# 使用说明书

目录

[使用说明书 1](#_Toc16595)

[产品亮点 3](#_Toc8146)

[常见问题 3](#_Toc8462)

[功能点 3](#_Toc2908)

[1. 文件 4](#_Toc26721)

[1.1. 数据文件自动发现 4](#_Toc3001)

[1.2. 常见格式，双击运行 4](#_Toc2029)

[1.3. 特殊格式，单独解析 4](#_Toc24653)

[1.4. 多个文件，一起解析 4](#_Toc12585)

[1.5. 解析 5](#_Toc11724)

[1.6. 保存 5](#_Toc12974)

[2. 数据集 5](#_Toc6745)

[3. 表格 6](#_Toc3155)

[3.1. 表头排序 6](#_Toc22441)

[4. 编辑 7](#_Toc32074)

[4.1. 撤回 7](#_Toc29485)

[4.2. 恢复 7](#_Toc9252)

[4.3. 删除行 7](#_Toc17547)

[4.4. 删除列 7](#_Toc18400)

[4.5. 重命名 7](#_Toc21749)

[4.6. 复制列 8](#_Toc23146)

[4.7. 拆分列 8](#_Toc11804)

[4.8. 替换值 10](#_Toc2114)

[4.9. 比较列 12](#_Toc13775)

[4.10. 修改值 12](#_Toc1563)

[4.11. 合并文件 12](#_Toc9099)

[5. 清洗 13](#_Toc5299)

[5.1. 元数据 13](#_Toc18304)

[5.2. 行去重 15](#_Toc13369)

[5.3. 合并词 16](#_Toc22239)

[5.4. 停用词 16](#_Toc3055)

[5.5. 过滤行 17](#_Toc20327)

[6. 分析 17](#_Toc28664)

[6.1. 切分词 17](#_Toc21643)

[6.2. 词频统计 17](#_Toc32364)

[6.3. 共现分析 18](#_Toc15023)

[6.4. 相似比较 19](#_Toc22051)

[6.5. 分组统计 20](#_Toc22850)

[6.6. 特征提取 20](#_Toc15584)

[6.7. 图表配置 20](#_Toc17096)

[7. 窗口 20](#_Toc30652)

[7.1. 保存布局 20](#_Toc14860)

[7.2. 恢复布局 20](#_Toc28675)

[8. 参数配置 20](#_Toc13847)

[9. 操作日志 20](#_Toc20451)

[10. 帮助 20](#_Toc4762)

# 产品亮点

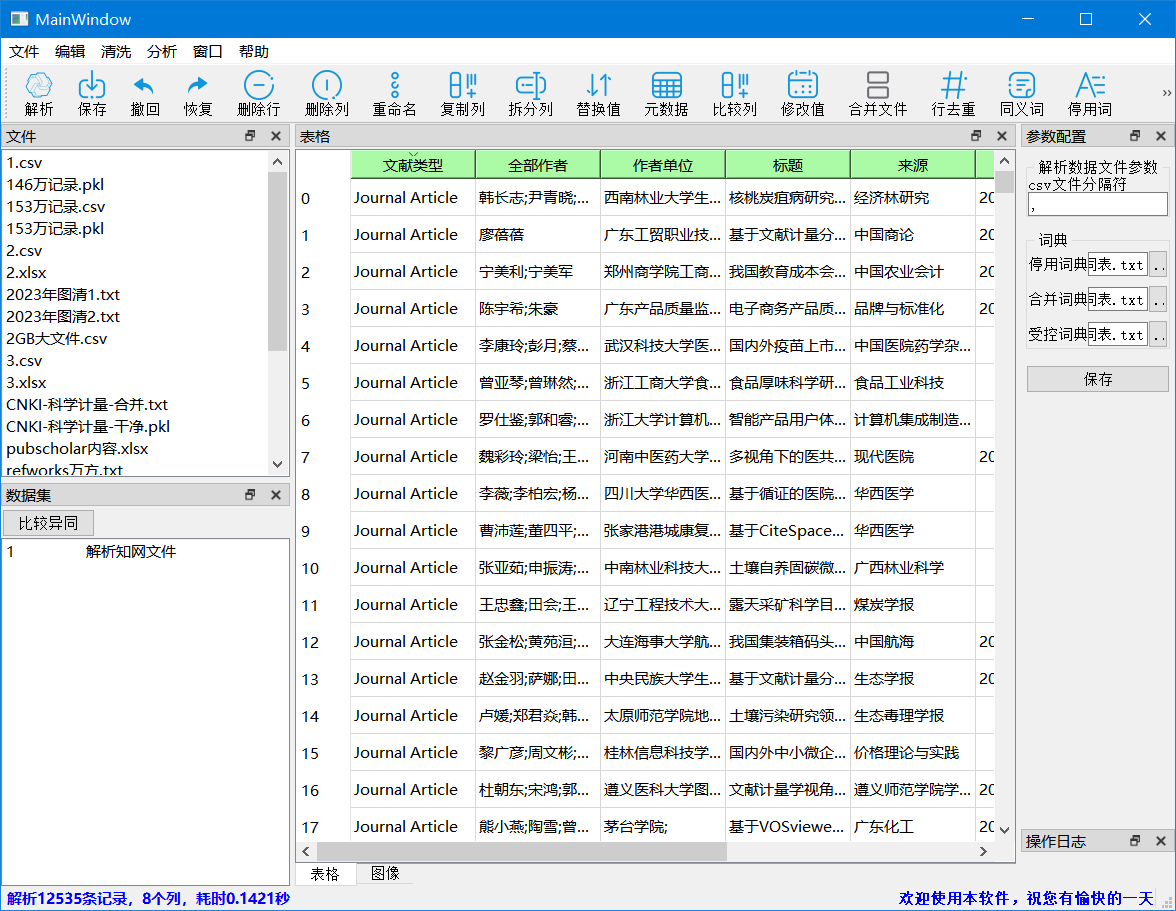
1. 支持各种自定义的文件格式，比如自己整理的文本文件、excel文件。
2. 能够处理大的数据量，在32位操作系统下，几秒钟内完成1GB文件或者100多万条记录的处理。常见的处理操作，在毫秒完成。
3. 适合对文本内容进行清洗、处理、分析。

# 常见问题

1. 数据文件放在哪里？ 答：放到datafiles文件夹中，软件会自动发现添加的文件。
2. 能否处理同构或异构的数据？ 答：可以。点击共现分析。

# 功能点

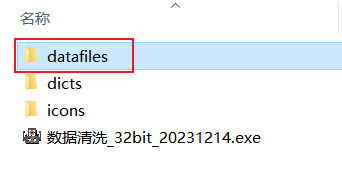
下面介绍的所有操作，都会在左下角提示。蓝色字体表示提示信息。红色字体表示错误信息。



## 文件

### **数据文件自动发现**

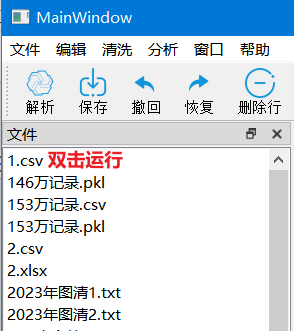
在datafiles文件夹中添加文件、删除文件、修改文件后，软件的数据文件列表自动更新。



### **常见格式，双击运行**

对于文件后缀是csv、xls、pkl的文件，双击就可以显示在右侧表格中。

自己制作的csv文件、excel文件，也可以双击时解析。



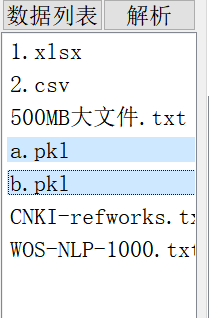
### **特殊格式，单独解析**

需要注意的是：知网、维普、万方、知网专利，需要下载refworks格式，不解析其他格式。

对于知网、维普、万方、wos等txt格式，需要单独解析。点击上面的解析按钮。

### **多个文件，一起解析**

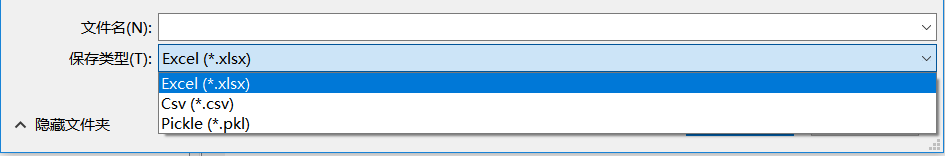
按住Ctrl，使用鼠标单击，可以选中多个文件，单击解析，一起处理。



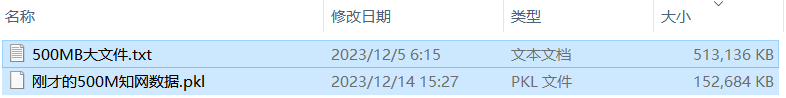
### 解析

对于知网、维普、万方、知网专利、WOS等格式，需要单独解析。

### 保存

处理完数据后，可以保存数据。支持的格式有三种，分别是Excel、Csv和Pickle。

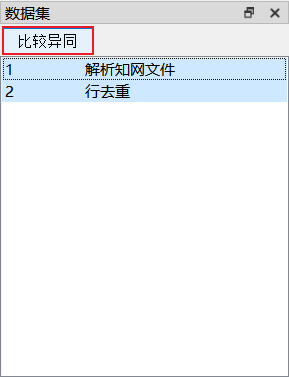
Pickle格式是一种特殊的压缩格式，500MB的文件，使用Pickle格式保存，大小是152MB。



对于大尺寸的数据文件，建议使用Pickle保存。

## 数据集

表格中数据的每一次变化，都会在数据集中保留上一次操作的数据集。



选中两个数据集，可以比较异同。方便我们对比每次操作的结果。

## 表格

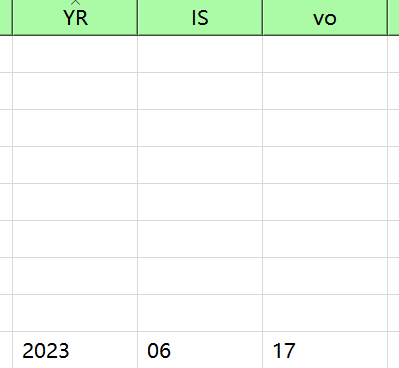
### 表头排序

点击表头可以按照当前列排序，升序和降序，切换。

升序降序

#### **场景1**

如果某一列含有空值，可以点击排序，让空值显示在最上面，方便继续处理。



#### **场景1**

比如执行复制列后，新增了一列，但是不想要了，点击“撤销”就恢复到复制列之前的状态。真的非常方便好用。



## 编辑

### 撤回

在进行文本处理的过程中，如果发现处理的效果不好，或者处理错了，就可以使用撤回功能。真的非常方便好用。

### 恢复

### **删除行**

可以删除一行，也可以同时删除多行。

### **删除列**

可以删除一列，也可以同时删除多列。

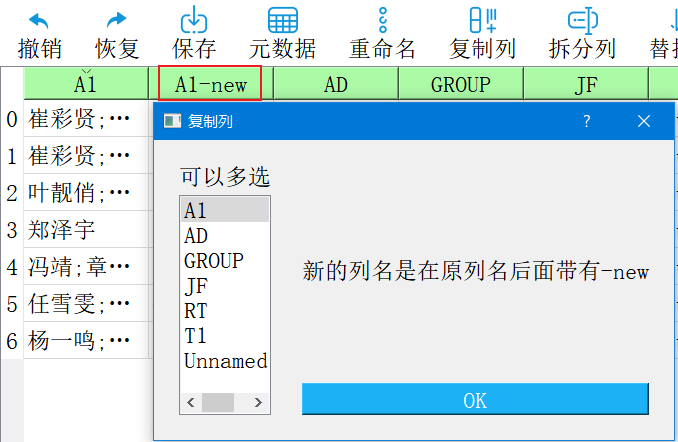
### 重命名

对表格中的列名重新命名，方便自己理解和查看。



### **复制列**

如果想对某一列处理，但是又想保留当前列的内容。可以复制一列，然后修改。方便左右比较，查看修改效果。



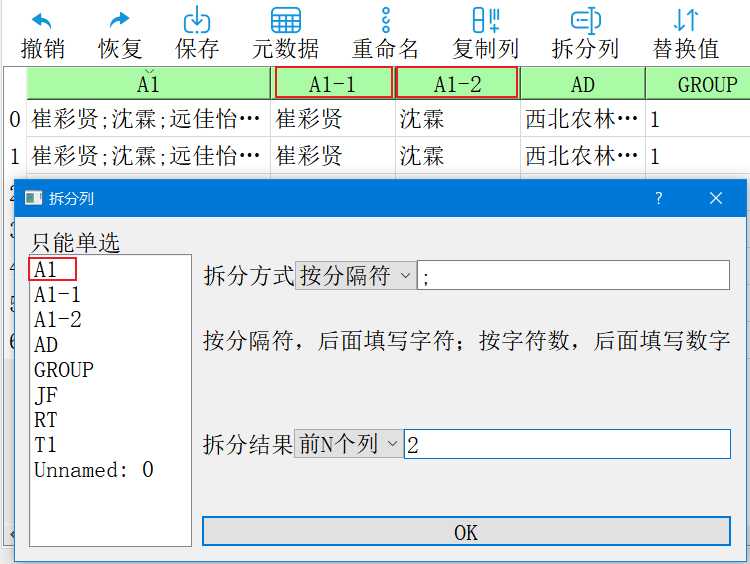
### **拆分列**

比如拆分第一、第二作者，就可以按照下面的方式设置。

拆分方式分为按分隔符、按字符数。【一定注意是否英文字符】

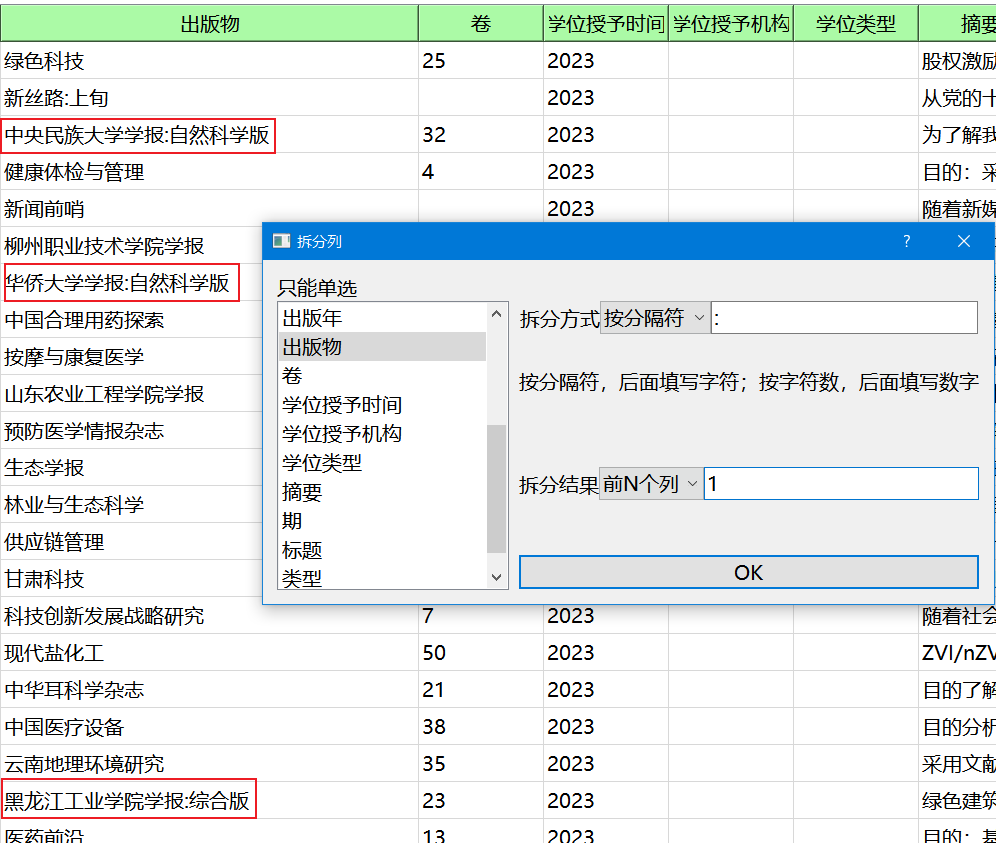
拆分结果包括提取前N个列或者第N列。

#### **场景1**



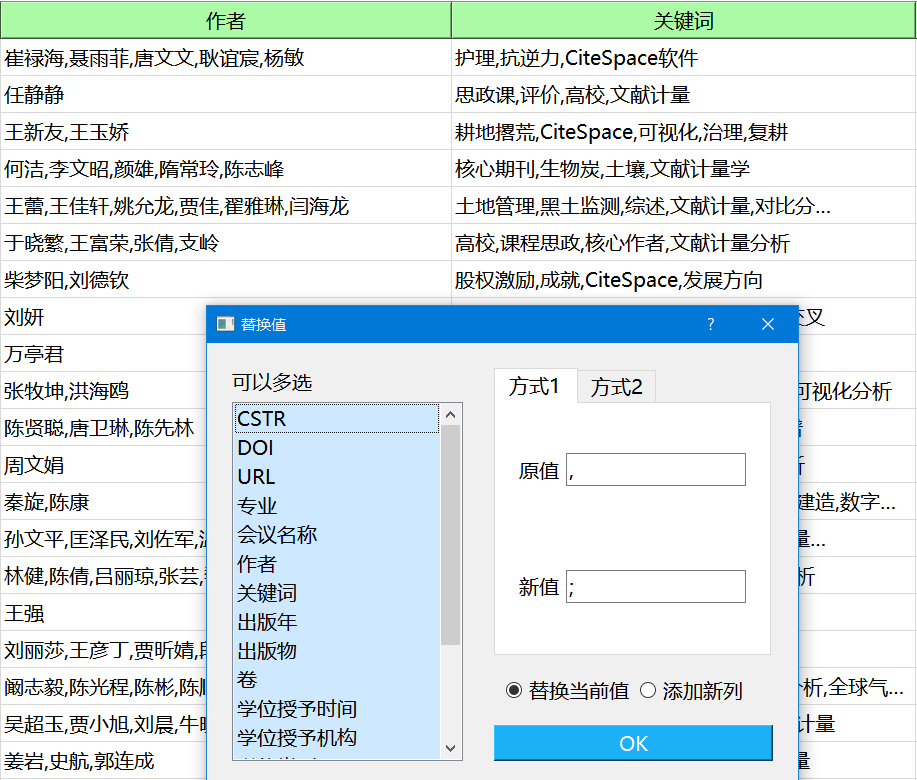
#### **场景2**

比如下面的刊物名称含有冒号，我们只想要冒号前面的内容，可以使用该功能。



### **替换值**

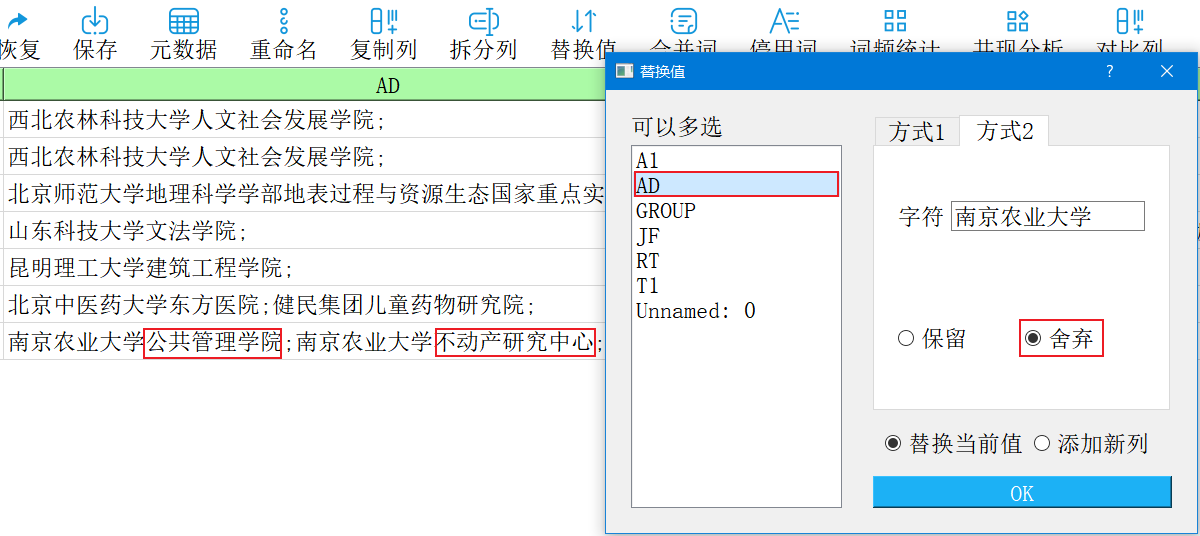
方式1：比如pubscholar.cn下载的题录信息，作者和关键词的分隔符是逗号，我们想全部改为分号，就可以使用方式1。



方式2：对某个词语处理的时候，可以保留这个词，也可以舍弃这个词。

#### **场景1**

比如打算分离二级机构。我们只统计南京农业大学下属的各个机构，就可以使用替换值。从而对所有行的内容进行处理。



如果有大量的词需要处理，使用下面的合并词功能。

### **比较列**

如果有一列是修改之前的值，有一列是修改后的值，可以使用该功能进行对比分析，查看变化。如果有不同的值，使用灰色标出。下面的例子是前面拆分列后，通过对出版物和出版物-1进行对比，可以查看修改的变化，非常方便。



如果恢复以前的颜色，去掉灰色，选中该列，点击“恢复颜色”。

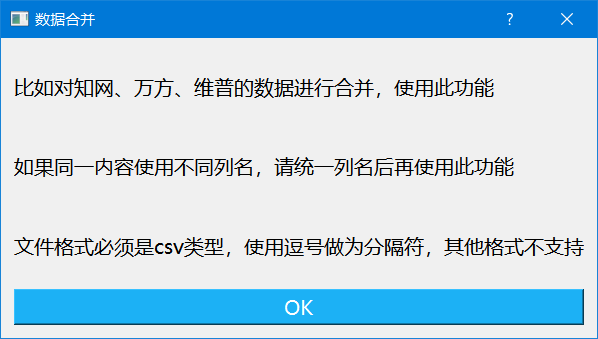
### **修改值**

表格内容的修改，跟Excel是类似的，双击修改。离开自动保存。

### **合并文件**

用于对知网、万方、维普的数据进行合并。

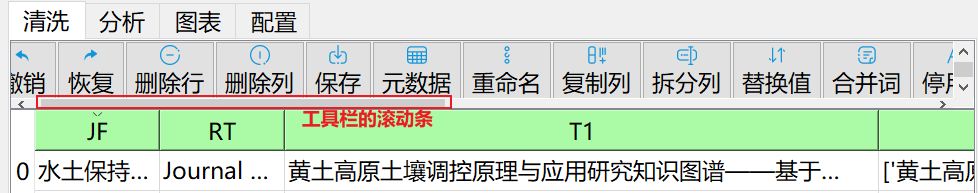
先分别解析这三种数据，然后保存为csv格式。最后使用该功能合并。



合并后去重，可以使用“行去重”功能。

## 清洗

清洗是对文本文件，进行预处理。功能有一些跟wps类似，但是会不断添加文本处理的方法。



功能太多，工具栏有滚动条，可以查看更多的功能按钮。

### 元数据

元数据是描述数据的分布特性。

**数据统计**是对每一列分别进行分析，四个指标分别是：

* 总数：表示有多少个值，包含空值。就是通常理解的有多少行数据。
* 唯一：指的是唯一值的数量，可以查看数据的分布。比如一个表有1000行，性别这一列的唯一就是2，因为性别只有男和女。
* 众频：指的是众数的频数，可以知晓众数的占比。众数指的是出现最多的值的出现次数，就是众频。
* 空值：就是空值有多少行。在数据清洗的时候，可以反复查看该值是否为零。空值为零，表示该列没有空的内容。



**数据分布**指的是每一列的数据分布情况。包括两个表格。

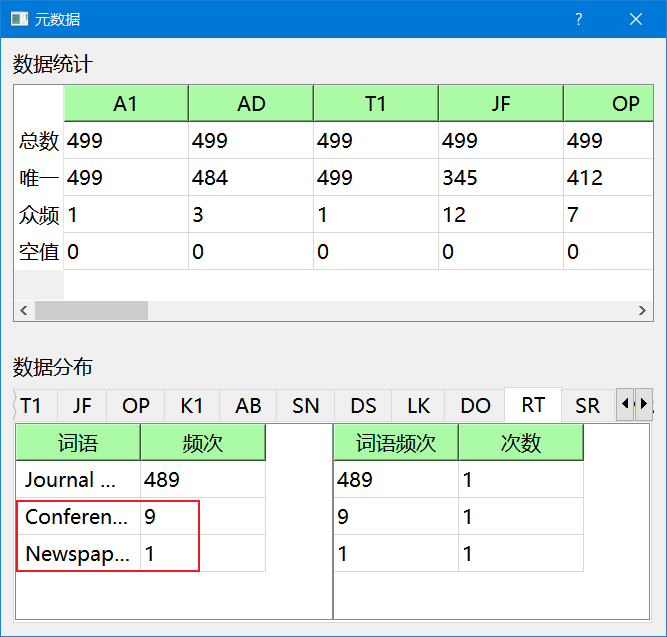
左侧表格，指的是每个词语的出现次数，即频次。

右侧表格指的是左侧频次的出现次数。上图右侧表格第一行表示在左侧表格中有20个词语出现1次。第二行表示在左侧表格中有5个词语出现2次。

#### 场景1

我们按照主题下载了很多题录，想只保留期刊，删除图书、报纸等题录。

使用数据分布，查看RT就很容易知道有多少不需要的内容。

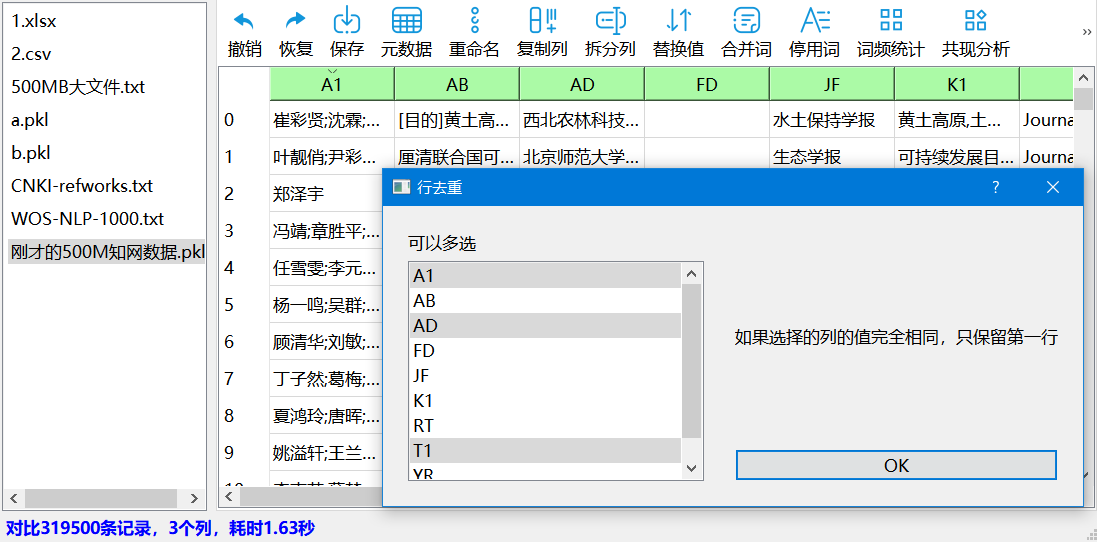


### **行去重**

如果有多个数据来源的文件，可能存在内容重复。那么，可以使用该方法进行去重。

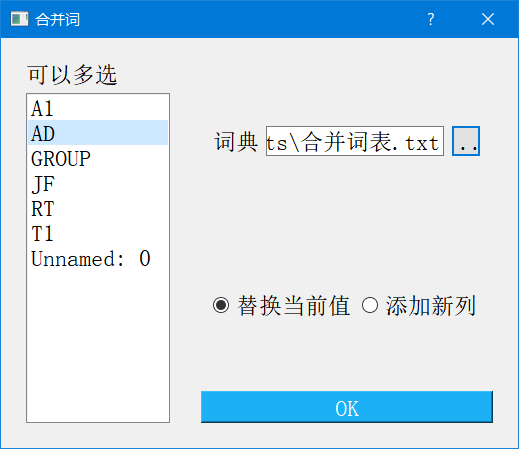
需要选择按照哪些列判断重复。

下面是按照A1、AD、T1三个列判断文献是否重复。31万条记录，用1.63秒完成。



### **合并词**

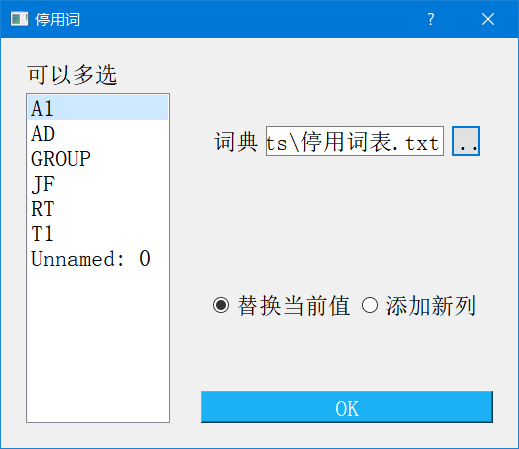
如果有大量的词需要替换掉，比如同义词、近义词，用一个词表示，可以使用该功能。



需要使用dicts文件夹中的合并词表。该词表，需要自己维护。

### **停用词**

停用词指的是编辑部寄语之类的内容，需要删除。



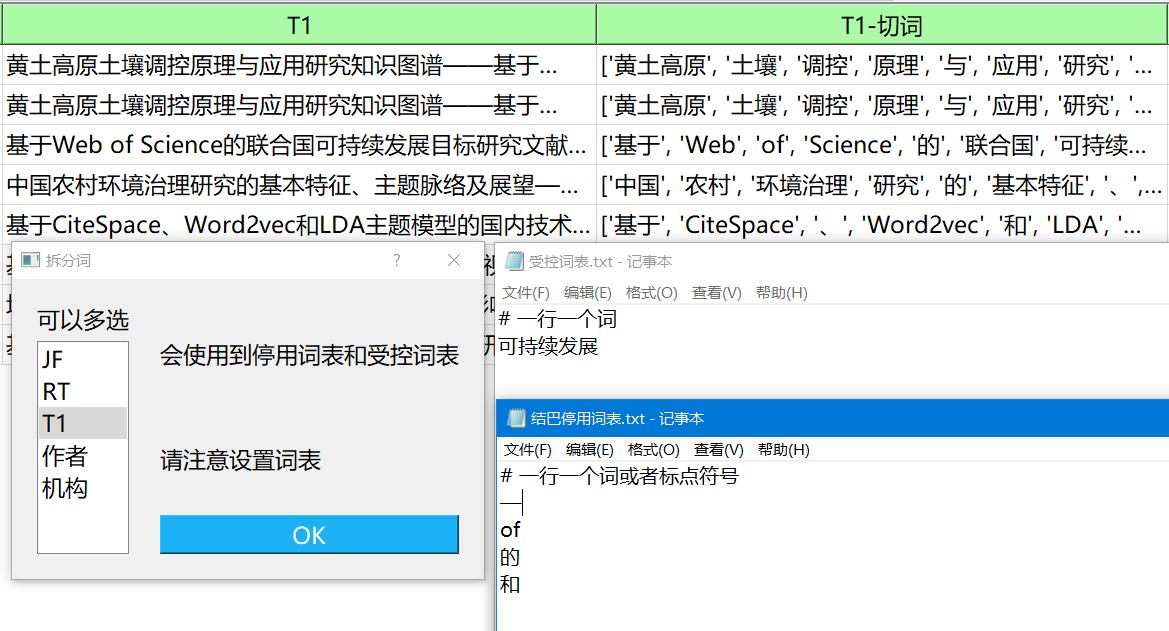
停用词表位于dicts文件夹，需要自己维护。

### **过滤行**

## 分析

### **切分词**

使用jieba进行切词，需要用到停用词表对结果进行过滤，需要用到受控词表。这都需要在配置中设置。

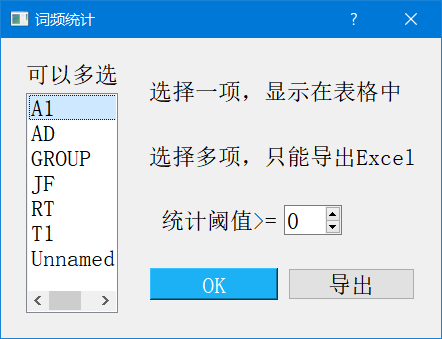


### **词频统计**

词频统计，可以对作者、机构、关键词、发表年份、期刊等等进行词频统计。

如果选择多个列，那么无法在表格中展示，只能导出到Excel中。

统计阈值指的是词频的数量，可以对结果进行过滤。



上面选择的是A1作者，下面是效果。

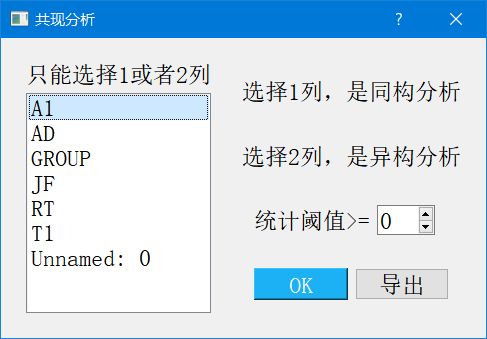


### **共现分析**

共现分析，可以实现作者共现、机构共现、关键词共现等，属于1模矩阵。

也可以实现作者-机构矩阵、作者-期刊矩阵等等，属于2模矩阵。

所以，大家可以发挥想象力。



对A1作者进行共现分析，下面是效果图。

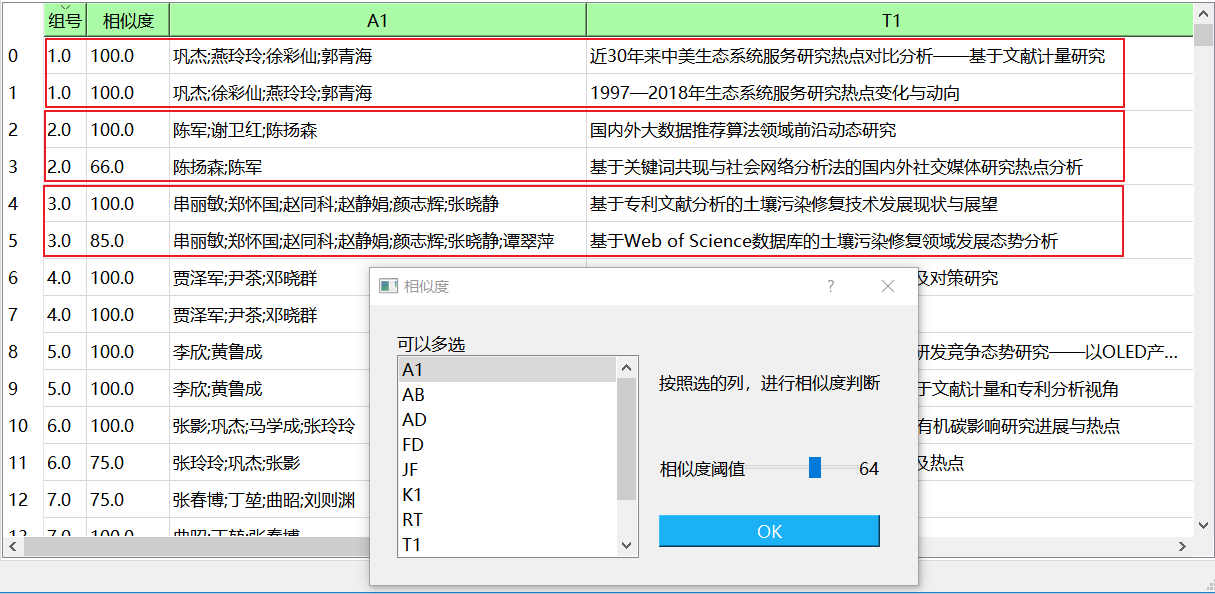


### **相似比较**

先介绍一下原理。下面的例子是按照A1作者进行杰卡德相似度比较，然后进行分组。

从截图中，可以发现1、2、3组的作者，应该是研究团体。

也可以按照其他列进行相似度判断。



### **分组统计**

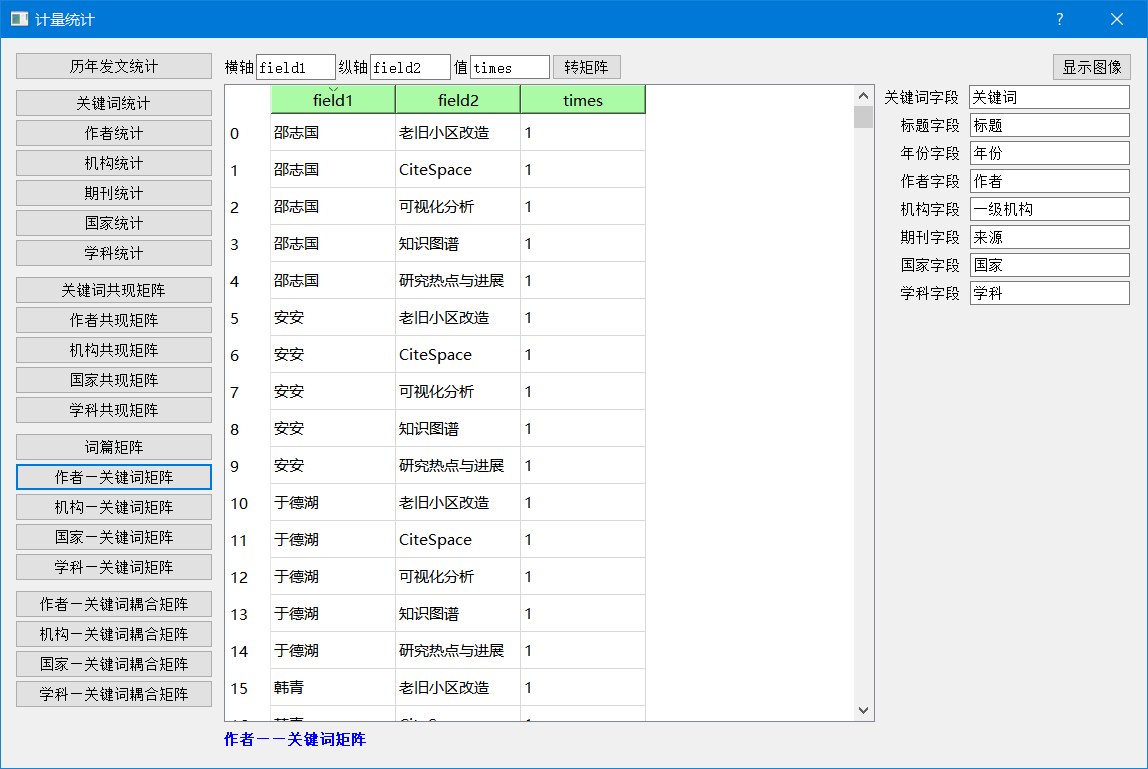
### **特征提取**

## 计量

### **计量统计**

下图是文献计量、科学计量各项指标的大综合。

左侧是指标集合，中间是计量产生的数据，右侧是各字段的对应名称。



运行成功后，下面状态栏以蓝色字体显示。如果错误，以红色字体显示。

### 转矩阵

列表可以转为矩阵



### **显示图表**

## 参数配置

## 操作日志

## 帮助