2020年春季学期信息检索课程project

项目计划

针对美国2018年1月到2018年5月的财经新闻生成索引，以便由于搜索。

项目意义

1. 内容层面
    1. 更好地了解美国在2018年1月到2018年5月有关财经的重大事项
    2. 感兴趣的人也可以搜索有关特朗普等美国领导人的行踪，从而分析出特朗普在在短时间内的种种新闻，看看美国主流媒体对其的态度的变化，以及特朗普在在短时间内对美国经济等方面的影响。
    3. 对美国在2018年1月到2018年5月进行详细分析。黄仁宇在《万历十五年》一书中，仅仅通过分析明朝在万历十五年这一年的政治、经济的面貌就让读者对历史兴衰有了深刻了理解。我想，我们也可以通过对这五个月美国有关经济的新闻进行分析索引，对美国的体制和精神有一个较为清晰的认识，见微知著，进而理解美国当前的局势与种种现象。
2. 功能层面
    1. 包含若干个索引方法，包括布尔索引、临近检索属性检索等功能
    2. 有UI以便用户打开浏览器直接使用
3. 优化层面
    1. 将进行适当的缩减词库
    2. 可能对一些检索方式进行优化处理

选择的数据集及其链接

1. 数据集：US Financial News Articles
    1. 具体的数据来源：Bloomberg.com, CNBC.com, reuters.com, wsj.com, fortune.com
    2. 语言：English-only
    3. 存储形式：json文件
2. 链接： **https://www.kaggle.com/jeet2016/us-financial-news-articles**

实现方式

1. 通过SPIMI来分段加载数据
    1. 将其处理成倒排索引表
    2. 同时记录有关相应新闻的一些属性，如作者、标题等信息
2. 逐个实现一些检索方式
    1. 将实现布尔索引、临近检索、属性检索等检索方式
3. 设计用户界面（UI）
    1. 实现方式：HTML/css/JavaScript，有可能使用vue.js、bootstrap等成熟的前端框架
    2. 过程：用户输入关键词，选择检索方式，前端将信息发送给后端，然后后端返回结果
    3. 通信方式：WebSocket