天津大学
Tianjin University

数字人文领域的知识图谱构建技术

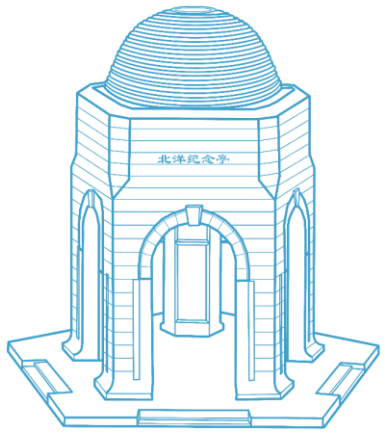
天津大学 智能与计算学部 人工智能学院

王　鑫

wangx@tju.edu.cn

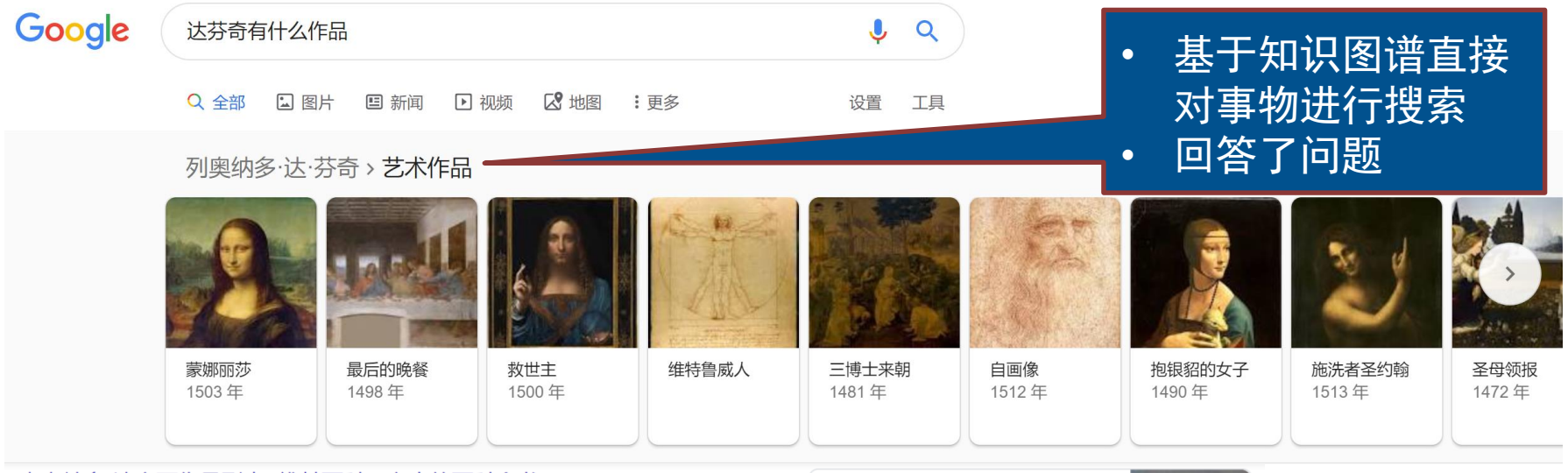
2020年5月25日 • IT4L2020

目录 CONTENTS



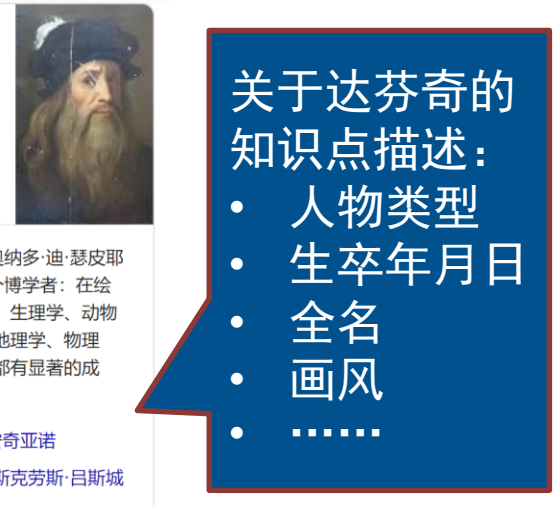
1 什么是知识图谱?

■ Google Knowledge Graph, 2012



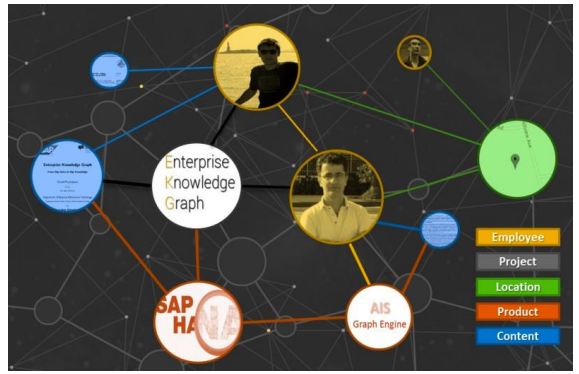
- 基于知识图谱直接对事物进行搜索
- 回答了问题

- 传统搜索
- 网页中关键字检索



关于达芬奇的知识点描述：

- 人物类型
- 生卒年月日
- 全名
- 画风
- ……



1 知识图谱：Linked Data


Anchiano

Village in Italy

Anchiano is a village in the comune of Vinci, Tuscany, central Italy. The village is known for the ancient Villa del Ferrale, with its chapel, Santi Antonio e Francesco. Wikipedia

Weather: 94°F (34°C), Wind W at 10 mph (16 km/h), 33% Humidity

Hotels: 3-star averaging $93. View hotels

Local time: Sunday 6:26 PM


Château du Clos Lucé

Website | Directions | Save

4.4 ★★★★½ 7,806 Google reviews

Castle in Amboise, France

BUY TICKETS

Leonardo da Vinci

Polymath

Leonardo di ser Piero da Vinci, more commonly Leonardo da Vinci, was an Italian polymath of the Renaissance whose areas of interest included invention, drawing, painting, sculpture, architecture, science, music, mathematics, engineering, literature, anatomy, geology, astronomy, botany, paleontology and cartography. Wikipedia

Born: April 15, 1452, Anchiano, Italy

Died: May 2, 1519, Château du Clos Lucé, Amboise, France

Artworks: Mona Lisa, The Last Supper, Salvator Mundi, MORE

On view: Ambrosian Library, Louvre Museum, MORE

Periods: High Renaissance, Early renaissance, Renaissance, Italian Renaissance, Florentine painting

Siblings: Giovanni Ser Piero, Guglielmo Ser Piero, MORE

Born

Died

Artworks

On view

Location

Mona Lisa

Painting by Leonardo da Vinci

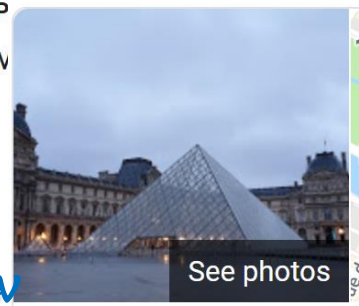
The Mona Lisa is a half-length portrait painting by the Italian Renaissance artist Leonardo da Vinci that has been described as "the best known, the most visited, the most written about, the most sung about, the most parodied work of art in the world." Wikipedia

Artist: Leonardo da Vinci

Dimensions: 2′ 6″ x 1′ 9″

Location: Louvre Museum (since 1797)

Created: 1503




Louvre Museum

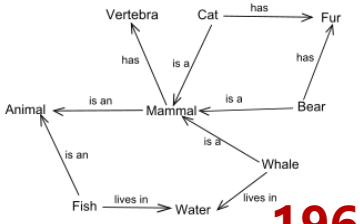
Website | Directions | Save

4.7 ★★★★½ 154,952 Google reviews

Museum in Paris, France

1 知识图谱：发展历史

■ 知识图谱的发展历史



1960s

语义网络

Semantic Networks

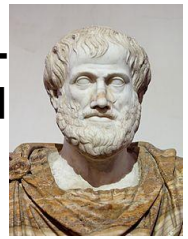
语义网络作为知识表示的一种方法被提出，主要用于自然语言理解领域

1980s

本体论

Ontology

哲学概念“本体”被引入到人工智能领域用来刻画知识



1989

Web

Tim Berners-Lee发明了万维网Linked Information System

The Semantic Web

从超文本链接到语义链接

1998

W3C Semantic Web

2006

链接数据

Linked Data

Tim突出强调语义网的本质是要建立开放数据之间的链接

知识图谱

Knowledge Graph

谷歌发布了其基于知识图谱的搜索引擎产品

2012

知识表示与知识库- Knowledge Representation / Knowledge Base

人工智能研究者陆续提出了大量知识表示的方法，如框架系统、产生式规则、描述逻辑等

知识图谱得益于Web的发展 (更多的是数据层面)，有着来源于KR、NLP、Web、AI多个方面的基因

1 知识图谱：人工智能的基石

■ 聪明的AI vs 有学识的AI

人的大脑依赖所学的知识进行思考、逻辑推理、理解语言

聪明的AI

感知

识别

判断

学习

有学识的AI

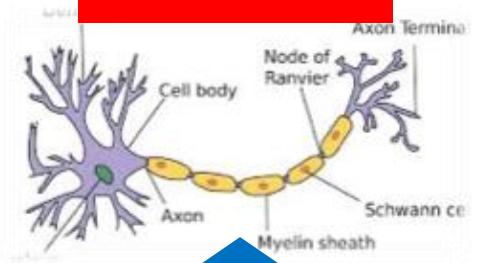
推理

思考

语言

推理

深度学习



联结主义

模拟人脑结构

知识图谱

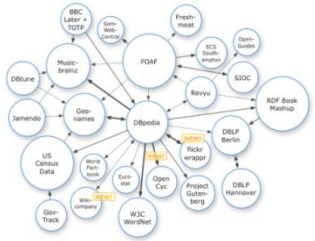
符号主义

模拟人的心智

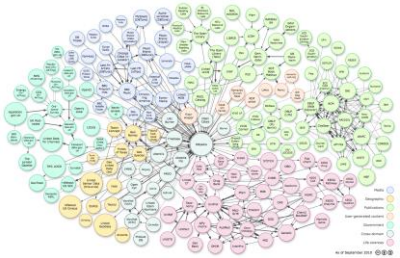
1 知识图谱：大规模 Linked Data

■ 大规模知识图谱数据的发布

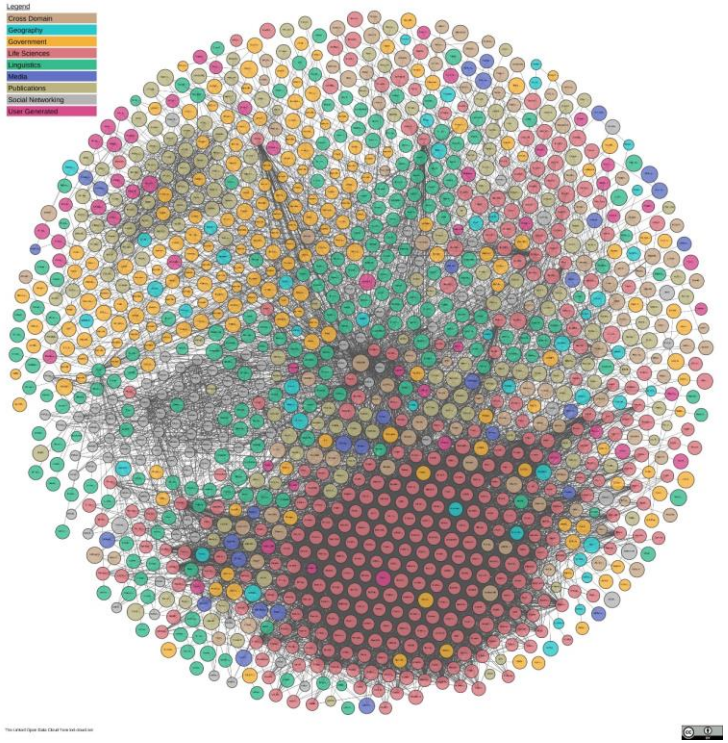
- 百万顶点（10^6）和上亿条边（10^8）已常见
- Linked Data



2007.10

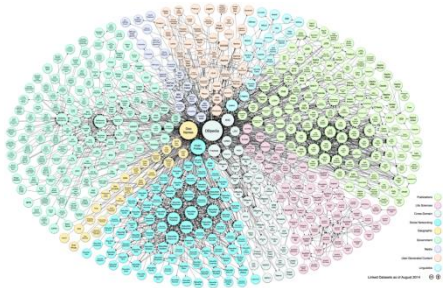


2010.09



2019.3, 1239 Datasets

百科知识 DBpedia
地理信息 LinkedGeoData
生物医学 PubMed
媒体出版 DBLP
政府数据 data.gov
……



2014.08

https://lod-cloud.net/

知识抽取 实现自动化构建大规模知识图谱的重要技术

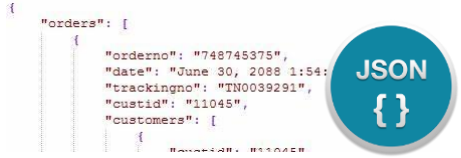
- 从不同来源、不同结构的数据中进行知识提取，存入知识图谱中

知识抽取的数据源

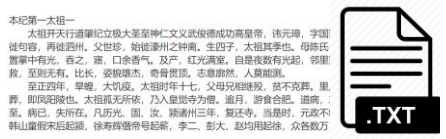
结构化数据

Company	Region	Product	Month	Sales $
ABC	EMEA	Product 1	Jan	1,000
ABC	EMEA	Product 2	Jan	1,010
ABC	EMEA	Product 3	Jan	1,020
ABC	EMEA	Product 4	Jan	1,030

半结构化数据



非结构化数据



苹果公司是一家美国跨国公司，总部位于美国加利福尼亚库比蒂诺，致力于设计、开发和销售消费电子、计算机软件、在线服务和个人计算机。苹果公司由史蒂夫·乔布斯、史蒂夫·沃兹尼克、罗纳德·韦恩创立于 1976 年 4 月 1 日，开发和销售个人计算机。该公司于 1977 年 1 月 3 日正式称为苹果电脑公司，并于 2007 年 1 月 9 日改名为苹果公司，反应其业务重点转向消费电子领域。

苹果公司	总部地址	美国加利福利亚库比蒂诺
苹果公司	创始人	史蒂夫·乔布斯
苹果公司	创始人	史蒂夫·沃兹尼克
苹果公司	创始人	罗纳德·韦恩
苹果公司	创立时间	1976年4月1日

知识抽取相关国际竞赛
Automatic Content Extraction, ACE
Knowledge Base Population，KBP

子任务

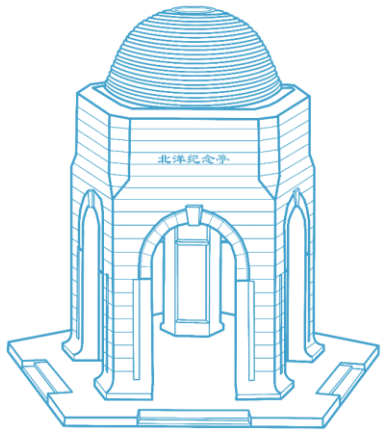
1. 命名实体识别 NER

- 检测：[苹果公司] 实体, [美国] 实体
- 分类：[苹果公司] 企业, [美国] 国家

2. 关系抽取 形成三元组

- [苹果公司] <所属国家> [美国]
- [苹果公司] <创始人> [史蒂夫•乔布斯]

目录 CONTENTS



天津大学 Tianjin University

实体抽取 命名实体识别 Named Entity Recognition, NER

- 实体抽取是知识抽取中最基本的任务

实体类别：人名 | 地名 | 组织 | 日期 | 时间 | 货币金额

例句：十四年冬十月，元丞相脱脱大败士诚于高邮，分兵围六合。

时间　人物　地名

步骤：
1. 从文本中识别和定位实体
2.将实体分类到预定义类别

命名实体识别方法

基于规则的方法

早期的命名实体识别方法主要采用人工编写规则的方式进行实体抽取，将规则与文本字符串进行匹配

规则集构造费时费力可移植性差

基于统计模型的方法

利用标注语料进行模型训练：隐马尔可夫模型：HMM；条件马尔可夫模型：CMM；最大熵模型：MEM；条件随机场模型：CRF

将命名实体识别作为序列标注问题处理

基于深度学习的方法

直接以文本词向量为输入，通过神经网络实现端到端的命名实体识别，不再依赖人工定义的特征　CNN　RNN

天津大学 Tianjin University

BIO序列标注法 Beginning-Inside–Outside

B-X	实体名称的起始字符
I-X	实体名称的中间字符
O	实体名称的外部字符
X	替换为具体的实体类别，如：人民PER 地名LOC 组织ORG

十	四	年	冬	十	月	,	元	丞	相	脱	脱
O	O	O	O	O	O	O	O	O	O	B-PER	I-PER

大	败	士	诚	于	高	邮	,	分	兵	围	六	合
O	O	B-PER	I-PER	O	B-LOC	I-LOC	O	O	O	O	B-LOC	I-LOC

- **人民日报数据集。**
  - 人名、地名、组织名三种实体类型
  - 1998：https://github.com/InsaneLife/ChineseNLPCorpus/tree/master/NER/renMinRiBao
  - 2004：https://pan.baidu.com/s/1LDwQjoj7qc-HT9qwhJ3rcA password: 1fa3

2 实体抽取：基于深度学习的方法

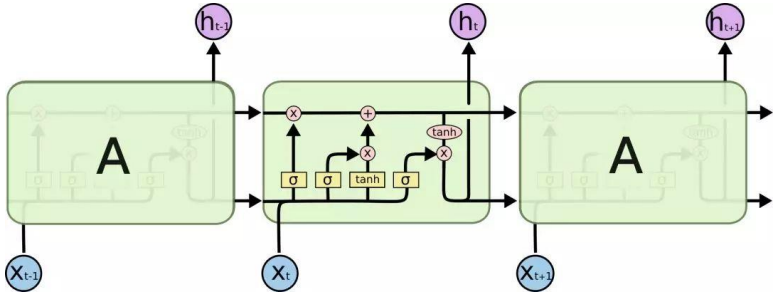
LSTM-CRF模型

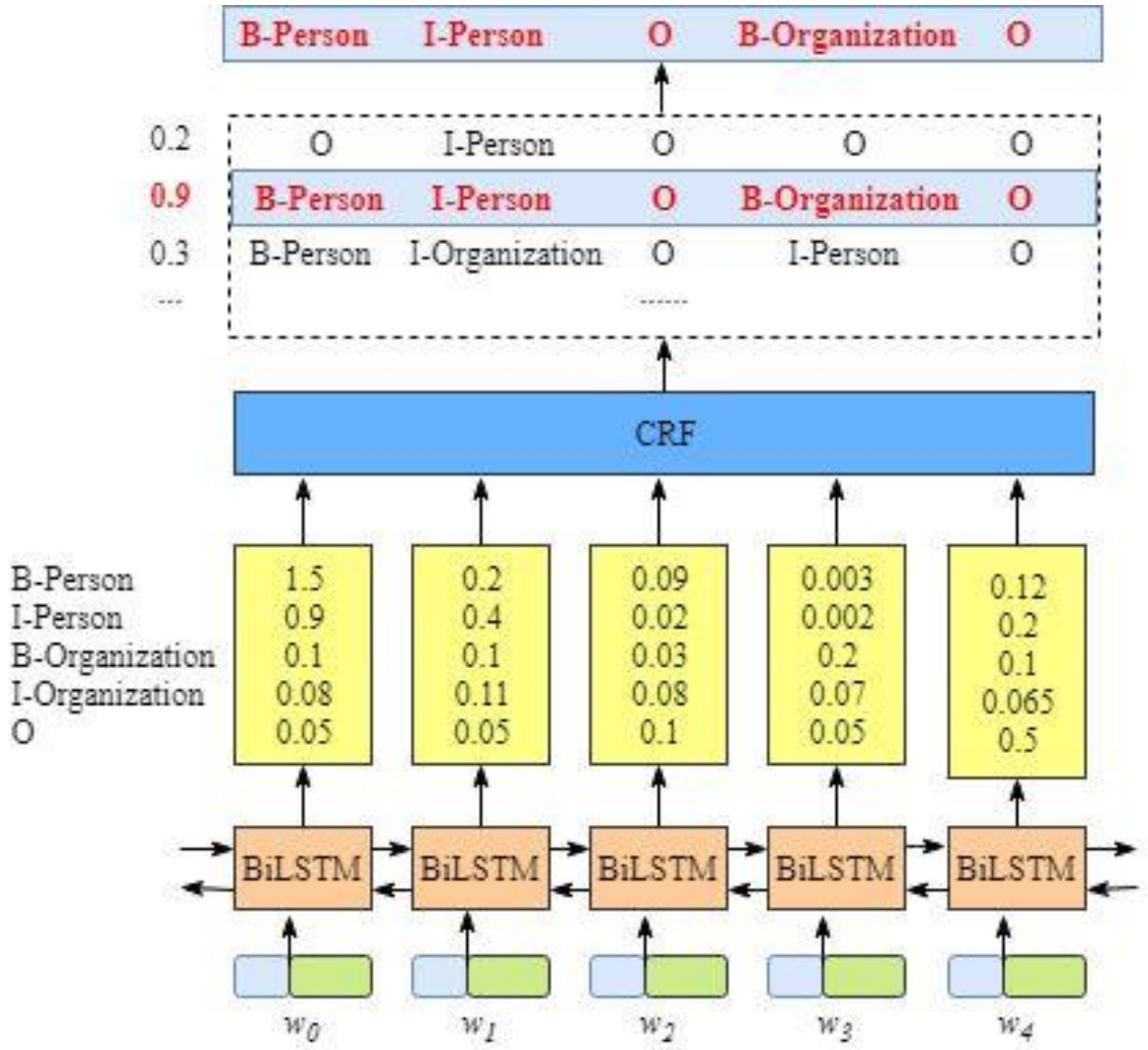
CRF层：条件随机场

通过CRF模型输出对每个词的标注结果
通过转移概率学习施加约束条件，保证预测结果有效

Bi-LSTM层：双向长短时记忆神经网络

神经网络结构在命名实体识别过程中扮演编码器的角色
基于输入以及词上下文信息，得到每个词的新向量表示



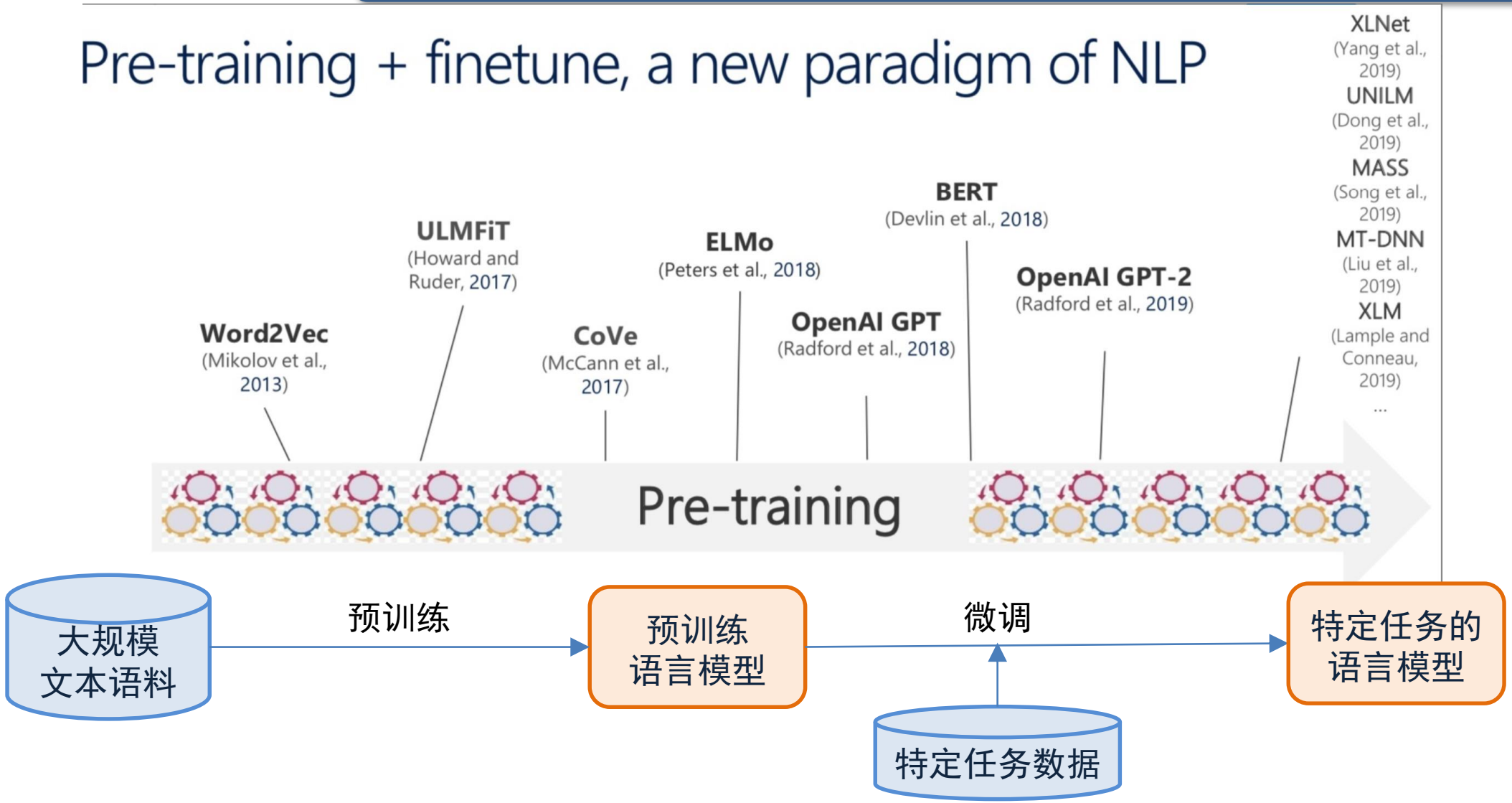


- Bidirectional LSTM-CRF Models for Sequence Tagging. 2015. https://arxiv.org/abs/1508.01991v1

2 NLP新范式：预训练语言模型

■ NLP新范式

预训练（Pretraining） + 微调（finetuning）



2 预训练语言模型：BERT

BERT Bidirectional Encoder Representation from Transformers

词向量模型

句子级别向量

sentence-level

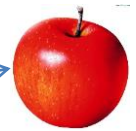
词级别向量 word-level

ELMo

上下文相关的动态向量

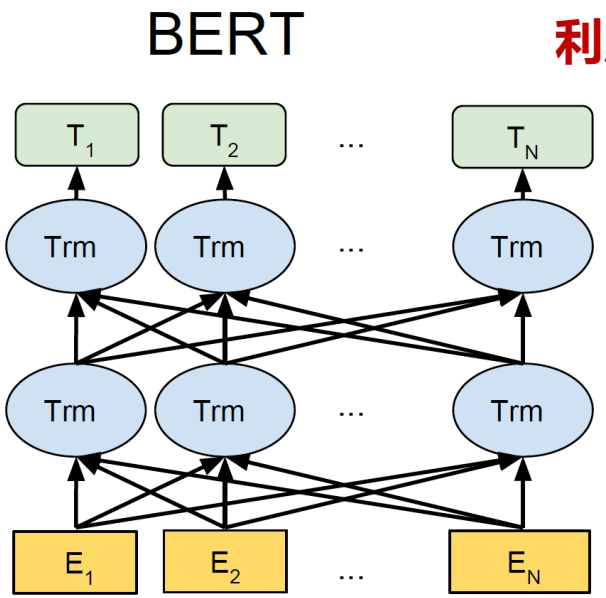
word2vec

上下文无关的静态向量



苹果 ？

泛化能力



利用大规模无标注语料训练

获得包含丰富语义信息的文本向量表示

Bi-LSTM

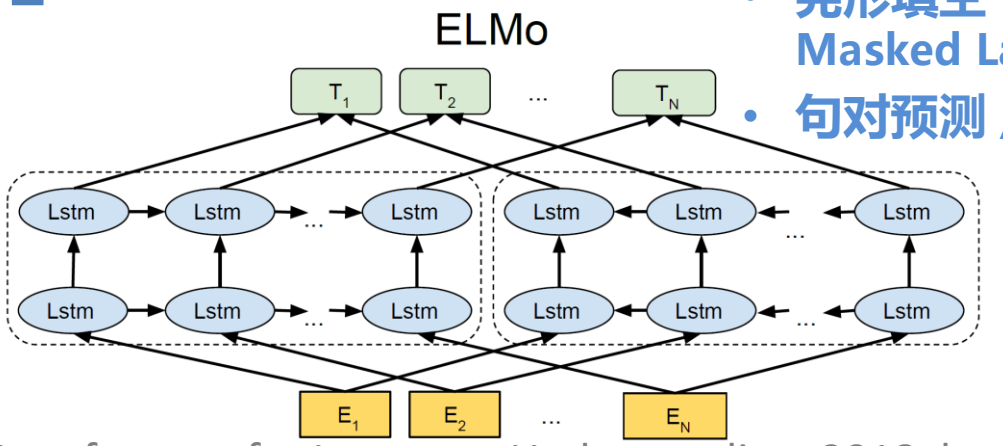
模型特点

- Transformer
- 句子级表示

CBOW
Skip-Gram

预训练任务

- 完形填空
  Masked Language Model
- 句对预测 / 预测下一个句子



Attention机制

- 鹊
- 鸿鹊之志

- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018. https://arxiv.org/abs/1810.04805

2 预训练语言模型：BERT

BERT

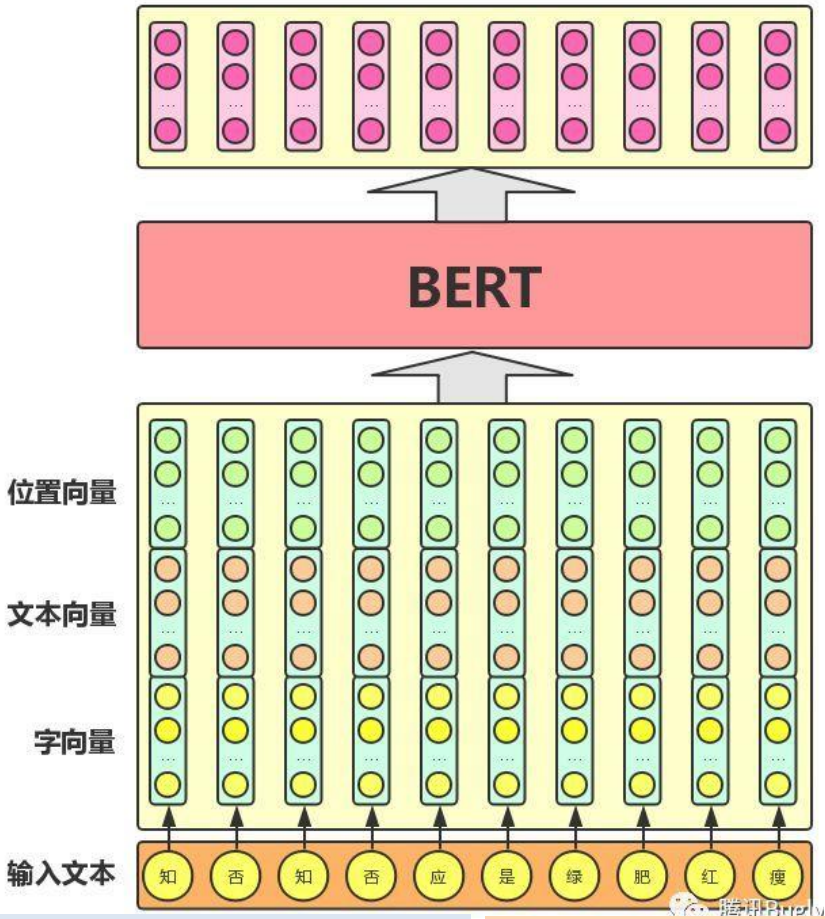
超越人类！ 在机器阅读理解测试SQuAD1.1中

F1提高到93.16分，**比人类表现高出近2.0分**

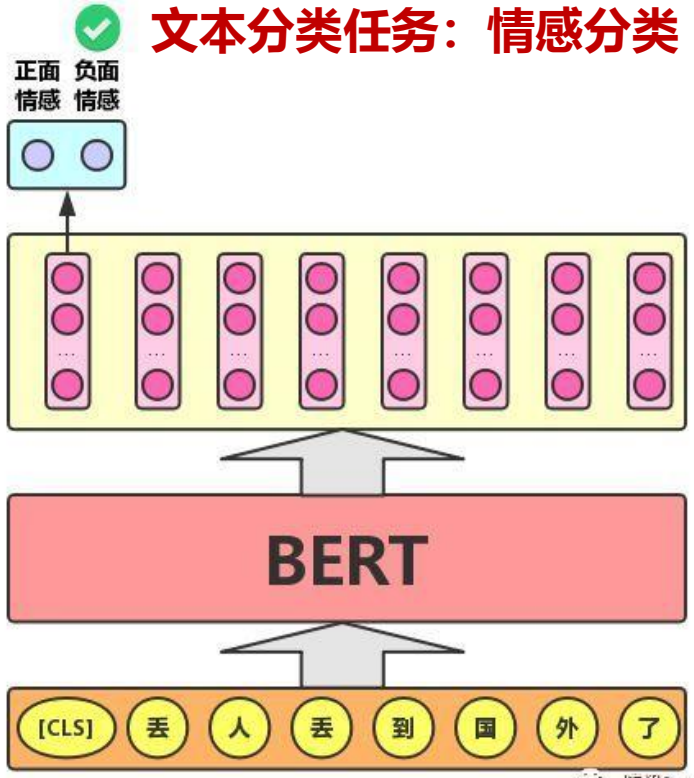
Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google A.I.*	87.433	93.160

刷新了11项NLP任务的最优结果！

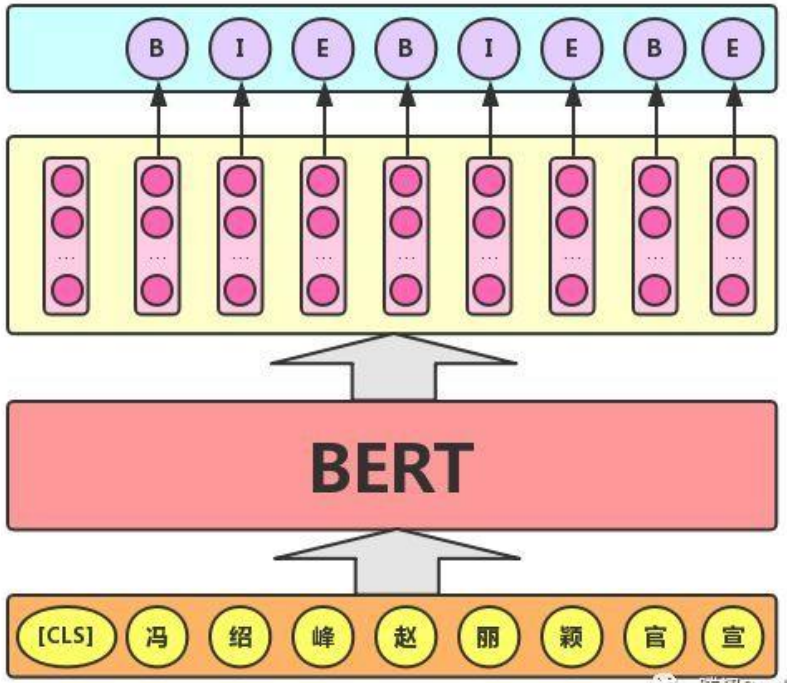
模型输入



文本分类任务：情感分类



序列标注任务：命名实体识别



Finetune标准四法 | 双句分类任务 | 单句分类任务 | 问答任务 | 单句标注任务

2 预训练语言模型：BERT

天津大学 Tianjin University

BERT Bidirectional Encoder Representation from Transformers

模型规模

- 数据：维基百科数据一共有33亿个词
- 参数：Base版：1亿个参数、12层、768-hidden
  Large版：3亿个参数、24层、1024-hidden
- 训练代价：Google用16个TPU集群（64块TPU）花4天时间训练Large版BERT

ymcui / Chinese-BERT-wwm

Watch 99 Star 3k

Code Issues 0 Pull requests 0 Actions Projects 0 Wiki Security 0 Insights

Pre-Training with Whole Word Masking for Chinese BERT（中文BERT-wwm系列模型） https://arxiv.org/abs/1906.08101

chinese-bert tensorflow pytorch bert nlp roberta bert-wwm bert-wwm-ext roberta-wwm rbt

google-research / bert

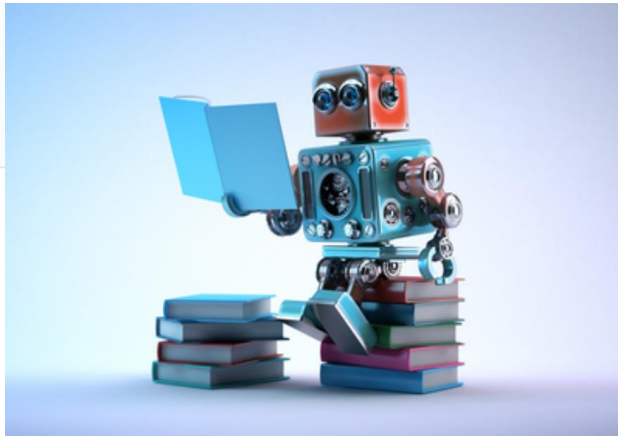
Watch 933 Star 22.9k Fork 6.4k

Code Issues 626 Pull requests 71 Actions Projects 0 Wiki Security 0 Insights

TensorFlow code and pre-trained models for BERT https://arxiv.org/abs/1810.04805

nlp google natural-language-processing natural-language-understanding tensorflow

111 commits 2 branches 0 packages 0 releases 27 contributors Apache-2.0



2 实体抽取：基于BERT的命名实体识别

天津大学 Tianjin University

基于 BERT 的中文命名实体识别

基于BERT-Base Chinese 中文预训练模型

微调Finetune训练：采用人民日报NER语料库

```
>>> model.fit(train_x, train_y, valid_x, valid_y, epochs=1)
327/327 [==============================] - 3334s 10s/step - loss: 2.1666 - accuracy: 0.9931 - val_loss: 120.2806 - val_accuracy: 0.9724
<tensorflow.python.keras.callbacks.History object at 0x7f0bb1276c50>
```

NER效果：迁移学习能力强！

十四年冬十月，元丞相脱脱大败士诚于高邮，分兵围六合。

['十/O', '四/O', '年/O', '冬/O', '十/O', '月/O', '，/O', '元/O', '丞/O', '相/O', **'脱/B-PER', '脱/I-PER'**, '大/O', '败/O', **'士/B-PER', '诚/I-PER'**, '于/O', **'高/B-LOC', '邮/I-LOC'**, '，/O', '分/O', '兵/O', '围/O', **'六/B-LOC', '合/I-LOC'**, '。/O']

二十年春二月，元福建行省参政袁天禄以福宁降。三月戊子，征刘基、宋濂、章溢、叶琛至。
夏五月，徐达、常遇春败陈 友谅于池州。

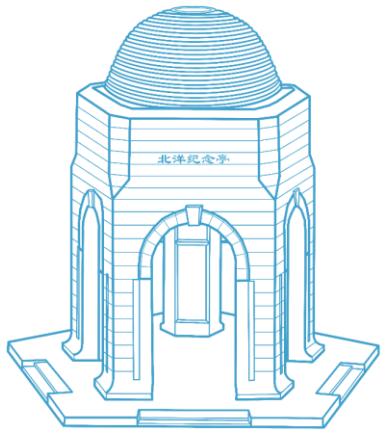
['二/O', '十/O', '年/O', '春/O', '二/O', '月/O', '，/O', '元/O', **'福/B-LOC', '建/I-LOC', '行/I-LOC', '省/I-LOC'**, '参/O', '政/O', **'袁/B-PER', '天/I-PER', '禄/I-PER'**, '以/O', **'福/B-LOC', '宁/I-LOC'**, '降/O', '。/O', '三/O', '月/O', '戊/O', '子/O', '，/O', '征/O', **'刘/B-PER', '基/I-PER'**, '、/O', **'宋/B-PER', '濂/I-PER'**, '、/O', **'章/B-PER', '溢/I-PER'**, '、/O', **'叶/B-PER', '琛/I-PER'**, '至/O', '。/O', '夏/O', '五/O', '月/O', '，/O', **'徐/B-PER', '达/I-PER'**, '、/O', **'常/B-PER', '遇/I-PER', '春/I-PER'**, '败/O', **'陈/B-PER', '/B-PER', '友/I-PER', '谅/I-PER'**, '于/O', **'池/B-LOC', '州/I-LOC'**, '。/O']

用于预测的Python代码

```
import kashgari
loaded_model = kashgari.utils.load_model('saved_ner_model')
loaded_model.predict([list("十四年冬十月，元丞相脱脱大败士诚
```

- Kashgari is a production-level NLP Transfer learning framework. https://github.com/BrikerMan/Kashgari

目录 CONTENTS



关系抽取 在识别出文本中的实体后，再抽取实体之间可能存在的关系

非结构化数据关系抽取方法

基于模板的关系抽取方法

例句1：[姚明]与妻子[叶莉]还有女儿……　模板1：[X]与妻子[Y]

例句2：[徐峥]老婆[陶虹]晒新写真……　模板2：[X]老婆[Y]

关系：夫妻　**优点**：构建过程简单

还可基于依存句法分析结果，在依存树上仅规则匹配　**缺点**：人工成本高、可维护性差、规则集小时召回率低

基于监督学习的关系抽取方法

预定义关系的类型➔人工标注数据➔设计关系识别所需特征➔选择分类模型➔训练模型➔进行模型评估

特征工程
- 轻量级：实体前后的词、实体类型
- 中量级：句子中语块序列的特征
- 重量级：实体间的依存关系路径

深度学习
- 不依赖特征工程，只需词向量、位置向量表示
- 流水线方法：CR-CNN、Att-CNN、Att-BLSTM
- 联合抽取方法：LSTM-RNN 实体和关系抽取相结合

标注成本高

基于弱监督学习的关系抽取方法

只利用少量的标注数据进行模型学习　远程监督方法：APCNN　Bootstrapping方法：KnowItAll、NELL

3 关系抽取：远程监督

远程监督 Distant Supervision

优点：减少模型对人工标注数据的依赖
缺点：假设会引入大量噪声，语义漂移

- 通过将外部知识图谱与非结构化文本对齐的方式自动构建大量训练数据

外部知识图谱

明成祖朱棣

皇后

仁孝文皇后徐氏

維基百科

假设：如果两个实体在外部知识图谱中存在某种关系，则包含两个实体的句子均表达了这种关系

用作训练正例

明史　卷五十四·志第三十　及成祖即位，册皇后徐氏，其制小异。皇帝皮弁服御华盖殿，翰林院官以诏书用宝讫，然后御殿，传制皇后受册。礼毕，翰林官以诏书授礼部官，部官奉诏书于承天门开读。皇帝还宫，率皇后具服先殿谒告毕。皇后具服于内殿，俟皇帝升座。赞官导诣拜位，行谢恩礼，乐作。八拜，兴，乐止毕。次日，

明史　卷一百十三·列传第一　成祖仁孝皇后徐氏，中山王达长女也。幼贞静，好读书，称女诸生。太祖闻后贤淑，召达谓曰"朕与卿，布衣交也。古君臣相契者，率为婚姻。卿有令女，其以朕子棣配焉"达顿首谢。

• M Mintz. Distant supervision for relation extraction without labeled data. ACL 2009.

3 知识抽取工具：DeepDive

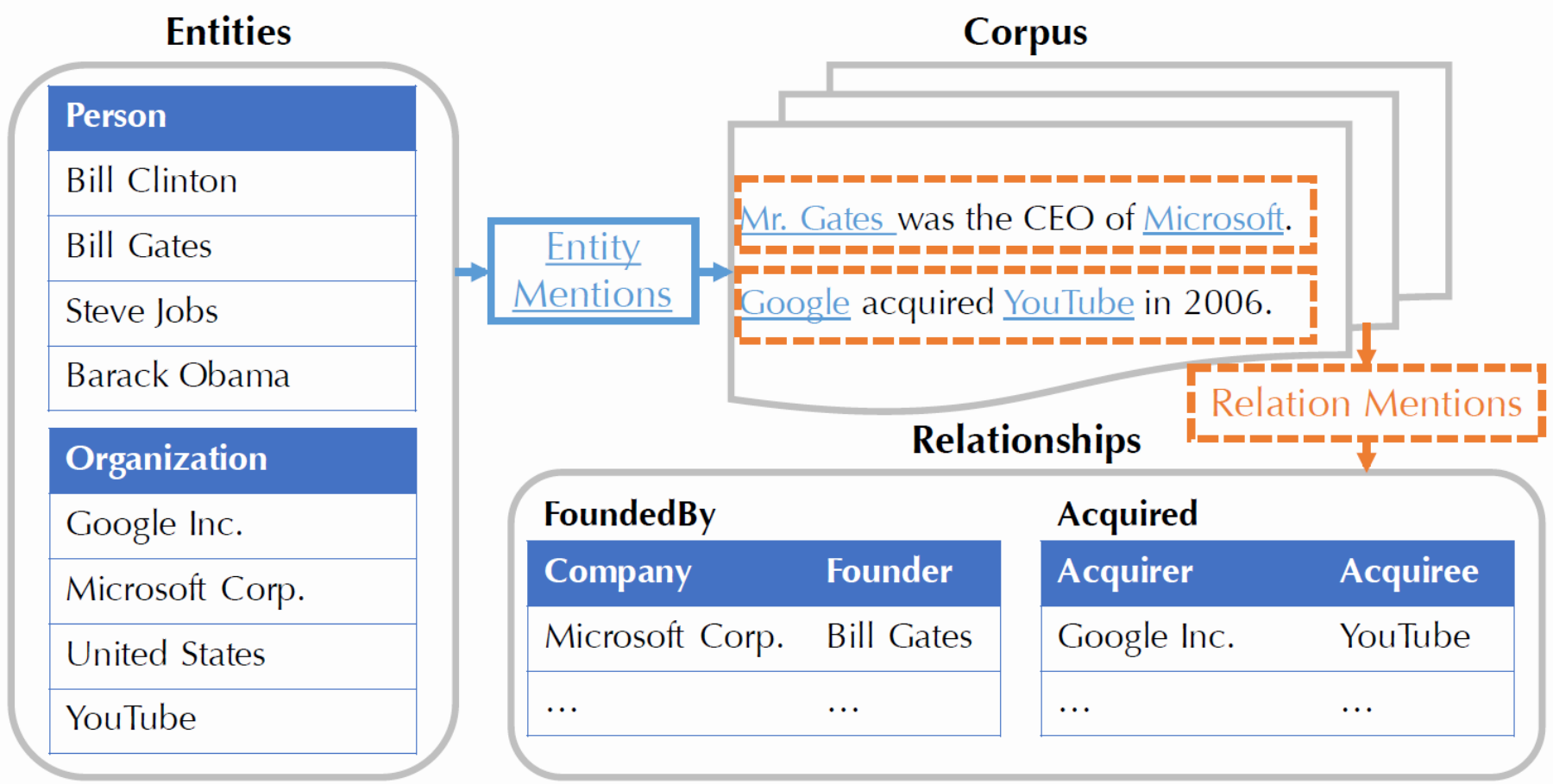
DeepDive

斯坦福大学InfoLab实验室开发的开源知识抽取系统

通过弱监督学习，从非结构化的文本中抽取结构化的关系数据

http://deepdive.stanford.edu/

1. 对文本进行分句
2. 分词并进行各种标注
3. 生成组合各种特征
4. 领域知识的集成
5. 远程监督学习
6. 使用因子图推理



OpenKG上的中文支持版本

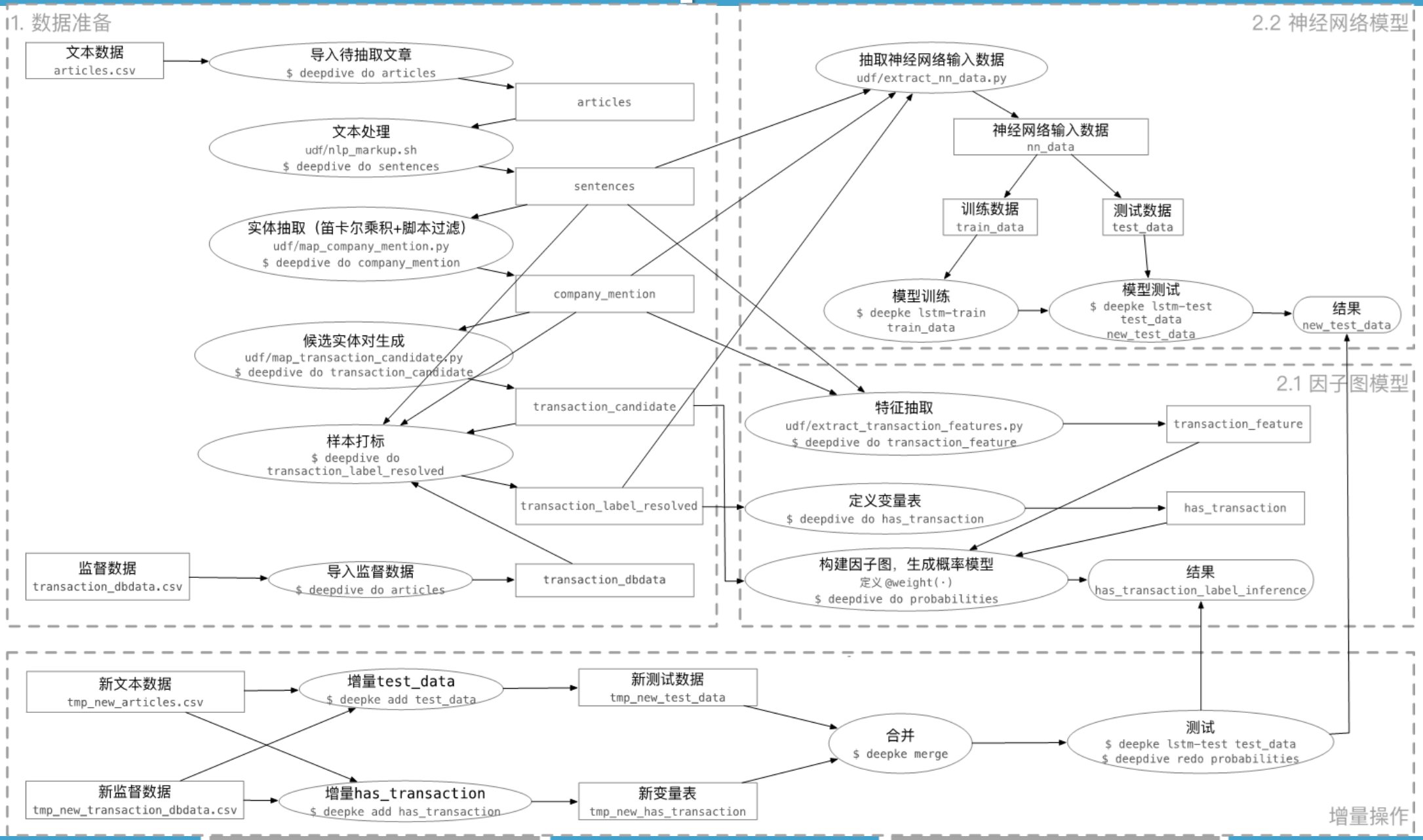
http://www.openkg.cn/tool/cn-deepdive

1. DeepDive: A Data Management System for Automatic Knowledge Base Construction. University of Wisconsin-Madison, 2015.
2. Incremental Knowledge Base Construction Using DeepDive. VLDB. 2015.

3 关系抽取工具：DeepDive

实践：

抽取上市公司中的股权交易关系



3 关系抽取工具：DeepDive

用Standford NLP模块进行文本处理

```
率，(，年率，)，向，公司，计付，理财，收益，，，实际，年化，收益率，4.5%，，，起止，期限，为，2015，年，5，月，25，日，至，2015，年，10，月
，8，日，。，公司，已，按期，收回，本金，和，利息，。，8，、，2015，年，5，月，25，日，，，公司，使用，25，000，万，元，人民币，购买，浦发，
银行，郑州，分行，“，利多多，对公，结构性，存款，2015，年，JG，525，期，”，理财，产品，，，该，理财，产品，为，保证，收益型，，，浦发，银
行，对，该，理财，产品，的，本金，提供，保证，承诺，，，并，按，协议，约定，的，投资，收益率，(，年率，)，，向，公司，计付，理财，收益，，，
实际，年化，收益率，3.6%，，，起止，期限，为，2015，年，5，月，27，日，至，2015，年，11，月，26，日，。，9，、，2015，年，6，月，29，日，
，，公司，使用，20，000，万，元，人民币，购买，广发，银行，郑州，黄河路，支行，“，名利双收，”，理财，产品，，，该，[illegible]，财，产品，为，保本，
浮动，收益型，，，广发，银行，对，该，理财，产，品，的，本金，提供，保证，承诺，，，并，按，协议，约定，的，投资，收益率，(，年率，)，向，
公司，计付，理财，收益，，，预期，年，收益率，:，3.0%，-4.5%，，，起止，期限，为，2015，年，6，月，29，日，至，2015，年，12，月，29，日，。
，10，、，2015，年，8，月，24，日，，，公司，使用，10，000，万，元，人民币，购买，光大，银行，郑州纬，二，路，支行，“，结构性，存款，”，理
财，产品，，，该，理财，产品，为，保证，收益型，，，光大，银行，对，该，理财，产品，的，本金，提供，保证，承诺，，，并，按，协议，约定，的，
投资，收益率，(，年率，)，向，公司，计付，理财，收益，，，预期，年，收益率，:，3.4%，，，起止，期限，为，2015，年，8，月，24，日，至，2016
，年，2，月，24，日，。，11，、，2015，年，8，月，25，日，，，公司，使用，8，000，万，人民币，购买，浦发，银行，郑州，分行，“，利，多多，对
公，结构性，存款，2015，年，JG，732，期，”，理财，产品，，，该，理财，产品，为，保证，收益型，，，浦，发，银行，对，该，理财，产品，的，本
金，提供，保证，承诺，，，并，按，协议，约定，的，投资，收益率，(，年率，)，向，公司，计付，理财，收益，，，实际，年化，收益率，3.45%，，，
起止，期限，为，2015，年，8，月，25，日，至，2016，年，2，月，26，日，。，12，、，2015，年，9，月，25，日，，，公司，使用，13，400，万，元，
人民币，购买，交通，银行，河南省，分行，4，郑州，煤矿，机械，集团，股份，有限，公司，Zhengzhou，Coal，Mining，Machinery，Group，Co.，，，L
td，营业部，“，蕴通，财富，，日，增利，182，天，”，理财，产品，，，该，理财，产品，为，保证，收益型，，，交通，银，行，对，该，理财，产品，
的，本金，提供，保证，承诺，，，并，按，协议，约定，的，投资，收益率，(，年率，)，向，公司，计付，理财，收益，，，实际，年化，收益率，3.55
%，，，起止，期限，为，2015，年，9，月，25，日，至，2016，年，3，月，25，日，。，13，、，2015，年，9，月，28，日，，，公司，使用，15，600，
万，元，人民币，购买，中国，银行，郑州，秦岭路，支行，“，中银，保本，理财，-，人民币，按期，开放，”，理财，产品，，，该，理财，产品，为，保
证，收益型，，，中，国，银行，对，该，理财，产品，的，本金，提供，保证，承诺，，，并，按，协议，约定，的，投资，收益率，(，年率，)，向，公
司，计付，理财，收益，，，预期，年，收益率，:，3.4%，，，起止，期限，为，2015，年，9，月，28，日，至，2016，年，3，月，28，日，。，14，、，
2015，年，10，月，16，日，，，公司，使用，10，000，万，元，人民币，购买，光大，银行，郑州纬，二，路，支行，“，结构性，存款，”，理财，产品，
，，该，理财，产品，为，保证，收益型，，，光大，银行，对，该，理财，产，品，的，本金，提供，保证，承诺，，，并，按，协议，约定，的，投资，
收益率，(，年率，)，向，公司，计付，理财，收益，，，预期，年，收益率，:，3.4%，，，起止，期限，为，2015，年，10，月，16，日，至，2016，年，
4，月，16，日，。，15，、，2015，年，10，月，21，日，，，公司，使用，30，000，万，元，人民币，购买，民生，银行，郑州，分行，营业部，“，结构
性，存款，”，理财，产品，，，该，理财，产品，为，保证，收益型，，，民生，银行，对，该，理财，产，品，的，本金，提供，保证，承诺，，，并，按
，协议，约定，的，投资，收益率，(，年率，)，向，公司，计付，理财，收益，，，预期，年，收益率，:，3.65%，，，起止，期限，为，2015，年，10，
月，21，日，至，2016，年，04，月，21，日，。，16，、，2015，年，10，月，27，日，，，公司，使用，10，000，万，元，人民币，购买，浦发，银行，
郑州，金水，支行，“，利多多，对公，结构性，存款，2015，年，JG，908，期，”，理财，产品，，，该，理财，产品，为，保证，收，益型，，，浦发，银
行，对，该，理财，产品，的，本金，提供，保证，承诺，，，并，按，协议，约定，的，投资，收益率，，(，年率，)，向，公司，计付，理财，收益，，，
预期，年，化，收益率，:，3.6%，，，起止，期限，为，2015，年，10，月，27，日，至，2016，年，4，月，28，日，。，，五，、，备查，文件，第三，
loading dd_tmp_sentences: 0:44:06 1.27k [ 478m/s] ([   0 /s]), 煤矿，机械，集团，股份，有限，公司，董事会，二○一五年，十一月，二日]
loading dd_tmp_sentences: 0:44:06 2.68MiB [1.04KiB/s] ([   0 B/s])
```

3 关系抽取工具：DeepDive

特征提取结果

```
(python2) [root@VM_32_17_centos input]#  deepdive query '| 20 ?- transaction_feature(_, _, feature).'
                    feature
-----------------------------------------------
 IS_INVERTED
 INV_WORD_SEQ_[广通 股份 有限 公司]
 INV_LEMMA_SEQ_[广通 股份 有限 公司]
 INV_NER_SEQ_[ORG ORG ORG ORG]
 INV_POS_SEQ_[NR NN JJ NN]
 INV_W_LEMMA_L_1_R_1_[电]_[第七]
 INV_W_NER_L_1_R_1_[O]_[MISC]
 INV_W_LEMMA_L_1_R_2_[电]_[第七 届]
 INV_W_NER_L_1_R_2_[O]_[MISC MISC]
 INV_W_LEMMA_L_1_R_3_[电]_[第七 届 董事会]
 INV_W_NER_L_1_R_3_[O]_[MISC MISC MISC]
 INV_W_LEMMA_L_2_R_1_[中 电]_[第七]
 INV_W_NER_L_2_R_1_[O O]_[MISC]
 INV_W_LEMMA_L_2_R_2_[中 电]_[第七 届]
 INV_W_NER_L_2_R_2_[O O]_[MISC MISC]
 INV_W_LEMMA_L_2_R_3_[中 电]_[第七 届 董事会]
 INV_W_NER_L_2_R_3_[O O]_[MISC MISC MISC]
 INV_W_LEMMA_L_3_R_1_[2015-020 中 电]_[第七]
 INV_W_NER_L_3_R_1_[O O O]_[MISC]
 INV_W_LEMMA_L_3_R_2_[2015-020 中 电]_[第七 届]
(20 rows)
```

3 关系抽取工具：DeepDive

预测的公司间交易关系概率

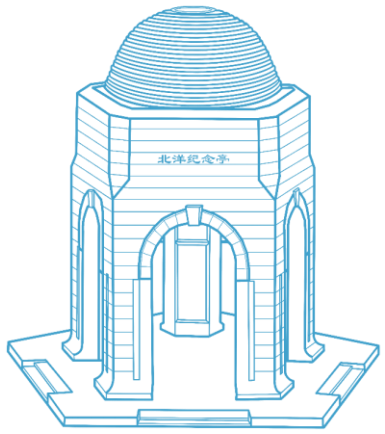
```
(python2) [root@VM_32_17_centos input]# deepdive sql "SELECT t1.p1_id, t1.p2_id, p1_name, p2_name, expectation FROM has_transaction_label_i
nference t1, transaction_candidate t2 WHERE t1.p1_id = t2.p1_id AND t1.p2_id = t2.p2_id
> ORDER BY expectation DESC LIMIT 20"
       p1_id        |       p2_id        |                p1_name                 |                p2_name                 | expectation
--------------------+--------------------+----------------------------------------+----------------------------------------+-------------
 1201743813_6_71_72 | 1201743813_6_48_53 | 铁投集团                               | 四川铁能电力开发有限公司               |       0.997
 1201743813_6_48_53 | 1201743813_6_71_72 | 四川铁能电力开发有限公司               | 铁投集团                               |       0.986
 1201738844_1_60_61 | 1201738844_1_35_39 | 合肥工投                               | 南京医药股份有限公司                   |       0.893
 1201738844_1_71_77 | 1201738844_1_35_39 | 南京医药合肥天星有限公司               | 南京医药股份有限公司                   |        0.66
 1201734457_18_13_18 | 1201734457_18_7_11 | 甘肃大有农业科技有限公司               | 甘肃天润薯业有限责任公司               |       0.533
 1201739153_25_12_18 | 1201739153_25_1_7  | 江苏汇鸿国际集团股份有限公司           | 江苏凤凰出版传媒集团有限公司           |       0.338
 1201739153_25_1_7  | 1201739153_25_12_18 | 江苏凤凰出版传媒集团有限公司           | 江苏汇鸿国际集团股份有限公司           |       0.323
 1201738844_1_35_39 | 1201738844_1_60_61 | 南京医药股份有限公司                   | 合肥工投                               |       0.313
 1201734457_18_7_11 | 1201734457_18_13_18 | 甘肃天润薯业有限责任公司               | 甘肃大有农业科技有限公司               |       0.302
 1201734460_1_53_54 | 1201734460_1_32_36 | 福普克药业                             | 福医药集团股份公司                     |       0.261
 1201743522_19_22_30 | 1201743522_19_10_18 | 中国电建集团昆明勘测设计研究院有限公司 | 中国水电顾问集团投资有限公司与公司     |       0.255
 1201743522_19_10_18 | 1201743522_19_22_30 | 中国水电顾问集团投资有限公司与公司     | 中国电建集团昆明勘测设计研究院有限公司 |       0.247
 1201734460_1_32_36 | 1201734460_1_53_54 | 福医药集团股份公司                     | 福普克药业                             |       0.215
 1201743522_28_22_26 | 1201743522_28_6_15 | 市镇开发建设有限公司                   | 中国电建集团华东勘测设计研究院有限公司 |       0.211
 1201747748_33_30_32 | 1201747748_33_21_27 | MiningMachineryGroup                   | 郑州煤矿机械集团股份有限公司           |       0.208
 1201743522_28_6_15 | 1201743522_28_22_26 | 中国电建集团华东勘测设计研究院有限公司 | 市镇开发建设有限公司                   |       0.195
 1201738844_1_35_39 | 1201738844_1_71_77 | 南京医药股份有限公司                   | 南京医药合肥天星有限公司               |       0.182
 1201747748_33_21_27 | 1201747748_33_30_32 | 郑州煤矿机械集团股份有限公司           | MiningMachineryGroup                   |       0.179
 1201743522_13_10_19 | 1201743522_13_4_8  | 中国电建集团成都勘测设计研究院有限公司 | 电建建筑集团有限公司                   |       0.155
 1201743522_13_4_8  | 1201743522_13_10_19 | 电建建筑集团有限公司                   | 中国电建集团成都勘测设计研究院有限公司 |       0.147
(20 rows)
```

3 关系抽取工具：DeepDive

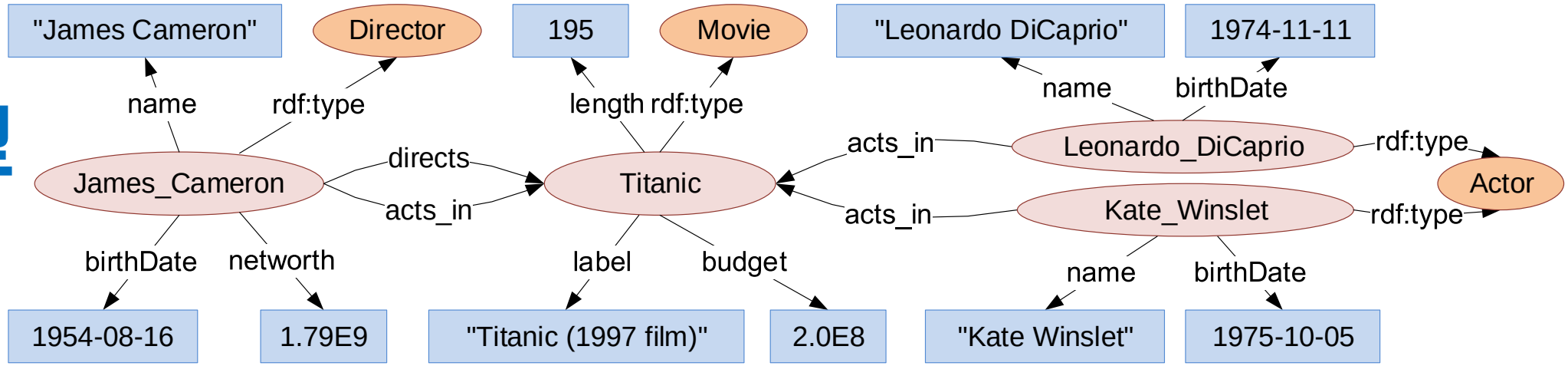
实践：抽取上市公司中的股权交易关系

下载链接：http://www.openkg.cn/tool/cn-deepdive

目录 CONTENTS



RDF图模型

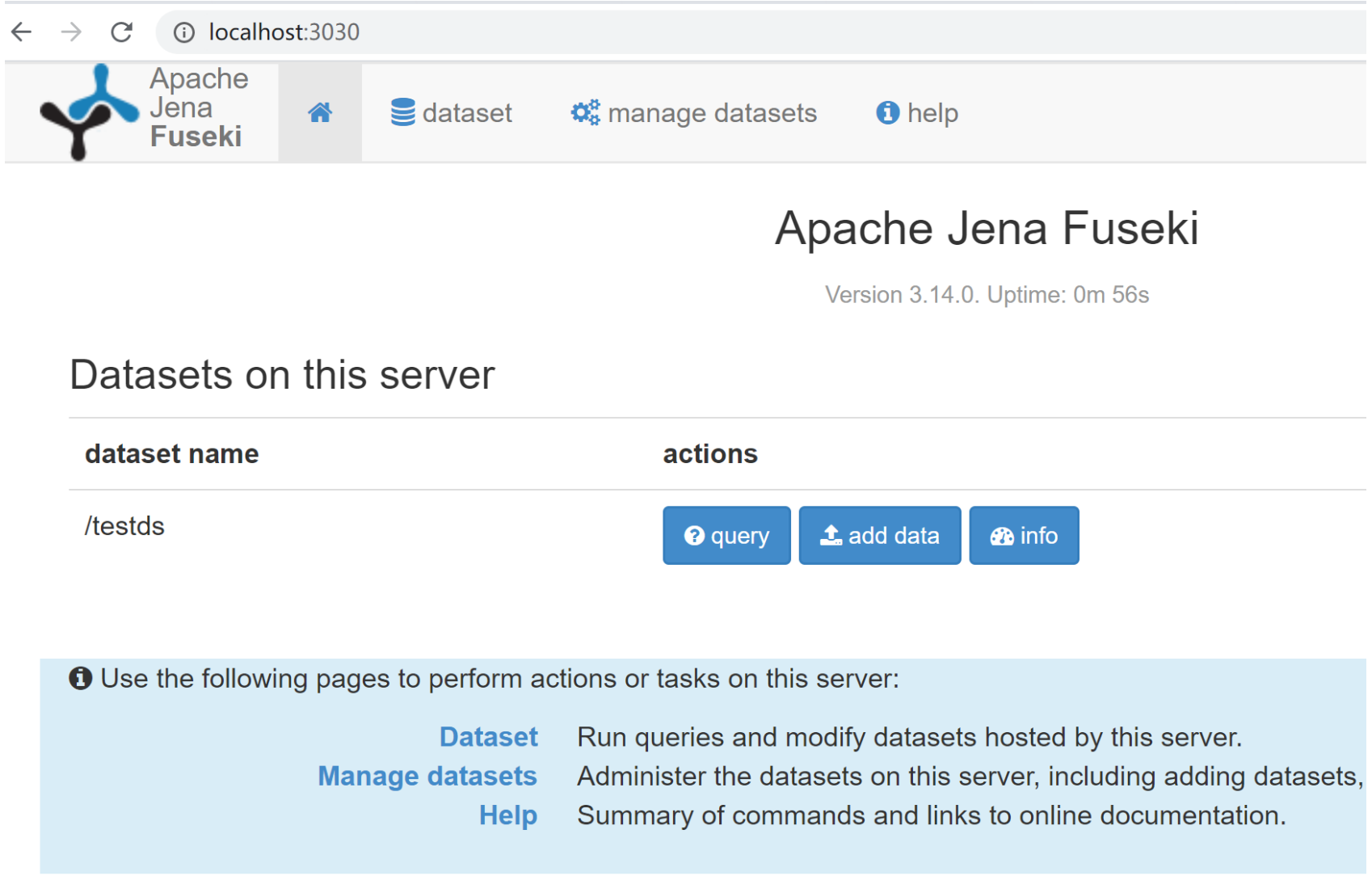


G = { (James_Cameron, **rdf:type**, **Director**),
(James_Cameron, birthDate, 1954-08-16),
(James_Cameron, directs, Titanic),
(Titanic, **rdf:type**, **Movie**),
(Titanic, budget, 2.0E8),
(Leonardo_DiCaprio, **rdf:type**, **Actor**),
(Leonardo_DiCaprio, birthDate, 1974-11-11),
(Kate_Winslet, **rdf:type**, **Actor**),
(Kate_Winslet, birthDate, 1975-10-05),
(James_Cameron, name, "James Cameron"),
(James_Cameron, networth, 1.79E9),
(James_Cameron, acts_in, Titanic),
(Titanic, label, "Titanic (1997 film)"),
(Titanic, length, 195),
(Leonardo_DiCaprio, name, "Leonardo DiCaprio"),
(Leonardo_DiCaprio, acts_in, Titanic),
(Kate_Winslet, name, "Kate Winslet"),
(Kate_Winslet, acts_in, Titanic) }

18条三元组

天津大学 Tianjin University

Jena Fuseki: Web Interface

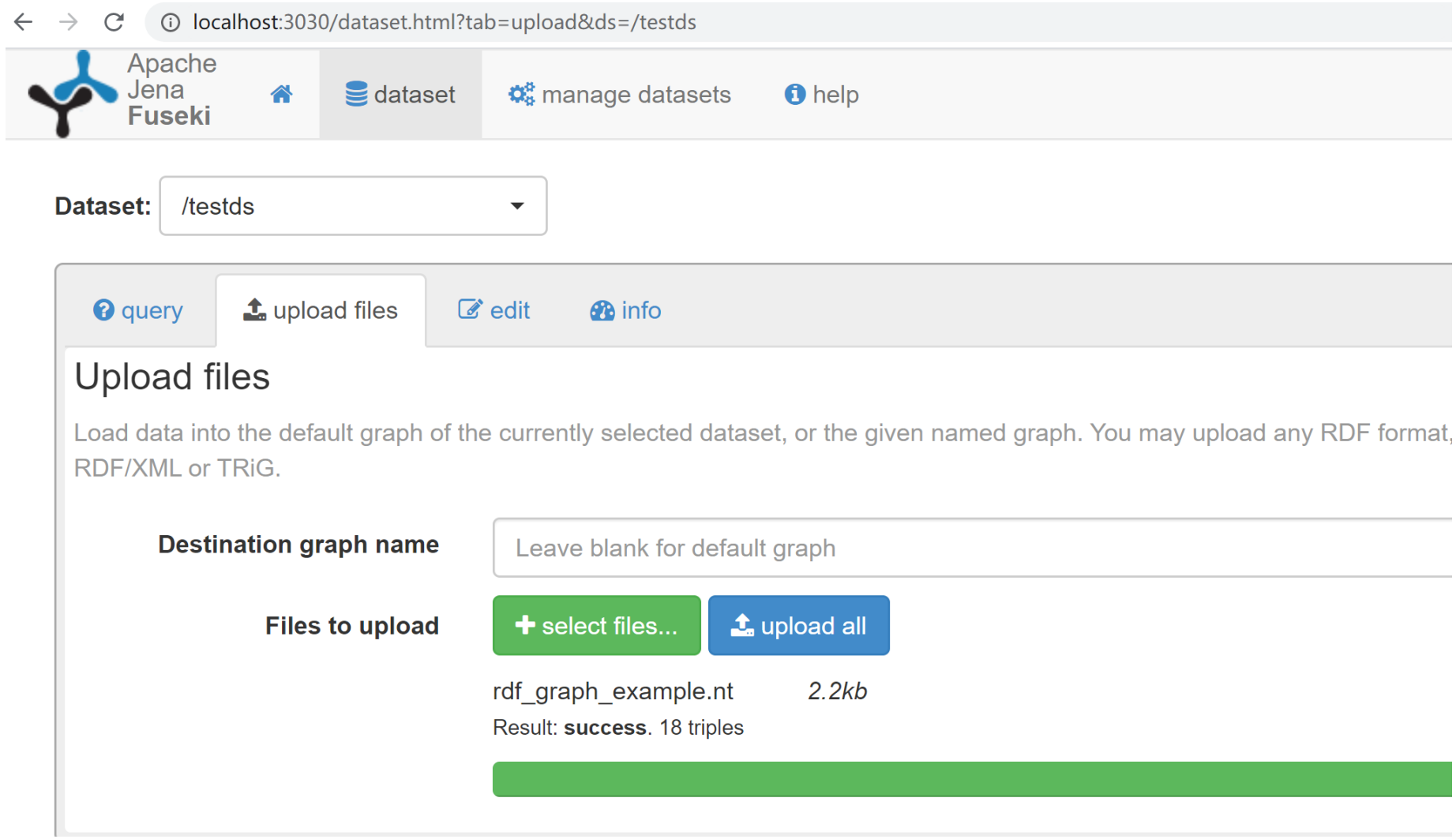


RDF Graph Example: In N-Triple Format

```
<http://dbpedia.org/resource/James_Cameron> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://dbpedia.org/ontology/Director> .
<http://dbpedia.org/resource/James_Cameron> <http://dbpedia.org/ontology/name> "James Cameron" .
<http://dbpedia.org/resource/James_Cameron> <http://dbpedia.org/ontology/birthDate> "1954-08-16"^^<http://www.w3.org/2001/XMLSchema#date> .
<http://dbpedia.org/resource/James_Cameron> <http://dbpedia.org/ontology/networth> "1.79E9"^^<http://www.w3.org/2001/XMLSchema#decimal> .
<http://dbpedia.org/resource/James_Cameron> <http://dbpedia.org/ontology/directs> <http://dbpedia.org/resource/Titanic> .
<http://dbpedia.org/resource/James_Cameron> <http://dbpedia.org/ontology/acts_in> <http://dbpedia.org/resource/Titanic> .
<http://dbpedia.org/resource/Titanic> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://dbpedia.org/ontology/Movie> .
<http://dbpedia.org/resource/Titanic> <http://www.w3.org/2000/01/rdf-schema#label> "Titanic (1997 film)" .
<http://dbpedia.org/resource/Titanic> <http://dbpedia.org/ontology/budget> "2.0E8"^^<http://www.w3.org/2001/XMLSchema#decimal> .
<http://dbpedia.org/resource/Titanic> <http://dbpedia.org/ontology/length> "195"^^<http://www.w3.org/2001/XMLSchema#integer> .
<http://dbpedia.org/resource/Leonardo_DiCaprio> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://dbpedia.org/ontology/Actor> .
<http://dbpedia.org/resource/Leonardo_DiCaprio> <http://dbpedia.org/ontology/name> "Leonardo DiCaprio" .
<http://dbpedia.org/resource/Leonardo_DiCaprio> <http://dbpedia.org/ontology/birthDate> "1974-11-11"^^<http://www.w3.org/2001/XMLSchema#date>
<http://dbpedia.org/resource/Leonardo_DiCaprio> <http://dbpedia.org/ontology/acts_in> <http://dbpedia.org/resource/Titanic> .
<http://dbpedia.org/resource/Kate_Winslet> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://dbpedia.org/ontology/Actor> .
<http://dbpedia.org/resource/Kate_Winslet> <http://dbpedia.org/ontology/name> "Kate Winslet" .
<http://dbpedia.org/resource/Kate_Winslet> <http://dbpedia.org/ontology/birthDate> "1975-10-05"^^<http://www.w3.org/2001/XMLSchema#date> .
<http://dbpedia.org/resource/Kate_Winslet> <http://dbpedia.org/ontology/acts_in> <http://dbpedia.org/resource/Titanic> .
```

天津大学 Tianjin University

Jena Fuseki: Upload Data

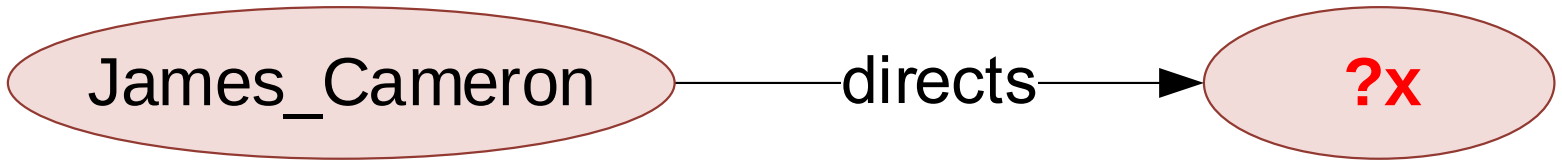


SPARQL

查询1: 查询 James_Cameron 执导的电影?

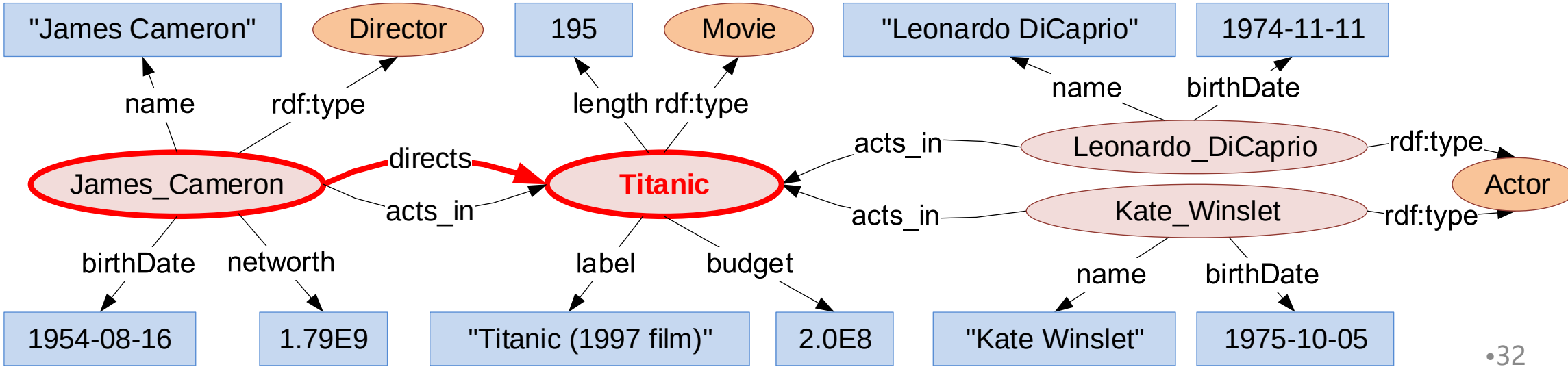
三元组模式

```
SELECT ?x
WHERE { dbr:James_Cameron dbo:directs ?x . }
```



查询结果

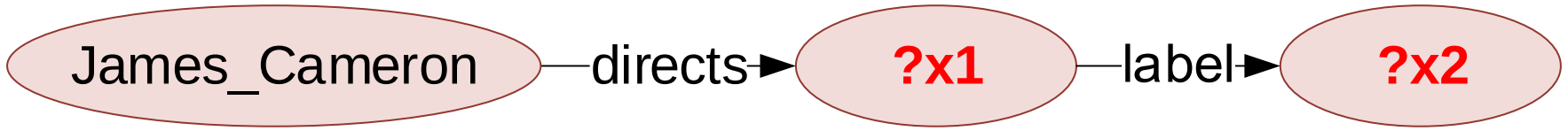
?x
Titanic



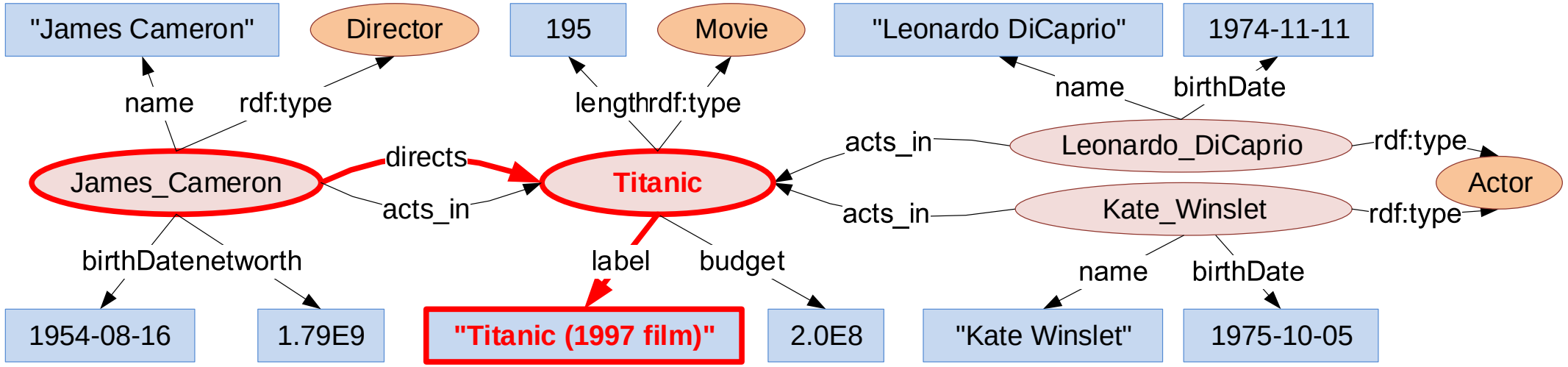
查询 2: 查询 James_Cameron 执导的电影及其片名?

基本图模式

```
SELECT ?x1 ?x2
WHERE { dbr:James_Cameron dbo:directs ?x1 . ?x1 rdfs:label ?x2 . }
```



链式 Chain



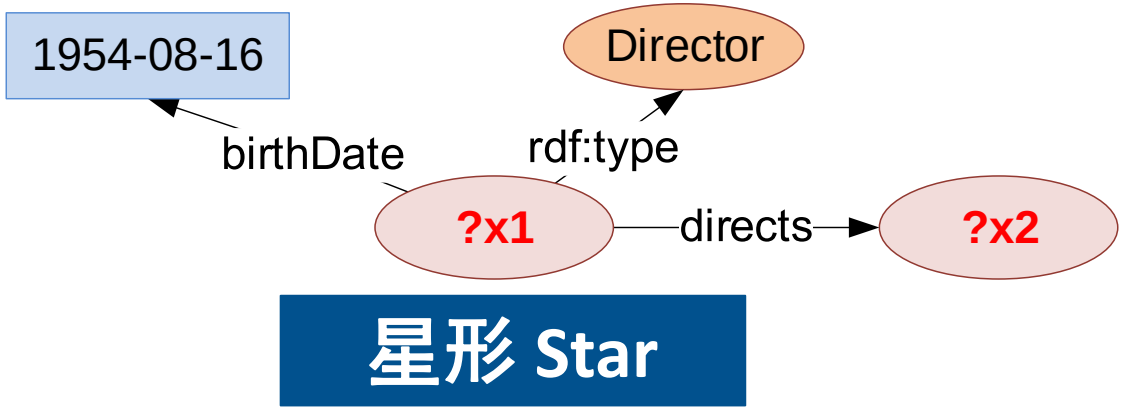
查询结果

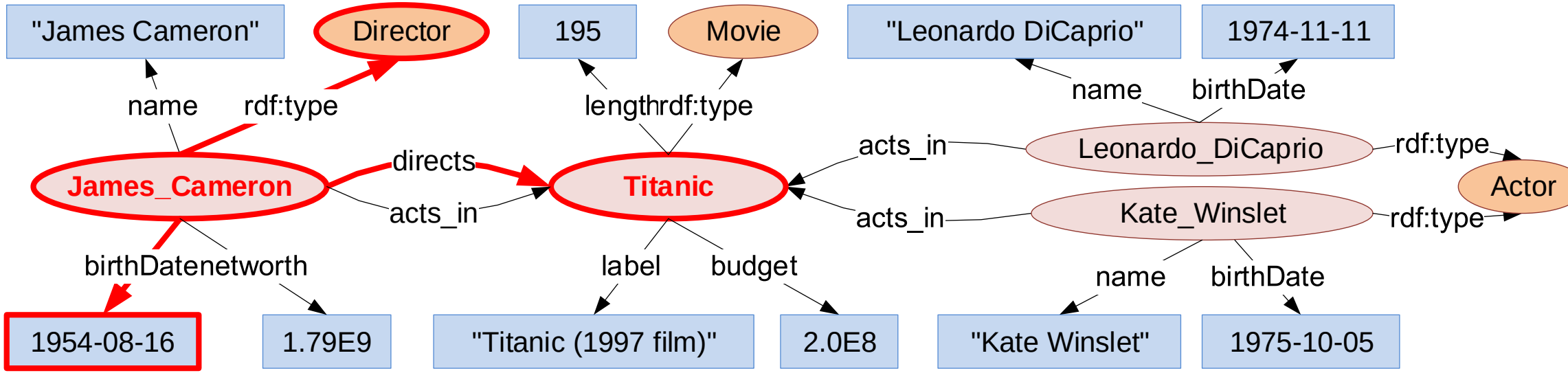
?x1	?x2
Titanic	"Titanic (1997 film)"

■ 查询 3: 查询1954-08-16出生的导演执导的电影?

基本图模式

```
SELECT ?x2
WHERE { ?x1 dbo:directs ?x2 .
        ?x1 rdf:type dbo:Director .
  ?x1 dbo:birthDate "1954-08-16"^^xsd:date .
        }
```

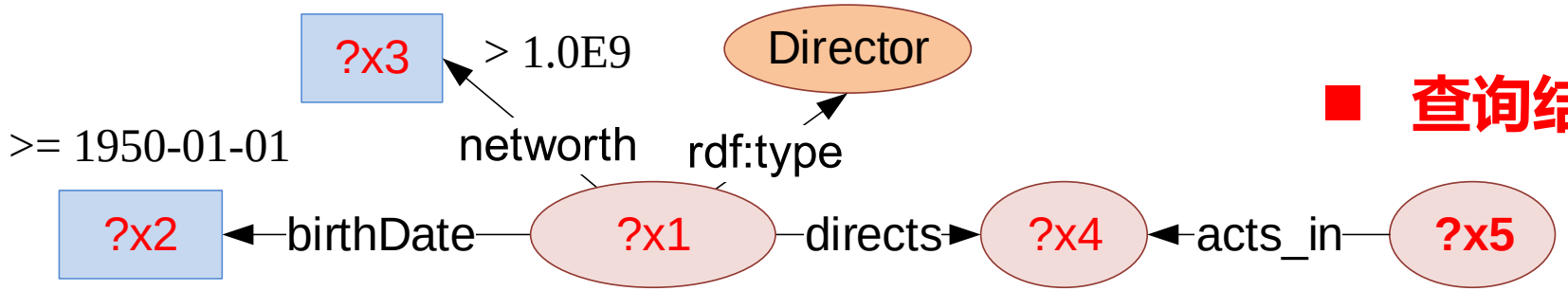




■ 查询 4: 查询1950年之后出生的资产大于1.0E9的导演执导的电影的出演演员?

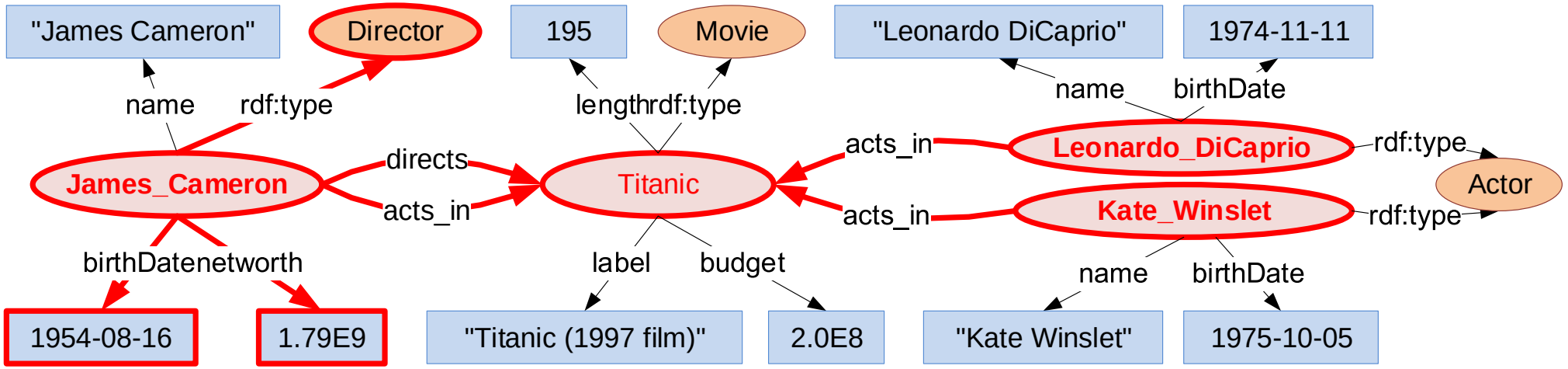
复杂图模式

```
SELECT ?x5
WHERE { ?x1 rdf:type dbo:Director . ?x1 dbo:birthDate ?x2 . FILTER (?x2 >= "1950-01-01"^^xsd:date)
        ?x1 dbo:networth ?x3 . FILTER (?x3 > 1.0E9)  ?x1 dbo:directs ?x4 . ?x5 dbo:acts_in ?x4 . }
```



■ 查询结果

?x5
Leonardo_DiCaprio
Kate_Winslet
James_Cameron



■ 查询 5: 查询具有多步“合作距离”的两名演员?

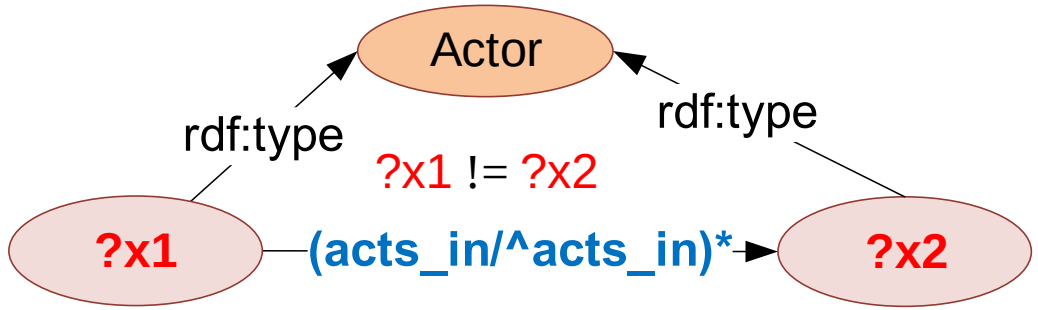
属性路径

```
SELECT ?x1 ?x2
WHERE {
?x1 (dbo:acts_in / ^dbo:acts_in)* ?x2 . FILTER (?x1 != ?x2)
?x1 rdf:type dbo:Actor . ?x2 rdf:type dbo:Actor . }
```

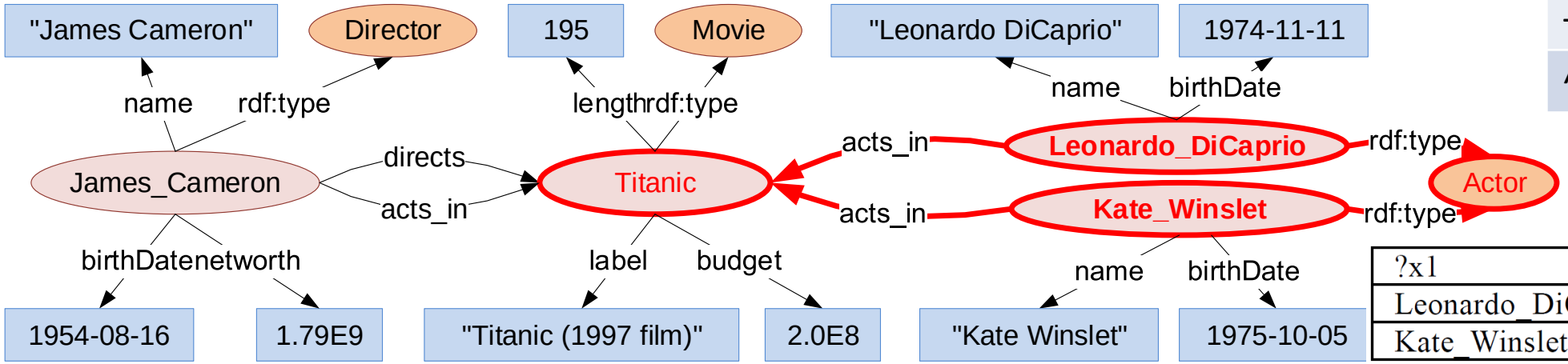
属性路径 property path: 导航式查询

正在表达式 regular expression:

(acts_in / ^acts_in)*



Op	语义 Semantics
/	path concatenation
|	Path union
*	Kleene closure
+	Kleene plus
^	Edge inverse

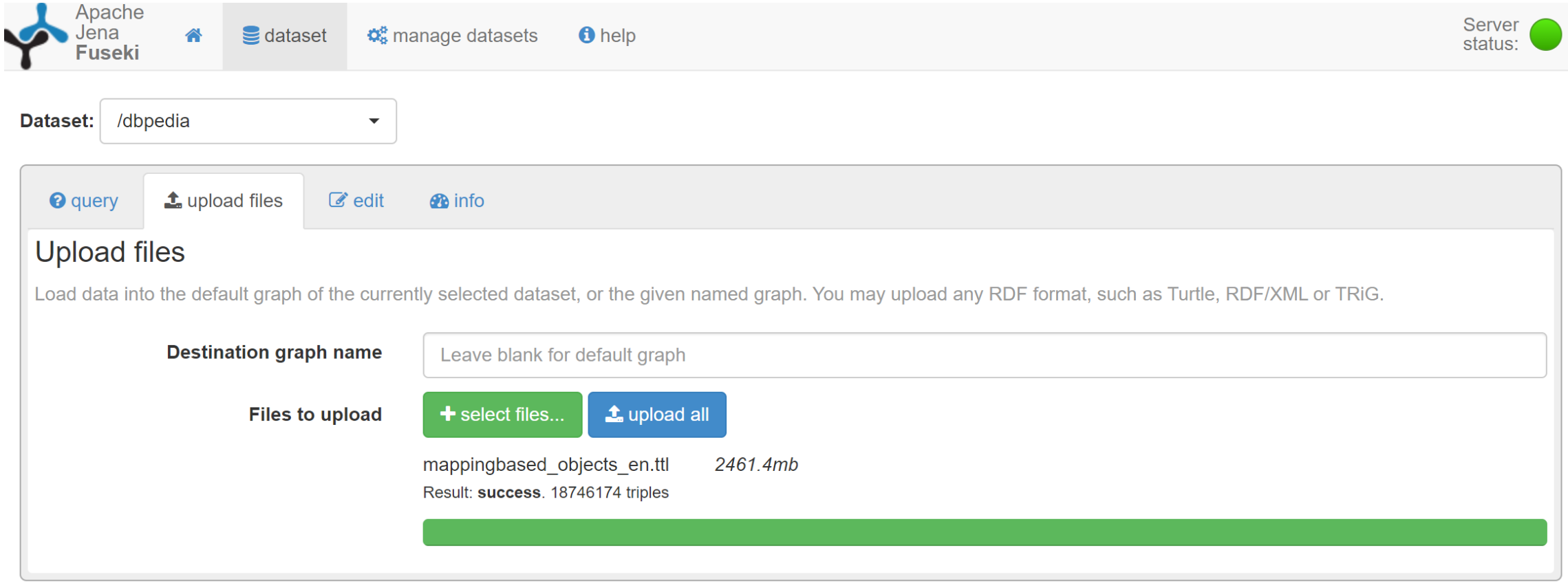


■ 查询结果

?x1	?x2
Leonardo_DiCaprio	Kate_Winslet
Kate_Winslet	Leonardo_DiCaprio

使用真实数据集：DBpedia

使用Jena Fuseki导入DBpedia-2016-10数据集mappingbased_objects_en.ttl



使用真实数据集：DBpedia

查询dbr:Tianjin_University所在城市（dbo:city）以及在同一城市的类型（dbo:type）为dbr:National_university的其他实体（dbr:Tianjin_University自身不包括在结果中）

SPARQL语句：

```
SELECT ?city ?university {
  dbr:Tianjin_University dbo:city ?city .
  ?university dbo:city ?city .
  FILTER (?university != dbr:Tianjin_University ) .
  ?university dbo:type dbr:National_university .
}
```

查询结果截图：

Showing 1 to 3 of 3 entries

Search:

	city	university
1	dbr:Tianjin	dbr:Nankai_University
2	dbr:Tianjin	dbr:Tianjin_Foreign_Studies_University
3	dbr:Tianjin	dbr:Tianjin_Crafts_and_Arts_Vocational_College

使用真实数据集：DBpedia

查找爱因斯坦dbr:Albert_Einstein的导师（dbo:doctoralAdvisor）、导师的导师以及导师的导师的导师。

SPARQL语句：

```
SELECT * {
  { dbr:Albert_Einstein dbo:doctoralAdvisor ?advisor1 . }
  UNION
  { dbr:Albert_Einstein dbo:doctoralAdvisor/dbo:doctoralAdvisor ?advisor2 . }
  UNION
  {dbr:Albert_Einstein dbo:doctoralAdvisor/dbo:doctoralAdvisor/dbo:doctoralAdvisor ?advisor3 . }
}
```

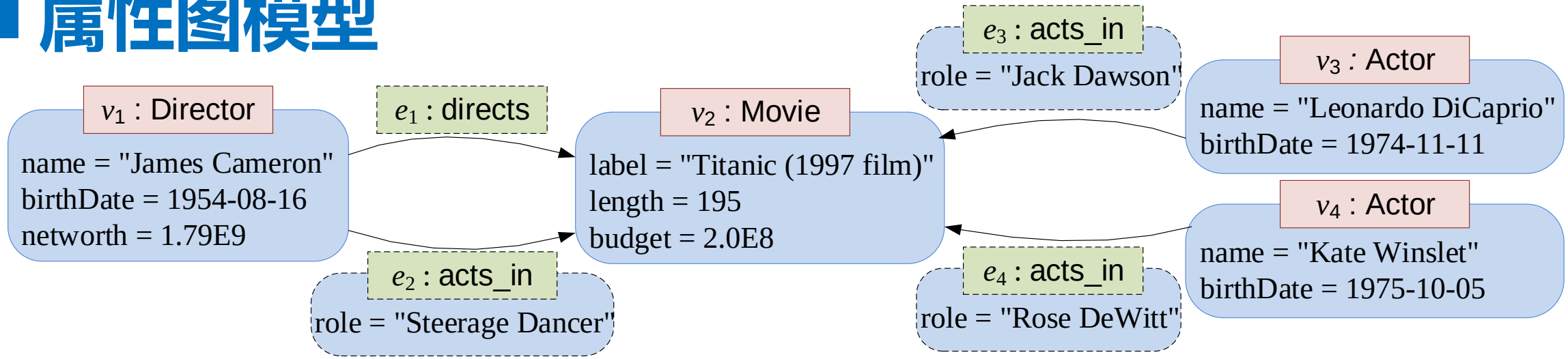
查询结果截图：

Showing 1 to 3 of 3 entries

Search:

	advisor1	advisor2	advisor3
1	dbr:Alfred_Kleiner		
2		dbr:Johann_Jakob_Müller	
3			dbr:Adolf_Fick

属性图模型



$V = \{v_1, v_2, v_3, v_4\}, \quad E = \{e_1, e_2, e_3, e_4\},$

$\rho(e_1) = (v_1, v_2), \quad \rho(e_2) = (v_1, v_2), \quad \rho(e_3) = (v_3, v_2), \quad \rho(e_4) = (v_4, v_2),$

$\lambda(v_1) = \text{Director}, \quad \lambda(v_2) = \text{Movie}, \quad \lambda(v_3) = \text{Actor}, \quad \lambda(v_4) = \text{Actor},$

$\lambda(e_1) = \text{directs}, \quad \lambda(e_2) = \text{acts\_in}, \quad \lambda(e_3) = \text{acts\_in}, \quad \lambda(e_4) = \text{acts\_in},$

$\sigma(v_1, \text{name}) = \text{"James Cameron"}, \quad \sigma(v_1, \text{birthDate}) = 1954\text{-}08\text{-}16, \quad \sigma(v_1, \text{networth}) = 1.79E9,$

$\sigma(v_2, \text{label}) = \text{"Titanic (1997 film)"}, \quad \sigma(v_2, \text{length}) = 195, \quad \sigma(v_2, \text{budget}) = 2.0E8,$

$\sigma(v_3, \text{name}) = \text{"Leonardo DiCaprio"}, \quad \sigma(v_3, \text{birthDate}) = 1974\text{-}11\text{-}11, \quad \sigma(v_4, \text{name}) = \text{"Kate Winslet"},$

$\sigma(v_4, \text{birthDate}) = 1975\text{-}10\text{-}05, \quad \sigma(e_2, \text{role}) = \text{"Steerage Dancer"}, \quad \sigma(e_3, \text{role}) = \text{"Jack Dawson"},$

$\sigma(e_4, \text{role}) = \text{"Rose DeWitt"}$

Neo4j: Create the example graph

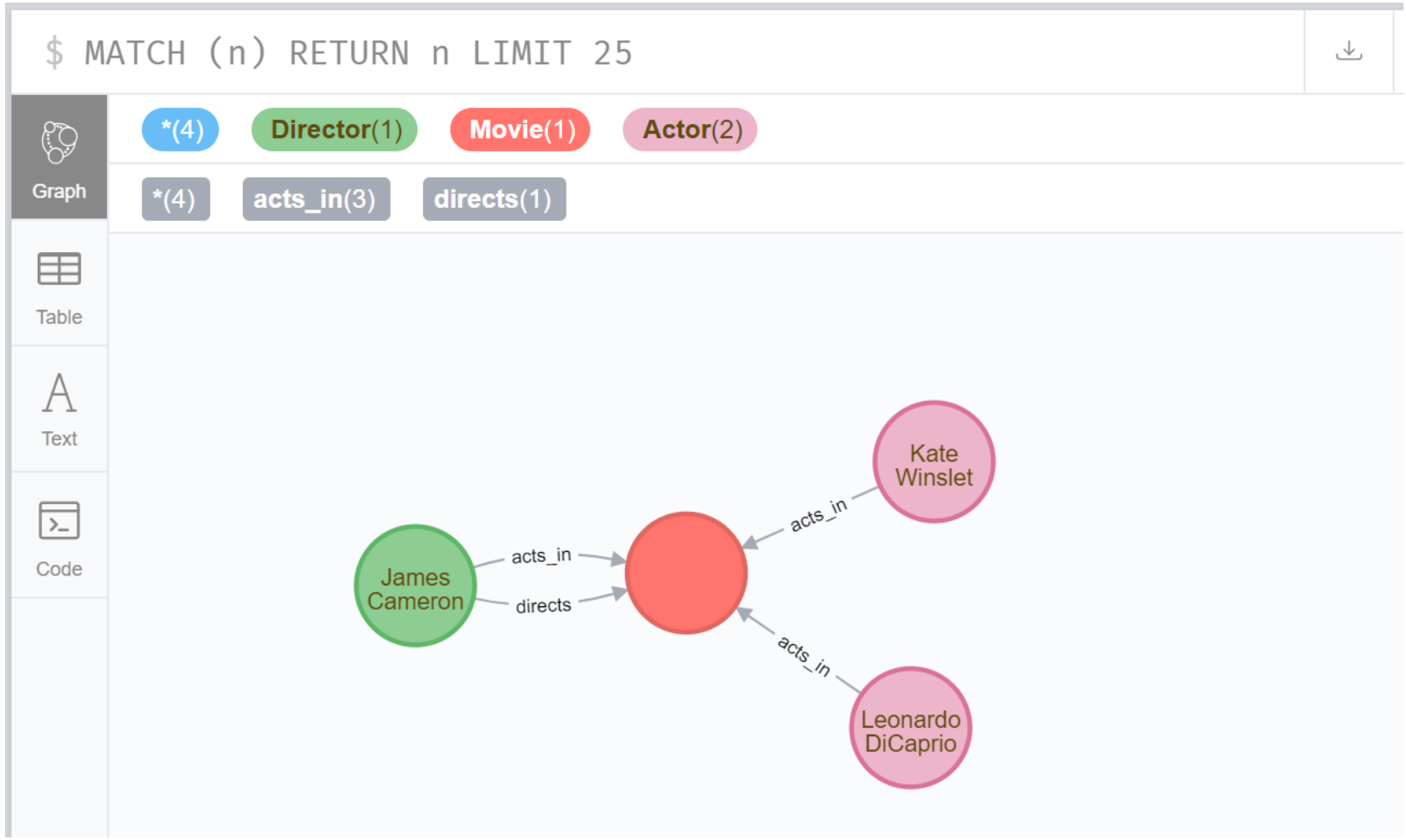
```
1 CREATE (v1:Director {name:"James Cameron", birthDate:"1954-08-16", networth:1.79E9})
2 CREATE (v2:Movie {label:"Titanic (1997 film)", length:195, budget:2.0E8})
3 CREATE (v3:Actor {name:"Leonardo DiCaprio", birthDate:"1974-11-11"})
4 CREATE (v4:Actor {name:"Kate Winslet", birthDate:"1975-10-05"})
5 CREATE (v1)-[e1:directs]→(v2)
6 CREATE (v1)-[e2:acts_in {role:"Steerage Dancer"}]→(v2)
7 CREATE (v3)-[e3:acts_in {role:"Jack Dawson"}]→(v2)
```

```
$ CREATE (v1:Director {name:"James Cameron", birthDate:"1954-08-16", networth:1.79E9}) CREATE (v…
```

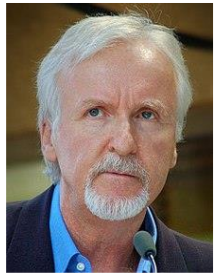
Table

Added 4 labels, created 4 nodes, set 13 properties, created 4 relationships, completed after 221 ms.

Neo4j: Create the example graph



图模式匹配



v_1 : Director

name = "James Cameron"
birthDate = 1954-08-16
networth = 1.79E9

e_1 : directs

v_2 : Movie

label = "Titanic (1997 film)"
length = 195
budget = 2.0E8

Node Relationship Node

```
MATCH (:Director {name: "James Cameron"}) -[:directs]-> (x:Movie) RETURN x, x.label
```

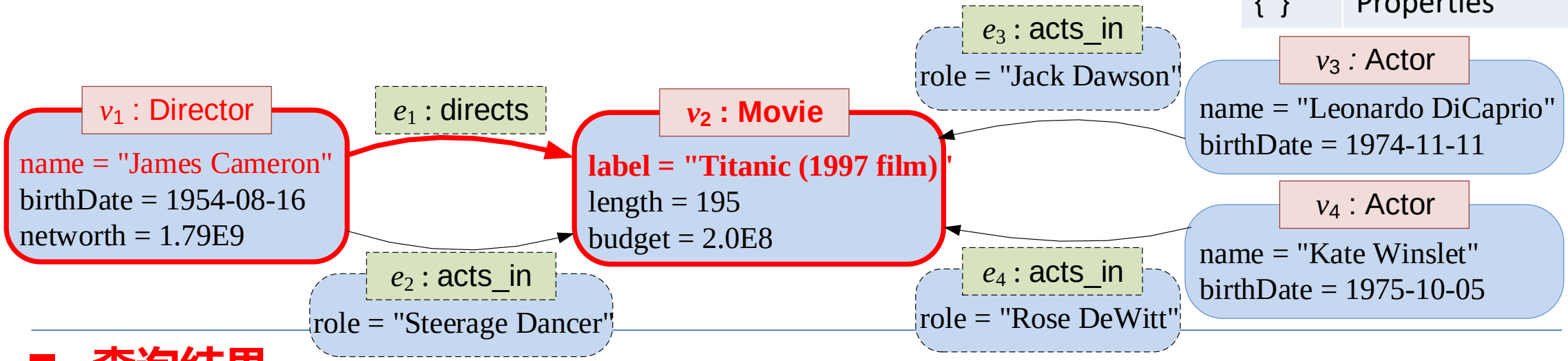
Label Property Variable

查询 6: 查询 James_Cameron 执导的电影及其片名?

SPARQL 查询 2

```
MATCH (:Director {name: "James Cameron"})-[:directs]->(x:Movie)
RETURN x, x.label
```

Op	语义 Semantics
( )	Node information
-[ ]->	Edge information
:	Node (edge) label
{ }	Properties



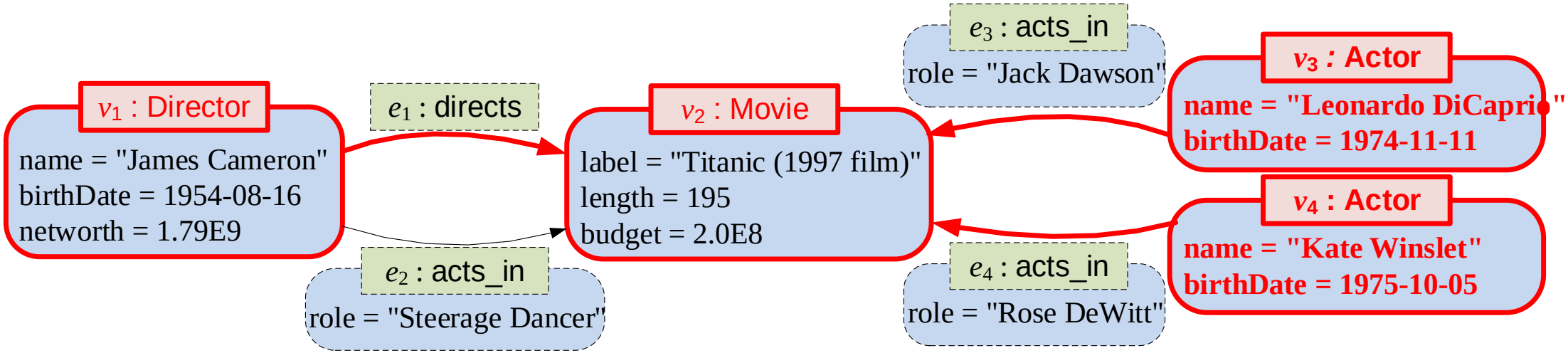
查询结果

x	x.label
{"length":195,"label":"Titanic (1997 film)","budget":200000000.0}	"Titanic (1997 film)"

查询 7: 查询1950年之后出生的资产大于1.0E9的导演执导的电影的出演演员?

SPARQL 查询 4

```
MATCH (x1:Director)-[:directs]->(:Movie)<-[:acts_in]-(x2:Actor)
WHERE x1.networth > 1.0E9 AND x1.birthDate >= date("1950-01-01")
RETURN x2
```



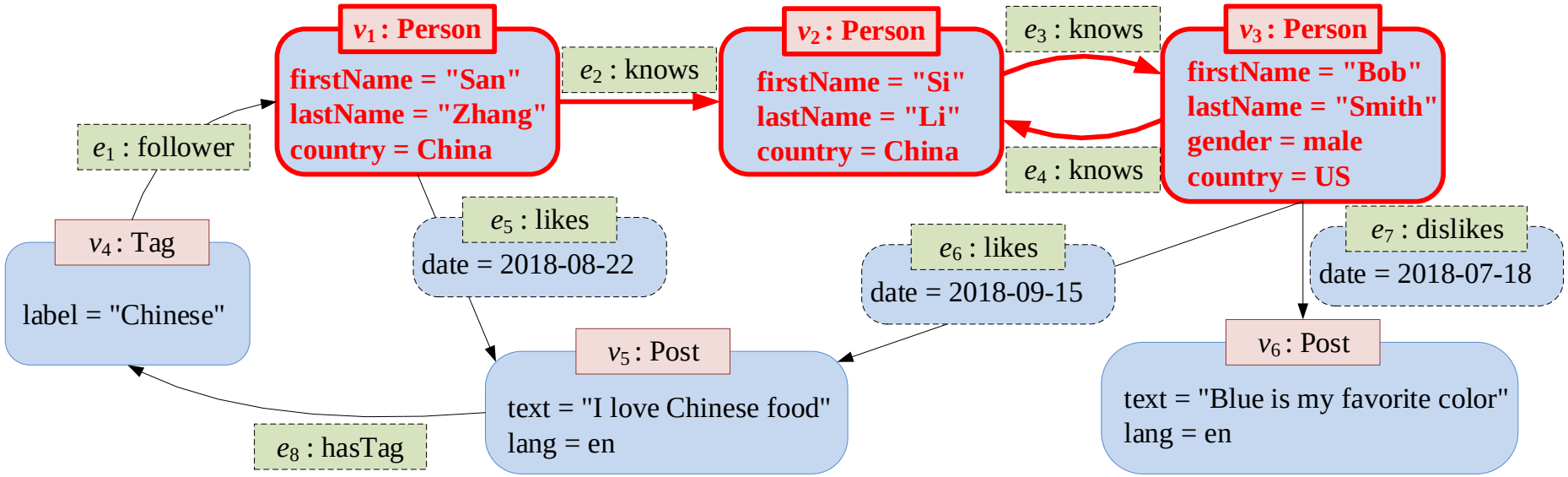
查询结果

x2
{"name":"Kate Winslet","birthDate":"1975-10-05"}
{"name":"Leonardo DiCaprio","birthDate":"1974-11-11"}

■ 查询 8: 查询“San Zhang”直接和间接认识的人?

```
MATCH (x1:Person)-[:knows*]->(x2:Person)
WHERE x1.firstName = "San" AND x1.lastName = "Zhang"
RETURN x1, x2
```

限制: 传递闭包算子*只能作用在单独一个边标签



■ **查询结果**

语义: 无重复边 no-repeated-edge

x1	x2
{"country":"China","firstName":"San","lastName":"Zhang"}	{"firstName":"Si","lastName":"Li","country":"China"}
{"country":"China","firstName":"San","lastName":"Zhang"}	{"firstName":"Bob","lastName":"Smith","country":"US","gender":"male"}
{"country":"China","firstName":"San","lastName":"Zhang"}	{"firstName":"Si","lastName":"Li","country":"China"}

使用真实数据集：IMDB

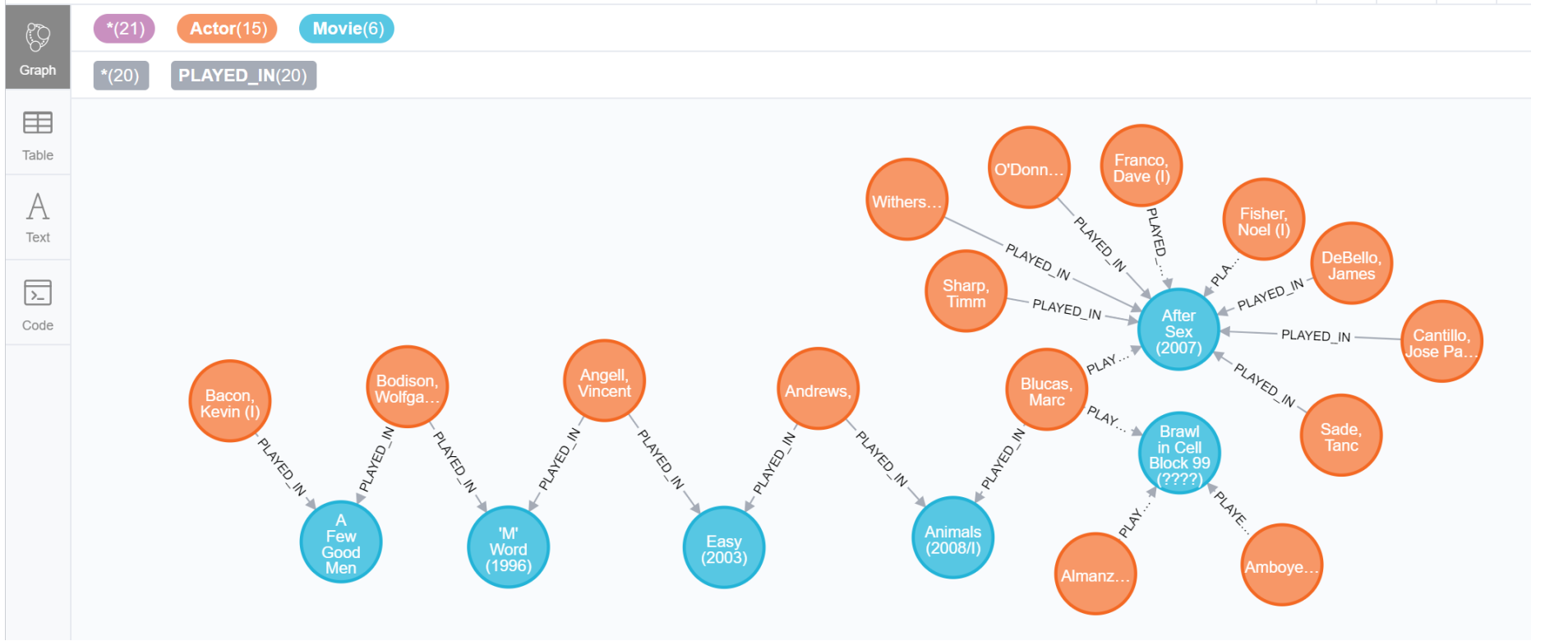
使用Neo4j导入IMDB电影知识图谱数据集actors.csv、movies.csv、roles.csv

查找“贝肯数”是5的10名演员，并输出由演员Kevin Bacon到这些演员的路径。

Cypher语句：

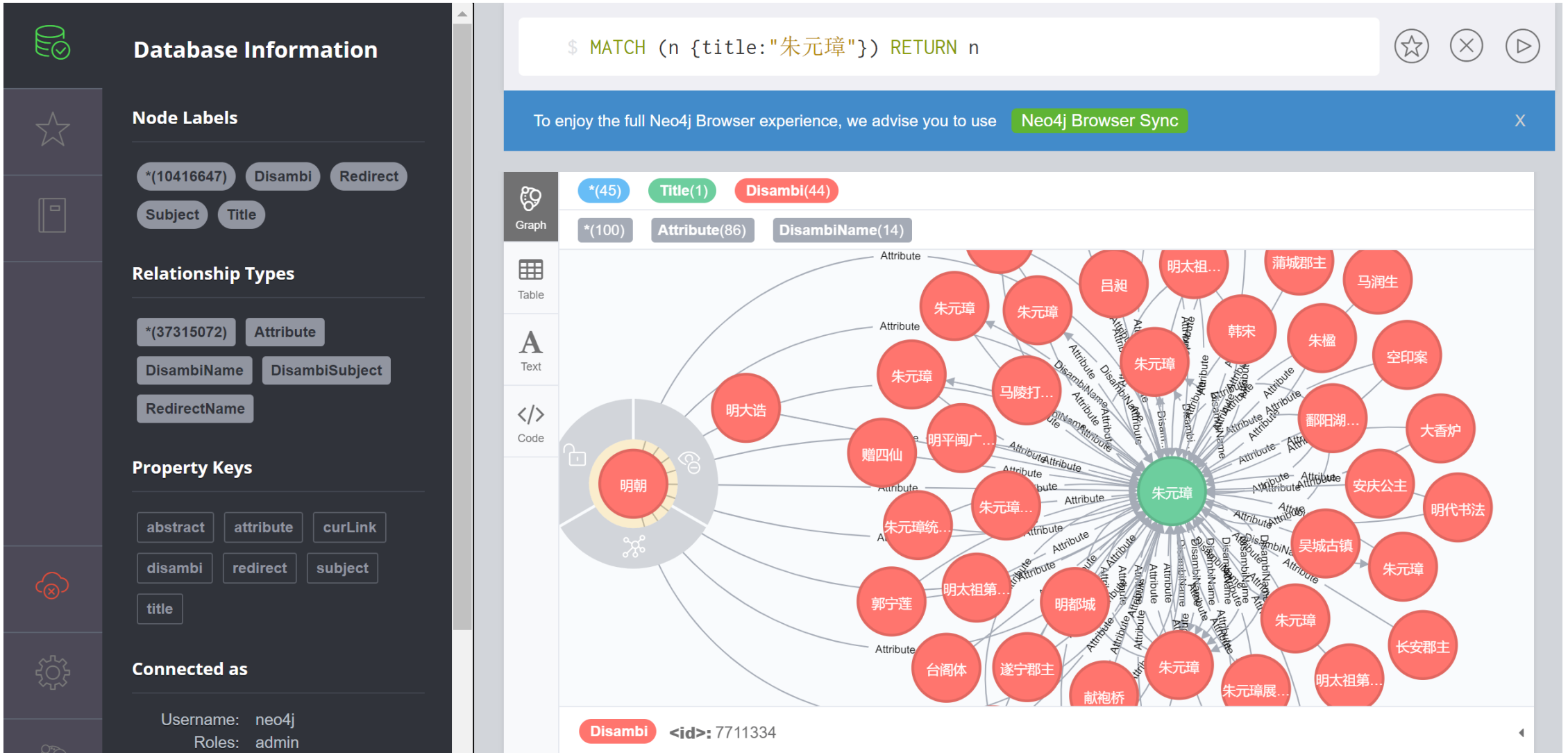
```
MATCH path=(:Actor{name:"Bacon, Kevin (I)"})-[:PLAYED_IN*10]-(:Actor)
RETURN path
LIMIT 10;
```

查询结果截图：

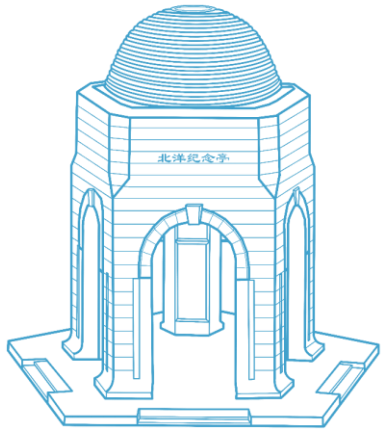


天津大学 Tianjin University

■ 使用真实数据集：百度百科爬取数据



数字人文知识图谱构想

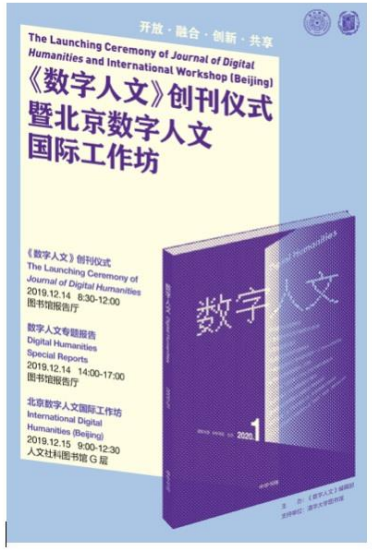


5 关于数字人文的思考

数字人文 Digital Humanities

数字人文，有时也被称为人文计算，它是针对计算与人文学科之间的交叉领域进行学习、研究、发明以及创新的一门学科。

中文名	数字人文	被称为	人文计算
外文名	Digital Humanities（DH）	性　质	一门学科
		属　于	方法论



- 提供人文学科领域的研究中的长期存在的问题提供新的研究方法
- 提升了人文研究工作效率，拓宽了研究空间，更重要地是给人文研究注入了新的研究方法和研究范式

胡韧奋等|让AI一口气读完《四库全书》会怎样?

章黄国学
北京师范大学章太炎黄侃学术研究中心

分享到:

2019-08-21 14:00:51　　字号: A- A A+　　来源: 张黄国学

中文信息学报 » 2019, **Vol. 33** » Issue (11): 57-63

语言分析与计算　　本期目录 | 过刊浏览 | 高级检索

引用本文:

俞敬松,魏一,张永伟. 基于BERT的古文断句研究与应用[J]. 中文信息学报, 2019, 33(11): 57-63.
YU Jingsong, WEI Yi, ZHANG Yongwei. Automatic Ancient Chinese Texts Segmentation Based on BERT.

基于BERT的古文断句研究与应用

俞敬松[1],魏一[1],张永伟[2]

1.北京大学 软件与微电子学院,北京 100871;
2.中国社会科学院 语言研究所,北京 100732

5 古汉语文本语料来源

殆知阁古代文献 16000种 20万卷 20亿字 4.8GB

https://github.com/garychowcmu/daizhigev20

5 古汉语文本语料来源

chinese-poetry 5.5 万首唐诗、26 万首宋诗、2.1 万首宋词

https://github.com/chinese-poetry/chinese-poetry

数据集

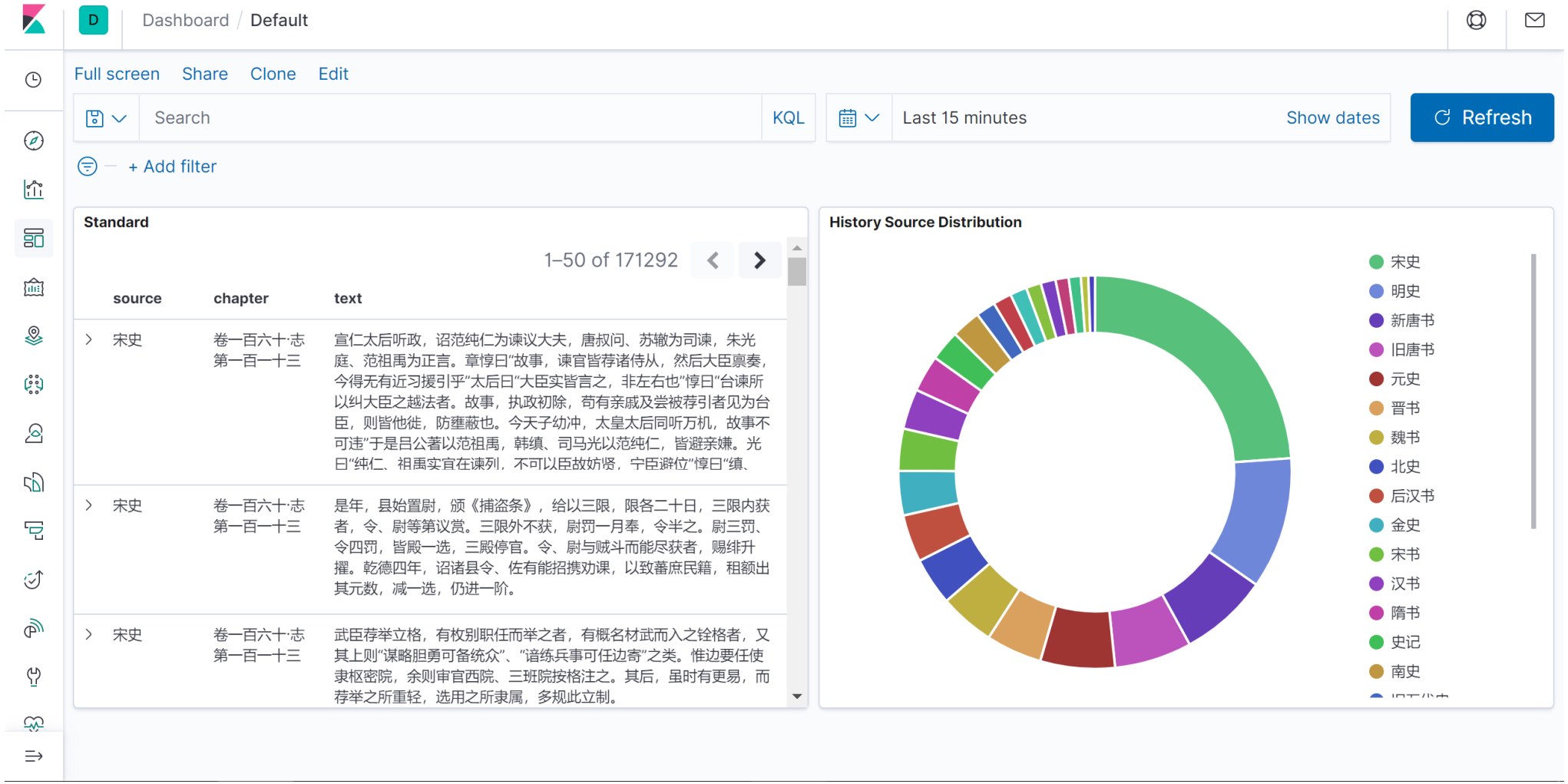
- 全唐诗 json
- 全宋诗 json
- 全宋词 ci
- 五代·花间集 wudai/huajianji
- 五代·南唐二主词 wudai/nantan
- 论语 lunyu
- 诗经 shijing
- 幽梦影 youmengying
- 四书五经 sishuwujing
- 蒙學 mengxue

```
[
    {
        "author": "太宗皇帝",
        "paragraphs": [
            "秦川雄帝宅，函谷壯皇居。",
            "綺殿千尋起，離宮百雉餘。",
            "連甍遙接漢，飛觀迴凌虛。",
            "雲日隱層闕，風煙出綺疎。"
        ],
        "title": "帝京篇十首 一",
        "id": "3ad6d468-7ff1-4a7b-8b24-a27d70d00ed4"
    },
    {
        "author": "太宗皇帝",
        "paragraphs": [
            "巖廊罷機務，崇文聊駐輦。",
            "玉匣啓龍圖，金繩披鳳篆。",
            "韋編斷仍續，縹帙舒還卷。",
            "對此乃淹留，欹案觀墳典。"
        ],
        "title": "帝京篇十首 二",
        "id": "13e72581-968b-457f-b381-a3b7d95b8b7c"
    },
```

天津大学 Tianjin University

基于ElasticSearch和Kibana的二十四史检索系统

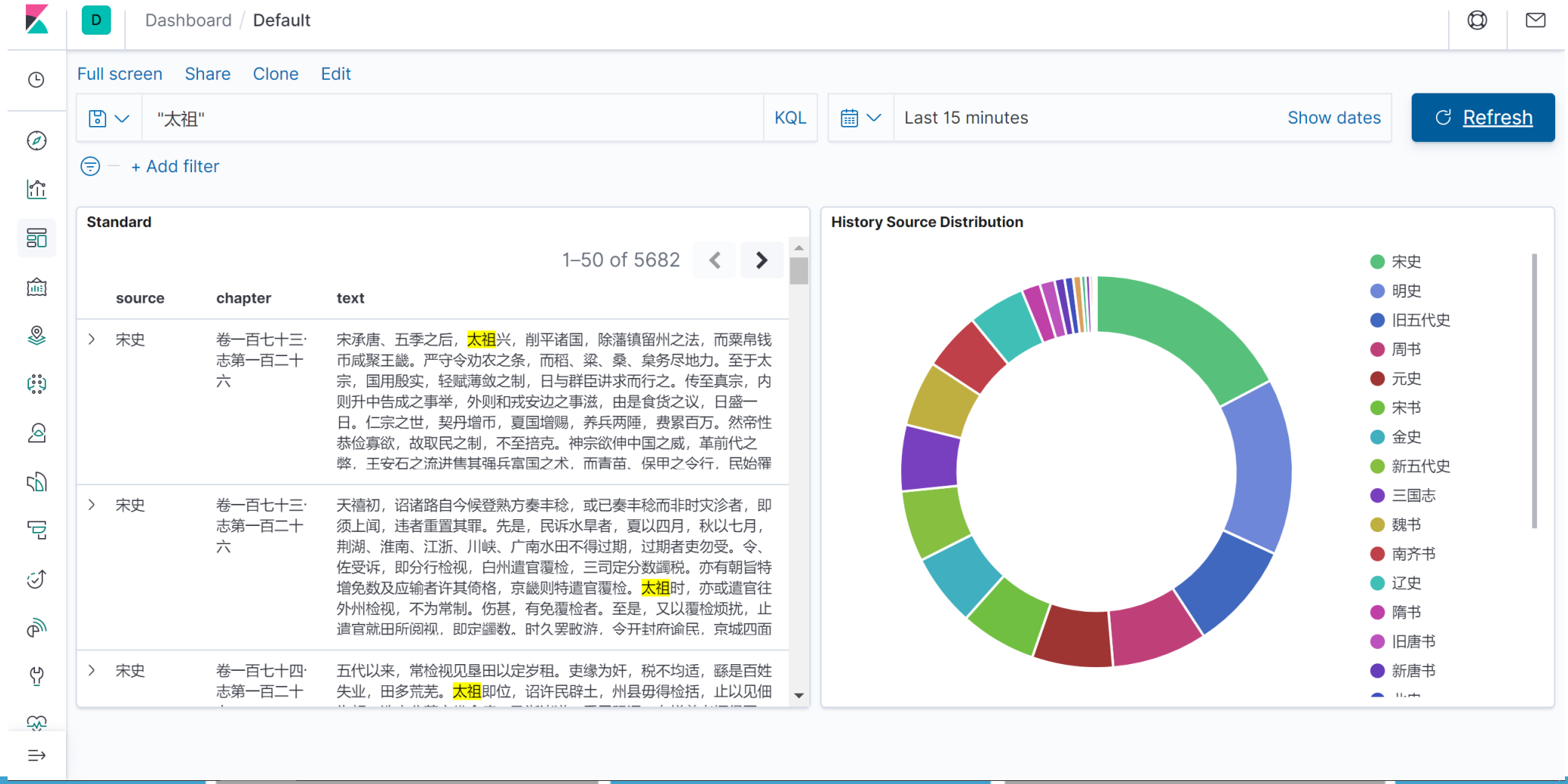
https://github.com/quzhi1/ChineseHistoricalSource



天津大学 Tianjin University

基于ElasticSearch和Kibana的二十四史检索系统

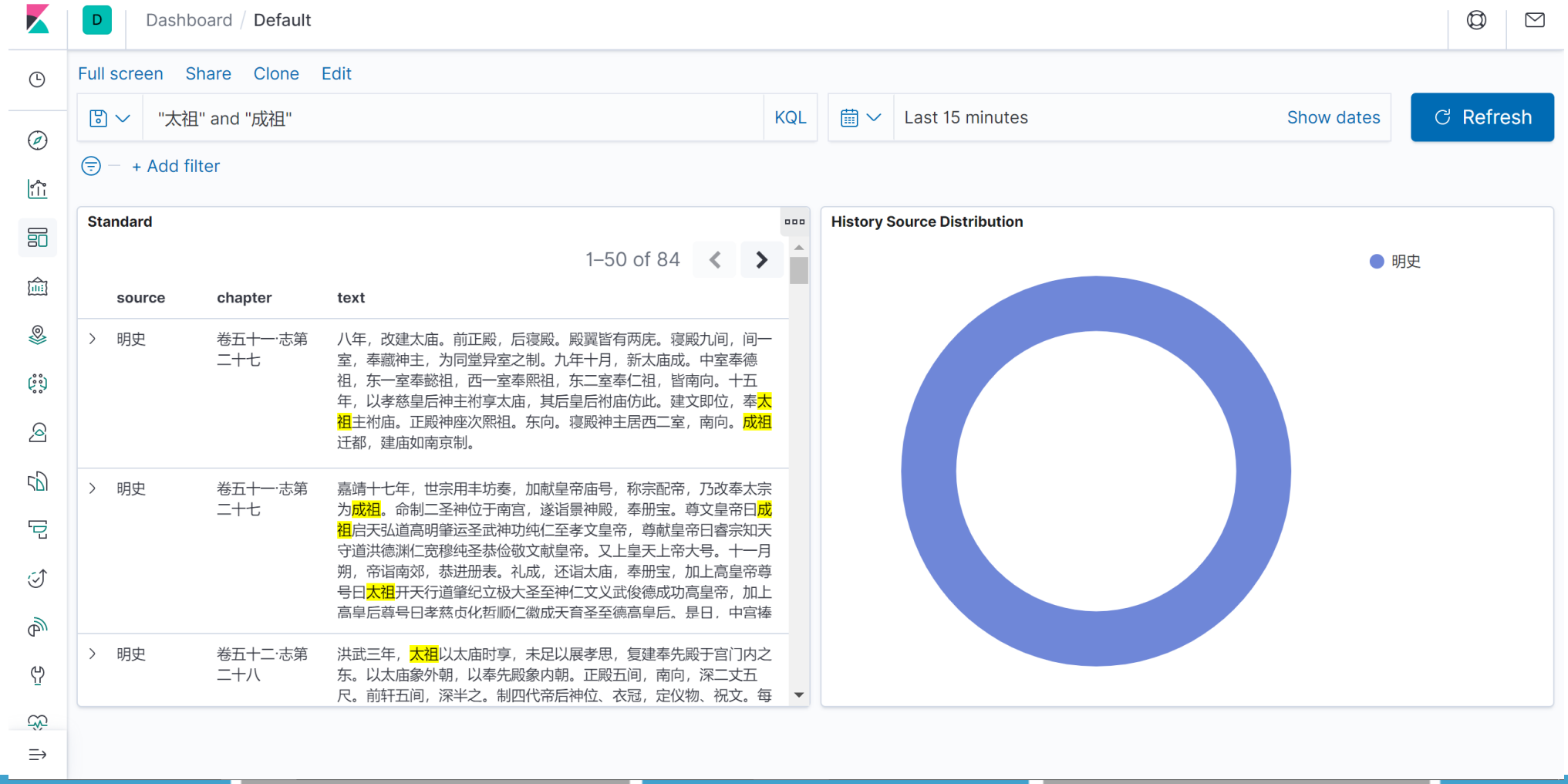
https://github.com/quzhi1/ChineseHistoricalSource



天津大学 Tianjin University

基于ElasticSearch和Kibana的二十四史检索系统

https://github.com/quzhi1/ChineseHistoricalSource



甲言 Jiayan 专注于古代汉语处理的NLP工具包

https://github.com/jiaeyan/Jiayan 支持文言词库构建、分词、词性标注、断句和标点

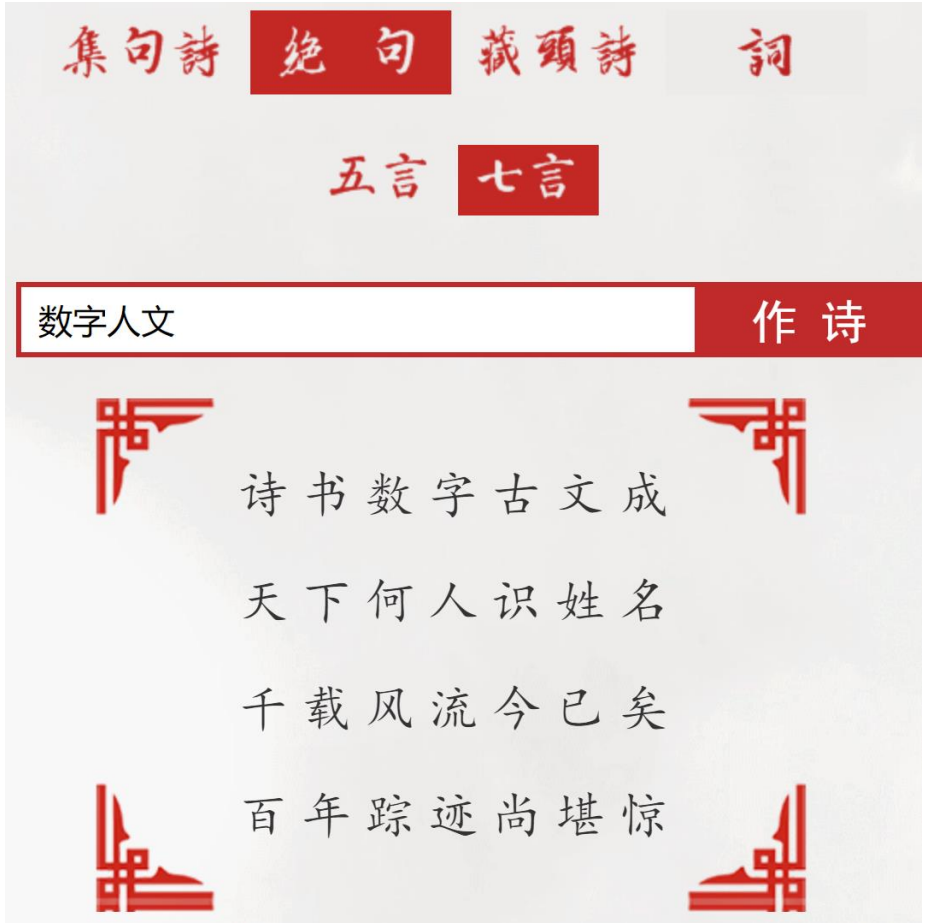
```
text = '十四年冬十月元丞相脱脱大败士诚于高邮分兵围六合'
lm = load_lm('jiayan.klm')
tokenizer = CharHMMTokenizer(lm)
words = list(tokenizer.tokenize(text))
print(words)
postagger = CRFPOSTagger()
postagger.load('pos_model')
print(postagger.postag(words))
```

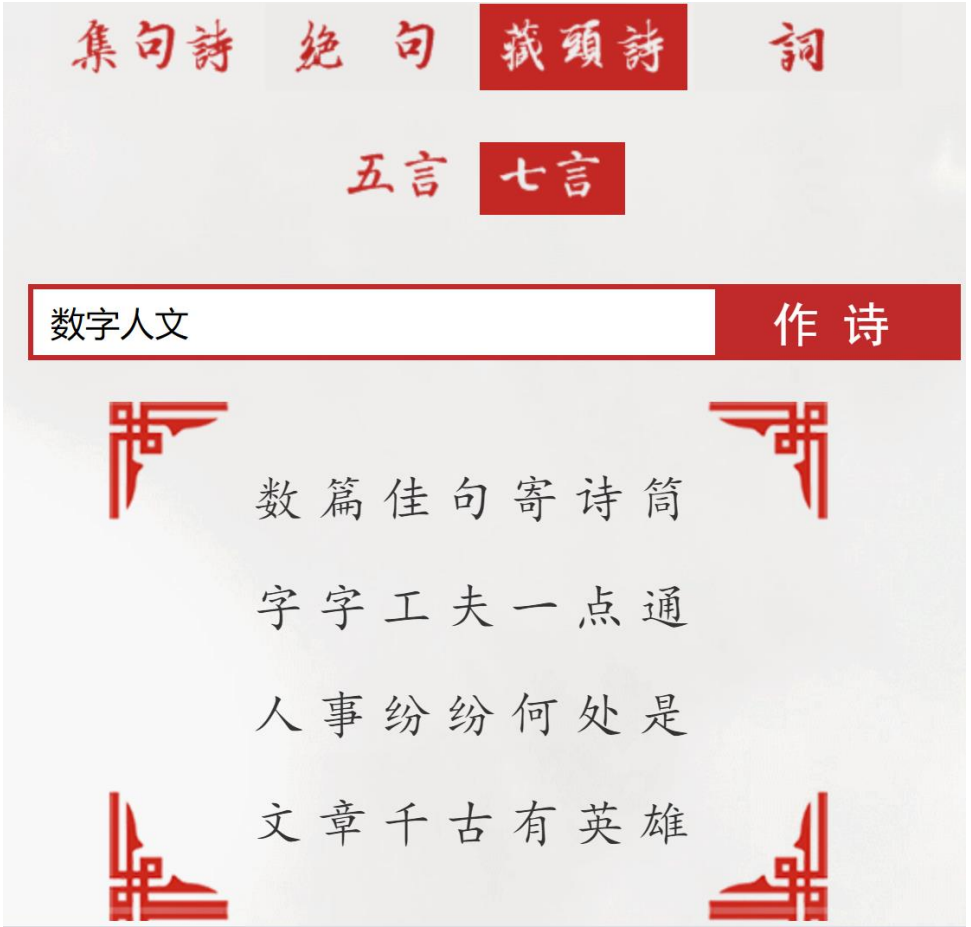
标记	词性
m	数字
n	一般名词
ns	地名
nt	时间
p	介词
q	量词
v	动词

分词	十四	年	冬十月	元	丞相	脱脱	大败	士诚	于	高邮	分兵	围	六合
结果	m	q	m	q	n	v	v	v	p	ns	n	v	v
正确	m	q	nt	n	n	n	v	n	p	ns	v	v	ns

5 NLP生成任务：诗词生成

- **九歌** THUNLP实验室
  https://github.com/THUNLP-AIPoet





未来研究构想

1. 古代汉语预训练语言模型 / BERT、XLNet
2. 历史文献知识图谱 / 二十四史知识图谱构建
3. 中医药知识图谱 / 医古文知识图谱构建
4. 数字人文知识图谱可视化 / 可视分析
5. 大规模数字人文知识图谱的数据集成与发布

5 知识图谱：方法、实践与应用

1. 知识图谱概述
2. 知识图谱表示与建模
3. **知识存储**
4. **知识抽取与知识挖掘**
5. 知识图谱融合
6. 知识图谱推理
7. 语义搜索
8. 知识问答
9. 知识图谱应用案例

[知识图谱：方法、实践与应用-实践工具Tutorial资源汇总](http://www.openkg.cn/dataset/kg-book)

谢谢大家