SpY第三次会议记录 3.19

参会人：袁晨 王柏霏 尹天禹 张瑞林 杜周纬 赵培炎

会议时间：3月19日 21:30-22:30

会议地点：腾讯会议

迭代一回顾与优化

总共多少数据返回

模糊搜索

- 找网上的资料

其他新闻页面的选择与爬取

- 换成自动爬取，就可以解决问题
- 需要url识别是否重复

迭代二讨论

> CICD

> 选取一种文本方法以实现新闻分类功能，可以仅使用新闻标题作为分类特征字 段，也可以结合新闻内容作为特征，类别细化为以下14个分类：财经、彩票、 房产、股票、家居、教育、科技、社会、时尚、时政、体育、星座、游戏、娱乐。将新闻的分类结果存入到数据库中“类别”的字段中。

使用大模型对新闻进行分类构造训练集，利用机器学习方法（随机森林，svm，Xgboost，朴素贝叶斯）对新闻进行分类

🌟 文本方法是指利用文本数据进行处理和分析的方法。在新闻分类任务中，文本方法是指使用文本数据（如新闻标题和/或新闻内容）作为输入，通过一系列的文本处理和机器学习技术来实现新闻分类的功能。下面介绍一种常见的文本方法来实现新闻分类：基于机器学习的文本分类方法。

基于机器学习的文本分类方法是指使用机器学习算法来训练分类模型，通过学习文本数据的特征和类别标签之间的关系，实现对新闻进行分类的功能。以下是一种基本的步骤：

1. 数据预处理：首先，需要对新闻数据进行预处理。这包括文本清洗（如去除标点符号、特殊字符等）、分词（将文本拆分为单词或词语）、去除停用词（如常见的无意义词语）等。

2. 特征提取：接下来，从预处理后的文本数据中提取有用的特征。在这种情况下，可以选择使用新闻标题作为分类特征。可以使用词袋模型（Bag-of-Words）或 TF-IDF（Term Frequency-Inverse Document Frequency）等技术将文本转换为数值表示。
3. 数据划分：将提取的特征和对应的类别标签划分为训练集和测试集。训练集用于训练分类模型，测试集用于评估分类模型的性能。
4. 模型训练和评估：选择适合的机器学习算法（如朴素贝叶斯、支持向量机、随机森林等）进行模型训练。使用训练集进行模型训练，并使用测试集评估模型的分类性能，如准确率、精确率、召回率等。
5. 分类预测：经过模型训练后，可以使用训练好的模型对新的新闻数据进行分类预测。将新闻数据输入模型，模型将输出预测的类别标签。
6. 结果存储：将分类结果存储到数据库中的相应字段，如"类别"字段。

值得注意的是，以上只是一种基本的文本分类方法，根据具体需求和数据集特点，还可以进行模型优化、特征工程和调参等进一步的改进。

- 大模型

> 对分类完成后的新闻进行数据分析，必需包括当日的新闻数量，每周的新闻数量，当日每个类别的新闻数量这三个指标，另需额外自拟三个其他指标。

- 每个类别的新闻相比昨日的变化量?

1. 关键词频率：识别并计算当日新闻中出现频率最高的关键词，以了解新闻报道的焦点和关注点。 1
    a. 需要爬取
    b. 词云
2. 新闻来源分布：分析当日新闻的来源分布，例如不同新闻机构或网站的报道比例，以了解新闻的多样性和报道的可靠性。
3. 地理分布：分析当日新闻报道涉及的地理位置分布，例如国家、城市或地区，以了解地域间的新闻报道差异和关注点。
    a. 使用地图
    b. 来自于实体识别
1. 新闻报道立场分析：对新闻报道进行立场分析，例如识别新闻报道中的倾向性或偏见，并计算正面、负面和中立报道的比例。新闻报道的媒体倾向性：评估不同媒体对新闻报道的倾向性或偏见，例如通过分析报道的词汇和语调来确定。
    a.
2. 人物关注度：识别并计算当日新闻报道中涉及的人物的关注度，例如政治领导人、商界名人或社会活动家。

> 实体识别： 通过分析新闻文本，可以识别出人物、地点、组织等实体，并在知 识图谱中以节点的形式表示。
>
> 关系抽取： 提取新闻中描述的实体之间的关系，例如合作关系、竞争关系、家 庭关系等，将这些关系以边的形式添加到知识图谱中。
>
> 将获取到的实体及其关系三元组存储到neo4j数据库中，使用ECharts等可视化工具将某篇新闻的实体关系图谱展示到前端界面上。
>
> 尝试使用大语言模型对新闻内容中的实体及其之间的关系进行提取。

> （加分点）事件提取： 从新闻中提取关键事件，并将这些事件表示为节点，同时通过边表示事件之间的时间序列关系，构建事件的时间线

动态自动爬取标准

- 每12h爬取一次

任务分工

1. 计划文档 张瑞林
2. 会议记录文档 分工情况 袁晨
3. 需求文档 王柏霏
4. 体系设计 张瑞林（mysql+ne04j）赵培炎(es)
5. 总共多少数据返回 袁晨
6. 模糊搜索 赵培炎
7. 买服务器 尹天禹 王柏霏
8. 后端CICD 尹天禹
    a. Spring、mysql、
    b. rabitmq
9. 前端CICD 王柏霏
10. 实体识别关系抽取 张瑞林 赵培炎 杜周纬
11. 做文本分类 杜周纬
12. 前端初步构建 袁晨

截止日期到3月26日

其他

课程项目申报