# 1、系统概述

**1.1、系统简介**

**提供可视化特定的社群信息（以股市资讯为例）的词频和日期关系，提供给用户发掘宏观事件之间驱动关系的辅助工具。通过统计分词并统计词频，并以时间为词频的关联特征生成词频与日期的二维统计图。另外生成制定关键词之间的简单卷积结果图表。以以上二者作为可视化信息辅助用户完成事件驱动的决策。**

**强调只是辅助决策，并非确定关联性。**

**1.2、系统目标**

功能目标：

1. **用户输入一段时间，给出可视化图表：**
   1. **一段时间内以最热为排序关键词，给出词频为与日期关系。**
   2. **一段时间内以热度变化最快为排序关键词，给出词频与日期关系**
   3. **一段时间内热度变化最快词频或最热词频的前5个之间两两比对，以相关性强弱排序给出比对两个词频单位样值响应。**
2. **用户根据自己判断，输入两个关键词，系统能反馈给出，两个词频各自和时间的关系和两个词频之间单位样值响应，共三个图。**

**1.3、系统运行环境**

Win10

MySQL数据库部署

**1.4、开发环境**

**Python 2.7.15  
MySQL 8.0带python connector  
Matlab 2018a带extern**

**其中Python的额外库文件**

unicodecsv

scrapy

docx

pdfminer

setuptools

MySQL-python

mysql

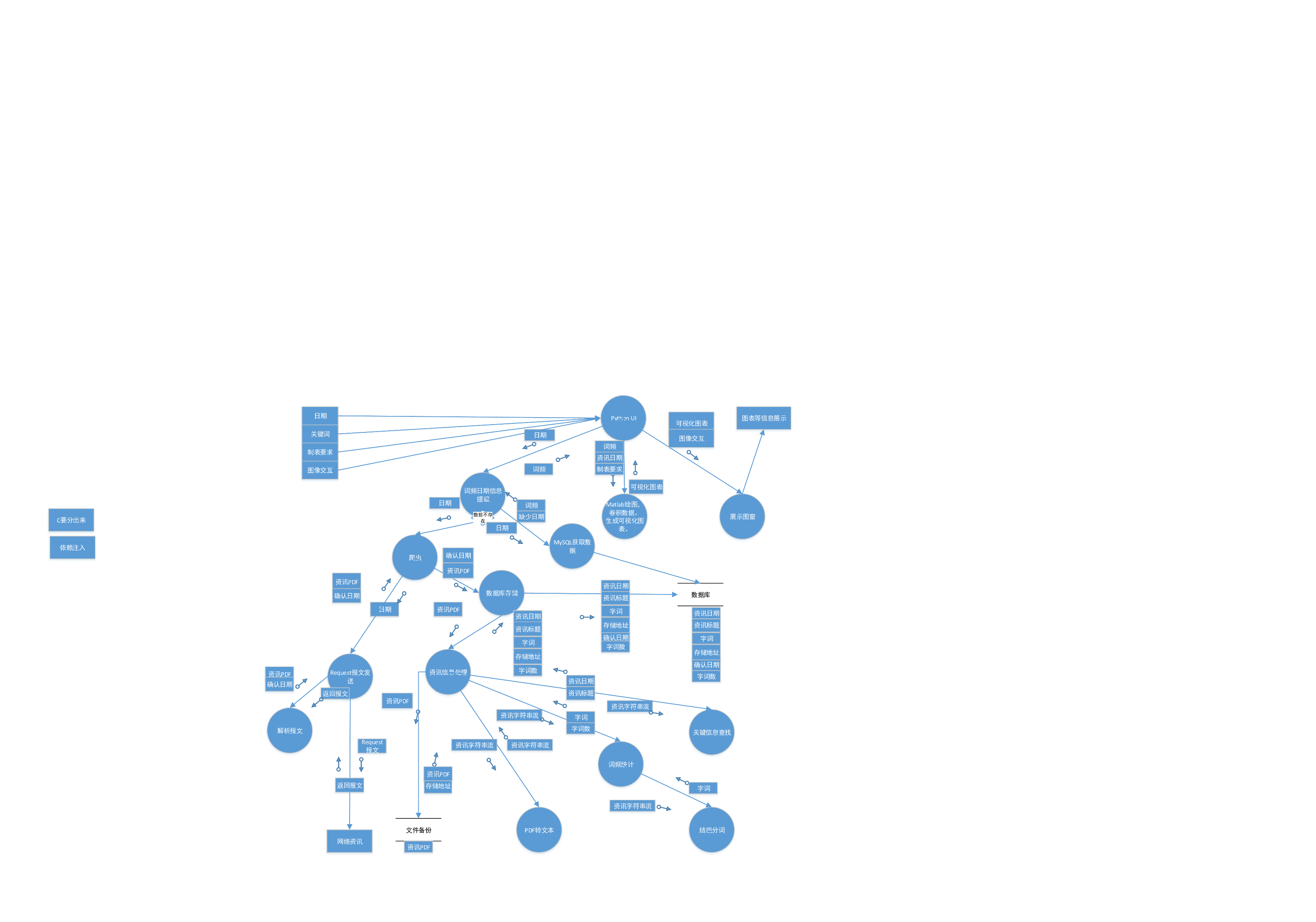
beautifulsoup4

request

# 2、总体结构设计

**2.1、软件结构**

**概要设计后的数据流图**

****

**2.2、设计思想**

**PythonUI是用GUI语言写的一个前台程序，可以捕捉用户操作，获取用户输入信息同时把获得的图表输出信息展示出来，并能被用户进行一定程度的操作。也作为主控函数进行如下步骤。**

**①根据用户输入和行为提供不同模式的展示信息。**

**②展示的信息来源于主控中预设的界面UI和图表信息。**

**③图表信息会从matlab对收集的数据进行处理，运算。**

**④.1收集的信息以词频信息为主，信息从数据库中获取。但是大多情况下，程序需要补充词频信息。**

**④.2通过调用爬虫函数获取新的资讯讯息。分为爬取，筛选，存储三步骤。**

**I）根据用户需求，发送报文并解析，实现资讯爬取，爬取完毕后会返回确认日期，告诉数据库当天资讯已经完整。**

**II）在筛选中预先存储资讯文件，顺便把资讯（pdf）转化为字符串，之后将字符串进行分词处理，之后对词语进行统计过滤。**

**III）最后将获取的词频信息打包存在数据库中。关注的信息只有日期和词语及其对应的统计次数。流图中大量的数据信息只是为后期提升提供支撑。**

**自此收集的信息的获取便完成。**

**另外，步骤④占用资源少，但是因为联网费时，所以独立分出线程。**

**程序中与Matlab和MySQL数据库对接时，采用线程通讯办法，即程序先打开二者的线程，然后通过二者嵌入式开发的指令与这两个线程通讯并完成操作以实现功能。**

# 3、模块设计

数据结构

|  |  |  |
| --- | --- | --- |
| 名称 | 含义 | 数据类型 |
| 日期① | 指定日期区间段，进行数据分析。 | string |
| 关键词② | 指定两个关键词，进行数据分析。 | string |
| 制表要求③ | 寄存图表的分析周期的当前状态。 | short |
| 图像交互④ | 用户的控制信息，包括移动窗口。 | 触发信号 |
| 资讯PDF⑤ | 大多资讯的原材料，以文件路径为标识。 | .PDF |
| Request报文⑥ | 请求数据的报文 | string |
| 返回报文⑦ | 请求数据后返回的报文 | string |
| 存储地址⑧ | 资讯最终存储地址 | string |
| 资讯字符串流⑨ | 资讯PDF转化后的字符串，作为分析的关键 | string |
| 词频⑩ | 收集的数据，作为数据分析的关键元素。  作为数组存在，其成员为日期，字词，字词数。 | [string，  string,  int] |
| 资讯日期⑾ | 当前统计字词来自的资讯的日期 | string |
| 字词数⑿ | 字词统计的数量 | int |
| 字词⒀ | 当前统计的字词 | string |
| 可视化图表⒁ | 数据处理后，并以规定制成的图表 | 图表 |
| 图表等信息展示⒂ | 最后展示给用户的画面 | NA |
| 缺少日期⒃ | 数据库的数据如果不全，会返回缺少数据的日期 | string |
| 确认日期⒄ | 当某日的资讯都抓取完毕，返回确认 | string |
| 资讯标题⒅ | 资讯标题 | string |
| 字词⒆ | 某个字词 | string |
| 字词数⒇ | 某个字词对应的字词数 | int |

模块

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 名称 | 功能 | 输入接口 | | | 输出接口 |
| PythonUI | 主控面板，交互窗口 | 由UI控件获得日期①  关键词②  制表要求③  图像交互④ | | | 向展示图窗输入  可视化图表⒁ |
|  | | | | | |
| 展示图窗 | 布局显示信息 | PythonUI输入  图像交互④  可视化图表⒁ | | | 图表等信息展示⒂展示在显示屏幕上 |
| Matlab数据绘图 | 对收集到的词频信息，并根据制表要求进行指定卷积等数据操作，最后生成图表 | PythonUI输入  词频⑩  制表要求③  *与matlab线程通信* | | | 返回  可视化图表⒁ |
| 词频日期信息提取  用这个标识子模块的关系top | 目的是获取词频信息，是模块Matlab数据绘图的 | PythonUI接入  日期①  MySQL获取数据返回:  词频⑩  缺少日期⒃  爬虫获取数据返回:  确认日期⒄ | | | 日期①输入爬虫  日期①输入  MySQL获取数据  词频⑩返回PythonUI |
| 词频日期信息提取模块与子模块的调用关系 | | | | | |
| MySQL获取  top-R | 数据库获取词频子模块  从数据库里提取信息 | | 词频日期信息提取接入 日期①  *接入 数据库数据服务* | 返回 词频⑩  返回 缺少日期⒃ | |
| 爬虫  top-L | 网络获取词频子模块  该模块相当于子控制模块，功能以其子模块执行  功能 获取信息，存储信息 | | 词频日期信息提取接入 日期① | 返回 确认日期⒄  *把获取的数据存储在数据库中* | |
|  | | | | | |
| Request报文发送  top-L-L | 爬虫子模块  获取网络资讯  本模块发送报文，  解析报文功能分离出来 | | 爬虫接入 日期①  *连接网络*  从网络资讯获取  返回报文⑦  从解析报文获取  资讯PDF⑤ | *连接网络*  *向外发送报文*  输入 返回报文⑦ 到解析报文模块  返回资讯PDF⑤ | |
| 解析报文  top-L-L-D 终端了 | 解析报文 | | Request报文发送接入资讯PDF⑤ | 返回 资讯PDF⑤ | |
| 数据库存储  top-L-R | 爬虫子模块  分析资讯PDF，将其解析为可以被使用的字符串 | | 爬虫接入  资讯PDF⑤  确认日期⒄  资讯信息处理返回  资讯日期⑾  资讯标题⒅  字词⒆  字词数⒇  存储地址⑧ | *输入数据库*  资讯日期⑾  资讯标题⒅  字词⒆  字词数⒇  存储地址⑧  确认日期⒄ | |
|  | | | | | |
| 资讯信息处理  top-L-R-D | PDF到字词的主控 | | 数据库存储  接入资讯PDF⑤  从文件备份获取  存储地址⑧  从PDF转文本获取  资讯字符串流⑨  从词频统计获取  字词⒆  字词数⒇  从关键信息查找获取  资讯日期⑾  资讯标题⒅ | 返回  存储地址⑧  资讯字符串流⑨  字词⒆  字词数⒇  资讯日期⑾  资讯标题⒅ | |
| 文件备份  top-L-R-D-1 | 备份文件 | | 资讯信息处理  接入资讯PDF⑤ | 完成存储在指定文件夹  返回存储地址⑧ | |
| PDF转文本  top-L-R-D-2 | PDF转化为字符串 | | 资讯信息处理  接入资讯PDF⑤ | 返回资讯字符串流⑨ | |
| 词频统计  top-L-R-D-3 | 把字符串的字词，和字词数完成统计 | | 资讯信息处理  接入资讯字符串流⑨  从结巴分词获取  字词⒆ | 返回  字词⒆  字词数⒇  输入结巴分词  资讯字符串流⑨ | |
| 关键信息查找  top-L-R-D-4 | 提取资讯日期资讯标题 | | 资讯信息处理  接入资讯字符串流⑨ | 返回  资讯日期⑾  资讯标题⒅ | |
| 结巴分词  top-L-R-D-3-D | 资讯字符串流切分成字词 | | 词频统计接入  资讯字符串流⑨ | 返回  字词⒆ | |

# 4、数据库与数据结构设计

**4.1、 数据库及数据表**

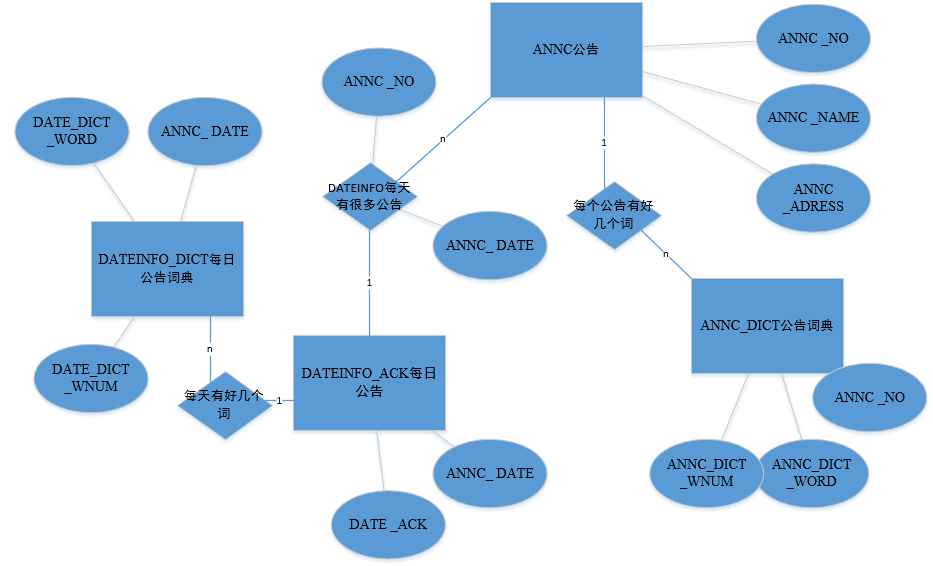
数据库名：STOCK

数据表：

|  |  |  |  |
| --- | --- | --- | --- |
| 表一：DATEINFO（含义Date每天的信息） | | | |
| 名称 | 含义 | 数据类型 | 存储长度 |
| ANNC\_ DATE | 该公告的日期 | DATE | 3 |
| ANNC \_NO | 该公告标号 | INT UNSIGNED | 4 |
| 表二：DATEINFO\_DICT（含义Date Dict 每天的词典信息） | | | |
| ANNC\_ DATE | 该公告的日期 | DATE | 3 |
| DATE\_DICT \_WORD | 该日期的某个词 | VARCHAR | 32 |
| DATE\_DICT \_WNUM | 该日期的某个词的数量 | INT UNSIGNED | 4 |
| 表三:DATEINFO\_ACK（含义：确认每天的信息是否收全） | | | |
| ANNC\_ DATE | 该公告的日期 | DATE | 3 |
| DATE \_ACK | 确认日期，意味着今日的公告爬取完了 | SMALLINT | 1 |
| 表四：ANNC（含义Announcement每个公告信息） | | | |
| 名称 | 含义 | 数据类型 | 存储长度 |
| ANNC \_NO | 该公告标号 | INT UNSIGNED | 4 |
| ANNC \_NAME | 该公告名字 | VARCHAR | 255 |
| ANNC \_ADRESS | 该公告的地址 | VARCHAR | 255 |
| 表五：ANNC\_DICT（含义Announcement Dict每个公告词典信息） | | | |
| ANNC \_NO | 该公告标号 | INT UNSIGNED | 4 |
| ANNC\_DICT \_WORD | 该公告的某个词 | VARCHAR | 32 |
| ANNC\_DICT \_WNUM | 该公告的某个词的数量 | INT UNSIGNED | 4 |

下划线的属性是主码

**4.2、 数据结构设计**



**4.3、 数据存储设计**

存储长度，类型见上表格，

存储位置、保密方面不考虑

时间效率暂不考虑，如果有同学想改进，只需要先考虑把字词标号，可以大幅缩短检索速度。

# 5、接口设计

**5.1、外部接口**

**MySQL数据库服务**

**Matlab线程**

键盘输入，鼠标控制，显示屏显示。

网络连接及其协议。

**5.2、内部接口**

**特别说明：对输入数据库的接口在“数据库存储”模块**

**数据库输出接口只在“MySQL获取数据”模块**

**网络接口只在“Request发送报文”模块中**

其余见数据流图和模块设计

# 6、其他设计

兼容处理，设置config文件，让用户可以配置自己的数据库信息

规定：汉字处理统一用UTF-8编码

存在多线程处理：

线程1:UI监听交互模块+Matlab数据处理模块+图像展示模块

线程2：词频日期信息提取模块

独立线程：1.Matlab线程

2.MySQL服务