1、任务概述

**1.1、目标**

**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。**

**提供统计分词并统计词频，以时间为特征信息的方法。**

**提供可视化信息以辅助用户完成事件驱动的决策。**

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，两个词频和时间的关系和两个词频单位样值响应，三个图。**

**希望作为数据挖掘的学习跳板。对大量数据流有有效操作行为。所以要求注意数据的访问性，程序可读性，兼容性。**

**1.2系统（或用户）的特点**

**系统特点：**

* **输入量较大，数据处理量大，须部署数据库。**
* **需要联网以获取资讯。**

**用户特点：**

**通常使用频度低，成周期化，即一周一次，一个月一次。面向人群多为政治，经济专业。好奇心重的人是本软件的最终用户，因此要赋予用户能自行处理数据的权利，即注意数据的访问性，程序可读性，兼容性。**

**1.3、假定和约束**

**1.开发工时大约80h,个人优化能力有限，偏重人机交互，相对舍去反馈信息的专业度**

**2.最终用户不为专业人士，弱化详细的数据处理为针对性数据处理。好奇心重的人是本软件的最终用户。**

**3.通过环境安装导读文本，间接提高兼容性。**

**4.以事件驱动明显的股市作为实现实例**

**5.默认用户偏好周期为一天，使用频次为一天一次，使用预期时间为5分钟。**

**6.默认使用用户使用PC端进行交互。**

2、需求规定

**2.1、软件功能说明**

**给予用户可视化信息。**

**具体功能有两大部分：**

1. **用户输入一段时间，用户输入时间区间能获取该范围的资讯，处理后给给出可视化图表：**
   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。（用户可以调整周期为，一天，三天，一周，一个月，不同周期对应的日期区间要调研）**
   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（用户可以调整周期为，一天，三天，一周，一个月，不同周期对应的日期区间要调研。同时最快的定义为：周期时间内词频之和，取matlab三次样条插值的模拟斜率乘以变化量乘以log2基数）**
   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，根据单位样值响应的性质，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，两个词频和时间的关系（周期可调），两个词频单位样值响应。**
3. **用户可以自己查自己想要的关键词的词频与日期的关系图。**

**细节： 数据库存储，用户可以自行确认数据流向**

**词频为与日期关系图，时间区间可以调节**

**友好的人机交互**

**进阶方面：数据预处理(以下功能不在实现范围内)**

**如果更换爬取信息，用户可以较快服用词频统计及处理功能。**

**理想功能：软件有一定的记忆功能和学习能力。**

**2.2、对功能的一般性规定**

**与时间有关的图表：可以调整区间长度**

**词频最快的定义为：[周期时间内词频之和，取matlab三次样条插值的模拟斜率]\*[变化量]\*log2[基数]**

**词频最热定义为：[周期内词频]/[某倍数周期内的词频数]**

**2.3、用户界面**

**Python的图形交互页面。**

**控制执行的控件：按钮:功能1“词频分析”；功能2“关键词联系挖掘”；修改排序方式的触发按钮。**

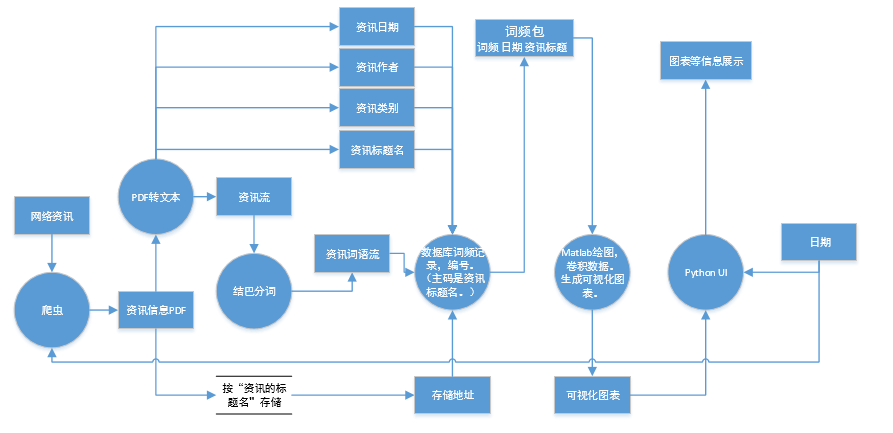
**滑块：调节时间区间长度的滑块；浏览滑动页面的滑块。**

**图层：放置生成的可视化图表的图层；放置调试信息，方便调试，方便提供用户提示的图层。**

**可能会制作菜单**

**2.4、对性能的一般性规定**

数据流图



**2.4.1、精度**

**输入的日期精确到日(可能会修改到小时)。**

**爬取的对象多为pdf和doc。**

**2.4.2、时间特性要求**

**爬取并统计时间不限**

**用户主动获取关键词词频关系图表，应在5s内获得响应**

**2.4.3、灵活性**

**只要系统环境正确，软件里的异常处理能保证程序能继续执行。**

**2.4.4、输入输出要求**

|  |  |  |
| --- | --- | --- |
|  | **输入1** | **输出1** |
| **数据名** | **日期** | **时间区间内数据分析后的可视化图表** |
| **数据类型** | **字符串** | **折线图** |
| **格式** | **YYYYMMDD** | **纵轴为词频数，横轴为日期，表头为对应词语，纵轴轴下限取0，横轴取长度为周期15倍** |
| **数值范围** | **20000101到[今日]** | **N/A** |
| **精度** | **日** | **N/A** |
| **示例** | **20170101 [20170331] 然后用触发按钮“词频分析”控制输出**  **然后可以通过交互按钮来调节时间区间，设定热度周期，选择排序标准（最热，热度变化最快，最有关系）** | **此图不为最终结果，仅供参考**  **按热度变化或热度最高筛选出的表，按定义高低排序，以下示例了3表排序**  **按定义判断出的最有关系的两个关键词进行处理显示，以下为一个图，理论是多个图排序放置** |
|  | **输入2** | **输出2** |
| **数据名** | **关键词** | **时间区间内数据分析后的可视化图表** |
| **数据类型** | **字符串** | **折线图** |
| **格式** | **已知词语** | **纵轴为词频数，横轴为日期，表头为对应词语，纵轴轴下限取0，横轴取长度为周期15倍** |
| **数值范围** | **N/A** | **N/A** |
| **精度** | **N/A** | **N/A** |
| **示例** | **中日韩自贸区钓鱼岛然后用触发按钮“关键词联系挖掘”控制输出**  **然后可以通过交互按钮来调节时间区间，热度周期** | **此图不为最终结果，仅供参考**  **两个关键词进行处理显示** |

**2.5、数据管理能力要求**

**存储空间推荐需求2GB，运存2GB，1MB/s的网速。**

**2.6、故障处理要求**

**爬取文件，没有返回值。提示用户确认网络。**

**数组里的数据过多导致越界。设置堆栈大小常量，异常时直接把常量翻倍。**

**2.7、其他专门要求**

**以方便用户使用为主(可补充性、易读性、可靠性)，适当舍去算法的实用性。**

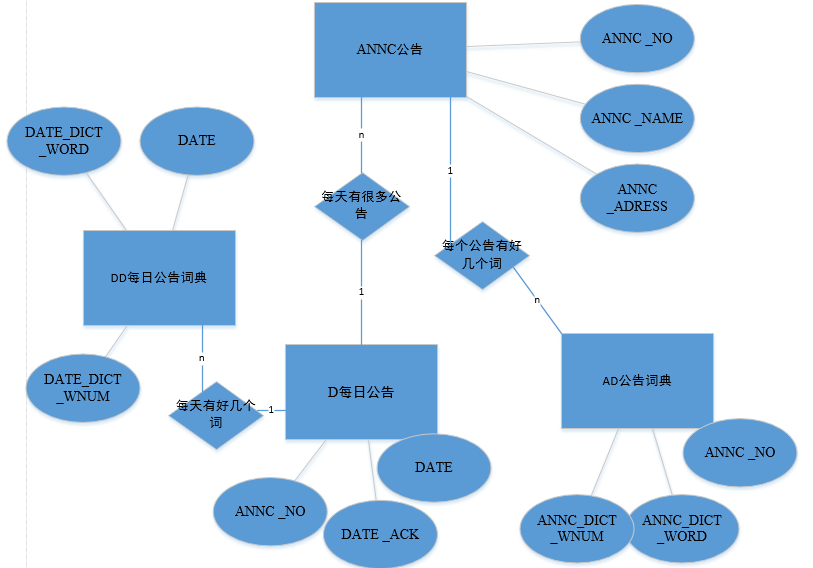
**数据库与数据结构设计**

数据表：数据表：

|  |  |  |  |
| --- | --- | --- | --- |
| 表一：DATE（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 |
| DATE \_ACK | 确认日期，意味着今日的公告爬取完了 | SMALLINT | 1 |
| 表二：DATE\_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：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 |
| 表四：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

**E-R图**



3、运行环境规定

**3.1、设备**

**Windows10及基本外设**

**3.2、支撑软件**

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

**3.3、接口**

**安装MySQL 8.0自带python connector  
Matlab 2018a自带extern**

**使用Pythonimport MySQLdbimportMatlab**

**这样MySQL和Matlab的各种方法，都可以在python程序中实现。**

**程序中，python爬取数据并分词然后通过MySQL存在数据库中，MySQL作为数据存储中心同时作为数据筛选器，python在用户操作后用MySQL筛选数据，经过运算同时调用Matlab库函数实现高等数学算法。**

**3.4、控制**

**键盘输入日期**

**鼠标点击操作交互页面，实现预设功能**

# 4、尚需解决的问题

* **词频最快的定义为：[周期时间内词频之和，取matlab三次样条插值的模拟斜率]\*[变化量]\*log2[基数]。该定义需要优化。**
* **不同周期对应的日期区间要调研。**