



# 面向中文的足球领域知识图谱

第三组

答辩人：李雨辰

小组成员：徐丹颖，黄振鹏，张静涵，  
李雨辰，杨喆，王兆阳

# 目 录



1

信息抽取（杨喆，徐丹颖，王兆阳，黄振鹏）

2

本体搭建（李雨辰，张静涵）

3

知识查询（杨喆，徐丹颖，王兆阳，黄振鹏，李雨辰）

4

可视化（张静涵）





Fact Extraction

信息抽取



# 半结构化数据

- 利用request+xpath经典方法进行wiki和百度百科的半结构化爬取



百度百科



维基百科

## 半结构化数据

```
{ '马德里竞技': { '全名': 'Club Atlético de Madrid',  
  '绰号': 'Los Colchoneros (床单军团) Los Rojiblancos (红白军团)',  
  '成立': '1903年4月26日, \u200b118年前\u200b (1903-04-26) [1]',  
  '城市': '西班牙马德里',  
  '主场': '万达大都会球场',  
  '容纳人数': '68,456[2]',  
  '拥有者': '米盖尔·安赫尔·希尔·马林 (英语: Miguel Ángel Gil Marín) (51%) 伊丹·奥佛 (英语: Idan Ofer) (30%) 恩里克·塞雷佐 (英语: Enrique Cerezo) (19%) [3][4][5]',  
  '主席': '恩里克·塞雷佐 (英语: Enrique Cerezo)',  
  '主教练': '迭戈·西蒙尼',  
  '联赛': '西班牙足球甲级联赛',  
  '2020-21': '西甲, 第 1 位'},  
  '皇家马德里': { '全名': 'Real Madrid Club de Fútbol皇家马德里足球俱乐部',  
    '简称': '皇马/银河舰队',  
    '绰号': 'Los Blancos (白色的) Los Merengues (蛋白甜饼) Los Vikings (维京人) Los Casa Blanca (白宫) Madridistas (马德里人)',  
    '成立': '1902年3月6日',  
    '城市': '西班牙, 马德里',  
    '主场': '圣地亚哥·伯纳乌球场',  
    '容纳人数': '85,454人',  
    '主席': '弗洛伦蒂诺·佩雷斯' }
```

提取infobox进行类型推断(type inference)



# 半结构化数据

- 定位球员个人百科

	number	name	country	birthday	location
0	1	马克-安德烈·特尔施特根	德国	1992-04-30	门将
1	2	塞尔吉尼奥·德斯特	美国	2000-11-03	后卫
2	3	杰拉德·皮克	西班牙	1987-02-02	后卫
3	4	罗纳德·阿劳霍	乌拉圭	1999-03-07	后卫
4	5	塞尔吉奥·布斯克茨	西班牙	1988-07-16	中场
5	6	里基·普奇	西班牙	1999-08-13	中场
6	7	奥斯曼·登贝莱	法国	1997-05-15	前锋
7	8	达尼·阿尔维斯	巴西	1983-05-06	后卫
8	9	孟菲斯·德佩	荷兰	1994-02-13	前锋
9	10	安苏·法蒂	西班牙	2002-10-31	前锋
10	11	阿达马·特拉奥雷	西班牙	1996-01-25	前锋
11	12	马丁·布雷思韦特	丹麦	1991-06-05	前锋
12	13	内托	巴西	1989-07-19	门将
13	14	尼科·冈萨雷斯	西班牙	2002-01-03	中场
14	15	克莱芒·朗格莱	法国	1995-06-17	后卫
15	16	佩德里	西班牙	2002-11-25	中场
16	17	卢克·德容	荷兰	1990-08-27	前锋

球员信息



# 非结构化数据

- 对维基百科的非结构化文本进行信息提取，以补充infobox中缺失的信息
- 百科在描述类似事物时常采用比较一致的表达方式，故采用正则表达式是比较高效的方式

## 正则表达式模板举例

### 成立日期

'成立于([0-9]{4}年)[,.]\*',  
'([0-9]{4}年)[成创立][,.]'

### 位置

'位于(西班牙\S{3,25})的\S\*俱乐部[,.]'

### 简称

'简称(\S{1,10})[,.]'

### 球队主场名称

r'主场场馆为(\S\*)球场',  
r'球[会队]主场([a-zA-Z\s]\*)',  
r',以(\S\s\*)作? 为主场'

### 球队主场容纳量

'容纳(\d\*,?\d\*)(人|观众)'

...





最终数据

subject	predicate	object
加的斯足球	成立时间	1910年
埃尔切足球	成立时间	1923年
塞维利亚足	成立时间	1905年
奥萨苏纳足	成立时间	1920年
巴塞罗那足	成立时间	1899年
格拉纳达足	成立时间	1931年
比利亚雷亚	成立时间	1923年
毕尔巴鄂竞	成立时间	1898年
瓦伦西亚足	成立时间	1919年
皇家比戈拿	成立时间	1923年
皇家社会足	成立时间	1909年
皇家西班牙	成立时间	1900年
皇家贝蒂斯	成立时间	1909年
皇家马德里	成立时间	1902年
皇家马略卡	成立时间	1916年
莱万特足球	成立时间	1909年
赫塔费足球	成立时间	1976年
阿拉维斯足	成立时间	1921年
马德里巴列	成立时间	1924年
马德里竞技	成立时间	1903年
加的斯足球	位置	西班牙安达鲁西亚自治区加的斯
埃尔切足球	位置	西班牙巴伦西亚自治区埃尔切
塞维利亚足	位置	西班牙安达卢西亚自治区首府塞维利亚
奥萨苏纳足	位置	西班牙纳瓦拉自治区首府潘普洛纳
巴塞罗那足	位置	西班牙巴塞罗那市
格拉纳达足	位置	西班牙安达鲁西亚自治区格拉纳达省省会格拉纳达
比利亚雷亚	位置	西班牙巴伦西亚自治区卡斯特利翁省比利亚雷亚尔市
毕尔巴鄂竞	位置	西班牙北部巴斯克自治区比斯开省毕尔巴鄂市

瓦伦西亚	主要竞争对手	比利亚雷亚尔
奥萨苏纳足	容纳	18375
巴塞罗那足	容纳	100000
皇家社会足	容纳	32076
皇家西班牙	容纳	40500
皇家贝蒂斯	容纳	52500
皇家马略卡	容纳	26500
赫塔费足球	容纳	17000
阿拉维斯足	容纳	19500
加的斯足球	球队主场	拉蒙·德卡兰萨球场
埃尔切足球	球队主场	曼努埃尔·马丁内斯·巴雷罗球场
塞维利亚足	球队主场	拉蒙·桑切斯·皮斯胡安球场
奥萨苏纳足	球队主场	为埃尔·萨达尔球场
巴塞罗那足	球队主场	诺坎普球场
格拉纳达足	球队主场	卡梅内斯球场
比利亚雷亚	球队主场	陶瓷球场
毕尔巴鄂竞	球队主场	新圣马梅斯球场
瓦伦西亚足	球队主场	梅斯塔利亚球场
皇家比戈拿	球队主场	巴莱多斯球场
皇家社会足	球队主场	阿诺埃塔球场
皇家西班牙	球队主场	埃尔普拉特球场
皇家贝蒂斯	球队主场	为洛佩拉球场
皇家马德里	球队主场	伯纳乌球场
皇家马略卡	球队主场	伊比利亚之星体育场
莱万特足球	球队主场	瓦伦西亚城市球场

球队信息





# 知识对齐

- 百度与维基名称通过球员号码作为标识符进行中英文名称对齐

球员列表				
号码	姓名	国籍	生日	场上位置
1	马库斯·安德烈·特施特根	德国	1992-04-30	门将
2	塞尔吉·诺伊施	美国	2000-11-03	后卫
3	迭戈·洛佩斯	西班牙	1987-02-02	后卫
4	多纳特·阿普雷	乌拉圭	1999-03-07	后卫
5	塞尔吉奥·布斯克茨	西班牙	1988-07-16	中场
6	基基·博斯	西班牙	1989-08-13	中场
7	奥斯曼·登贝莱	法国	1997-05-15	前锋
8	达尼·阿尔维斯	巴西	1983-05-06	后卫
9	盖雷斯·贝尔	荷兰	1994-02-13	前锋
10	安苏·法蒂	西班牙	2002-10-31	前锋
11	阿达马·特拉奥雷	西班牙	1996-01-25	前锋
12	马丁·布雷斯韦特	丹麦	1991-08-05	前锋
13	内托	巴西	1989-07-19	门将
14	厄科·利萨雷斯	西班牙	2002-01-03	中场

百度百科

号码	国籍	姓名	位置	出生日期	加盟年份	转会费	转会费
1	德国	马库斯·安德烈·特施特根 (Marc-André ter Stegen)	门将	1992年4月30日 (29岁)	2014年	门兴格拉德巴赫	1,200万欧元
12	美国	塞尔吉·诺伊施 (Sergiño Dest)	后卫	2000年11月03日 (23岁)	2019年	芝加哥火焰	3,500万欧元
2	西班牙	迭戈·洛佩斯 (Diego López)	后卫	1987年02月02日 (35岁)	2009年	皇家马德里	2,400万欧元
3	乌拉圭	多纳特·阿普雷 (Donat Arap)	后卫	1999年03月07日 (23岁)	2020年	巴塞罗那	500万欧元
4	法国	奥斯曼·登贝莱 (Ousmane Dembélé)	前锋	1997年05月15日 (25岁)	2017年	巴黎圣日耳曼	1,500万欧元
5	西班牙	塞尔吉奥·布斯克茨 (Sergio Busquets)	中场	1988年07月16日 (33岁)	2008年	巴塞罗那	500万欧元
6	西班牙	基基·博斯 (Kike Bosch)	中场	1989年08月13日 (31岁)	2014年	巴塞罗那	500万欧元
7	法国	奥斯曼·登贝莱 (Ousmane Dembélé)	前锋	1997年05月15日 (25岁)	2017年	巴黎圣日耳曼	1,500万欧元
8	巴西	达尼·阿尔维斯 (Dani Alves)	后卫	1983年05月06日 (38岁)	2012年	巴塞罗那	500万欧元
9	荷兰	盖雷斯·贝尔 (Gareth Bale)	前锋	1994年02月13日 (28岁)	2010年	热刺	3,500万欧元
10	西班牙	安苏·法蒂 (Ansu Fati)	前锋	2002年10月31日 (18岁)	2019年	巴塞罗那	500万欧元
11	西班牙	阿达马·特拉奥雷 (Adama Traoré)	前锋	1996年01月25日 (24岁)	2017年	巴塞罗那	500万欧元
12	丹麦	马丁·布雷斯韦特 (Martin Brunsby)	前锋	1991年08月05日 (29岁)	2019年	巴塞罗那	500万欧元
13	巴西	内托 (Neto)	门将	1989年07月19日 (30岁)	2014年	巴塞罗那	500万欧元
14	西班牙	厄科·利萨雷斯 (Eco Lizarazu)	中场	2002年01月03日 (17岁)	2022年	巴塞罗那	500万欧元

维基百科



# 知识对齐

- 百度与领域专业知识库借助百度翻译API与difflib字符串相似度匹配模块进行中英文名称对齐

球员列表				
序号	姓名	国籍	生日	场上位置
1	马库斯·安德烈·特纳特	德国	1992-04-30	门将
2	塞尔吉·格诺特	美国	2000-11-03	后卫
3	迭戈·阿隆索	西班牙	1997-02-02	后卫
4	多明戈·阿隆索	乌拉圭	1999-03-07	后卫
5	塞尔吉·格诺特	西班牙	1998-07-16	中场
6	塞尔吉·格诺特	西班牙	1999-08-13	中场
7	奥斯曼·登贝莱	法国	1997-05-15	前锋
8	达尼·阿尔瓦雷斯	巴西	1993-05-06	后卫
9	塞尔吉·格诺特	荷兰	1994-02-13	前锋
10	安苏·法蒂	西班牙	2002-10-31	前锋
11	阿达马·特雷泽盖	西班牙	1996-01-25	前锋
12	马丁·布雷斯韦特	丹麦	1991-08-05	前锋
13	内托	巴西	1989-07-19	门将
14	厄科·利萨雷斯	西班牙	2002-01-03	中场

百度百科

Standard Stats 2021-2022 Barcelona: La Liga																															
Share & Export <span>▼</span> Glossary <span>▼</span> Toggle Per 90 Stats													Scroll Right For More Stats <span>▼</span> Switch to Widescreen View <span>▼</span>																		
Player		Nation	Pos	Age	Playing Time			Performance						Per 90 Minutes					Expected					Per 90 Minutes							
					MP	Starts	Min.	90s	Gls	Ass	G-PK	PK	Pts	Att	CrdY	CrdR	Gls	Ass	G+A	G-PK	G+A+PK	xG	xG-PK	xA	xG+PK	xG+A+PK	ppg	ppg-PK	ppg xA	ppg+PK	ppg+xA
	Terence Kongolo	ESP	MF	33-200	28	28	2,478	27.3	1	0	1	0	0	0	0.04	0.00	0.04	0.04	0.04	0.0	0.0	1.7	2.3	0.02	0.00	0.00	0.02	0.02			
	Marcelino García Blázquez	ESP	DF	34-340	27	27	2,430	27.0	0	0	0	0	0	0	0.00	0.00	0.00	0.00	0.00	0.0	0.0	0.0	0.0	0.00	0.00	0.00	0.00	0.00			
	David Pérez	ESP	DF	35-067	25	25	1,584	22.0	1	0	1	0	0	0	0.05	0.00	0.05	0.05	0.05	2.1	2.1	0.1	2.1	0.04	0.00	0.10	0.04	0.04			
	Jordi Alba	ESP	DF	33-009	22	22	1,824	21.4	1	0	1	0	0	0	0.05	0.00	0.05	0.05	0.05	3.0	3.0	0.1	3.0	0.04	0.00	0.04	0.04	0.04			
	Antonio G. Ruiz	MEX	MF	24-335	24	22	1,833	20.4	3	1	3	0	0	0	0.15	0.15	0.29	0.15	0.29	4.3	4.3	0.6	4.9	0.21	0.13	0.34	0.21	0.21			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo	ESP	MF	37-046	25	25	1,584	17.7	3	4	3	0	0	0	0.11	0.00	0.04	0.11	0.34	3.7	3.7	0.7	4.4	0.13	0.13	0.33	0.13	0.13			
	David Pardo																														

## 最终数据

	球衣号码	姓名	国籍	生日	职位	所属球队	英文名
0	1	霍埃尔·罗夫莱斯	西班牙	1990.6.17	门将	皇家贝蒂斯足球俱乐部球员	Rodrigo
1	2	马丁·蒙托亚	西班牙	1991.4.14	后卫	球员	Martin Montoya
2	3	埃德加·冈萨雷斯	西班牙	1997.4.1	后卫	皇家贝蒂斯足球俱乐部	Edgar Gonzalez
3	4	保罗·阿库奥库	科特迪瓦	1997.12.20	中场	NaN	Nabil Fekir
4	5	马克·巴特拉	西班牙	1991.1.15	后卫	NaN	Mark batra
5	6	比克托·鲁伊斯	西班牙	1989.1.25	后卫	NaN	Bictor Ruiz
6	7	胡安米	西班牙	1993.5.20	前锋	NaN	Willian José
7	8	纳比勒·费基尔	法国	1993.7.18	中场	NaN	Aitor Ruibal
8	9	博尔哈·伊格莱西亚斯	西班牙	1993.1.17	前锋	NaN	Andrés Guardado
9	10	塞尔希奥·卡纳莱斯	西班牙	1991.2.16	中场	NaN	Youssef Sabaly
10	11	克里斯蒂安·特略	西班牙	1991.8.11	前锋	NaN	Kike Hermoso
11	12	威廉·若泽	巴西	1991.11.23	前锋	NaN	Juanmi
12	13	鲁伊·席尔瓦	葡萄牙	1994.2.7	门将	NaN	Marc Bartra
13	14	威廉·卡瓦略	葡萄牙	1992.4.7	中场	NaN	William Carvalho
14	15	亚历克斯·莫雷诺	西班牙	1993.6.8	后卫	NaN	Alex Moreno
15	16	赫尔曼·佩泽拉	阿根廷	1991.6.27	后卫	NaN	Herman pezela
16	17	华金	西班牙	1981.7.21	中场	NaN	Hua Jin
17	18	安德烈斯·瓜尔达多	墨西哥	1986.9.28	中场	NaN	Cristian Tello
18	19	埃克托·贝列林	西班牙	1995.3.19	后卫	NaN	Ektor berelin
19	20	迭戈·莱内斯	墨西哥	2000.6.9	前锋	NaN	Diego Raines
20	21	吉多·罗德里格斯	阿根廷	1994.4.12	中场	NaN	Guido Rodriguez

## 球员信息





# Ontology Building

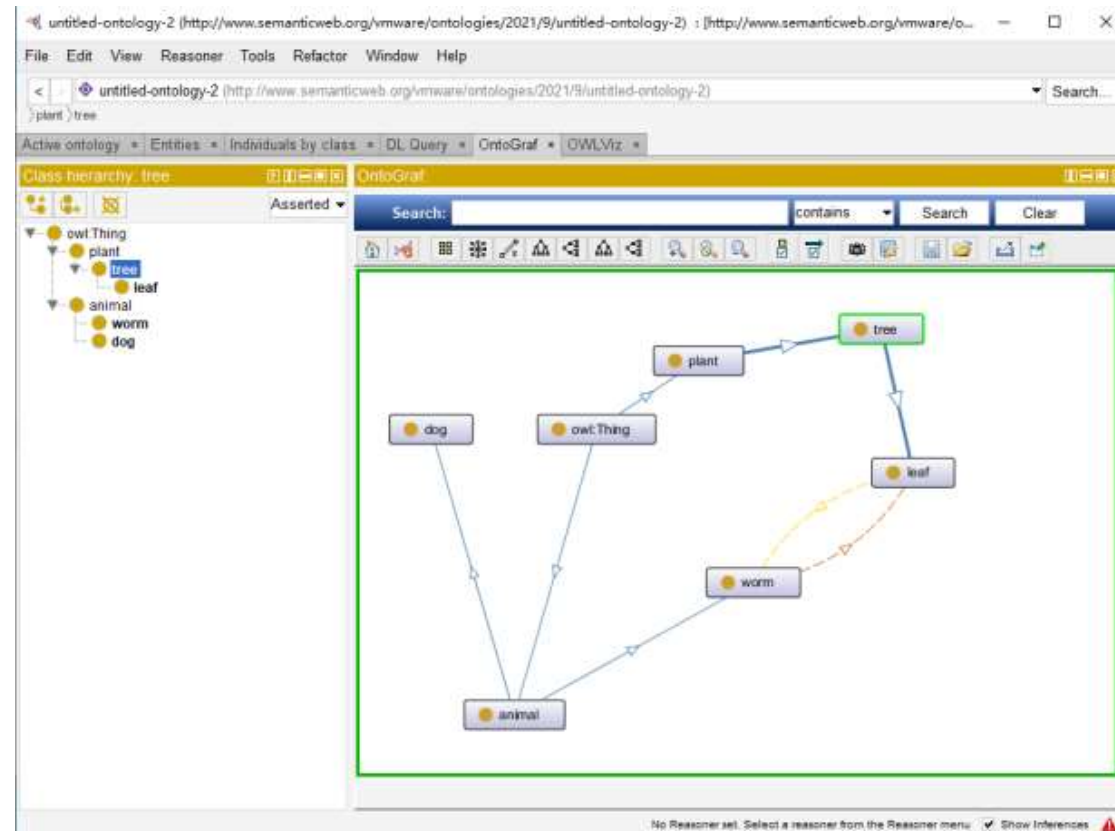
本体搭建



# Protégé



- 斯坦福大学医学院生物信息研究中心基于Java语言开发的本体编辑和知识获取软件
- 本体开发工具，基于知识的编辑器，属于开放源代码软件
- 主要用于语义网中本体的构建，是语义网中本体构建的核心开发工具
- 提供了本体概念类，关系，属性和实例的构建



# Excel数据导入Protégé

Excel数据

奥萨苏纳足球俱乐部							
	A	B	C	D	E	F	G
1	球衣号码	name	国籍	生日	职位	所属球队	英文名
2	1	塞尔希奥·埃雷拉	西班牙	1993.6.5	门将	奥萨苏纳足球俱乐部球员	Sergio Herrera
3	2	纳乔·比达尔	西班牙	1995.1.24	后卫	球员	Roberto Torres
4	3	胡安·克鲁斯	西班牙	1992.7.28	后卫	奥萨苏纳足球俱乐部	Darko Brašanac
5	4	乌奈·加西亚	西班牙	1992.9.3	后卫		Kike Barja
6	5	戴维·加西亚	西班牙	1994.2.14	后卫		Unai Dufur
7	6	奥耶尔·圣胡尔霍	西班牙	1986.5.25	中场		José Ángel
8	7	霍恩·蒙卡约拉	西班牙	1998.5.13	中场		Horn moncajola
9	8	达尔科·布拉沙纳茨	塞尔维亚	1992.2.12	中场		Javier Ontiveros
10	9	奇米·阿维拉	阿根廷	1994.2.6	前锋		Chimi Avila
11	10	罗伯托·托雷斯	西班牙	1989.3.7	中场		Roberto Torres
12	11	基克·巴尔哈	西班牙	1997.4.4	前锋		Juan Cruz Armada
13	12	豪梅·格劳	西班牙	1997.5.5	中场		Nacho Vidal
14	13	胡安·佩雷斯	西班牙	1996.7.15	门将		Aridane Hernández
15	14	鲁文·加西亚	西班牙	1993.7.14	前锋		Reuven Garcia
16	15	若纳斯·拉马略	安哥拉	1993.6.10	后卫		Unai García
17	16	科特	西班牙	1989.9.5	后卫		Juan Pérez

Generated Axioms

Cellfie generates 218 axioms:

- Individual: 阿里达内·埃尔南德斯
- Individual: 霍恩·蒙卡约拉
- Individual: 马努·桑切斯
- Individual: 鲁文·加西亚
- 奥萨苏纳足球俱乐部球员 SubClassOf 球员
- 乌奈·加西亚 Type 奥萨苏纳足球俱乐部球员
- 伊尼戈·佩雷斯 Type 奥萨苏纳足球俱乐部球员
- 卢卡斯·托罗 Type 奥萨苏纳足球俱乐部球员
- 哈维尔·蒙托亚 Type 奥萨苏纳足球俱乐部球员

[View Log](#)

Cancel Add to a new ontology Add to current ontology



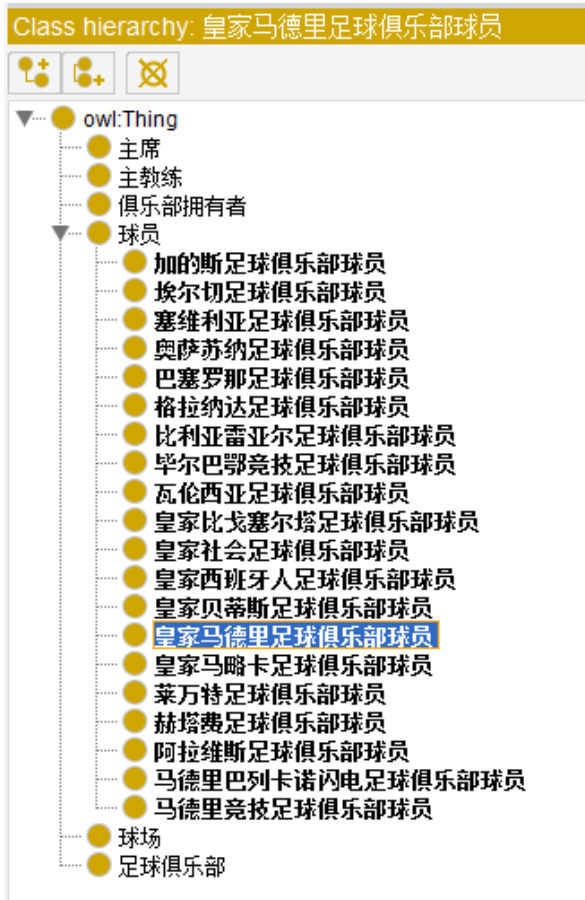
创建规则  
保存为json文件

✓	Sheet Name	Start Column	End Column	Start Row	End Row	Rule
✓	奥萨苏纳足球俱乐部	A	G	1	+	Class:@F2 SubClassOf:@F3
✓	奥萨苏纳足球俱乐部	A	G	2	+	Individual:@B* Types:@F2 Facts:@A1 @A* (xsd:integer) Facts:@C1 @C* Facts:@D1 @D* Facts:@E1 @E* Facts:@G1 @G* Facts:@F1(ObjectProperty) @F4

- 阿拉维斯足球俱乐部
- 埃尔切足球俱乐部
- 奥萨苏纳足球俱乐部
- 巴塞罗那足球俱乐部
- 比利亚雷亚尔足球俱乐部
- 毕尔巴鄂竞技足球俱乐部
- 格拉纳达足球俱乐部



# Protégé构建知识图谱



本体分类



Description: 皇家马德里足球俱乐部球员

Equivalent To +

SubClass Of +

- 球员

General class axioms +

SubClass Of (Anonymous Ancestor)

Instances +

- 伊斯科
- 加雷斯·贝尔
- 卡塞米罗
- 卡里姆·本泽马
- 卢卡·约维奇
- 卢卡·莫德里奇
- 卢卡斯·巴斯克斯
- 埃当·阿扎尔
- 埃德尔·米利唐
- 安德里·卢宁
- 托尼·克罗斯
- 爱德华多·卡马温加
- 米格尔·古铁雷斯
- 纳乔·费尔南德斯
- 维尼修斯·儒尼奥尔
- 罗德里戈
- 蒂博·库尔图瓦
- 费兰·门迪
- 费德里科·巴尔韦德

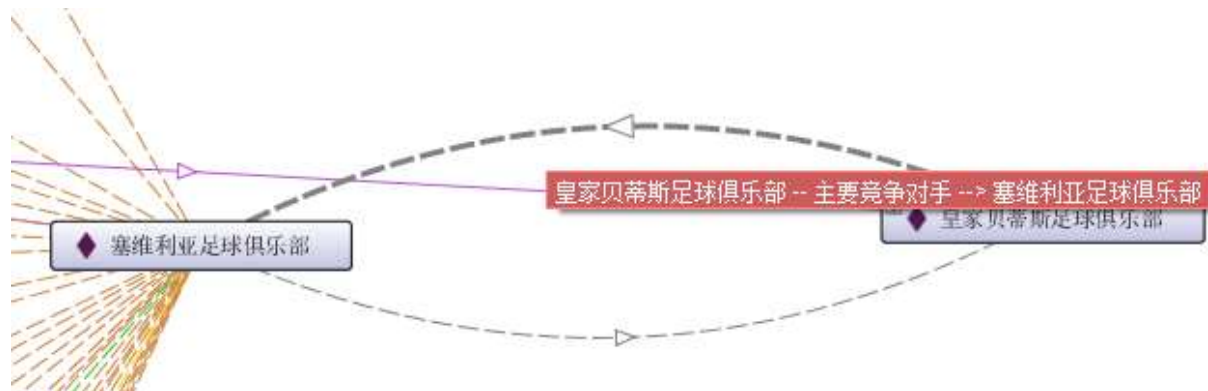
创建实例



# Protégé构建知识图谱



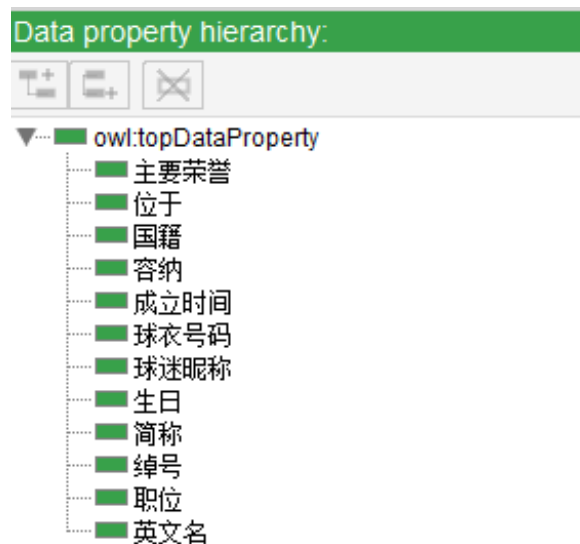
关系创建



关系实例



# Protégé构建知识图谱



数据属性



数据实例





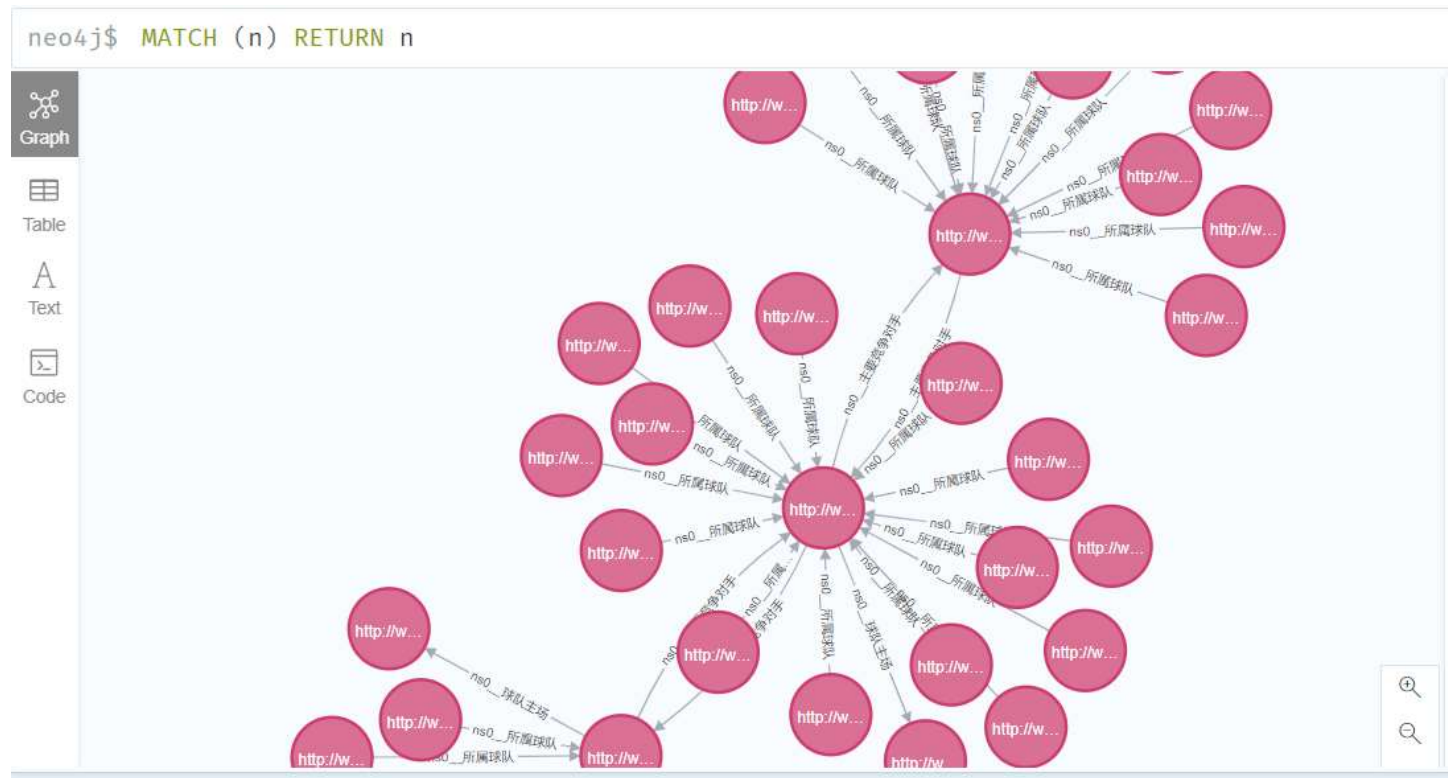
Knowledge Querying

知识查询

# Neo4j图数据库




- 高性能图形数据库
- 将结构化数据存储在网络上而不是表中
- 一个嵌入式的、基于磁盘的、具备完全的事务特性的Java持久化引擎
- 一个高性能的图引擎，具有成熟数据库的所有特性



# 图谱数据文件转换

owl格式图谱数据



 rdf2rdf-1.0.1-2.3.1

Supported extensions	
rdf, rdfs, owl, xml	RDF/XML
nt	N-Triples
ttl	Turtle
n3	N3
trig, xml	TriX
trig	TriG



rdf格式图谱数据

```
<!-- http://www.semanticweb.org/think/ontologies/Football.owl#主要竞争对手 -->
```

```
<owl:ObjectProperty rdf:about="http://www.semanticweb.org/think/ontologies/Football.owl#主要竞争对手"/>
```

```
<!-- http://www.semanticweb.org/think/ontologies/Football.owl#就任主席 -->
```

```
<owl:ObjectProperty rdf:about="http://www.semanticweb.org/think/ontologies/Football.owl#就任主席"/>
```

```
<!-- http://www.semanticweb.org/think/ontologies/Football.owl#所属球队 -->
```

```
<owl:ObjectProperty rdf:about="http://www.semanticweb.org/think/ontologies/Football.owl#所属球队"/>
```

```
<!-- http://www.semanticweb.org/think/ontologies/Football.owl#执教 -->
```



```
<rdf:Description rdf:about="http://www.semanticweb.org/think/ontologies/Football.owl#主要竞争对手">  
  <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#ObjectProperty"/>  
</rdf:Description>
```

```
<rdf:Description rdf:about="http://www.semanticweb.org/think/ontologies/Football.owl#就任主席">  
  <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#ObjectProperty"/>  
</rdf:Description>
```

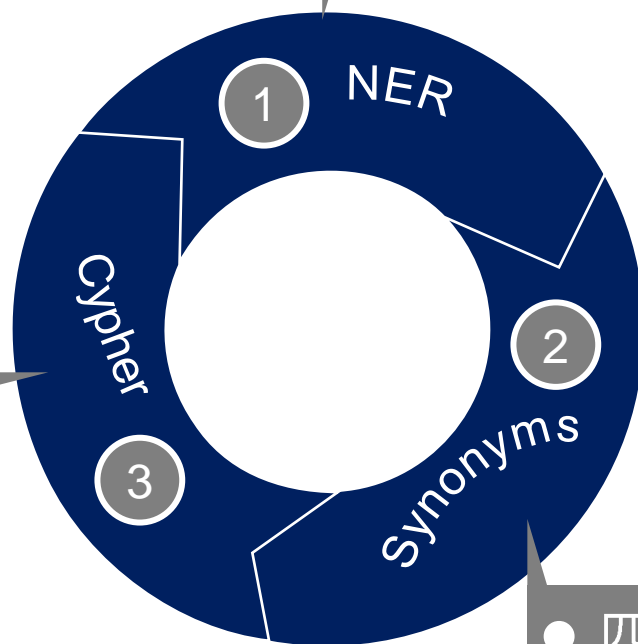
```
<rdf:Description rdf:about="http://www.semanticweb.org/think/ontologies/Football.owl#所属球队">  
  <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#ObjectProperty"/>  
</rdf:Description>
```



# 自然语言查询

- 将命名实体识别（NER）和 Neo4j查询语句（Cypher）结合
- 实现自然语言转换知识图谱查询功能

- 匹配问题转Cypher查询模板



- BiLSTM+CRF模型

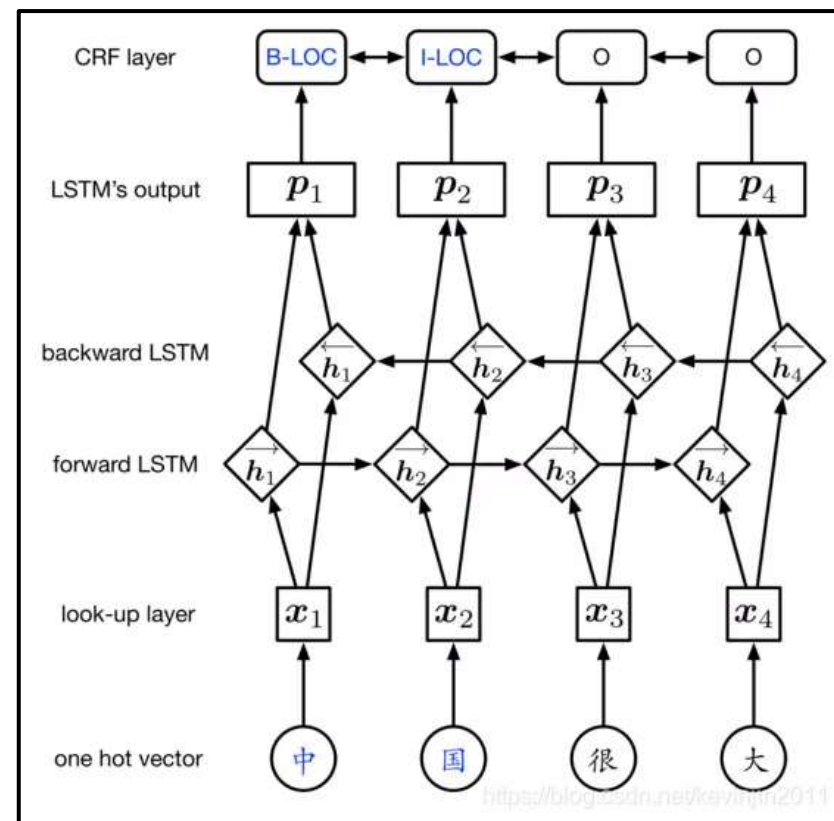
- 匹配查询实体与问题





## BiLSTM+CRF模型

- BiLSTM-CRF输入是词向量，输出每个单词预测的序列标注
- 第一步：单词输入，单词进入look-up layer层，使用CBOW、Skip-gram或者glove模型映射为词向量
- 第二步：词向量进入BiLSTM层，通过学习上下文的信息，输出每个单词对应于每个标签的得分概率
- 第三步：所有的BiLSTM的输出将作为CRF层的输入，通过学习标签之间的顺序依赖信息，得到最终的预测结果



BiLSTM+CRF模型



# 自然语言查询

- Synonyms

- 用于自然语言理解的很多任务：文本对齐，推荐算法，相似度计算，语义偏移，关键字提取，概念提取，自动摘要，搜索引擎等
- 对进行了命名实体识别后的结果与问题模板的关键词进行匹配

词语	2016词林改进版	知网	Synonyms	人工标准
"轿车", "汽车"	0.82	1.0	0.73	0.98
"宝石", "宝物"	0.83	0.17	0.71	0.96
"旅游", "游历"	1.0	1.0	0.59	0.96
"男孩子", "小伙子"	0.81	1.0	0.88	0.94
"海岸", "海滨"	0.94	1.0	0.68	0.93
"庇护所", "精神病院"	0.96	0.58	0.64	0.90
"魔术师", "巫师"	0.85	0.58	0.66	0.88
"中午", "正午"	1.0	1.0	0.81	0.86
"火炉", "炉灶"	0.98	0.58	0.85	0.78

Synonyms

```
1 import synonyms
2 # print('国籍: ', synonyms.nearby("国籍"))
3 print(synonyms.compare('简称', '绰号'))
```

0.197



```
1 get_entity_info('马略卡的昵称是什么')
MATCH (a:entity{name:"马略卡"})
RETURN a.绰号

'MATCH (a:entity{name:"马略卡"})\nRETURN a.绰号'
```

同义词对齐



# 自然语言分词

```
In [5]: 1 text = u'西甲的第一名是谁'
        2 lac_result = lac.run(text)
        3 lac_result
```

```
Out[5]: [['西甲', '的', '第一名', '是', '谁'], ['nz', 'u', 'm', 'v', 'r'], [3, 0, 2, 0, 1]]
```

```
In [6]: 1 texts = [u'马德里竞技俱乐部在西甲里面排名多少', '皇家贝蒂斯足球俱乐部的排名和皇家马德里俱乐部的排名那个比较好']
        2 lac_results = lac.run(texts)
        3 lac_results
```

```
Out[6]: [['马德里竞技俱乐部', '在', '西甲', '里面', '排名', '多少'],
          ['ORG', 'p', 'nz', 'f', 'v', 'r'],
          [3, 0, 3, 1, 2, 1]],
          [['皇家贝蒂斯足球俱乐部', '的', '排名', '和', '皇家马德里俱乐部', '的', '排名', '那个', '比较好'],
          ['ORG', 'u', 'vn', 'c', 'ORG', 'u', 'vn', 'r', 'a'],
          [2, 0, 2, 0, 3, 0, 2, 1, 2]]]
```

```
In [7]: 1 text = '塞维利亚俱乐部的出生年月是什么时候?'
        2 lac_result = lac.run(text)
        3 lac_result
```

```
Out[7]: [['塞维利亚俱乐部', '的', '出生', '年月', '是', '什么时候', '?'],
          ['ORG', 'u', 'vn', 'n', 'v', 'n', 'w'],
          [3, 0, 2, 2, 0, 2, 0]]
```



## 自然语言转Cypher查询模板

- 查询一个entity的所有信息：（例：甲的出生年月？）

```
MATCH (a:entity{name:' 甲 '})
```

```
RETURN a.property
```

- 查询与a有对应关系的b：（例：甲隶属于哪个球队？）

```
MATCH (a:entity{name: '甲' }) - [:relation]-(b)
```

```
RETURN b
```

- 与entity有关的某一类关系值：（甲球队的所有球员？）

```
MATCH (a:entity{name:甲}) - [r:relation]-(b)
```

```
RETURN r
```

- 查询a与b的关系：（甲和乙是什么关系？）

```
MATCH (a:entity1{name: '甲' }) - [r:relation]-
```

```
(b:entity2{name: '乙' })
```

```
RETURN r
```



# 自然语言转Cypher语句

- 查询一个entity的所有信息
- 例：皇家马略卡足球俱乐部的昵称是什么？

```
1 def get_entity_info(sentence):
2     lac_res = lac.run(sentence)
3     # print(lac_res)
4     max_sim = -1
5     max_n = ''
6     # print(list(enumerate(lac_res[2])))
7     for index, i in enumerate(lac_res[2]):
8         if i > 0:
9             # print(index)
10            tmp_n, tmp_sim = get_closest_prop(lac_res[0][index])
11            if tmp_sim > max_sim:
12                max_sim = tmp_sim
13                max_n = tmp_n
14    entity = ''
15    for index, i in enumerate(lac_res[1]):
16        if i == 'PER' or i == 'LOC' or i == 'ORG':
17            entity = lac_res[0][index]
18    # print(entity, max_n)
19    SPARQL_SEN = 'MATCH (a:`ns0__足球俱乐部` {uri:"http://www.semanticweb.org/think/ontologies/Football.owl#' + entity + '"})\nRETURN a.ns0__绰号'
20    print(SPARQL_SEN)
21    return SPARQL_SEN
```

```
1
2 ql=get_entity_info('皇家马略卡足球俱乐部的昵称是什么')
3 print(ql)
```

```
MATCH (a:`ns0__足球俱乐部` {uri:"http://www.semanticweb.org/think/ontologies/Football.owl#皇家马略卡足球俱乐部"})
RETURN a.ns0__绰号
MATCH (a:`ns0__足球俱乐部` {uri:"http://www.semanticweb.org/think/ontologies/Football.owl#皇家马略卡足球俱乐部"})
RETURN a.ns0__绰号
```



# 自然语言转Cypher语句

```
1 #接口
2 lac = LAC(mode='rank')
3 aim_list = ['主要荣誉','位于','国籍','容纳','成立时间','球衣号码','球迷昵称','生
4 l_list = ['主要荣誉','位于','国籍','容纳','成立时间','球衣号码','球迷昵称','生日
5 r_list = ['主要竞争对手','现任主席','所属球队','执教','球队主场','赞助商',
6 def question_part(s,aim_list):
7     key_index = [] #关键词汇下标
8     simi_list = [] #对比aim_list相似度表
9     lac_result = lac.run(s)
10    max_index = 0
11    temp_list = []
12    for i in lac_result[2]:
13        if(i!=0):
14            key_index.append(i)
15    for i in range(len(aim_list)):
16        for j in range(len(key_index)):
17            temp_list.append(synonyms.compare(aim_list[i],lac_result[0][j]))
18            simi_list.append(temp_list)
19            temp_list = []
20    #定位最相似下标
21    temp = simi_list[0][0]
22    for i in range(len(simi_list)):
23        for j in range(len(simi_list[i])):
24            if(simi_list[i][j]>temp):
25                temp = simi_list[i][j]
26                max_index = i
27    # print(lac_result)
28    # print(simi_list)
29    return aim_list[max_index]
```



```
def which_template(q_part, sentence):
    temp = 0
    lac_res = lac.run(sentence)
    # print(lac_res)
    entity_list = []
    for index, i in enumerate(lac_res[1]):
        if i == 'PER' or i == 'LOC' or i == 'ORG':
            entity_list.append(lac_res[0][index])
    if(len(entity_list)>1):
        get_relation_between(sentence) #q3
    else:
        for i in range(len(l_list)):
            if(q_part == l_list[i]):
                get_entity_info(sentence) #q1
                temp = 1
    if(temp == 0):
        get_entity_relation(sentence) #q2
```

决定使用何种模板

NER2Cypher接口封装





# 自然语言转Cypher语句

## 测试

```
1 #aim_list已固定, 只需修改输入的sentence
2 sentence1 = '皇家马略卡足球俱乐部的昵称是什么'
3 sentence2 = '卡尔洛·安切洛蒂在哪个球队执教'
4 sentence3 = '皇家贝蒂斯足球俱乐部和塞维利亚足球俱乐部是什么关系?'
5 q_part = question_part(sentence1, aim_list)
6 which_template(q_part, sentence1)
7 q_part = question_part(sentence2, aim_list)
8 which_template(q_part, sentence2)
9 q_part = question_part(sentence3, aim_list)
10 which_template(q_part, sentence3)
```

```
MATCH (a:`ns0__足球俱乐部` {uri:"http://www.semanticweb.org/think/ontologies/Football.owl#皇家马略卡足球俱乐部"})
```

```
RETURN a.ns0__绰号
```

```
MATCH (a:`ns0__主教练` {uri:"http://www.semanticweb.org/think/ontologies/Football.owl#卡尔洛·安切洛蒂"}) -[:`ns0__执教`]->(b)
```

```
RETURN b
```

```
MATCH (a:`ns0__足球俱乐部` {uri:"http://www.semanticweb.org/think/ontologies/Football.owl#皇家贝蒂斯足球俱乐部"})-[r]-(b:`ns0__足球俱乐部` {uri:"http://www.semanticweb.org/think/ontologies/Football.owl#塞维利亚足球俱乐部"})
```

```
RETURN r
```

## 三种问题测试





# Py2neo

```
1 import json
2 from py2neo import Graph, Node, Relationship, NodeMatcher, Subgraph
3
4 graph = Graph('http://localhost:7474', auth=("neo4j", "Lyc1111yc"))
5
```

```
1 matcher = NodeMatcher(graph)
```

```
1 graph.run(q1)
```

"LosBermellones"

```
1 graph.run(q2)
```

```
{
  "identity": 1083,
  "labels": ["Resource", "ns0__足球俱乐部", "owl__NamedIndividual"],
  "properties": {
    "ns0__简称": "皇马",
    "ns0__主要荣誉": "西班牙足球甲级联赛冠军(34次)",
    "ns0__位于": "西班牙首都马德里",
    "ns0__成立时间": "1902年",
    "ns0__球迷昵称": "美凌格",
    "uri": "http://www.semanticweb.org/think/ontologies/Football1.owl#皇家马德里足球俱乐部"
  }
}
```

```
1 graph.run(q3)
```

```
{
  "identity": 1039,
  "start": 1278,
  "end": 1164,
  "type": "ns0__主要竞争对手",
  "properties": {}
}
{
  "identity": 912,
  "start": 1164,
  "end": 1278,
  "type": "ns0__主要竞争对手",
  "properties": {}
}
```

Python终端直接查询Neo4j数据



## cypher语句查询

- “皇家马略卡足球俱乐部的昵称是什么？” ➡

```
MATCH (a:`ns0__足球俱乐部`  
`{uri:"http://www.semanticweb.org/think/ontologies/Football.owl#皇家马略卡足球俱乐部"})  
RETURN a.ns0__绰号
```

The screenshot shows the Neo4j Cypher query interface. The query entered is: `neo4j$ MATCH (a:`ns0__足球俱乐部`{uri:"http://www.semanticweb.org/think/ontologies/Football.owl#皇家马略卡足球俱乐部"}) RETURN a.ns0__绰号`. The interface has a sidebar with three options: Table (selected), Text, and Code. The main area displays a table with one column, `a.ns0__绰号`, and one row with the value `"LosBermellones"`. The status bar at the bottom indicates: "Started streaming 1 records after 6 ms and completed after 6 ms."

a.ns0__绰号
"LosBermellones"



## cypher语句查询

- “卡尔洛·安切洛蒂在哪个球队执教？” ➡

```
MATCH (a:`ns0__主教练`  
`{uri:"http://www.semanticweb.org/think/ontologies/Football.owl#卡尔洛·安切洛蒂"}) - [:`ns0__执教`]->(b)  
RETURN b
```

The screenshot shows the Neo4j Cypher query interface. The query entered is:

```
neo4j$ MATCH (a:`ns0__主教练`{uri:"http://www.semanticweb.org/think/ontologies/Football.owl#卡尔洛·安切洛蒂"}) - [:`ns0__执教`]->(b)  
RETURN b
```

The interface includes a sidebar with icons for Graph, Table, Text, and Code. The main area displays a graph visualization with a central node labeled "http://w..." and several edges connecting it to other nodes. The right sidebar shows the "Node Properties" for the selected node, including:

- Resource: ns0\_\_足球俱乐部
- owl\_\_NamedIndividual
- <id>: 1083
- ns0\_\_主要荣誉: 西班牙足球甲级联赛冠军(34次)
- ns0\_\_位于: 西班牙首都马德里
- ns0\_\_成立时间: 1902年
- ns0\_\_球迷昵称: 美凌格
- ns0\_\_简称: 皇马
- uri: http://www.semanticweb.org/think/ontologies/Football.owl#皇家马德里足球俱乐部



## cypher语句查询

- “皇家贝蒂斯足球俱乐部和塞维利亚足球俱乐部是什么关系？”



```
MATCH (a:`ns0__足球俱乐部`  
`{uri:"http://www.semanticweb.org/think/ontologies/Football.owl#皇家贝蒂斯足球俱乐部"}`)-[r]-(b:`ns0__足球俱乐部`  
`{uri:"http://www.semanticweb.org/think/ontologies/Football.owl#塞维利亚足球俱乐部"}`)  
RETURN r
```

neo4j\$ MATCH (a:`ns0\_\_足球俱乐部`{uri:"http://www.semanticweb.org/think/ontologies/Football.owl#皇家贝蒂斯足球俱乐部"}...

Table

1

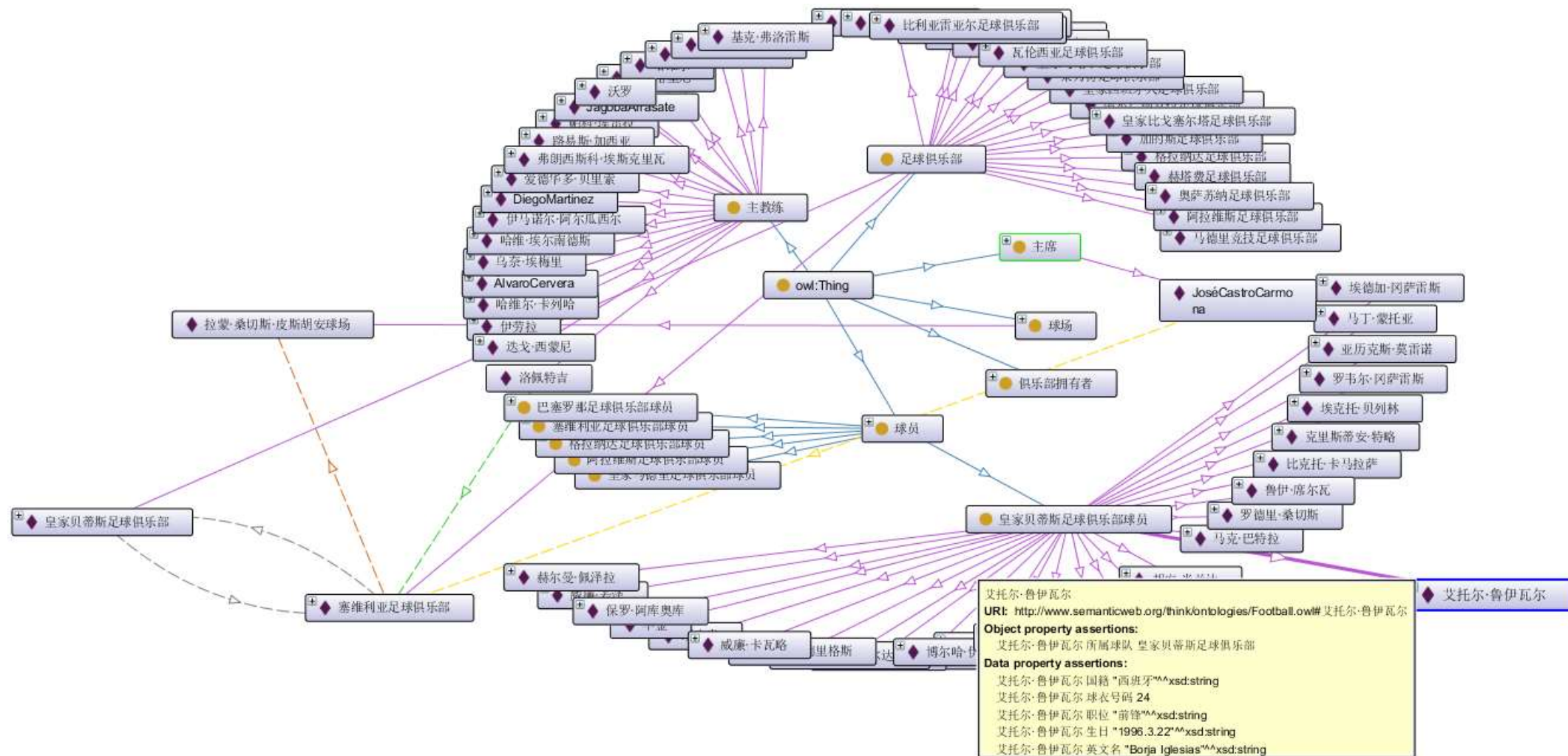
```
{  
  "identity": 1039,  
  "start": 1278,  
  "end": 1164,  
  "type": "ns0__主要竞争对手",  
  "properties": {  
  
  }  
}
```



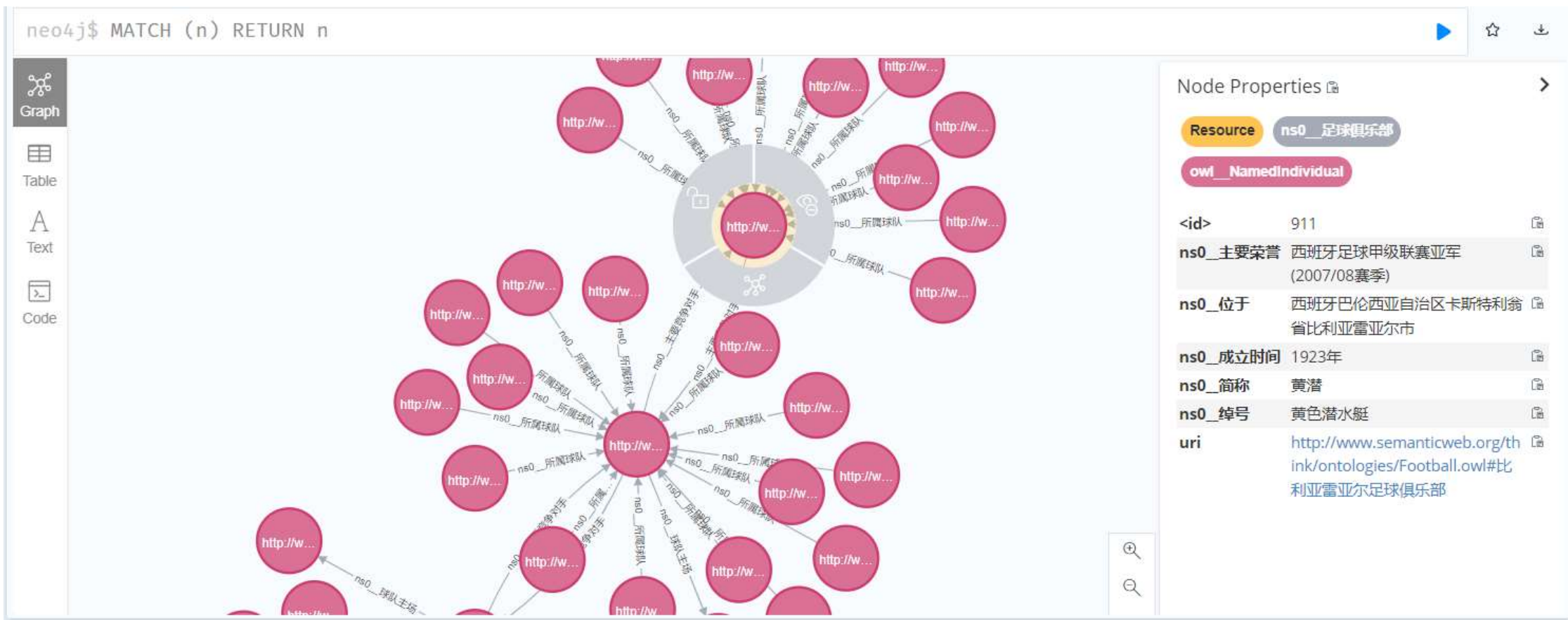


# Graduation thesis 可视化

# Protégé图谱—可视化



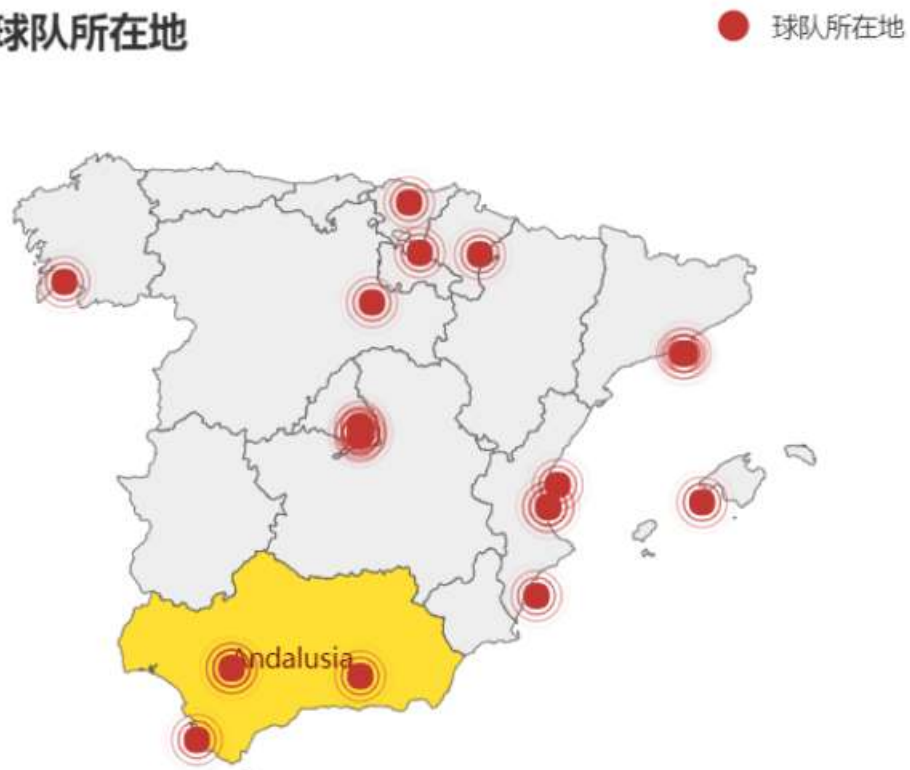
# Neo4j图数据





# 网页动态可视化

西甲全部球队所在地



西甲全部球队所在地





# 感谢您的观看

答辩人： 第三组