1. **科学文献管理系统（难度★★★★）**

科研工作者的日常工作离不开查阅科学文献，并对其中的信息进行分析、筛选、挖掘和管理。请你为科研工作者设计一个管理系统，提高科学文献的管理效率。

* **目标用户**：科研工作者。
* **数据配置**：请通过以下方法下载数据文件dblp.xml.gz.

<http://dblp.uni-trier.de/xml/dblp.xml.gz>

将该数据文件解压后，其中包含一个dblp.xml文件。该文件由科学文献的记录序列组成，记录的格式如下所示：

<article mdate="2002-01-03" key="persons/Codd71a">

<author>E. F. Codd</author>

<title>Further Normalization of the Data Base Relational Model.</title>

<journal>IBM Research Report, San Jose, California</journal>

<volume>RJ909</volume>

<month>August</month>

<year>1971</year>

<cdrom>ibmTR/rj909.pdf</cdrom>

<ee>db/labs/ibm/RJ909.html</ee>

</article>

每个记录对应一篇文章，其中包含对作者，题名，发表杂志，卷号，出版时间等的详细说明。请基于该数据，设计能满足后述功能的文献管理系统。

**注：1）dblp.xml的大小超过1G，所以不要直接点击打开该文件。需要通过命令行命令’more’ 或者自行编程查看。**

**2） 编程过程中，不允许使用数据库系统。需要自己建立管理文献数据的数据结构。**

* **功能要求**：

F1. （必须实现） 基本搜索功能。输入作者名，能展示该作者发表的所有论文信息。输入完整的论文的题目，能展示该论文的其他相关信息

F2. （必须实现） 相关搜索。输入作者名，能展示于该作者有合作关系的其他所以作者。

F2. （必须实现） 作者统计功能。输出写文章最多的前100名作者。

F3. (扩展实现) 热点分析功能。分析每一年发表的文章中，题目所包含的单词中，出现频率排名前10的关键词。

F4. (扩展实现) 部分匹配搜索功能。给定若干个关键字，能快速搜索到题目中包含该关键字的文章信息

F5. (扩展实现) 聚团分析。作者之间的合作关系可以看成是一个图，每个作者对应一个顶点，任两个作者之间如果存在合作关系，则在两个顶点之间建立连边。这个图中的每一个完全子图我们称为一个聚团（所谓完全子图指的是该子图的任意顶点都和该子图的其他顶点有连边，完全子图的顶点个数称为该完全子图的阶数），请统计整个图中各阶完全子图的个数。

F6. (扩展实现) 可视化显示。通过图形化界面，展示作者之间合作关系图及其相关文章信息。

F1-F2

solving：可以在初始化的时候使用sax解析文件，该xml文件使用iso-8895-1编码，一个字符大小等于一个字节。一、获取每个记录的title和文件位置偏移量记录在Hashtable中，文件偏移量=字节数，偏移量可以在使用RandomAccessFile读取的时候边判断边读取；二、同时获取每个作者和论文信息，也记录在另一个Hashtable中，键是作者名，值是一个偏移量链表（后续根据偏移量可以迅速获取论文的全部信息），而且根据这个Long类型链表，也就可以根据size()方法快速统计该作者发表的论文数，也就很容易统计发表论文数前100的作者；三、获取作者与作者之间的关系，存储在无向图中；基于链表的无向图，每个作者一个链表，该作者有几个作者就有几个链表节点，每个链表节点就存储合作者的姓名以及他们合作的论文题目，形式类似于：

假如作者1、2、3、4、5一起写了论文A，0、1一起写了B，那么就是

0-〉1B

1-〉0B-〉2A-〉3A-〉4A-〉5A

2-〉1A-〉3A-〉4A-〉5A

3-〉……

4-〉……

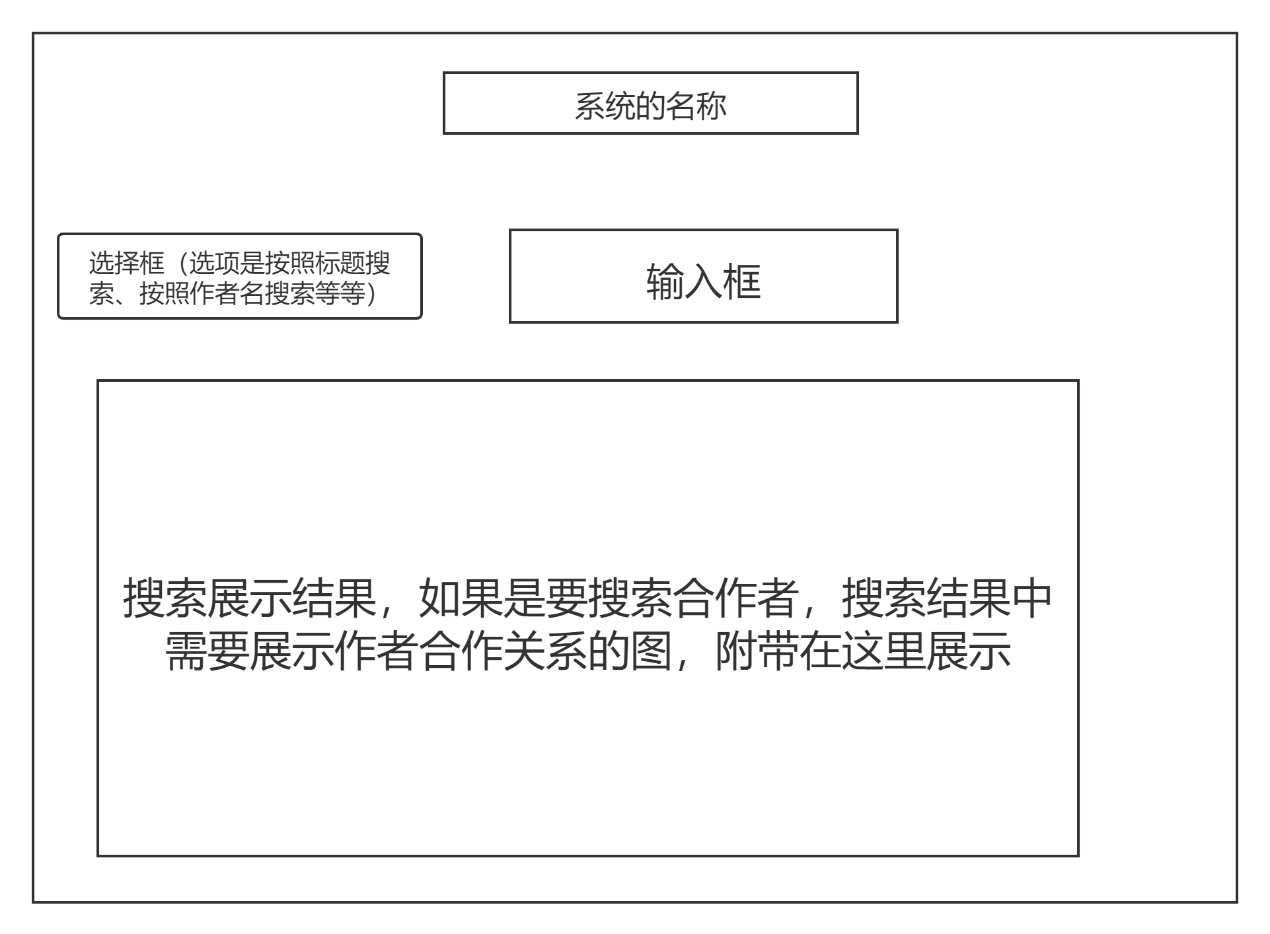
5-〉……

存title的话，就是可以完善些功能。查找作者1的时候，就可以查询出他跟作者2有合作，合作的论文是A，然后又可以根据这个题目A，查找出他们合作的这个论文的全部信息。

考虑在链表上查找作者的时候可能太慢，这里可以在外层再封装一个hashtable，键存作者名，值存链表的引用。

根据偏移量（偏移字节数）快速定位跳转到相应位置读取文件可以用RandomAccessFile类的api实现。

可视化实现部分考虑swing或者网页的形式展示



F3-F5 等待……