

组会汇报

陈钊杰
专业:计算数学

October 17, 2023

目录

1

代码调试

- ChatGLM大模型在金融领域的相关应用
- 序列分类任务
- k线图编码以后的进行降维可视化结果
- 下一步的计划

LSTM模型与自然语言模型进行序列分类对比实验

- ① 关于FinGLM问答系统的技术框架
- ② 关于预测，这周主要做了以下几个测试：
 - transformer模型,历史节点:25,单一模型建模
 - transformer模型添加lstm,历史节点:25,单一模型建模
 - lstm添加注意力的模型,历史节点:25,单一模型建模
 - transformer模型,历史节点:25,共享模型建模
 - lstm添加注意力的模型,历史节点:25,单一模型建模
 - transformer模型添加lstm,历史节点:25,单一模型建模
- ③ 使用Word2Vec方法对所有数据集中的字符进行训练以后，构造新的一些字符组合，并对每个字符的高维向量进行降维，然后进行可视化。其中按照频率的顺序，把高频率的放在最前面

任务简介

以ChatGLM2-6B模型为中心制作一个问答系统，回答用户的金融相关的问题

➤ 初级：数据基本查询（40分）

- 参赛者需要利用提供的ChatGLM2-6B开源模型和上市公司年报原始数据，并以此为基础创建信息问答系统。系统需能够解决基本查询，如：某公司2021年的研发费用是多少？等问题。

➤ 中级：数据统计分析查询（30分）

- 在初级阶段的基础上，参赛者需要进行金融数据的统计分析和关联指标查询。系统需基于各类指标，提供问题和答案，如：某公司2021年研发费用增长率为多少？等问题。

➤ 高级：开放性问题（30分）

- 如：某公司2021年主要研发项目是否涉及国家创新领域，如新能源技术、人工智能等？

解决问题示例

任务简介 - 评测指标

➤ 问题: 2019年中国工商银行财务费用是多少元?

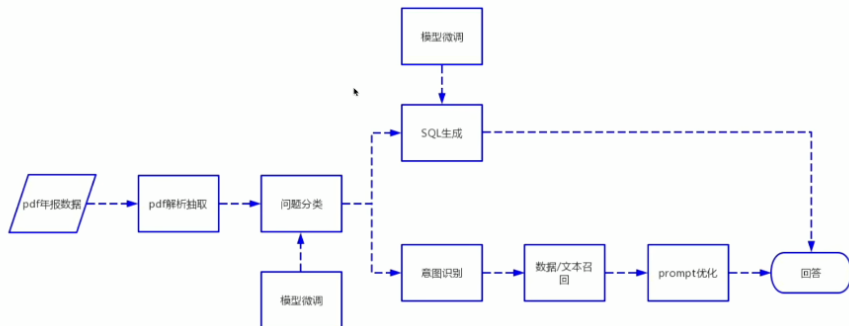
- **关键词:** 财务费用、2019
- **prompt:** "财务费用": "12345678.9元"
- **答案:** 12345678.9元
- **示例答案:** ["2019年中国工商银行财务费用是12345678.9元。", "2019年工商银行财务费用是12345678.9元。", "中国工商银行2019年的财务费用是12345678.9元。"]

➤ 评测计算示例:

- **答案一:** 工商银行2019年财务费用是12345678.9元。
- **most similar sentences:**
 - ✓ 2019年工商银行财务费用是12345678.9元。(Score: 0.9915)
 - ✓ 中国工商银行2019年的财务费用是12345678.9元。(Score: 0.9820)
 - ✓ 2019年中国工商银行财务费用是12345678.9元。(Score: 0.9720)
- **评分:** $0.25 + 0.25 + 0.9915 \times 0.5 = 0.9958$ 分。
- **评分解释:** prom_answer正确、包含所有key_word、相似度最高0.9915。

整体流程

方案介绍 - 整体流程



主要思路

- ① 对于问题分类中部分问题主要分为阅读理解题,计算题和信息检索题.
 - 通过微调语言模型,使其能够将问题进行编号归类
- ② 根据编号对应的问题类型选择合适的工具进行解决
 - ① 信息检索编号:
 - 比如问题如果是信息检索,尝试使用了NL2SQL系统(旨在将人类自然语言查询转换为数据库查询语言),但是效果不佳
 - 通过使用chatgpt协助生成各式的提问模板,得到训练集,然后再进行模型的微调,使得语言模型能以精确的SQL语言回答问题!
 - ② 其他意图识别:
 - 核心就是构建数据集,至于构建方法,他们通过提取问题的关键词方法进行回答对应的问题

Table: 准确率结果(无特殊说明历史数据点均选取为25)

模型	平均准确率
随机初始化词向量, 单一模型建模	27.26%
Word2Vec方法, 单一模型建模	27.36%
Word2Vec方法, 历史数据点:100, 单一模型建模	27.40%
Word2Vec方法, 共享模型建模	29.49%
chatglm模型, 历史数据点:25	32.34%
LSTM引入注意力机制, 单一模型建模	32.73%
LSTM引入注意力机制, 共享模型建模	34.56%
transformer模型, 单一模型建模	33.61%
transformer模型, 共享模型建模	34.55%
Transformer-LSTM联合模型, 单一模型建模	32.93%
Transformer-LSTM联合模型, 共享模型建模	31.61%



[C B D A]

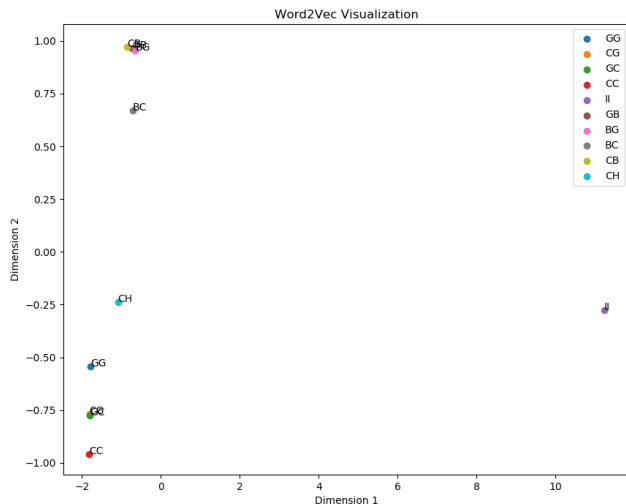


[G F H E]

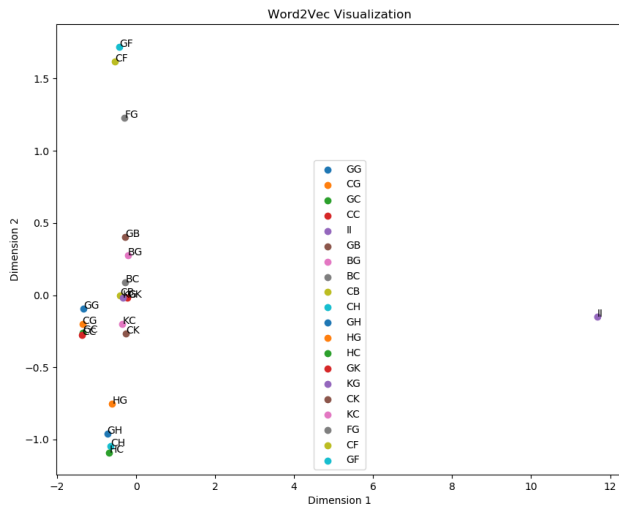


[I L J K]

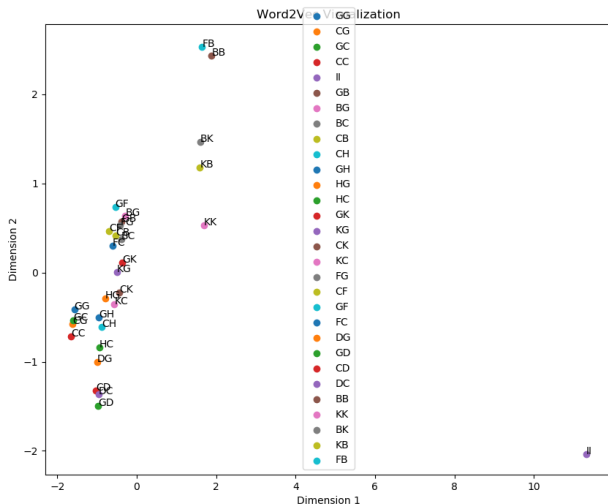
两字符组中数量占比前10的可视化结果



两字符组中数量占比前20的可视化结果



两字符组中数量占比前30的可视化结果



- 共享模型都准确度基本都高于单一模型的准确度
- LSTM模型引入注意力机制以后明显准确度有明显的上升，且超越了chatglm模型都准确度
- 在最高最低价接近的时候，开盘价和收盘价之间的变化紧密相关
- 从馒头科技的实验成果中可以看出语言模型都潜力很大，确实能够解决一些实际问题

下一步计划及相关问题

- 仔细查看FinGLM代码，学习如何联系数据库来完成检索任务

谢谢老师和同学们的聆听!