EL流程

- 1. 根据股票名称库,去百科爬取信息,建立知识库。【产出kb_data.json】【产出id_embedding.pkl】
 - a. 包含以下内容:
 - i. kb-data.json

```
1 {"entity_name":"", "entity_id":"", "ali":"", "abstract":"", else}
```

ii. id_embedding.pkl

```
1 {"entity_id":[list,1*768]}
```

2. 新闻库,【已知new.json】

```
1 新闻库包含:
2 {"text_id":"", "title":"", "text":"","stocks_name":""}
```

将新闻库的text输入到NER中,找到ORG(机构名),会得到

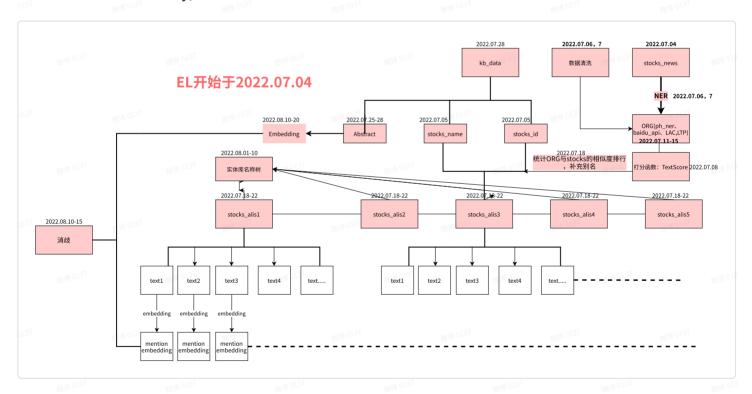
```
1 n个stocks_name: setSN: {A B C D}
2 m个ORG结果: setORG: {a b c d e f g 等}
3 分别计算A B C D 与SetORG的分数:
4 得到一个表格:
5
6
        Α
                 C
                      D
        1
            2
   а
        4
            3
9 c
        2
10
11 e
12
13 g
14 等
15
```

```
16
17 对A找到前几的SetORG的元素并计数。
18 例如
19 A: {d:d+=1, text: text.append(text_id)}
20 B: {b:b+=1, text: text.append(text_id)}
21 C: {a:a+=1, text: text.append(text_id)}
22 等
23 也就是找到跟A B C D最像的1个或者多个 并计数,通过大量样本的一个排序,会得到关于
24 A的一些ORG的排序结果,我们可以认定排名前多少的作为A的一个别名库。
```

并把这个结果放进kb_data.json的别名库中。【产出alis.json】【产出一个实体字典树】 alis.json:

```
1 {"stocks_id":"", "alsi:",["","","",....]}
```

3. 利用识别到的<entity,alis>对,训练一个消歧的模型。【产出消歧的model】



注: 粉色部分是已有的或者已做的。