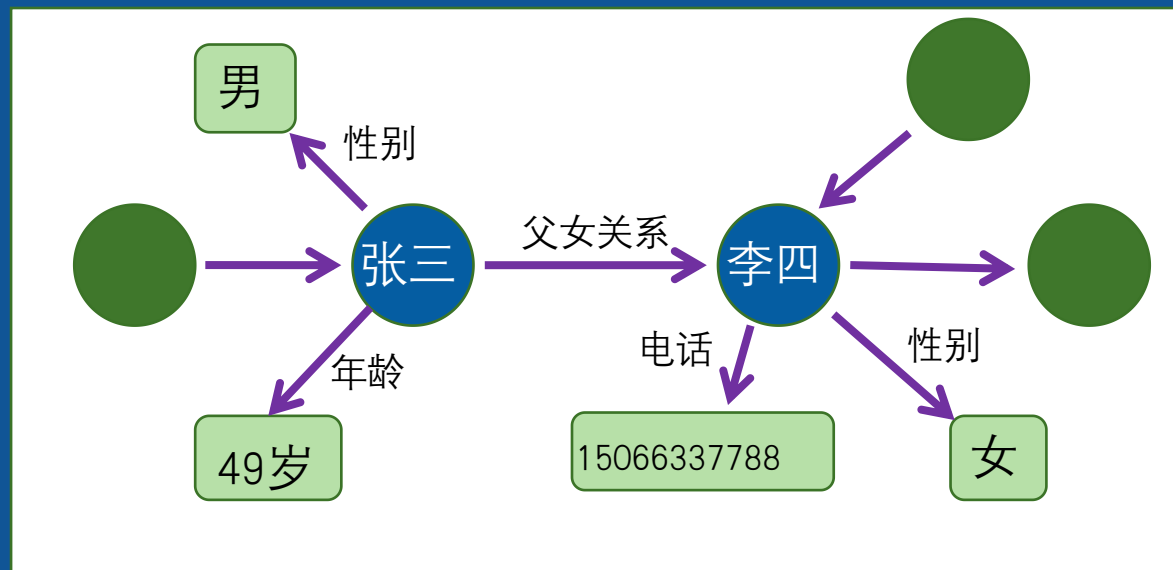
SPO三元组

International Resource Identifiers

(IRIs) , blank nodes 和 literals。

1.1、发展脉络

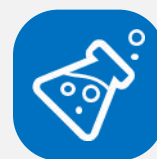
“A knowledge graph consists of a set of interconnected typed entities and their attributes.”



知识图谱

是由一些相互连接的实体和他们的属性构成的。

是由一条条知识组成，每条知识表示为一个 SPO 三元组



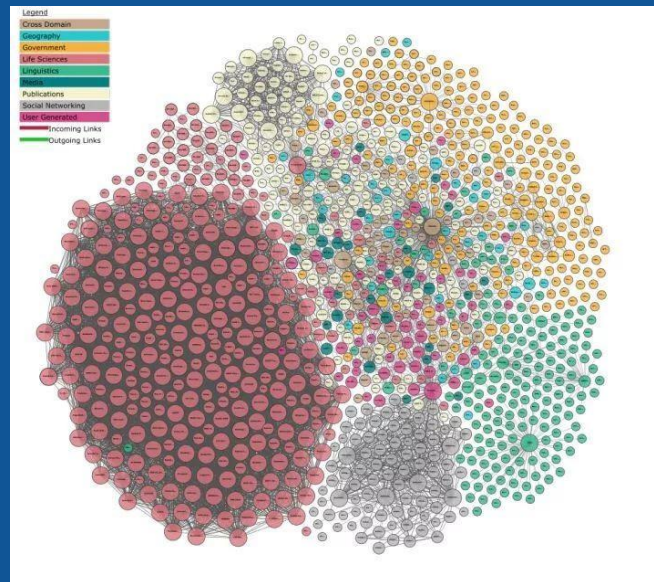
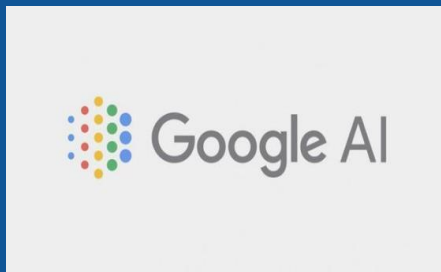
SPO三元组

知识图谱的核心

(实体，关系，实体)

(实体，属性，字面量)

1.1、发展脉络



聪明的AI

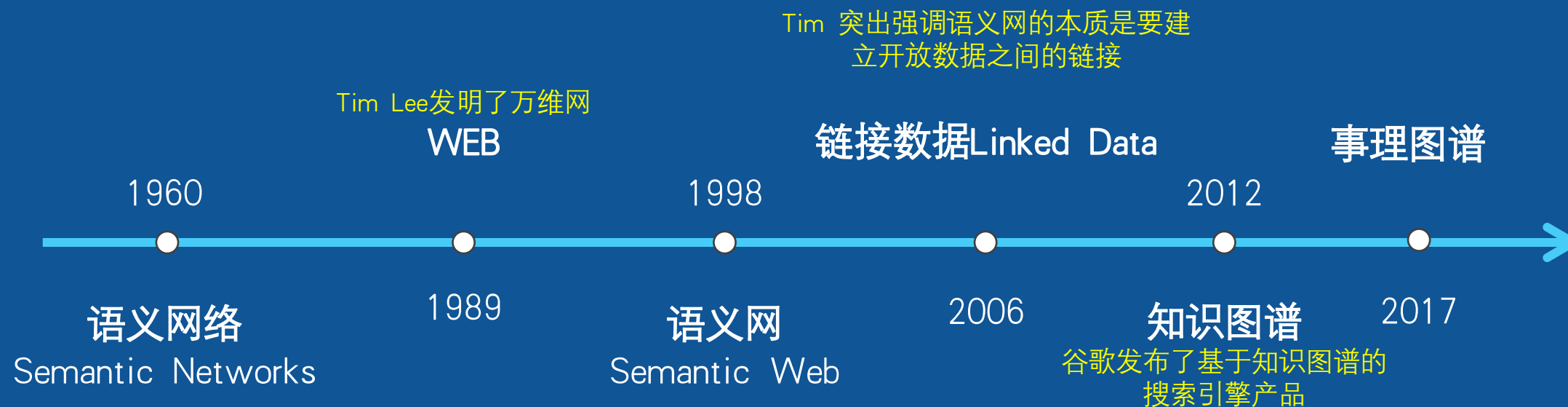
感知，识别，判断



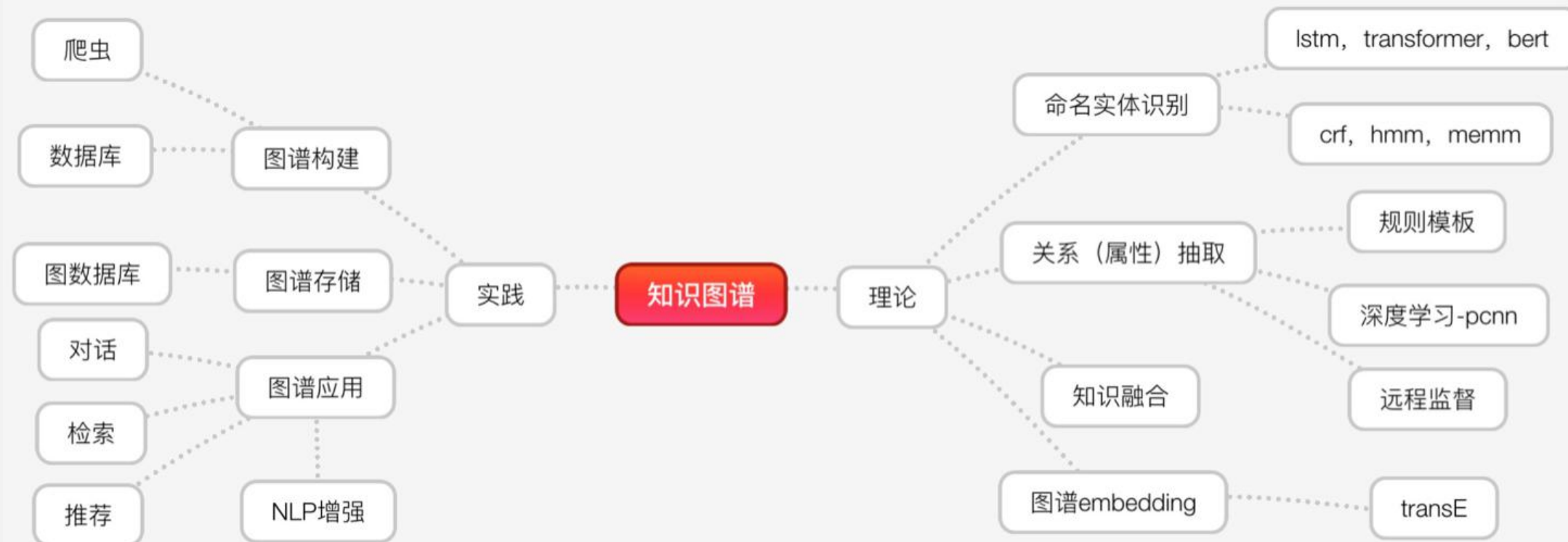
有学识的AI

思考，推理，语言

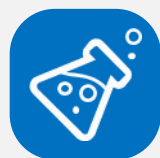
1.1、发展脉络



1.2、技术体系



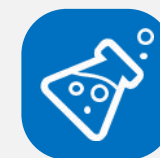
1.3、开源项目



Freebase



WordNet



webkb-2

1.4、商业案例——安全领域



- 成立于2004年，总部设在加州帕罗奥多。” Palantir” 就是《指环王》里那个能看到一切、穿越时空的水晶球。

1.4、商业案例——商品推荐

淘宝商品图谱



- 四个核心节点：商品、产品、品牌、条码。围绕着这四个节点进行扩展，最终形成知识图谱中实体的关系结构。
- 在线图数据库提供在线服务，毫秒级查询，
- 在线关系数据库，解决在图数据库中跨多个本体，长路径的查询响应慢的问题。
- 搜索引擎，支持模糊匹配，节点倒排索引。
- 缓存，数据模型（算法包）和数据分析。
- 离线关系数据库，存储全量数据。

1.4、商业案例——医疗图谱

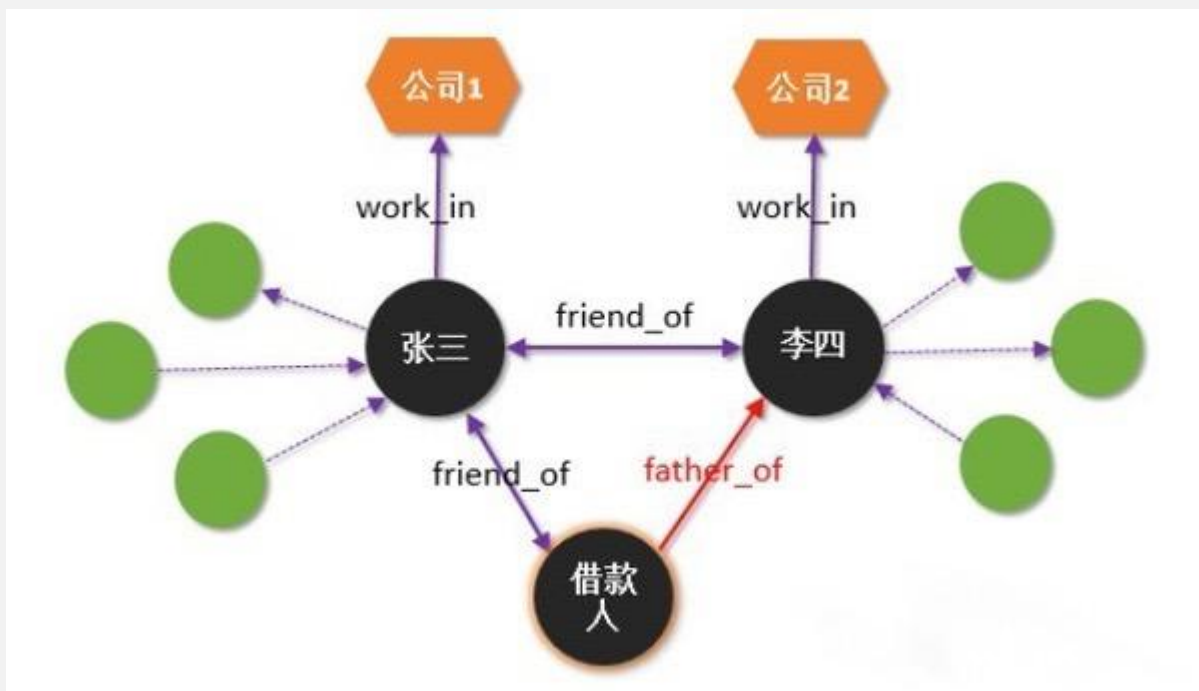


深度智耀

- 药物研发知识图谱
- 整合数百个开放数据源，PB级的大数据，通过机器学习技术，结合医药研发专家知识，自动提取医药实体、关系和属性，构建医药研发知识图谱

1.4、商业案例——金融风控

互联网金融



- 比如，借款人说跟张三是朋友关系，跟李四是父子关系。当我们试图把借款人的信息添加到知识图谱里的时候，“一致性验证”引擎会触发。引擎首先会去读取张三和李四的关系，从而去验证这个“三角关系”是否正确。很显然，朋友的朋友是父子关系的概率较低，所以存在不一致性的概率较高。

1.4、商业案例——金融风控

天眼查

- 整治了大量的公司信息。包括高管、股东、股权关系、风险事件。



1.4、商业案例——智能对话

智投研

中国移动4G 13:11

智投研

贵州茅台的盈利能力怎样?

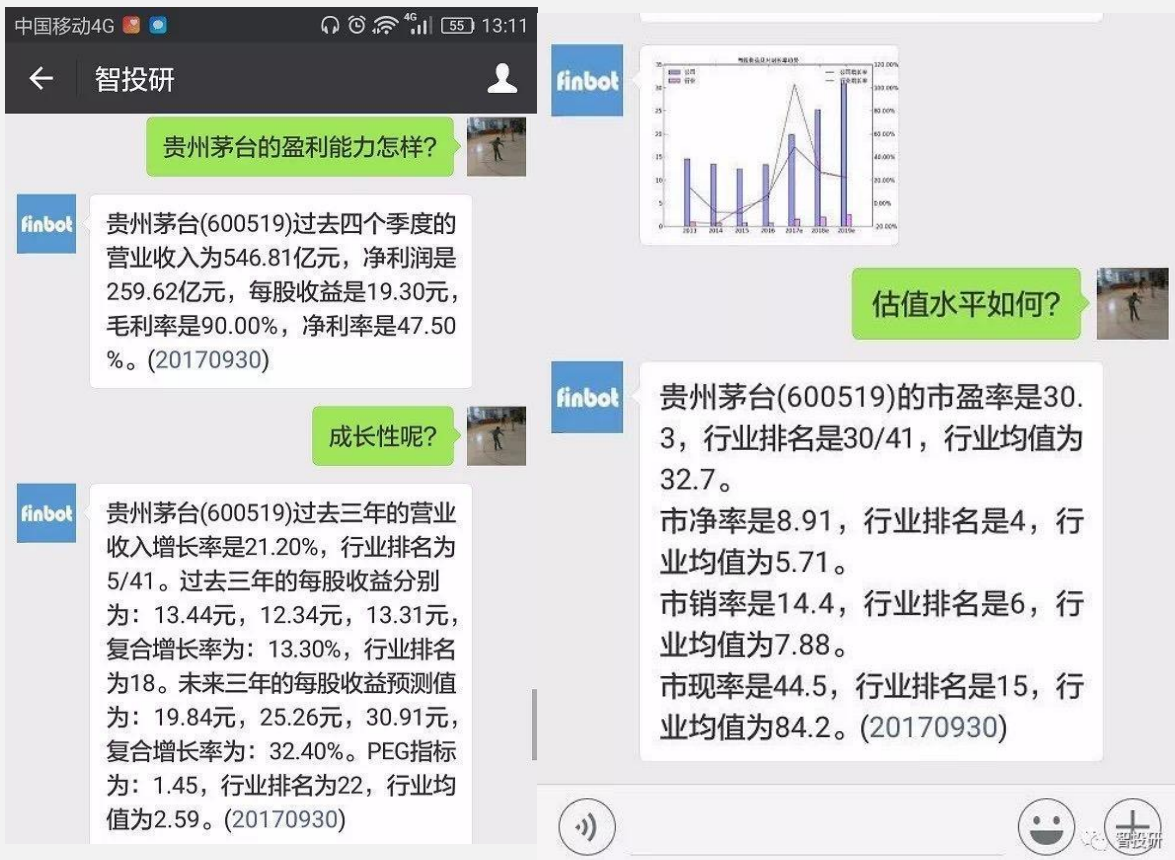
finbot 贵州茅台(600519)过去四个季度的营业收入为546.81亿元,净利润是259.62亿元,每股收益是19.30元,毛利率是90.00%,净利率是47.50%。(20170930)

成长性呢?

finbot 贵州茅台(600519)过去三年的营业收入增长率是21.20%,行业排名为5/41。过去三年的每股收益分别为:13.44元,12.34元,13.31元,复合增长率为:13.30%,行业排名为18。未来三年的每股收益预测值为:19.84元,25.26元,30.91元,复合增长率为:32.40%。PEG指标为:1.45,行业排名为22,行业均值为2.59。(20170930)

估值水平如何?

finbot 贵州茅台(600519)的市盈率是30.3,行业排名是30/41,行业均值为32.7。市净率是8.91,行业排名是4,行业均值为5.71。市销率是14.4,行业排名是6,行业均值为7.88。市现率是44.5,行业排名是15,行业均值为84.2。(20170930)

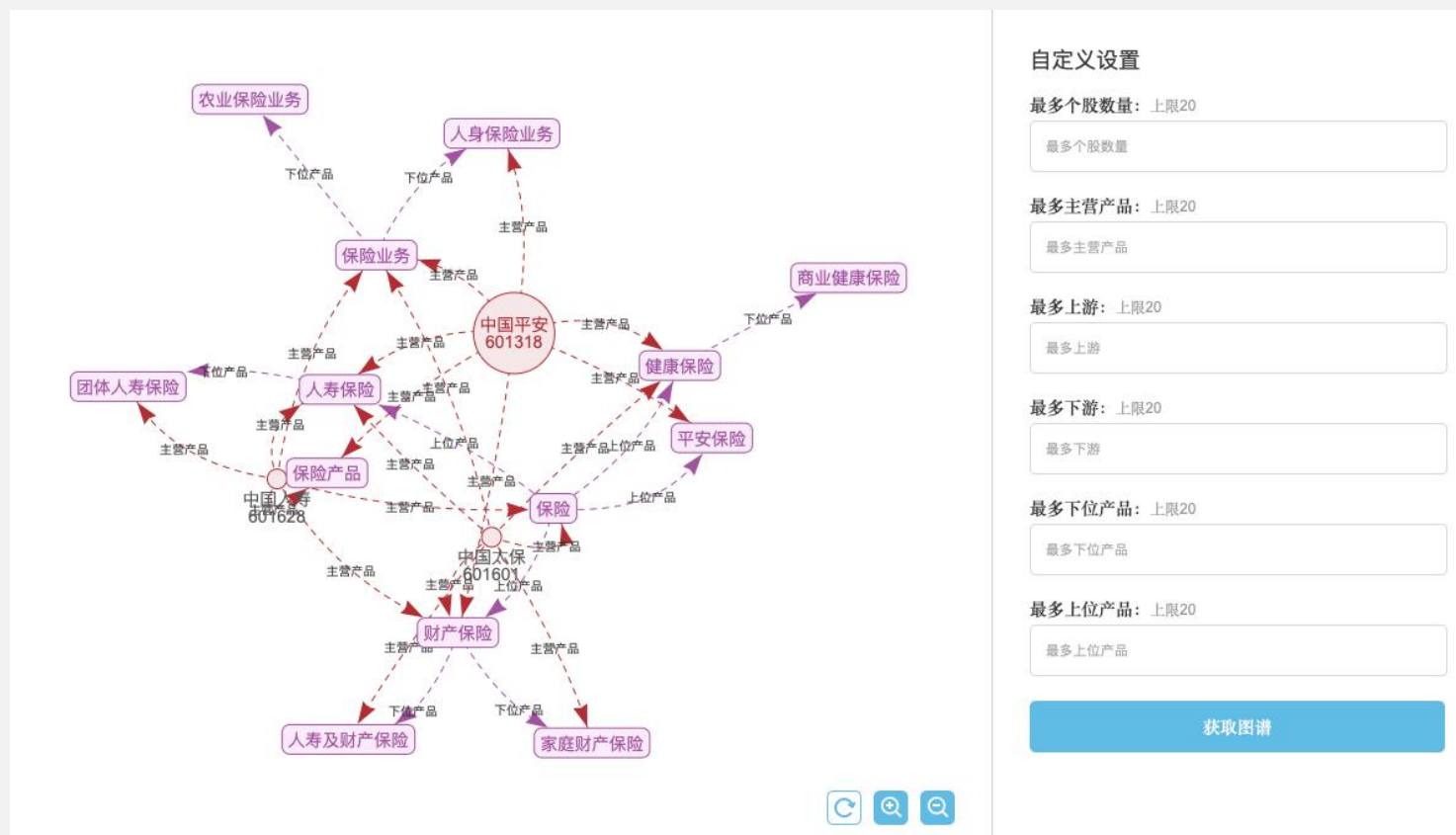


- 建立在金融知识图谱基础上的智能投资客服机器人。
- 相似的场景还有法律咨询。

1.4、商业案例——金融投资

爱问财

- 老牌互联网金融服务公司。
- 产业图谱





2

the part one
知识图谱实践



2.0、Schema设计

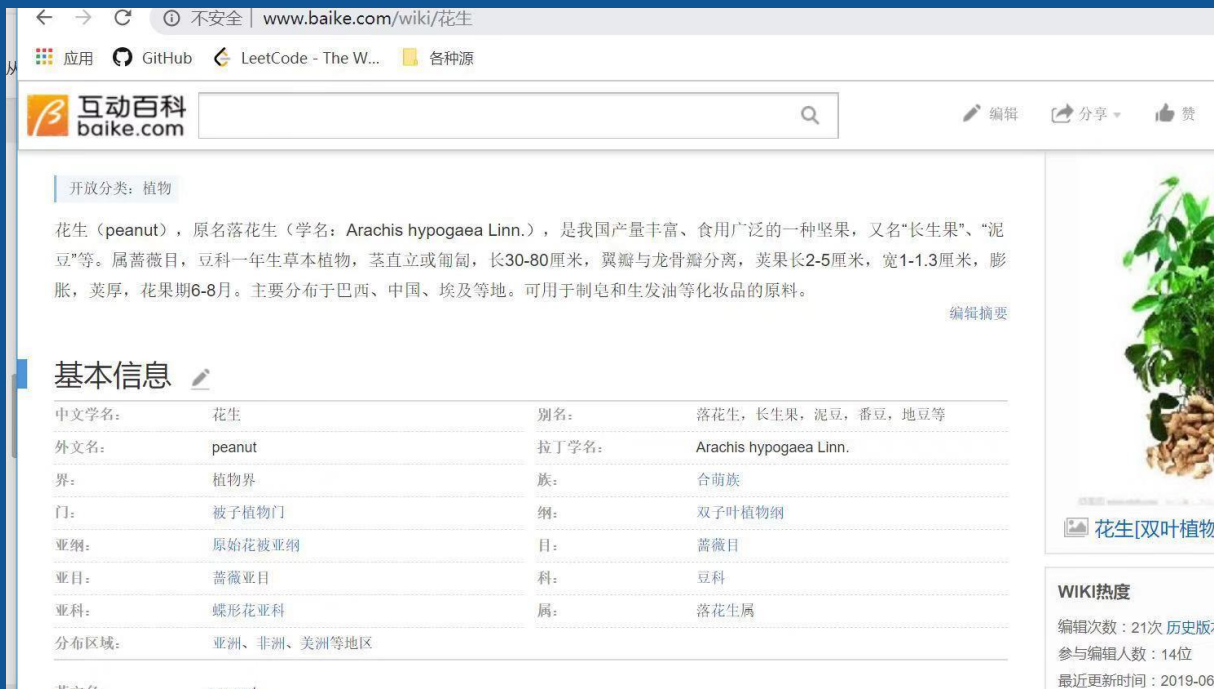
设计知识图谱的结构，要构建哪些类别的实体，实体有什么属性，实体间有什么关系。注意“本体”和“实体”

类型	名称	举例	属性
0	不合法	非具体实体 “文化”，“条件”	无
1	人物	“袁隆平” “周恩来”	生日，性别
2	机构	“农业部” “华东师范大学”	名称，联系方式
3	气候	“夏天” “温带季风气候”	描述
4	动植物 产品	“奶酪” “牛奶” “面粉”	上下级关系



2.1、获取数据

目标数据 => 目标网页 => 爬取

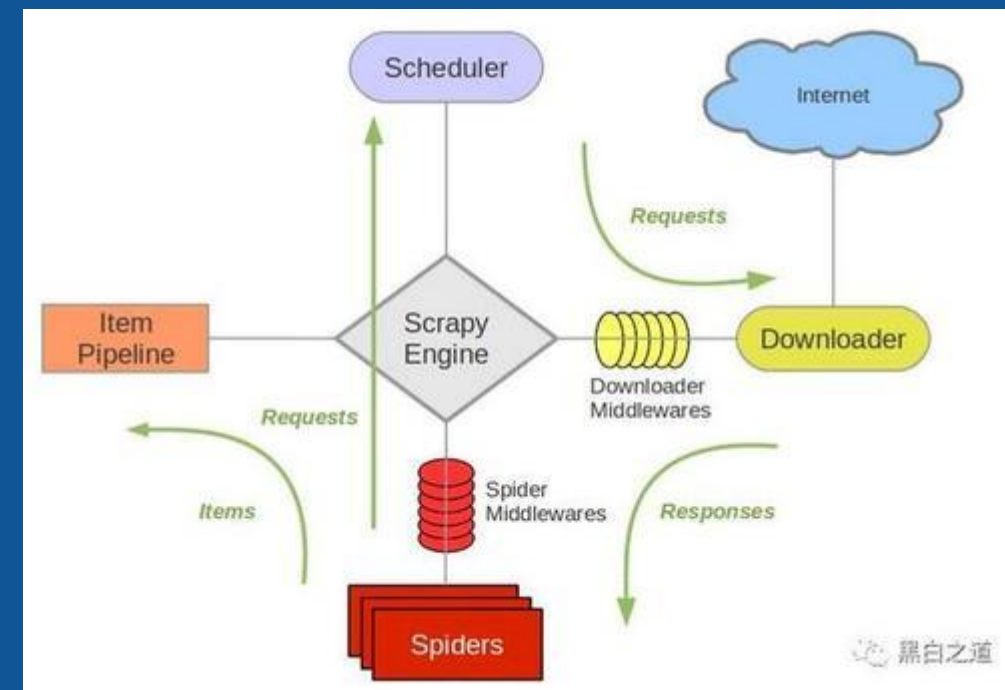


花生 (peanut), 原名落花生 (学名: *Arachis hypogaea* Linn.), 是我国产量丰富、食用广泛的一种坚果, 又名“长生果”、“泥豆”等。属蔷薇目, 豆科一年生草本植物, 茎直立或匍匐, 长30-80厘米, 翼瓣与龙骨瓣分离, 荚果长2-5厘米, 宽1-1.3厘米, 膨胀, 荚厚, 花果期6-8月。主要分布于巴西、中国、埃及等地。可用于制皂和生发油等化妆品的原料。

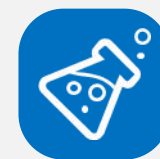
基本信息

中文学名:	花生	别名:	落花生, 长生果, 泥豆, 番豆, 地豆等
外文名:	peanut	拉丁学名:	<i>Arachis hypogaea</i> Linn.
界:	植物界	族:	合萌族
门:	被子植物门	纲:	双子叶植物纲
亚纲:	原始花被亚纲	目:	蔷薇目
亚目:	蔷薇亚目	科:	豆科
亚科:	蝶形花亚科	属:	落花生属
分布区域:	亚洲、非洲、美洲等地区		

WIKI热度
编辑次数: 21次 历史版本
参与编辑人数: 14位
最近更新时间: 2019-06-0



互动百科



scrapy

2.2、实体分类

开放分类: 植物 ← OpenType

花生 (peanut), 原名落花生 (学名: *Arachis hypogaea* Linn.), 是我国产量丰富、食用广泛的一种坚果, 又名“长生果”、“泥豆”等。属蔷薇目, 豆科一年生草本植物, 茎直立或匍匐, 长30-80厘米, 翼瓣与龙骨瓣分离, 荚果长2-5厘米, 宽1-1.3厘米, 膨胀, 荚厚, 花果期6-8月。主要分布于巴西、中国、埃及等地。可用于制皂和生发油等化妆品的原料。

编辑摘要

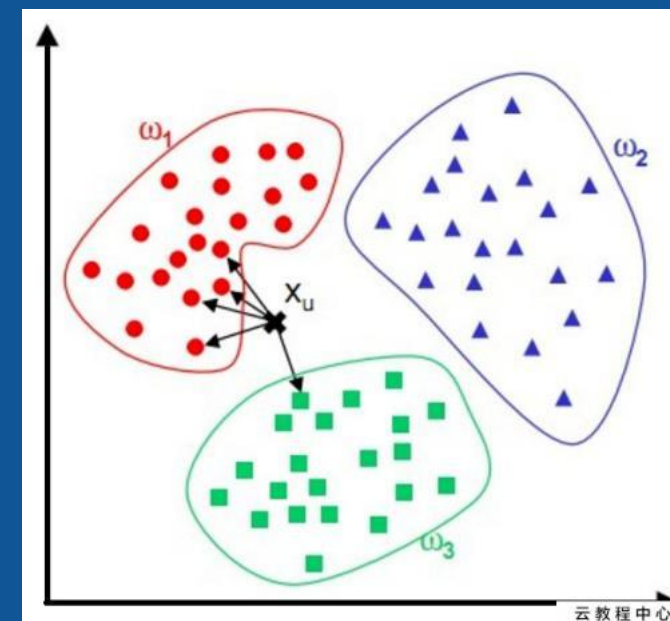
基本信息 ← BaseInfoList

中文学名:	花生	别名:	落花生, 长生果, 泥豆, 番豆, 地豆等
外文名:	peanut	拉丁学名:	<i>Arachis hypogaea</i> Linn.
界:	植物界	族:	合萌族
门:	被子植物门	纲:	双子叶植物纲
亚纲:	原始花被亚纲	目:	蔷薇目
亚目:	蔷薇亚目	科:	豆科
亚科:	蝶形花亚科	属:	落花生属
分布区域:	亚洲、非洲、美洲等地区		

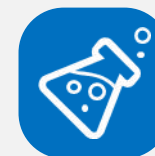
云教程中心

← detail

← BaseInfoValue



实体特征构造



KNN分类模型

2.3、数据存储

关系型数据库

优点：成熟，支持分布式

缺点：跨表繁琐



数据库对比

Neo4j

优点：性能强大，开源，易用性好

缺点：分布式支持一般



Titan--JanusGraph

优点：分布式

缺点：部署繁琐

操作繁琐



企业自研

优点：团队维护

功能响应快

缺点：成本，通用



2.4、图谱应用—KBQA

农业知识图谱 DEMO

open Github

实体识别

实体查询

关系查询

农业知识概览

农知问答

农业智能决策

输入问题:

阜阳市太和县适合种什么

提交

热门搜索:

崇明县适合种什么植物?

胡萝卜汁含有哪些营养成分?

中国的首都的气候类型是什么?

大豆的植物学分类

答案:

答案

山毛榉

核桃

阔叶树

蒙古栎

藤本

槭属

图谱演示:



问题答案

推理依据

1. 图谱简介

- 发展脉络
- 技术体系
- 应用领域

2. 图谱实践

- schema设计
- 生成数据
- 数据存储
- 问答应用

谢 谢 大 家