知识图谱问答平台

目录

CATALOG

1 目标问题和意义价值

2 设计思路与方案

3 实现

4 运行效果

5 创新与特色

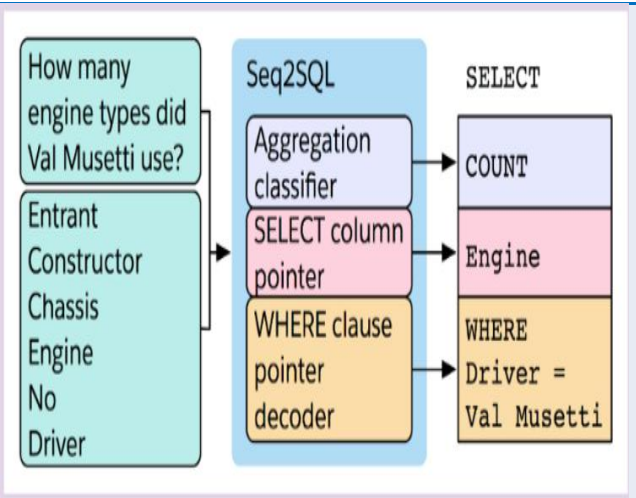
6 待改进的地方

7 程序文件结构

01

PART

智能问答系统简介



知识图谱问答平台是什么?

知识图谱问答平台可以充当数据库的智能接口，让不熟悉数据库的用户能够快速地找到自己想要的数据。

自助查询平台自动将自然语言转换为SQL， 用户不用了解SQL的具体语法，而是直接以自然语言的形式输入问题，然后平台会利用深度学习技术，将问题解析成SQL语言，进而在数据库进行查询。

问题产生的背景?

目前数据库依然是企业存储数据的主要方式，尽管有着各色的数据库，但访问和操作数据库的SQL是通用的。人性化的编程语言SQL为开发者在工作中访问数据库提供了便利，但同时也极大地限定了非专业用户按需查询数据库的场景和查询界限。随着人工智能取得突破进展，结合了人工智能相关技术为非专业用户查询数据库提供了新的思路。

有哪些特色和创新点?

利用深度学习模型将自然语言转换为SQL;

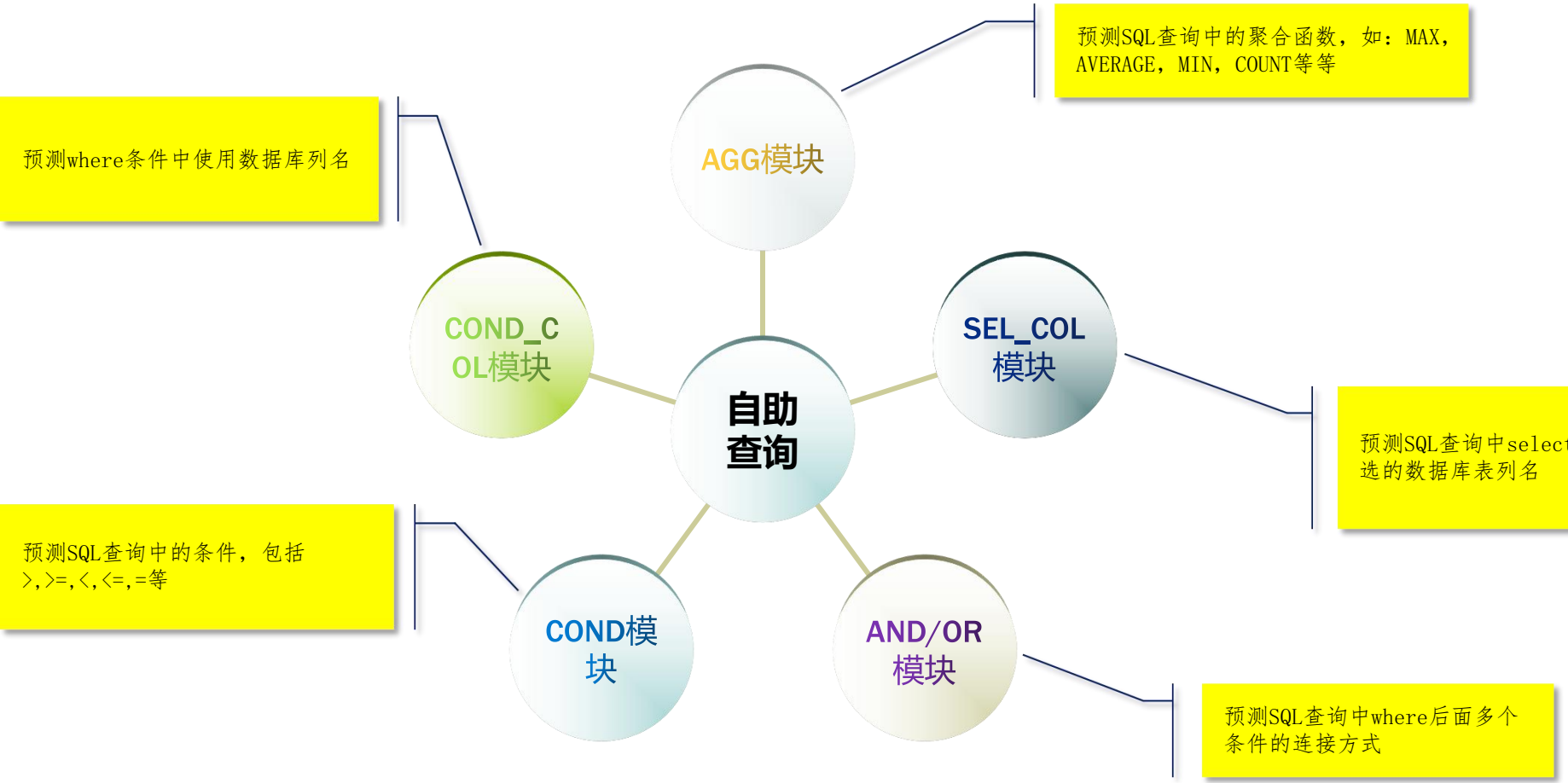
支持查询结果图形化显示;

支持对知识图谱数据进行查询;

能否带来真正的业务价值?

为了让非专业用户也可以按需查询数据库，当前流行的技术方案设计了基于条件筛选的专门界面，用户可以通过点选不同的条件来查询数据库。然而，通过界面操作极大地限制了数据库查询的使用场景和查询界限。

通过自助查询平台可以有效降低人机交互的距离和门槛，用户只需以自然语言的形式输入问题，平台自动帮你转换成可以执行的SQL语言，提供了极大地灵活性。


自助查询
AGG模块
预测SQL查询中的聚合函数，如：MAX，AVERAGE，MIN，COUNT等等
COND_COL模块
预测where条件中使用数据库列名
SEL_COL模块
预测SQL查询中select所选的数据库表列名
COND模块
预测SQL查询中的条件，包括>,>=,<,<=,=等
AND/OR模块
预测SQL查询中where后面多个条件的连接方式

本软件旨在利用国内上市公司信息构建一个问答系统，通过该系统用户可以查询上市公司常见信息。

查询内容涵盖：

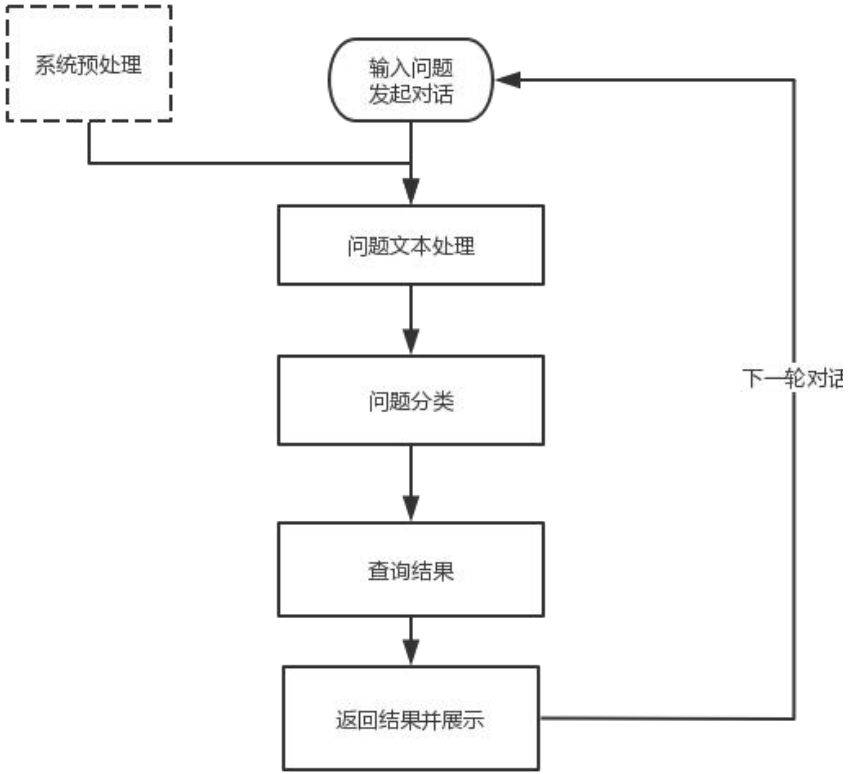
公司大股东
机构持股情况
上市公司高管人员
高管履历
公司或机构之间的联系
公司利润表
公司资产负债表
公司现金流量表
常见国家经济统计指标

02

PART

设计思路与方案

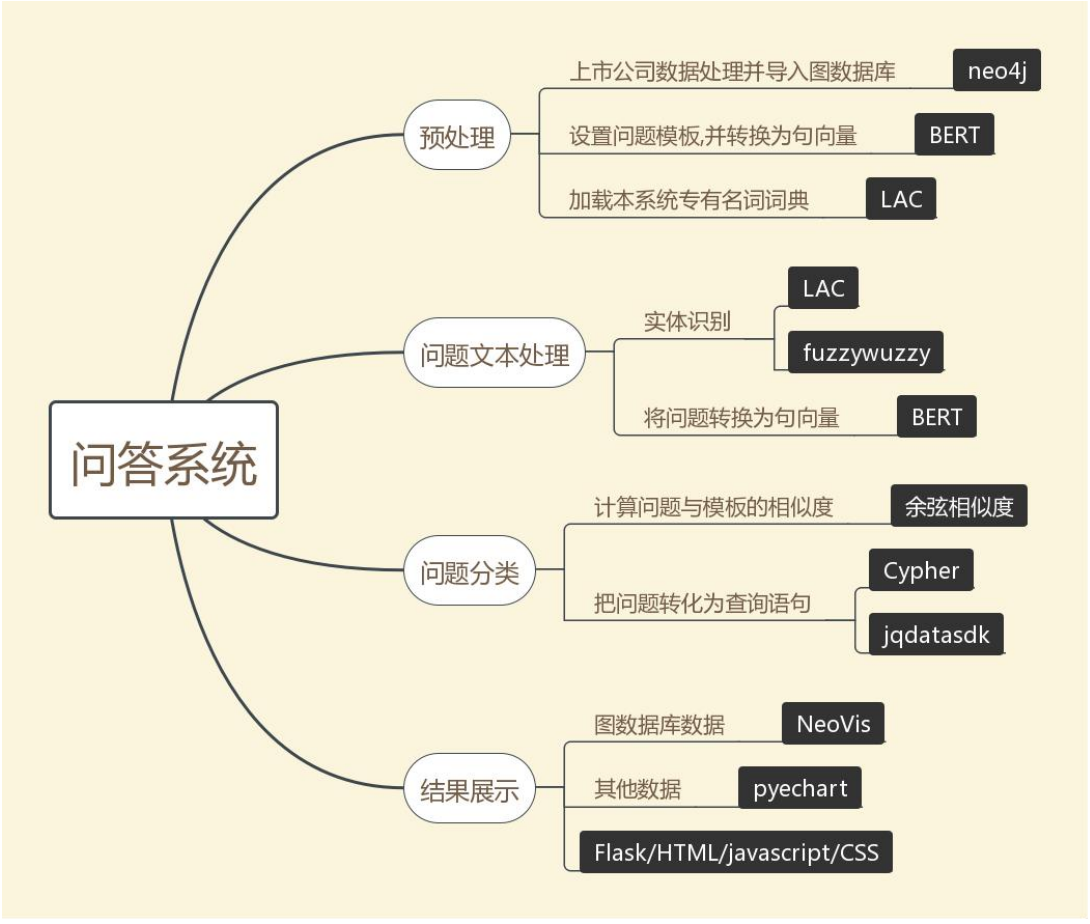
对话流程如图所示：



我们的设计思路是这样的：

该系统是一个检索型问答系统，系统以数据库的形式把上市公司相关信息进行存储，同时预先设置了一些问题模板，当用户提问时，系统判断该问题对应哪个模板，然后使用该模板对应的查询语句在数据库进行查询，将查询结果返回给用户，一轮对话结束。

对话流程中每个模块所用到的主要技术



系统预处理之数据源

本系统用到的数据源自三个不同地方：

1.上市公司股东，机构持股，公司高管，高管履历等信息，这类信息可以形成网络形状，比较适合用知识图谱的形式存储，因此存在图数据库neo4j中，这类数据也是本系统中占比最多的数据；

2.上市公司利润表，资产负债表和现金流量表，这类数据占用空间不大，直接使用聚宽提供的接口在线获取；

3.国家经济指标数据，例如GDP，人口等等，这部分数据来自国家统计局统计年鉴，处理后直接以dataframe的形式存储在本地。

所以，根据用户所提的问题不同，系统可能会查询不同的数据源。

系统预处理之问题模板

本系统是基于模板进行问答的，所以需要预先设置用户可能会问到的问题模板，所问的问题也只能在模板范围内，超出范围系统就无法识别。

系统支持的问题范围如下：

公司大股东
机构持股情况
上市公司高管人员
高管履历
公司或机构之间的联系
公司利润表
公司资产负债表
公司现金流量表
常见国家经济统计指标

系统预处理之问题模板

```
template0 = ['KG_ORG的股东有哪些','哪些公司入股了KG_ORG','KG_ORG的大股东','KG_ORG股东','KG_ORG的持股人']
template1 = ['KG_ORG持有哪些公司的股票','KG_ORG入股了哪些公司','KG_ORG持股公司','KG_ORG入股的公司']
template2 = ['PER在过的机构','PER简历','PER经历','PER工作学历经历']
template3 = ['PER任职过的公司','PER在哪些公司当过高管','PER上市公司任职经历','PER服务过的公司','PER管理过哪些公司']
template4 = ['KG_ORG的高管有哪些','KG_ORG管理层','KG_ORG管理人员','KG_ORG管理团队','管理KG_ORG的人员']
template5 = ['在ORG学习或工作过的人','ORG出来的人','在ORG待过的人']
template6 = ['KG_ORG和KG_ORG之间有什么关系','ORG和ORG之间的路劲','ORG和ORG关联','ORG和ORG关系','ORG和ORG的共同点'
template7 = ['查看KG_ORG的利润表','KG_ORG利润']
template8 = ['查看KG_ORG的资产负债表','KG_ORG资产']
template9 = ['查看KG_ORG的现金流量表','KG_ORG现金流']
```

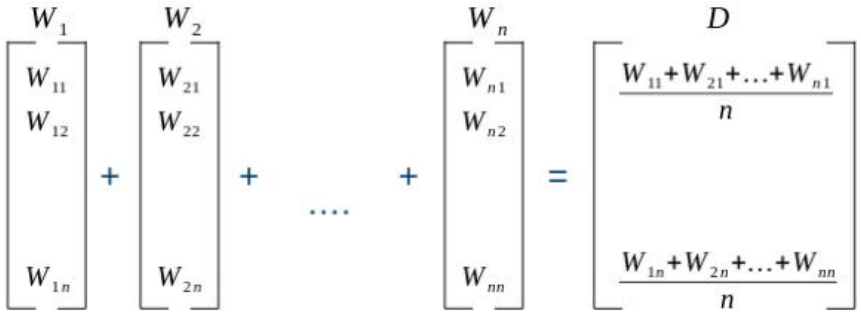
具体的模板如图所示，以**template0**为例，该模板的意思是查询某个公司的股东，该问题可能会有好几种不同的提问形式，首先把同一问题的可能表达方式都列出来，然后将每种形式都通过**BERT**转化为向量，再取它们向量的平均值(**Average word vectors**)作为该模板最终的向量表示。

图中的**ORG**，**KG_ORG**代表机构名，**PER**代表人名。

Average word vectors

上一页提到通过取几句话向量的平均值作为模板的表示，这一方法也经常被用来表示一段文本。

为了通过计算距离来实现近似搜索，我们需要将一个句子或者一篇文档表示为一个向量，这种方式叫做 Sentence Embedding 或者 Doc Embedding。其中最常用的一种模型是 Average word vectors 或者叫 Word Averaging Model(WAM)。其方法是得到将一个句子中每一个词向量在同一个 embedding 维度的值取平均值得到该句子在该嵌入维度的值。如图 5所示。



系统预处理之专有名词加载

LAC全称**Lexical Analysis of Chinese**，是百度自然语言处理部研发的一款联合的词法分析工具，实现中文分词、词性标注、命名实体识别等功能。

在系统接收到用户提问后，需要识别问题中提到的实体，比如公司名称，方便后续进行查询，本系统使用**LAC**进行命名实体识别。

由于**LAC**自身的词典可能并不包含我们会用到的所有专有名词，需要将我们自身的专有名词词典加载进**LAC**，这样**LAC**便会优先使用我们的词典。

比如，‘长城汽车’这一名词，可能**LAC**自身词典没有这个词，它就会把这个词识别为‘长城’，‘汽车’两个词，但我们的专有词典包含了‘长城汽车’，加载之后，这样系统就能正确识别‘长城汽车’了。

问题文本处理之命名实体识别

命名实体识别（Named Entity Recognition，简称NER），又称作“专名识别”，是指识别文本中具有特定意义的实体，主要包括人名、地名、机构名、专有名词等。

本系统中，命名实体识别被用来识别用户提问中提到的机构名称（包括公司名，学校名等等）或人名，后续在数据库中查询时，这些实体名会被当做查询条件。

举例：
用户输入：格力电器的股东有谁?

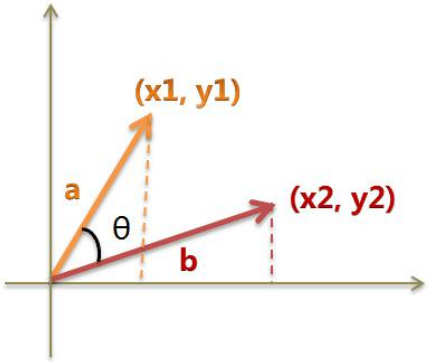
系统识别到格力电器是一个机构名，然后用‘ORG’替换它，ORG代表机构的意思

问题被转化为：ORG的股东有谁?

然后被转化后的问题去与预置的模板比较，判断应该使用哪个模板。

问题分类之利用余弦相似度查找对应问题模板

将模板和问题都转换为向量后，就可以计算问题和模板之间的余弦相似度，其实就是计算两个向量之间的夹角，夹角越小，相似度越高。



以二维空间的两个向量a和b为例，图中夹角θ越小，代表a和b越相似

例如我们计算被转化后的问题‘ORG的股东有谁’与系统中模板的相似度。发现与模板0的相似度最高，所以最终模板0作为我们的查询路径。

模板0	0.94
模板1	0.87
模板2	0.68
模板3	0.73
模板4	0.85
模板5	0.68
模板6	0.79
模板7	0.76
模板8	0.77
模板9	0.70

问题分类之从模板到查询语句

每个模板都已经事先写好了相应的查询语句，只需更改相应的查询条件即可，比如具体哪个公司。

例如前面的例子，我们找到模板0的查询语句，只需把条件中公司名替换成我们在问题中识别出来的公司名‘格力电器’即可，**... where company_name=‘格力电器’**。

返回结果并展示

系统执行完查询语句返回结果，并将结果在网页上进行展示，一轮对话就结束了。

03

PART

实现

数据存储，模型，前后端

数据存储：绝大部分数据存在neo4j中，neo4j是一个开源，免费的图数据库，适合存储知识图谱数据。

模型：系统主要用到的是深度学习模型BERT，用于将句子转换为向量。BERT是谷歌2018年推出的深度语言模型，推出后在多项自然语言处理任务上打破了记录。同时还用到了百度的LAC库用于命名实体识别，经试验LAC在中文命名实体识别上效果不错。

前后端：我们采用的是基于网页端的形式，所以前端主要使用HTML/javascript/CSS技术，后端使用的是flask框架。同时展示结果还使用了pyechart这个库，pyechart可以画出精美的图画。

04

PART

运行效果

查询某个公司的股东

问吧

请问有什么需要帮助的？

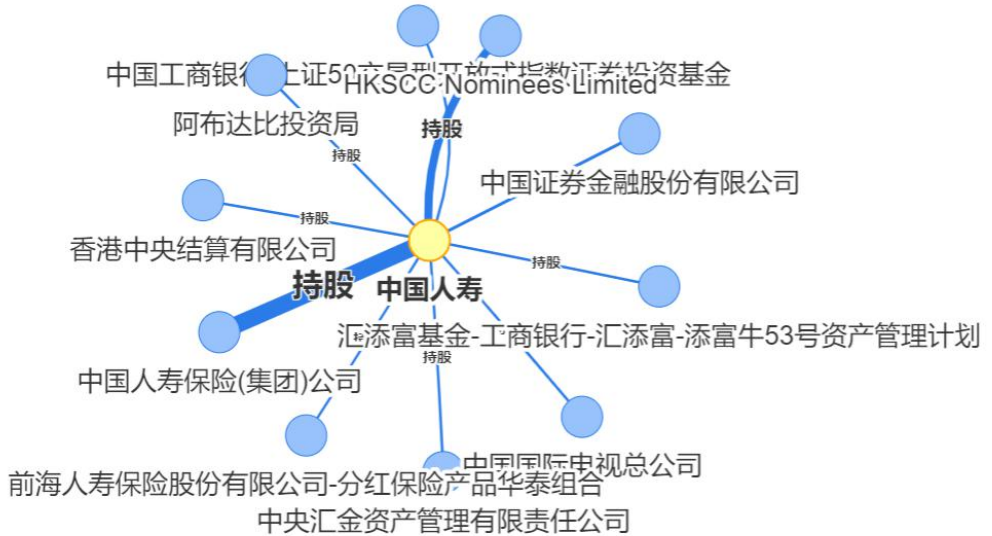
中国人寿的股东有哪些

中国人寿的股东有：
前海人寿保险股份有限公司-分红保险产品华泰组合
阿布达比投资局
香港中央结算有限公司
中央汇金资产管理有限责任公司
中国工商银行-上证50交易型开放式指数证券投资基金
中国国际电视总公司
中国证券金融股份有限公司
HKSCC Nominees Limited
汇添富基金-工商银行-汇添富-添富牛53号资产管理计划
中国人寿保险(集团)公司

请输入问题

Message

SEND



查询某个机构持股的公司，通过这个查询还能看到间接持股情况

问吧

中国人寿保险(集团)公司持股的公司有哪些

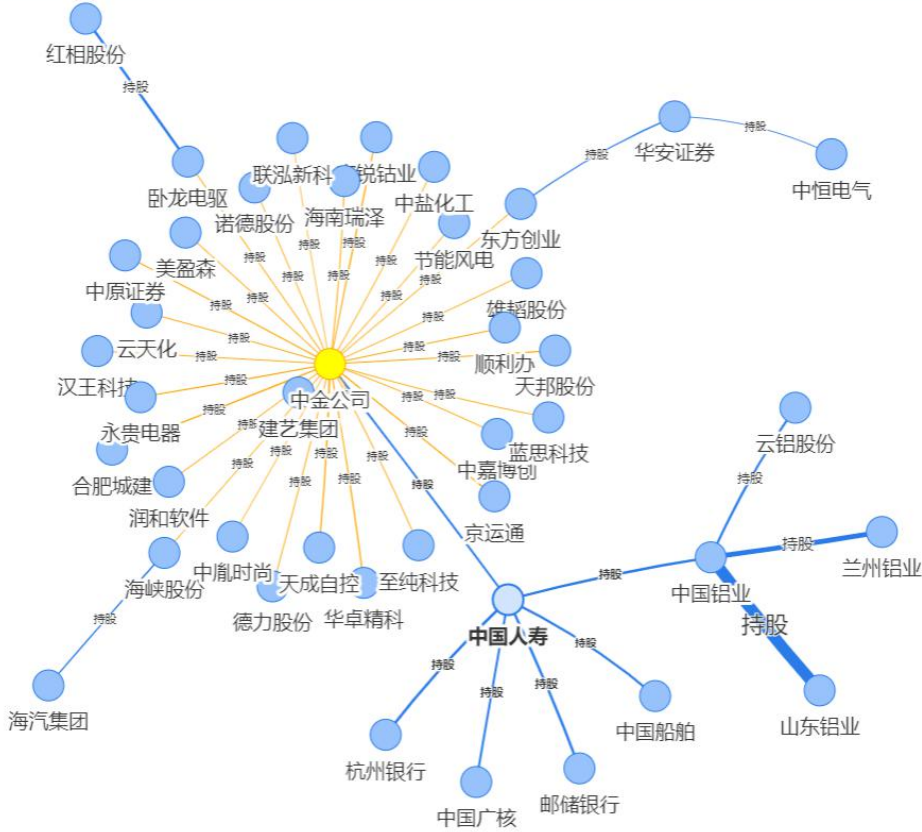
中国人寿持有的股票有：

节能风电
山东铝业
海汽集团
永贵电器
美盈森
天邦股份
中国铝业
建艺集团
海南瑞泽
中金公司
云铝股份
红相股份
邮储银行
至纯科技
中原证券
润和软件
中嘉博创
中恒电气

请输入问题

Message

SEND



激活

查询某个公司的高管团队

问吧

请问有什么需要帮助的？

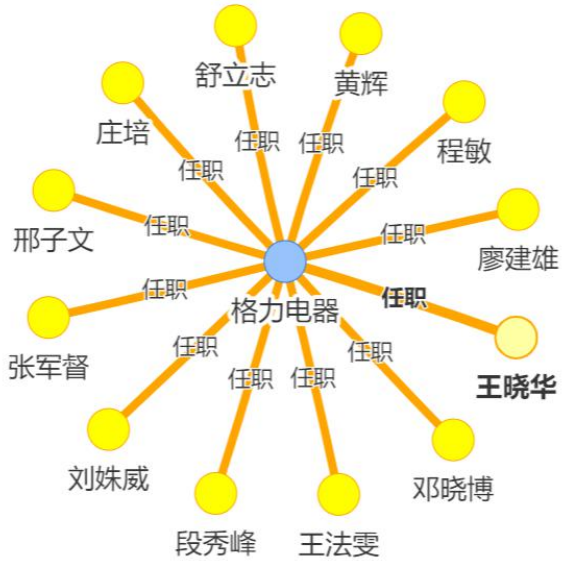
格力电器的高管团队

格力电器的高管团队包括：
王晓华
程敏
邢子文
黄辉
王法雯
邓晓博
段秀峰
舒立志
庄培
廖建雄
刘姝威
张军督

请输入问题

Message

SEND



查询某个人简历中待过的机构

问吧

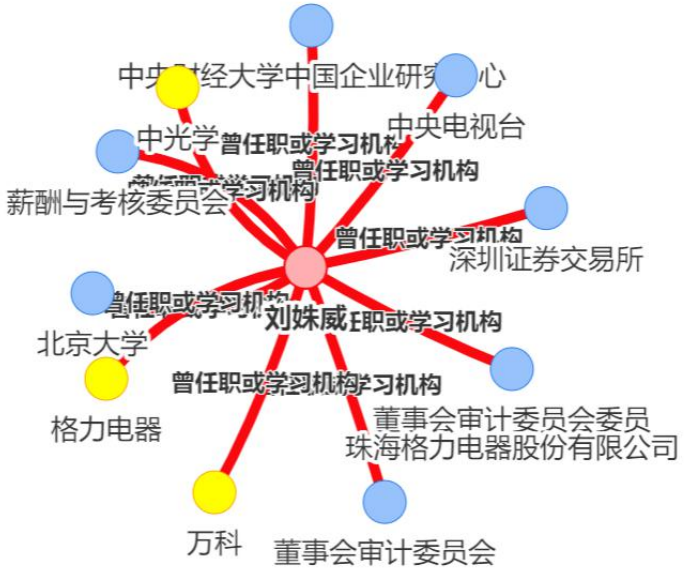
舒立志
庄培
廖建雄
刘姝威
张军胄

庄培的简历

刘姝威的简历

刘姝威待过的机构有：
董事会审计委员会
中央财经大学中国企业研究中心
中央电视台
中光学
薪酬与考核委员会
格力电器
北京大学
董事会审计委员会委员 珠海格力电器股份有限公司
万科
深圳证券交易所

请输入问题



查询在某个机构学习工作过的人

问吧

陈辉
谢发利

在中央电视台工作过的人

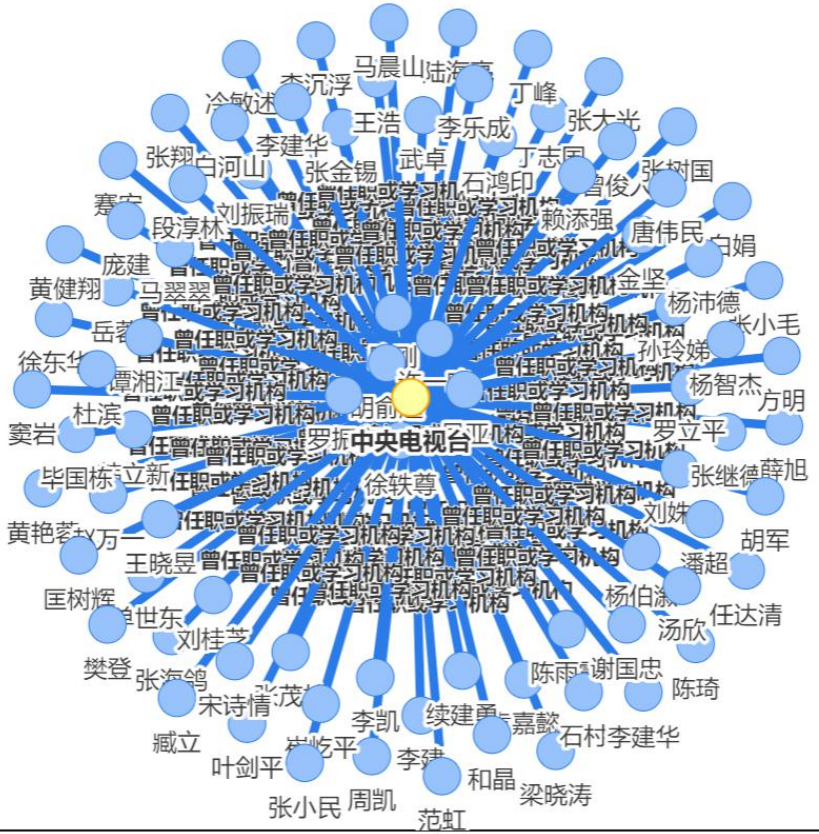
中央电视台，在该机构学习或工作过的人有：

梁晓涛
匡树辉
袁嘉懿
白娟
徐轶尊
张继德
韩立新
蹇宏
李沉浮
黄健翔
窦岩
冷敏述
石村
岳蓉
叶剑平
王晓昱

请输入问题

Message

SEND



查询某两个机构之间关系

问吧

请问有什么需要帮助的？

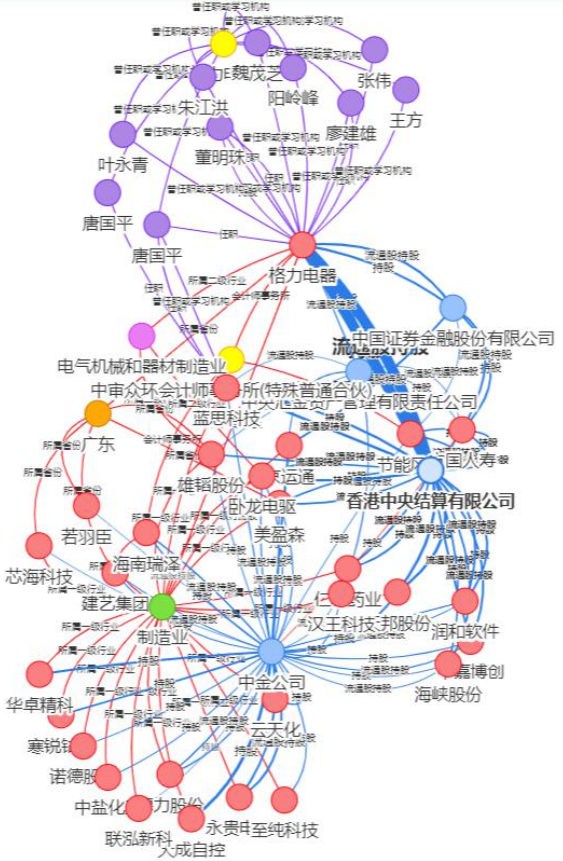
格力电器和中金公司的关系

ok

请输入问题

Message

SEND



查询某公司的利润表

问吧

请问有什么需要帮助的？

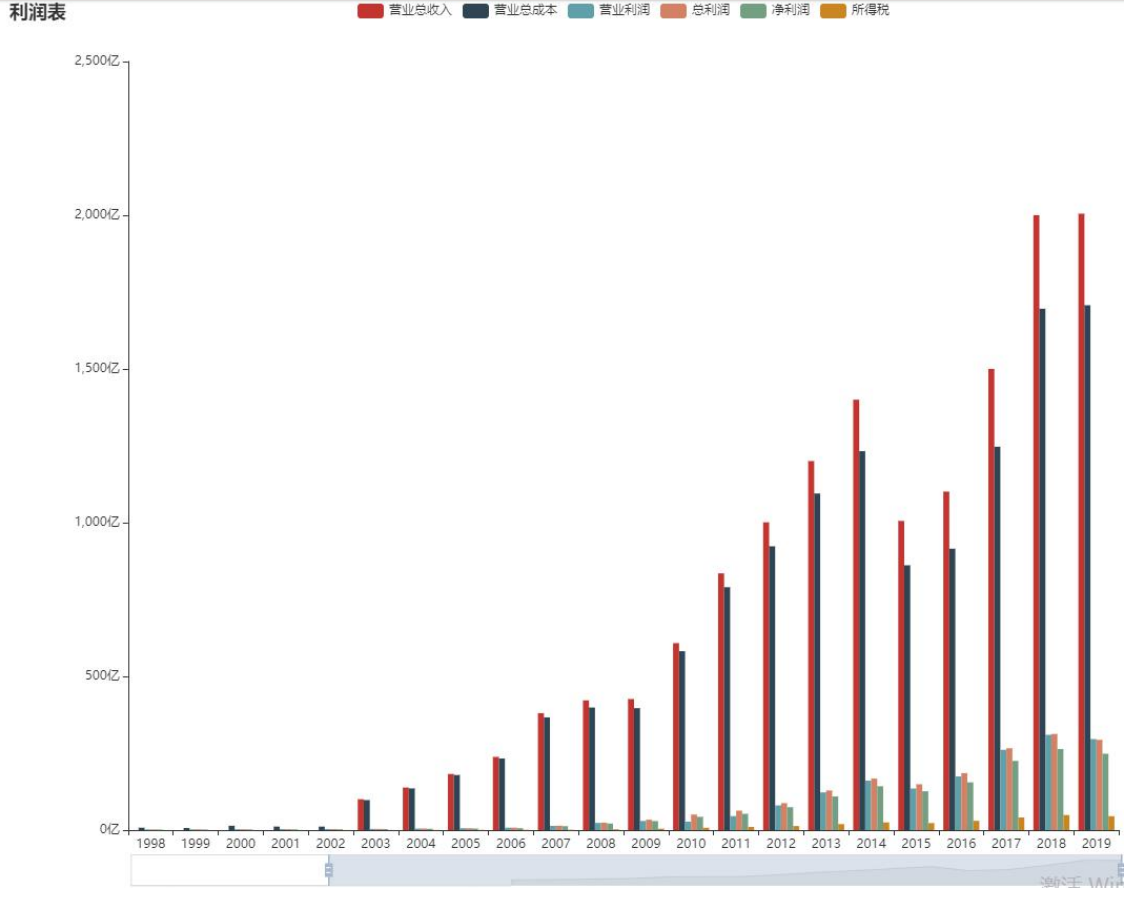
珠海格力电器股份有限公司的利润表

ok

请输入问题

Message

SEND



查询某公司的资产负债表

问吧

请问有什么需要帮助的？

珠海格力电器股份有限公司的利润表

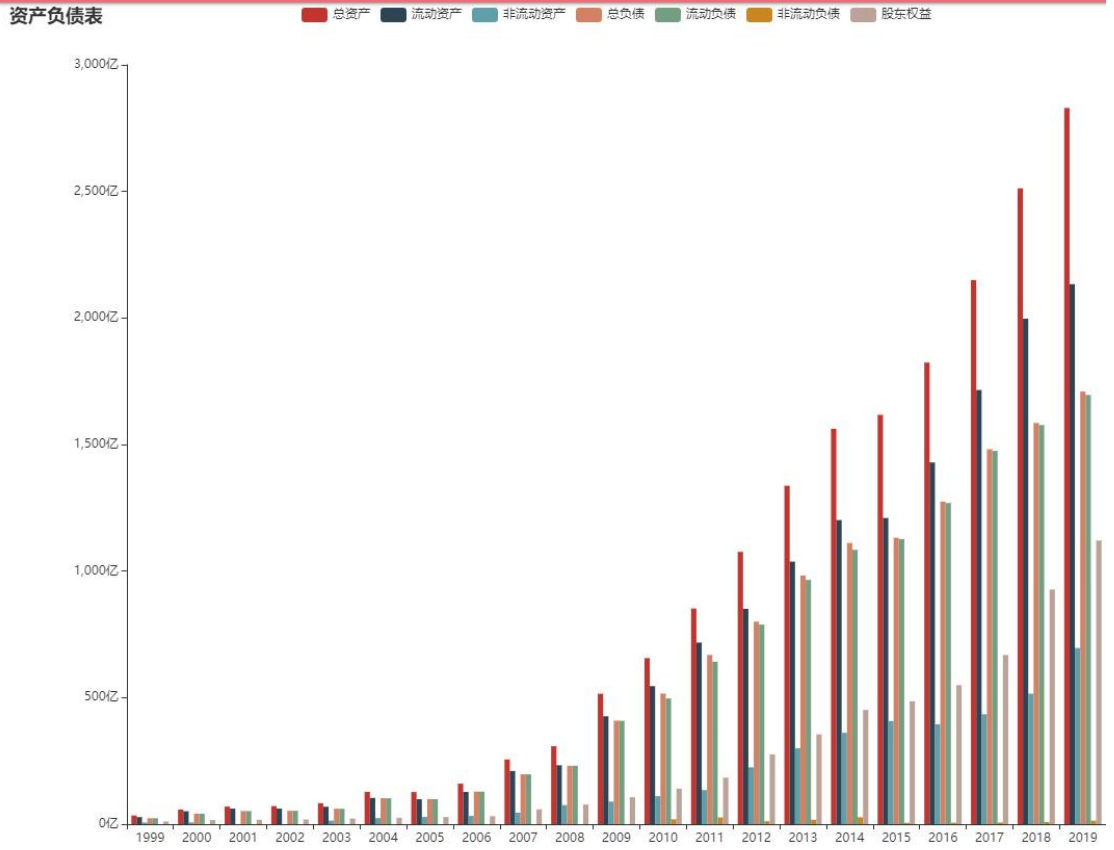
ok

珠海格力电器股份有限公司的资产负债表

ok

请输入问题

Message



查询某公司的现金流量表

问吧

请问有什么需要帮助的？

珠海格力电器股份有限公司的利润表

ok

珠海格力电器股份有限公司的资产负债表

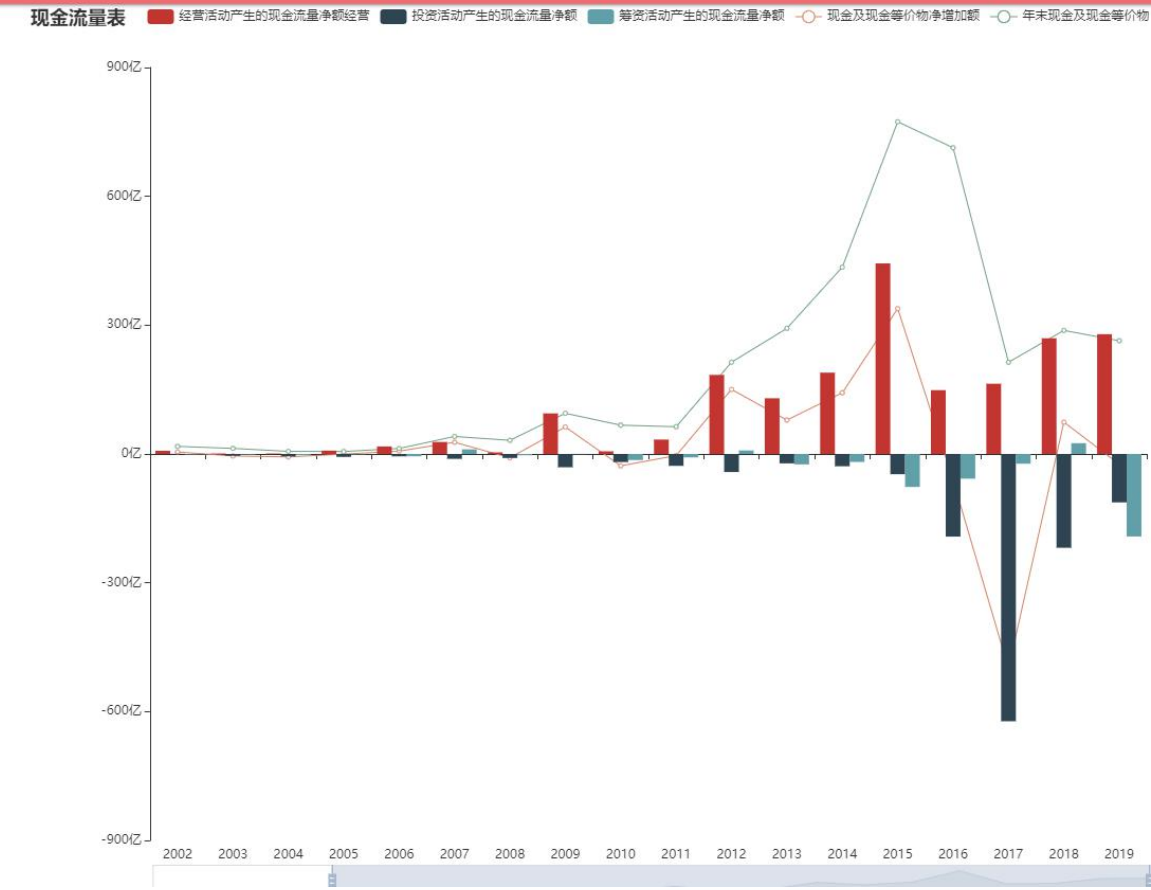
ok

珠海格力电器股份有限公司的现金流量表

ok

请输入问题

Message



查询国家经济统计指标

问吧

请问有什么需要帮助的？

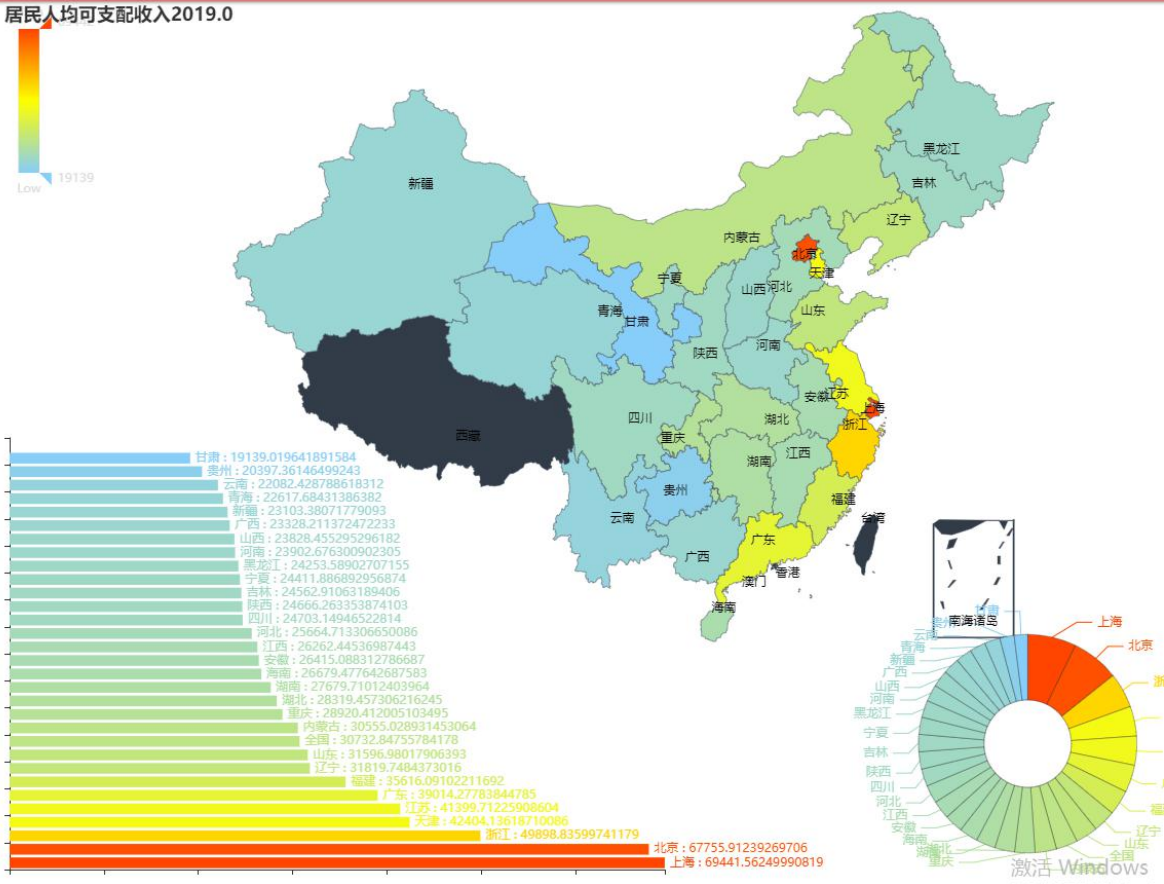
居民人均可支配收入2019

ok

请输入问题

Message

SEND



激活 Windows

05

PART

创新与特色

创新点一：
将模板与问题都转换为向量来进行匹配，实现‘软’匹配。

例如：‘格力电器的股东有谁’，‘谁持股了格力电器’，这两个问题意思类似，用户输入这两个问题，都可以找到同一个模板。

创新点二：
直接使用预训练好的**BERT**进行向量转换，这样每增加一个模板不用再次训练模型，节省时间。否则，如果把问题分类当做传统分类任务来做，每个模板当做一个类别，每增加一个模板就要训练一次模型。

06

PART

待改进的地方

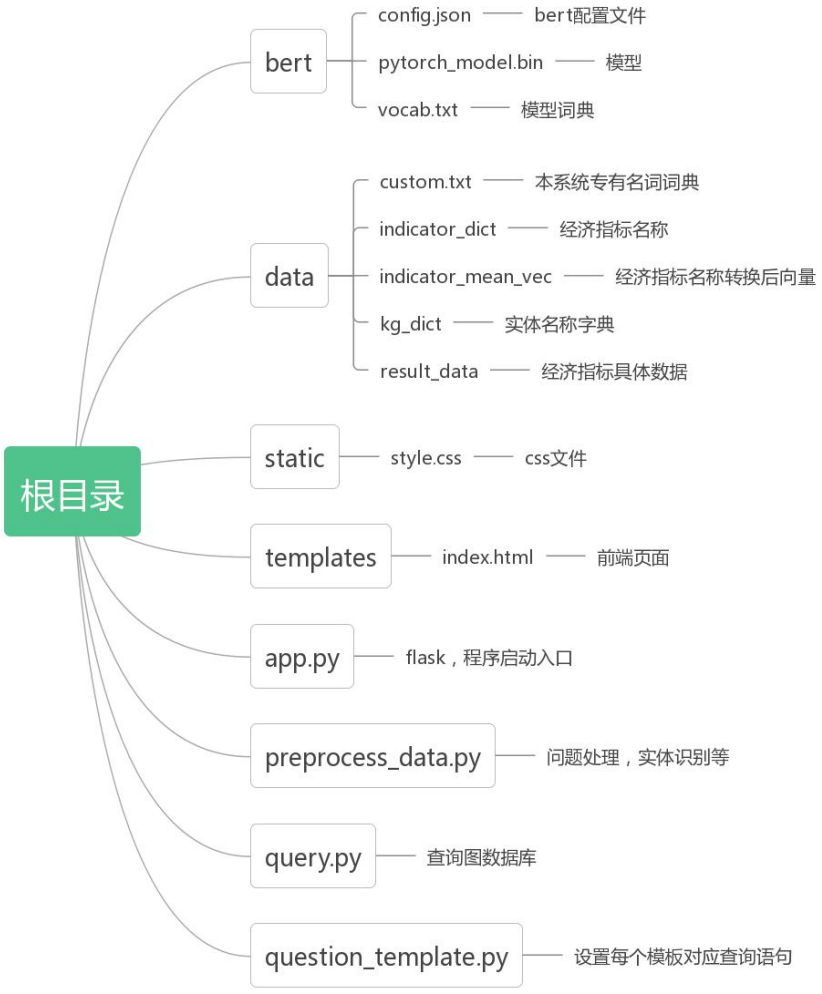
待改进点一：
速度较慢。**BERT**深度模型大，参数多，每次调用执行时都会比较慢；同时有部分数据是在线获取的，这也导致系统比较慢。

待改进点二：
直接使用预训练好的**BERT**虽然节省了训练时间，但由于模型没有针对我们的问答系统进行优化，可能导致问题识别率不够高。

07

PART

程序文件结构


根目录
bert
config.json
bert配置文件
pytorch_model.bin
模型
vocab.txt
模型词典
data
custom.txt
本系统专有名词词典
indicator_dict
经济指标名称
indicator_mean_vec
经济指标名称转换后向量
kg_dict
实体名称字典
result_data
经济指标具体数据
static
style.css
css文件
templates
index.html
前端页面
app.py
flask，程序启动入口
preprocess_data.py
问题处理，实体识别等
query.py
查询图数据库
question_template.py
设置每个模板对应查询语句