20190219进度报告

# 状态概述



总体进度有点落后.因为实际操作过程中发现配置服务器的工作,以及学习计划.因此计划产生变更.

# 已经完成的部分

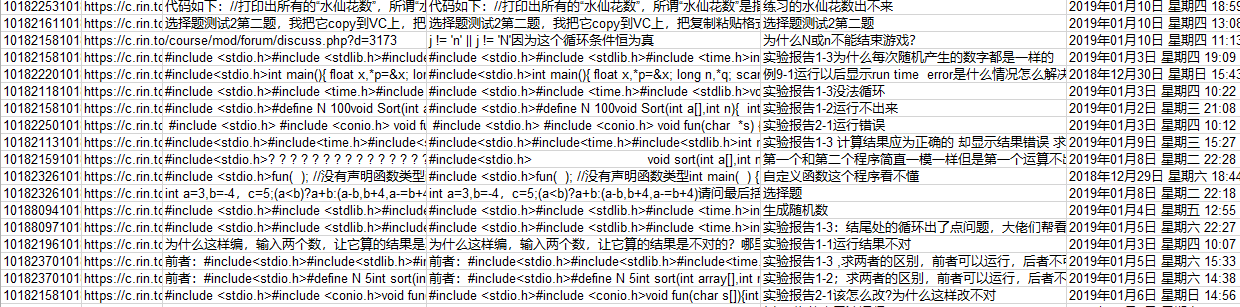
## 爬取教学网站”大学之学”的所有页面:

功能描述:用beautifulsoup4库自动将整个大学之学C语言程序设计的讨论区的全部内容爬下.

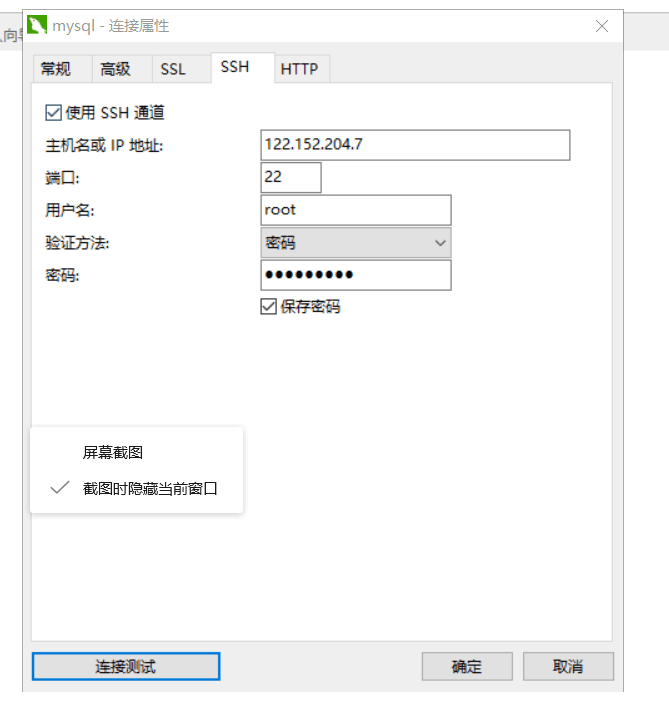
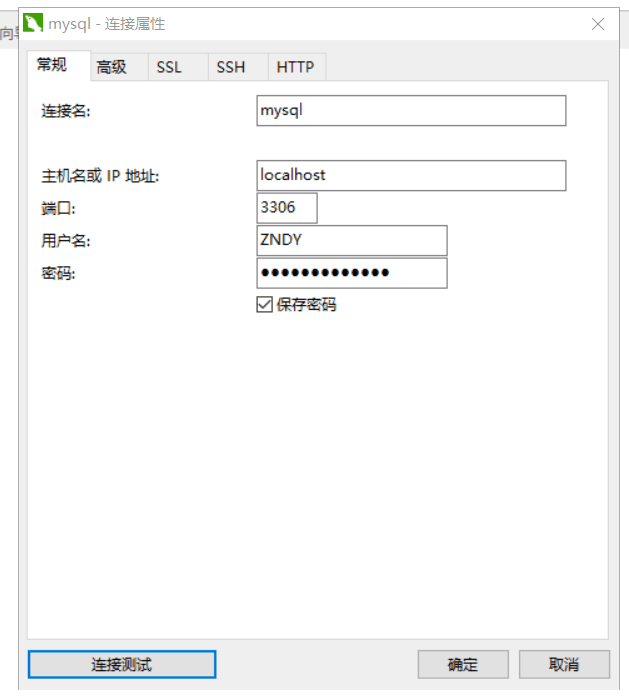
性能:1-2s一行

结果展示:

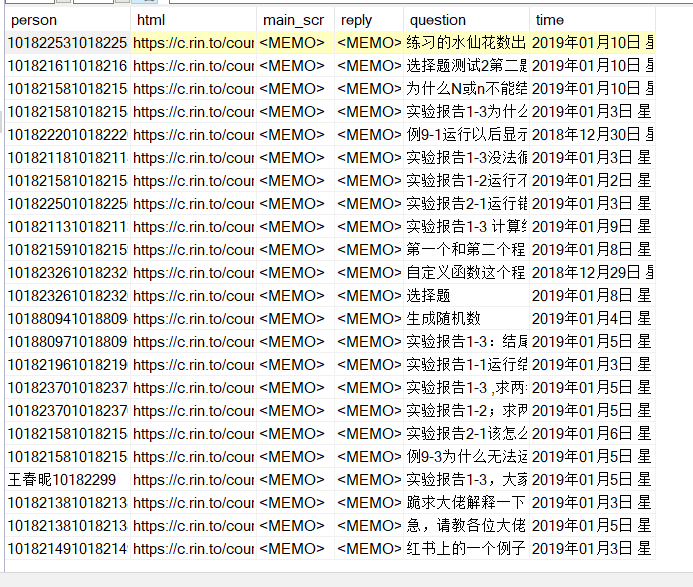
列名: 提问者 问题具体内容 回答文本 问题概述 提问时间



## MySQL数据库建立:

  
SSH密码zdx062625 数据库密码:ZNDY@ecust123

当前状态:



## 网络爬虫模块

爬取网站:百度知道 百度 知乎 CSDN

输入:检索关键词

输出:标题文本 摘要文本 URL

库:python的beautifulsoup4 Scrapy

已经写入服务器.

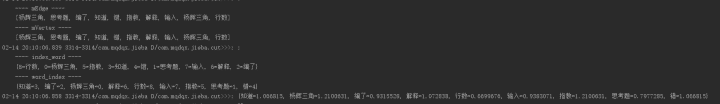
## 关键词提取模块

Java编写

算法:TextRank

输入:文本

输出:文本的分词 每个词的权值



性能:3s内完成

主要步骤:生成字典 生产结束字符表 计算词的发射率[[1]](#footnote-1) 用维特比算法找到概率最大的分词方法以及词性 用TextRank得到关键词的权值

[[2]](#footnote-2)

存在问题:

需要调节窗口大小

缺失文本预处理模块(接下来计划中解释)

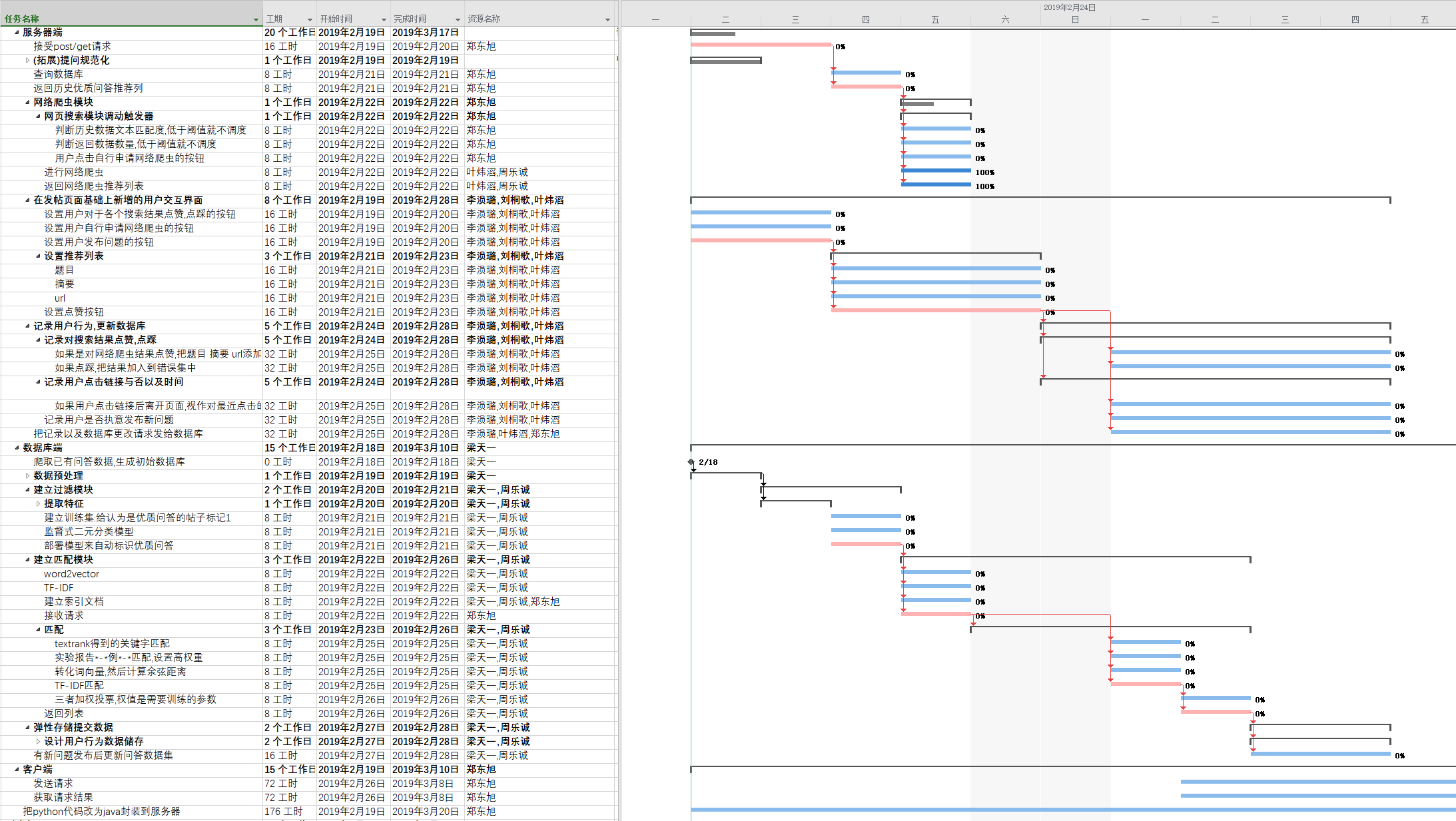
排序完选择的关键词数目

## 前端页面设计



虽然和计划的有点偏差,但是该模块可以用作问题发布时候使用.

# 计划变更



### 编程语言:

python>>python + java>>java(在服务器上都使用java)

### 数据库:

本地数据库>>阿里RDS的MySQL数据库

### 服务器:

处理服务器先是用centos,配python.然后ssl报错;然后换成Spring Boot框架,把python的爬虫模块用java写进去.

### 开发的工作分解结构(WBS)变更:



详细请查看html文件

解释:

1发现事情不是写写.py文件甚至在jupyter notebook操作就可以完成.现在需要面对的是服务器,MySQL数据库,网页客户端系统工程.因此变更整体框架.

2为了提高响应速度,改为java编程.因此对于python程序,需要添加”把python代码改为java封装到服务器”任务

3随着认识加深,更加深入领会nlp以及搜索引擎知识.因此添加:

文本预处理模块

匹配模块细化

# 技术困难:

## 索引文档的创建

目前索引文档要包含各个帖子所有词的TF-IDF值,词向量,特殊内容:例几.几/实验报告几.几/代码.

## 服务器框架问题

之前用centos,配python.然后ssl报错:

centos7.2服务器安装python-3.6.4

利用source编译安装，已安装openssl依赖

在执行服务端py时引入ssl，抛出no module named '\_ssl'

导致python只能转java编写.

1. 隐马尔科夫模型中中间隐含层状态转移到最外层可以观测到的状态的概率,采用统计词频的方式计算. [↑](#footnote-ref-1)
2. https://www.cnblogs.com/clover-siyecao/p/5726480.html [↑](#footnote-ref-2)