

周报

- 2020.3.18-2020.3.25
- 熊浩博

本周工作

个人项目

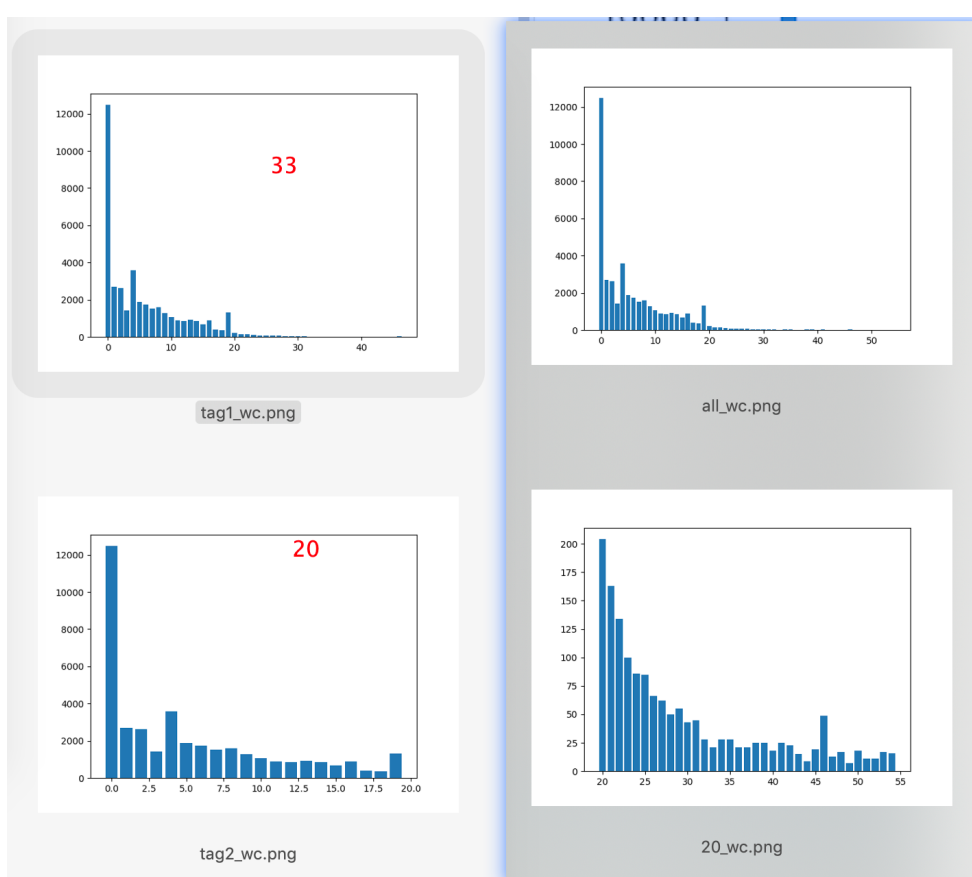
1. 找数据集

ECG

1. catboost进行超参数选择
想了一下，发现除了学习率没什么可以调的
2. 实现只用传统特征进行catboost的效果

| | train | val | test | |
|----------|--------------|-------|---------------|---------|
| catboost | 0.817 | 0.797 | 0.7329 | 2048,8 |
| catboost | 0.813 | 0.793 | 0.7275 | 2048,12 |
| catboost | 0.841 | 0.824 | 0.7631 | 5000,12 |
| | - | 0.825 | 0.7714 | |
| catboost | 0.847 | 0.830 | 0.7747 | 5000,8 |

1. 实现选取部分标签训练DeepNN
选了33，和20两种



(结果还在跑)

| | train | val | test | |
|-------------------|-------|-------|--|---------|
| DeepNN-33 | 0.921 | 0.858 | 0.8256 0.8317(33) | 5000,12 |
| DeepNN33-catboost | 0.934 | 0.853 | submit_tag33/ subA_2020032 51145.txt | 5000,12 |
| DeepNN-33 | | | | 5000, 8 |
| DeepNN33-catboost | | | | 5000, 8 |
| DeepNN-20 | | ing | | 5000, 8 |
| DeepNN20-catboost | | | | 5000, 8 |

openKE

找了下问题，改了bug（勉强算是bug）

- 问题
训练玩无法保存embedding
- 原因
numpy.ndarray的tolist()需要大量内存，无非得到满足，进程被系统killed
- 解决
 - 方法一：改用pickle保存numpy.ndarray

- 方法二：直接从model.vec.tf文件读取embedding

交流

1. 李美凝

3月17日：调用python版CRF算法，出现bug

3月18日：已经自行解决

2. 赵亦威

- 讨论了一下kg目前存在的缺陷

a. 部分信息设为属性未设为实体

b. 部分属性未拆分

如：个人信息：性别 年龄 学位

概念里：top3

c. 部分实体没有去重

如：Holder与person、company有重合

d. 外键信息没有利用

如：实际控制人应该对应一个person实例，现在知识图谱里只是一个股票的属性

- 结论：

a. 并未构建实体，仅根据ER图和表结构构建了从mysql数据库到neo4j的映射，没有先转为RDF格式等最后再画本体图，提前画了可能会变动

b. 对于第2个问题部分属性未拆分

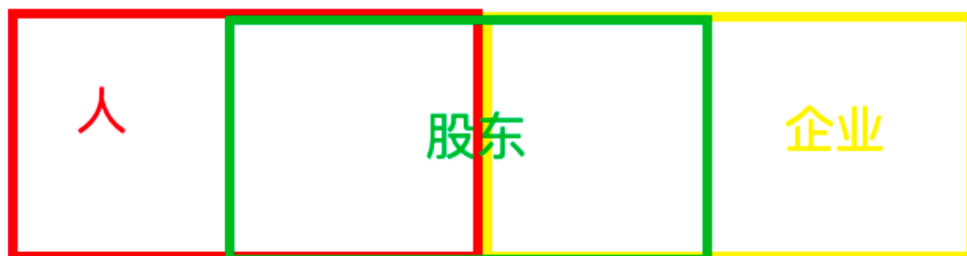
赵：之前没注意，可以再处理一下，构建的脚本等都有保留，可修改后重新构建neo4j数据库

c. 第1和第4是同一个问题

问题：部分属性对应的实体不存在，例如某上市企业的实际控制人是公司B，但B并不是上市企业，故不存在实体B。解决：不存在的企业可以创建，股票股东信息可以缺失。企业应包括上市企业和非上市企业。赵：好，去改一下

d. 针对第3个问题如Holder与person、company有重合

觉得是没有层次导致的，需要先明确层次关系，先放一下



e. 小的先改，大的先放一下，等本体出来再弄

3. 周琪丰

传达了一下，从年报里抽取公司的股东结构信息，与网页上的类似

十大流通股东

| 2019-12-31 | 2019-09-30 | 2019-06-30 | 2019-03-31 | 2018-12-31 | | |
|---|------------|------------|------------|------------|------|----------------------|
| 前十大流通股东累计持有：137.86亿股，累计占流通股比：71.04%，较上期变化：8863.91万股 ↑ | | | | | | |
| 机构或基金名称 | 持有数量(股) | 持股变化(股) | 占流通股比例 | 实际增减持 | 股份类型 | 持股详情 |
| 中国平安保险(集团)股份有限公司-集团本级-自有资金 | 96.19亿 | 不变 | 49.57% | 不变 | 流通A股 | 点击查看 |
| 香港中央结算有限公司 | 15.04亿 | ↑8199.89万 | 7.75% | ↑5.76% | 流通A股 | 点击查看 |
| 中国平安人寿保险股份有限公司-自有资金 | 11.86亿 | 不变 | 6.11% | 不变 | 流通A股 | 点击查看 |
| 中国平安人寿保险股份有限公司-传统-普通保险产品 | 4.40亿 | 不变 | 2.27% | 不变 | 流通A股 | 点击查看 |
| 中国证券金融股份有限公司 | 4.29亿 | 不变 | 2.21% | 不变 | 流通A股 | 点击查看 |

- 感觉没必要，学弟自己解释

下周安排

个人项目

1. 找数据
2. 想baseline的实现方式

ECG

1. 跑完实验——选取部分标签训练DeepNN
2. 看到一个名为 [ReZero](#) 的神经网络结构改进方法，可以试着改进top1
3. catboost进行超参数选择
再找找