1. 爬虫编写(知乎)
   1. py
   2. Scrapy
      1. 模拟登录
      2. xpath信息筛选
      3. 信息保存(数据库)

爬取内容:

1. 问题
   1. 问题ID (int)
   2. 问题名 (string)
   3. 问题所属分类ID列表 (int[])
   4. 问题所属分类ID列表对应名称列表 (string[])
   5. 问题详细说明 (string)
   6. 问题评论数目 (int)
      1. 选作 评论内容列表
   7. 问题关注者数量 (int)
   8. 问题被浏览次数 (int)
2. 答案
   1. 答案ID (int)
   2. 所属问题ID (int)
   3. 答案内容 (string)
   4. 答题者ID (int)
      1. 选作 答题者名称
      2. 选作 答题者自我介绍
   5. 答案点赞数 (int)
   6. 评论数目 (int)
      1. 选作 评论内容列表
   7. 发布或者修改时间
3. 信息提取
   1. 中文分词
      1. jieba
   2. 关键词提取
      1. tf-idf(基于频率)
      2. Topic Model(提取主题)
4. 模型设计
5. 模型训练
6. 验证
7. 软件开发