基于单词模型的知识图谱构建

Word Bubble

张羽戈
10165102136

2019 年 1 月 9 日

目标

- 作业要求: 知识图谱
  - 结点表示实体.
  - 实体与实体之间的关系.
  - 在已有关系的基础上实现「推理」.
- 具体任务
  - 找一个合适的领域.
  - 搜集这个领域中的实体和实体之间的关系数据.
  - 建立一套查询系统 (比如 Cypher).
  - 将查询结果展示出来 (比如 Neo4j Browser).

想法

- 实体: 英文单词.
- 关系: 英文单词之间的天然的关系.
  - 同义词.
  - 紧密相关的词.
  - 词组搭配 (葡萄配红酒).
  - 词根 (Word Family).
- 查询:
  - 一个单词的近邻. 有用!
  - 两个单词的距离. 无用!

数据准备

- 开始:
  - 牛津 3000 核心词汇表 (3839).
  - 学术单词表 (AWL 10 Sublists).
- 爬取 Macmillan Dictionary.
  - 省下了 API 昂贵的开销.
- 用网页中的单词链接加深搜索.
- 两次迭代: 14801 词条, 8576 词组, 3157 索引 (同义词 + 相关词).
- 使用 py2neo 将数据注入到 Neo4j.

一些数据: 度数最多的点

```
match (k:Word) with k, size((k)--()) as degree where
degree > 30 return k.name, degree
```

Word	Degree
get	165
put	115
around	69
day	60

一些数据: 连通性

```
match (u) return u.partition as partition, count(*) as
size order by size desc
```

CONNECTED!

一些数据: 最长最短路

- 需要求所有点对最短路, 使用 C++ 实现.
- 最大距离为 18.

Source	Target
Monday	disdain
Monday	fake
Monday	hypothesize

实体设计

- 单词 (具体词性).
  - Id, Name, Group, Part 以属性存储. Senses 以 JSON 存储.
  - 保留所有例句中的链接, 方便跳转.
  - 剔除空词条、无法处理的释义等.
- 抽象词条 (抽象词性): 没有具体的页面对应.
- 词组: 带空格的都算.
- 索引.

查询设计

- 使用 Javascript 调用 Neo4j REST API.
- 查询一个单词若干跳内的邻居.
  - 太多怎么办: 停用词处理.
  - 找不到怎么办: 警告信息.
  - 实体查询方法: 默认以 Name 查询, 可以指定字段
    → **翻译成 Cypher**.
  - 用户不会写怎么办: 跳转链接 (交互设计).
- 查询两个单词之间的路径.

难点: 可视化与交互设计 (1/4)

- 不使用 Neo4j Browser:
  - 没办法自定义.
  - 不开源, 无法嵌入.
  - 与查询联动较差.
- 实现本地应用: Electron.js.
- 使用 Material UI: React + Material-UI.
- 可视化类库: d3.js.

难点: 可视化与交互设计 (2/4)

- 模块:
  - 搜索框 (SearchBar).
  - 结点关系图 (Graph).
  - 最佳匹配词条 (Entry).
- 联动:
  - SearchBar $\leftrightarrow$ Graph.
  - SearchBar $\leftrightarrow$ Entry.
  - Graph $\leftrightarrow$ Entry.

难点: 可视化与交互设计 (3/4)

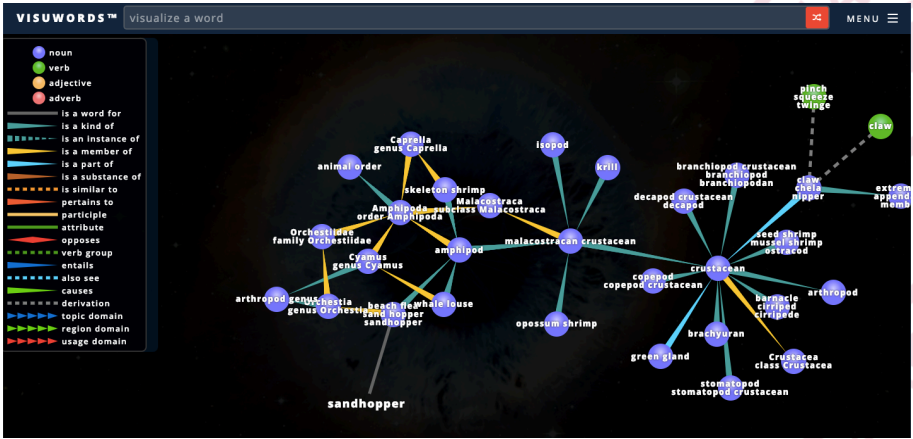
- 搜索框细节:
  - 搜索历史回溯.
  - 从上一次到这一次走了有多远.
- 词条内容:
  - 看起来不比词典差.

难点: 可视化与交互设计 (4/4)

- 图的实现:
  - 带碰撞检测的力导向图.
  - 词条的自动字体大小和自动换行设计.
  - 可自定义的类别颜色区分和类别隐藏.
  - 可自定义的自动缩放.
- 痛点: 在 React 上用 d3, DOM 操作无法回避, 且新增同步问题.


KEEP
CALM
AND
LET'S
TRY IT

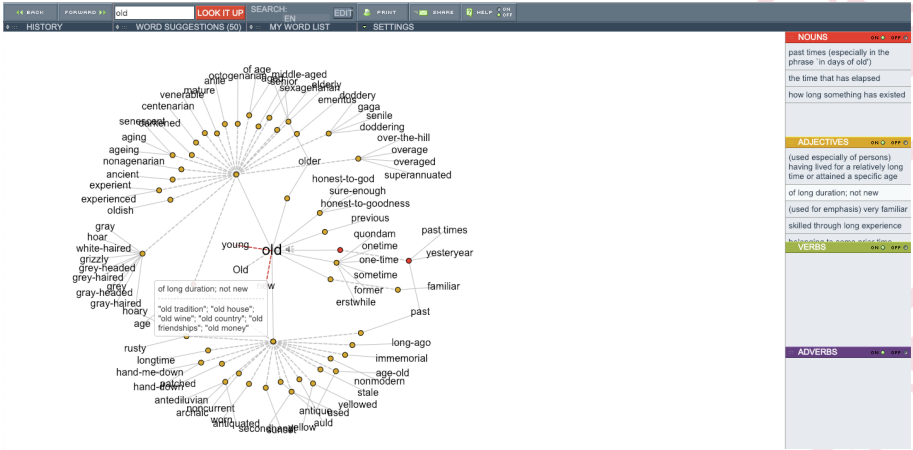
比较: Visuwords (1/2)



比较: Visuwords (2/2)

- 优点:
  - 图支持增量扩充.
  - 支持的关系种类多, 区分明确.
- 缺点:
  - 丑.
  - 没有「词典」的功能.

比较: Visual Thesaurus (1/2)



比较: Visual Thesaurus (2/2)

- 优点:
  - 释义分开列举, 有交互.
  - 历史、提示功能强大.
- 缺点:
  - 界面复古.
  - 付费.

比较: ConceptNet (1/2)

Synonyms	Related terms	Similar terms	Derived terms
Terms with this context	Antonyms	Distinct terms	Links to other sites
Word forms	Things with the property old	Etymologically related	Symbols of old
Context of this term	Derived from	Location of old	Etymological roots of "old"
old is a type of...	Things motivated by old	/r/NotHasProperty old	Etymologically derived terms
old is capable of...	Things located at old	Things made of old	Types of old

比较: ConceptNet (2/2)

- 优点:
  - 关系种类非常丰富.
  - 历史悠久的开源软件, 拥有上千 star.
- 缺点:
  - 原生可视化支持较弱（不支持?）.

说到开源, 我就想到……

- Word Bubble 会在结课后开源.

  https://github.com/ultmaster/

- 希望大家多多关注.