SpY第四次会议记录 3.25

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25日 15:00-16:00

会议地点：腾讯会议

迭代二进度分享

1. hanlp完成文本分类

- 类似形式情感分类

- 关键词提取

- 北京时间问题
- 选择统计关键词
- 统计情感
- 浙江大学实体关系研究
- https://github.com/zjunlp/DeepKE/blob/main/README_CNSCHEMA_CN.md
- 新闻摘要
  - 可以设置关键句数量

编号	中心句	权重
1	据DigiTimes报道，在上海疫情趋缓，防疫管控开始放松后，苹果供应商广达正在逐步恢复其中国工厂的MacBook产品生产。	100.00%
2	据供应链消息人士称，生产厂的订单拉动情况正在慢慢转强，这会提高MacBook Pro机型的供应量，并缩短苹果客户在过去几周所经历的延长交货时间。	44.86%
3	苹果上周表示，防疫措施和元部件短缺将继续使其难以生产足够的产品来满足消费者的强劲需求，这最终将影响苹果6月份的收入。	14.40%

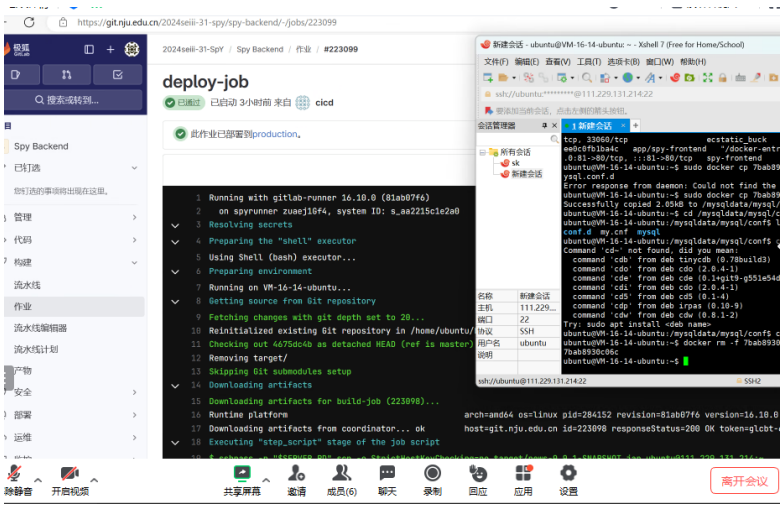
可视化

据DigiTimes报道，在上海疫情趋缓，防疫管控开始放松后，苹果供应商广达正在逐步恢复其中国工厂的MacBook产品生产。
据供应链消息人士称，生产厂的订单拉动情况正在慢慢转强，这会提高MacBook Pro机型的供应量，并缩短苹果客户在过去几周所经历的延长交货时间。仍有许多苹果笔记本用户在等待3月和4月订购的MacBook Pro机型到货，由于苹果的供应问题，他们的发货时间被大大推迟了。据分析师郭明錤表示，广达是高端MacBook Pro的唯一供应商，自防疫封控依赖，MacBook Pro大部分型号交货时间增加了三到五周，一些高端定制型号的MacBook Pro配置要到6月底到7月初才能交货。尽管MacBook Pro的生产逐渐恢复，但供应问题预计依然影响2022年第三季度的产品销售。苹果上周表示，防疫措施和元部件短缺将继续使其难以生产足够的产品来满足消费者的强劲需求，这最终将影响苹果6月份的收入。

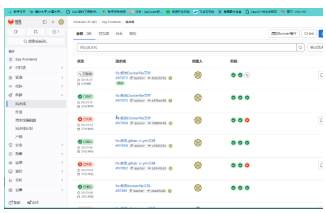
- deepke

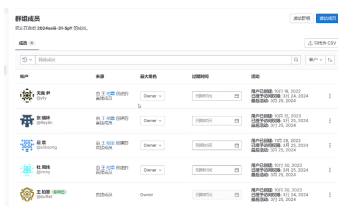
- 前端图表示例

- CICD
  - 做了镜像
  - 需要设为私有
  - 显示成功，但是数据库有点问题

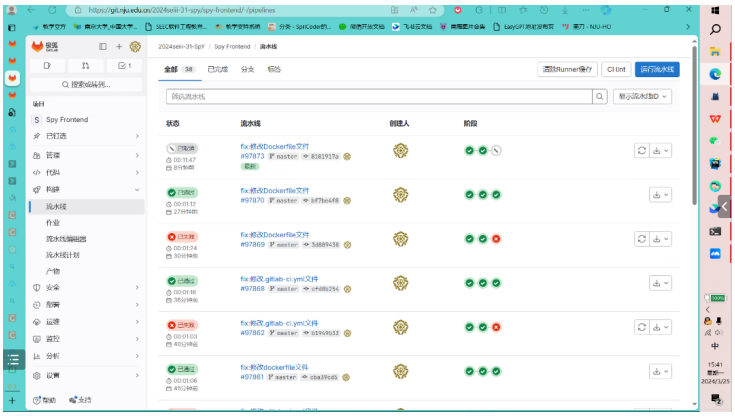


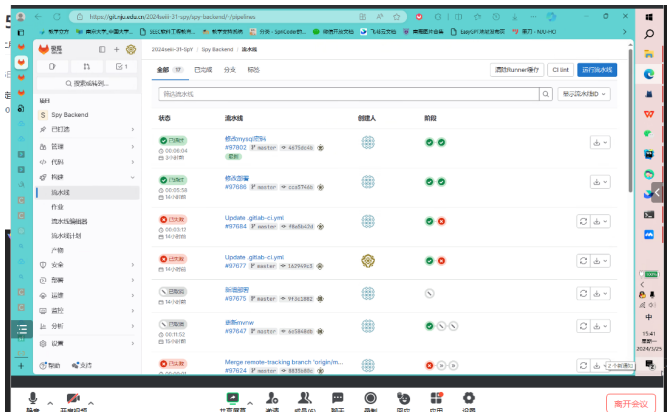
- 前端CICD
  - 尝试docker



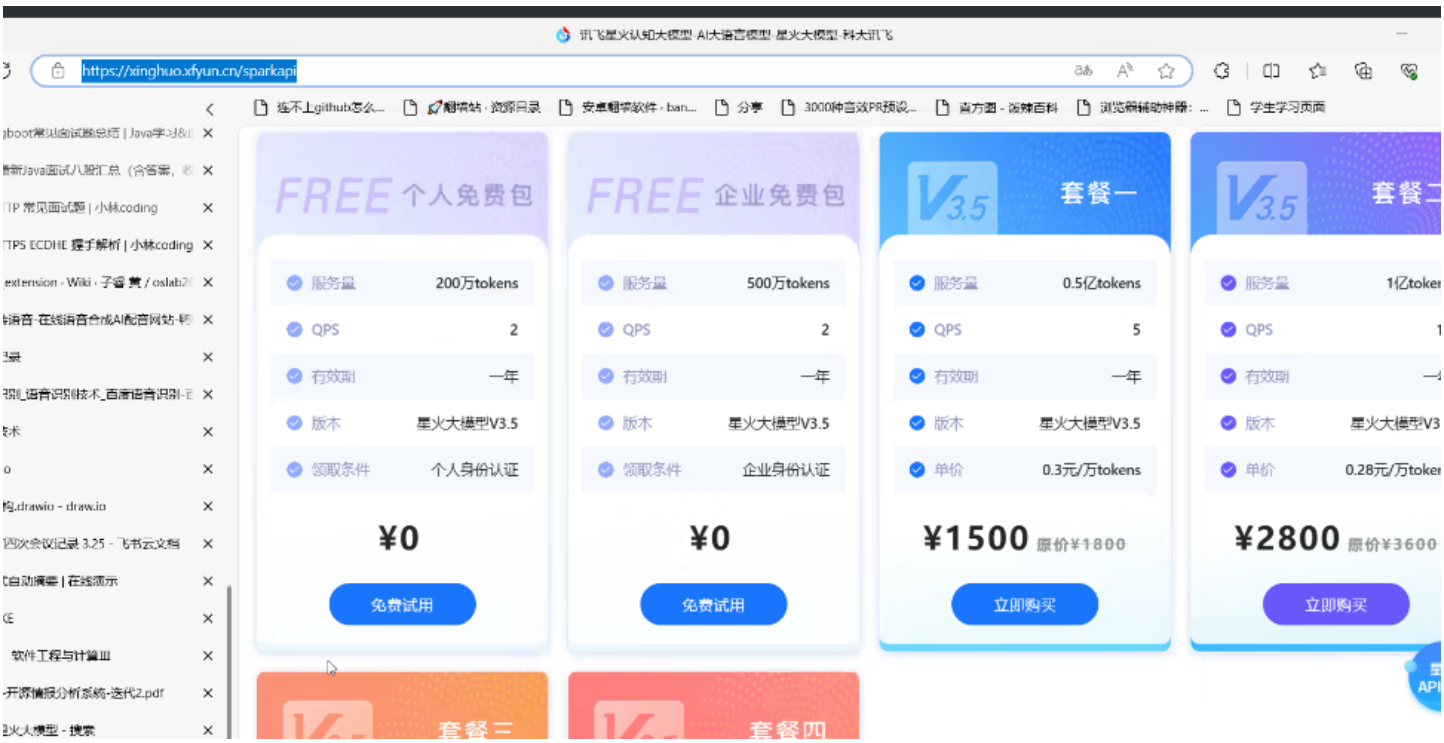



- CICD 成功截图





- neo4j数据库
  - 存结点和边
- ElasticSearch



https://xinghuo.xfyun.cn/sparkapi

迭代二下一步讨论

> 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。

> 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

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。1
3. 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。 2
4. 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。 2
    a. 使用地图
    b. 来自于实体识别
    c. 统计当日新闻，某个地点总共出现了多少次
    d. 可以只要求统计中国城市（国际）

> 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

任务分工

1. Elasticsearch模糊搜索 赵培炎
2. 连接讯飞大模型接口 杜周纬

3. 实体关系提取 杜周纬, prompt
    a. 系统
    b. 数据实际搞出来
4. Hanlp调用（https://hanlp.hankcs.com/demos/tok.html）java或python 尹天禹
    a. 数据实际搞出来
5. 接口文档和表修改 袁晨
6. 实现接口和表修改 杜周纬
7. 前端实现类别的新闻数量，昨日的变化量，当日的新闻数量，每周的新闻数量 王柏霏
8. 每日动态，新闻报道立场分析，关键词频率，实体关系 袁晨
9. neo4j存储节点和边，查询新闻实体节点和边 张瑞林

截止日期 4.1