使用说明书

目录

使用说明书 1
产品亮点 2
常见问题 2
功能点 2
1. 数据文件 3
1.1. 数据文件自动发现 3
1.2. 常见格式，双击运行 3
1.3. 特殊格式，单独解析 3
1.4. 多个文件，一起解析 4
2. 清洗 4
2.1. 表头排序 5
2.2. 撤回和恢复 5
2.3. 保存 6
2.4. 元数据 6
2.5. 重命名 8
2.6. 复制列 9
2.7. 拆分列 9
2.8. 替换值 11
2.9. 合并词 13
2.10. 停用词 13
2.11. 词频统计 14
2.12. 共现分析 14
2.13. 对比列 15
2.14. 修改值 16
2.15. 切分词 16
2.16. 行去重 17
2.17. 相似度 17
2.18. 删除行 18
2.19. 删除列 18
2.20. 数据合并 18

产品亮点

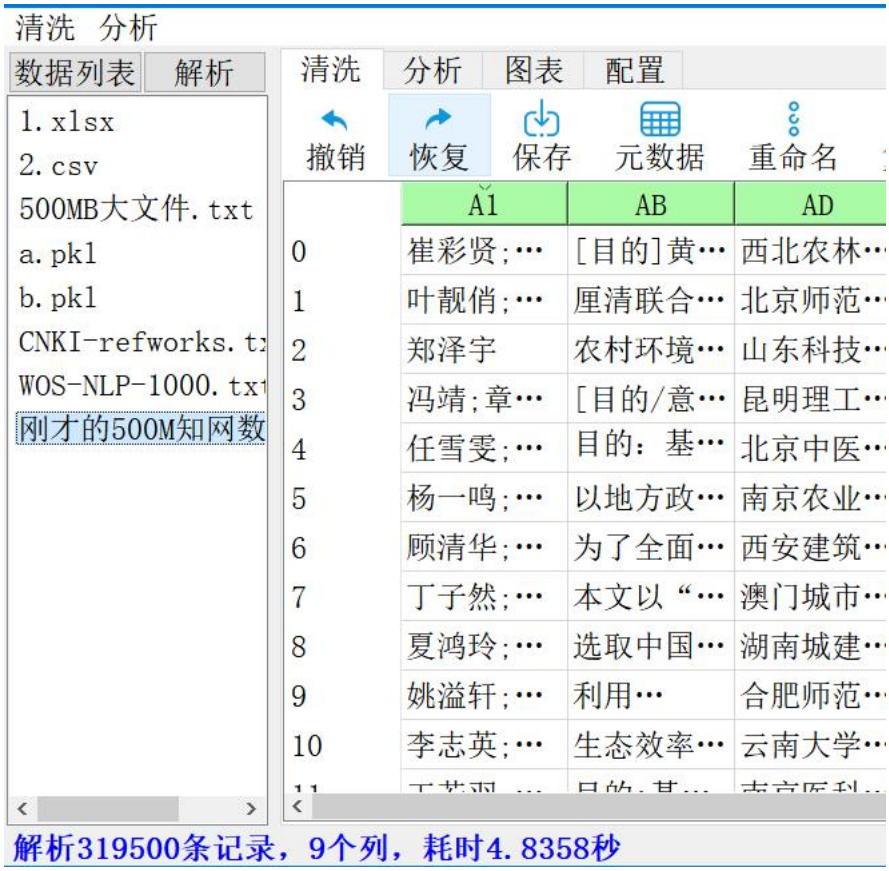
1. 支持各种自定义的文件格式，比如自己整理的文本文件、excel 文件。
2. 能够处理大的数据量，在 32 位操作系统下，几秒钟内完成 1GB 文件或者 100 多万条记录的处理。常见的处理操作，在毫秒完成。
3. 适合对文本内容进行清洗、处理、分析。

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

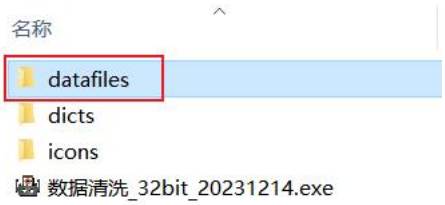
下面介绍的所有操作，都会在左下角提示。蓝色字体表示提示信息。红色字体表示错误信息。



1. 数据文件

1.1. 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。



1.2. 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，都可以解析。

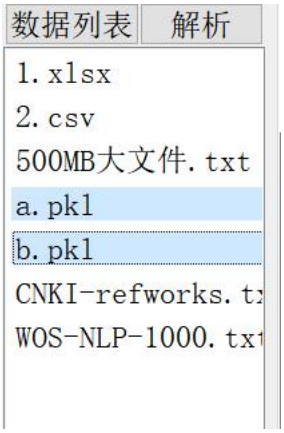
1.3. 特殊格式，单独解析

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 等格式，需要单独解析。点击上面的解析按钮。

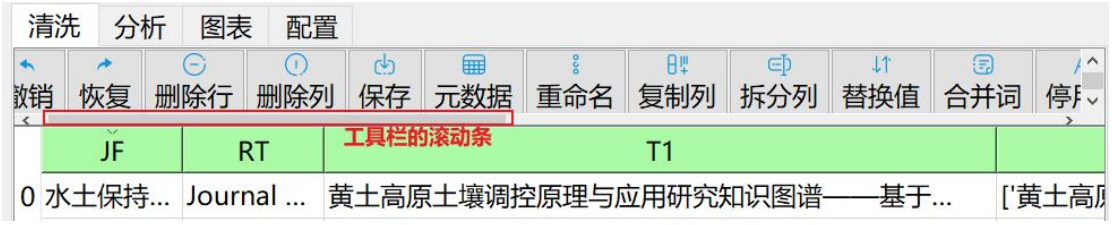
1.4. 多个文件，一起解析

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。



2. 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。

2.1. 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 

2.1.1.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。

2.2. 撤回和恢复

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。

2.2.1.场景 1

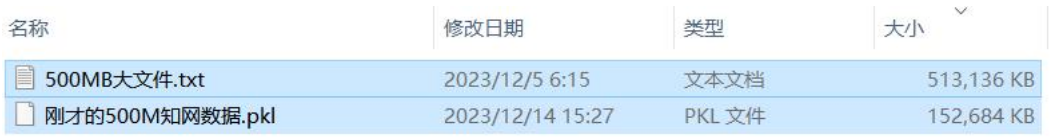
比如执行复制列后，新增了一列，但是不想要了，点击“撤销”就恢复到复制列之前的状态。真的非常方便好用。

	A1	A1-new	AD	GROUP
0	崔彩贤;…	崔彩贤;…	西北农林…	1
1	崔彩贤;…	崔彩贤;…	西北农林…	1
2	叶靓俏;…	叶靓俏;…	北京师范…	1
3	郑泽宇	郑泽宇	山东科技…	2
4	冯靖;章…	冯靖;章…	昆明理工…	2
5	任雪雯;…	任雪雯;…	北京中医…	3
6	杨一鸣;…	杨一鸣;…	南京农业…	3

2.3. 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。



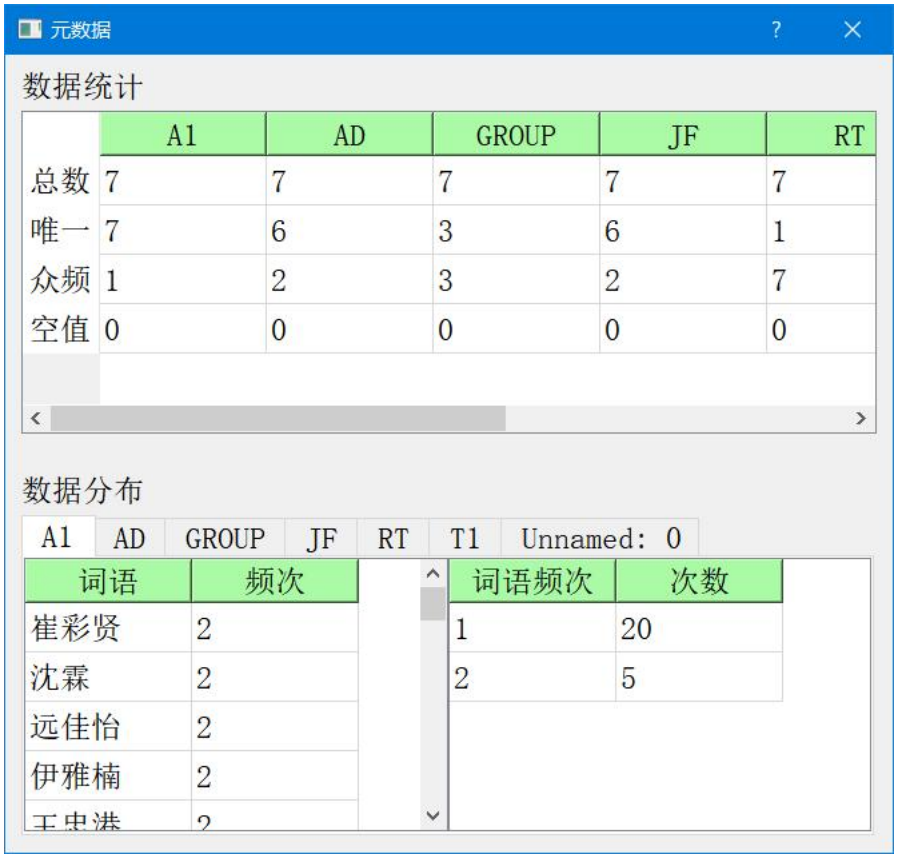
对于大尺寸的数据文件，建议使用 Pickle 保存。

2.4. 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：

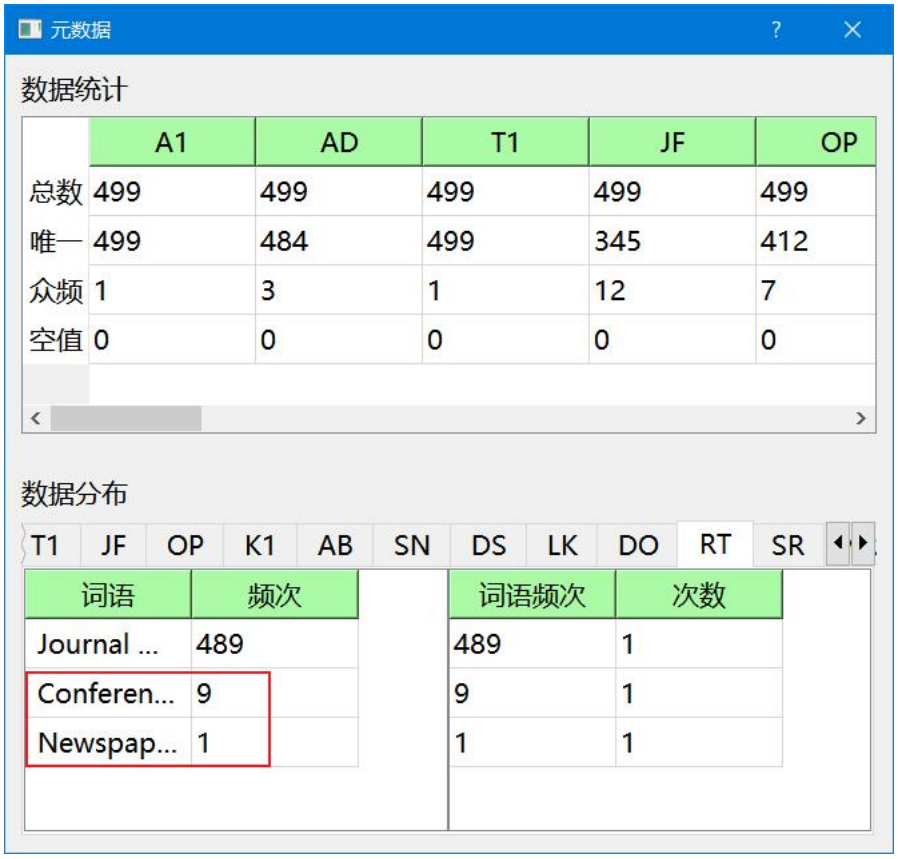
- 总数：表示有多少个值，包含空值。就是通常理解的有多少行数据。
- 唯一：指的是唯一值的数量，可以查看数据的分布。比如一个表有 1000 行，性别这一列的唯一就是 2，因为性别只有男和女。
- 众频：指的是众数的频数，可以知晓众数的占比。众数指的是出现最多的值的出现次数，就是众频。
- 空值：就是空值有多少行。在数据清洗的时候，可以反复查看该值是否为零。空值为零，表示该列没有空的内容。



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。

2.4.1.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。

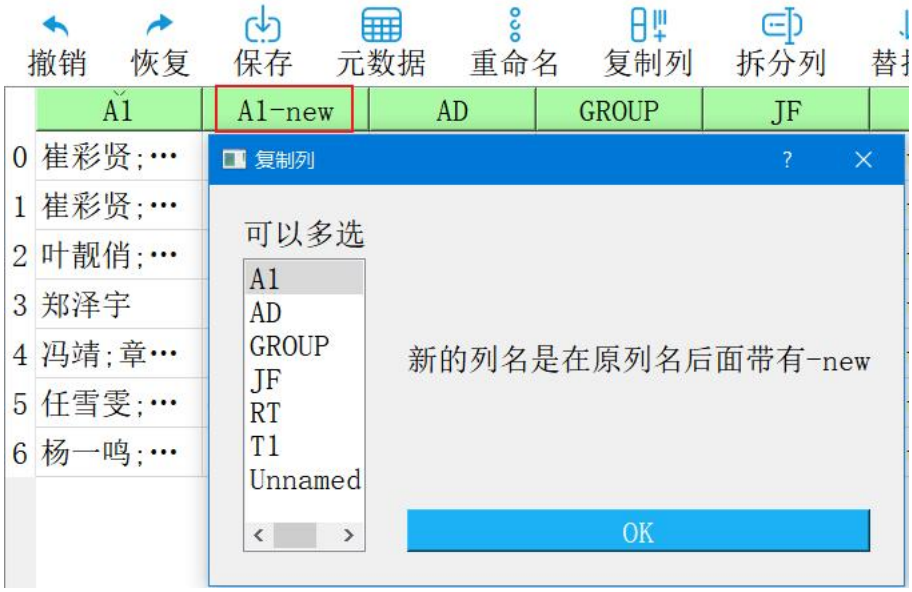


2.5. 重命名

对表格中的列名重新命名，方便自己理解和查看。

2.6. 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。



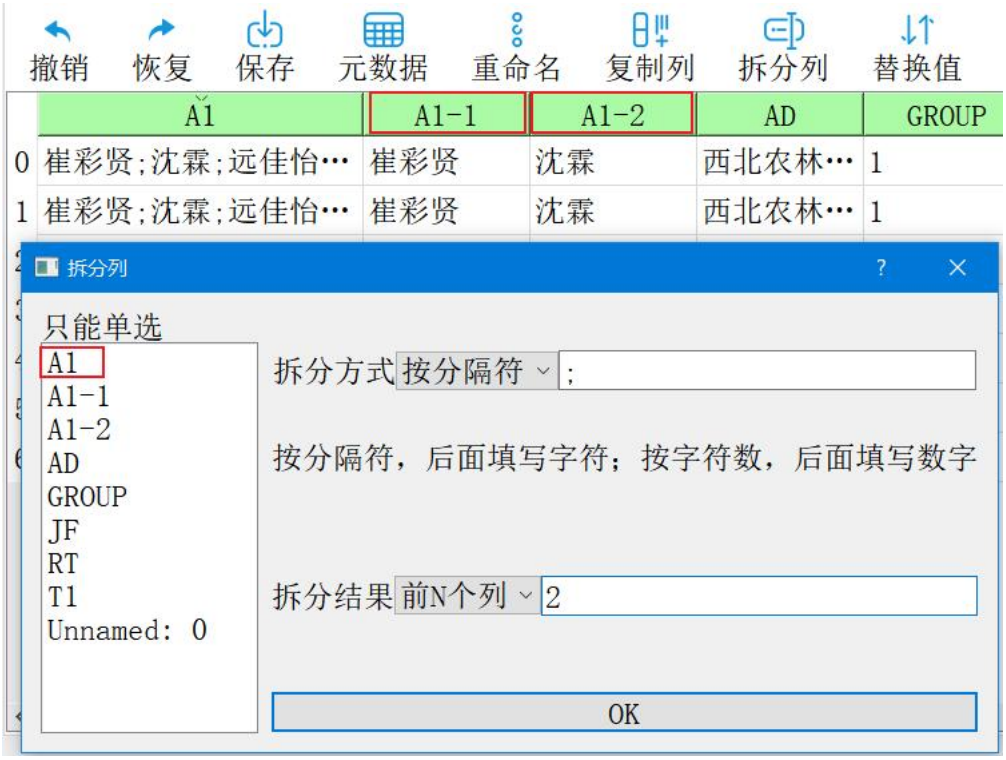
2.7. 拆分列

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。

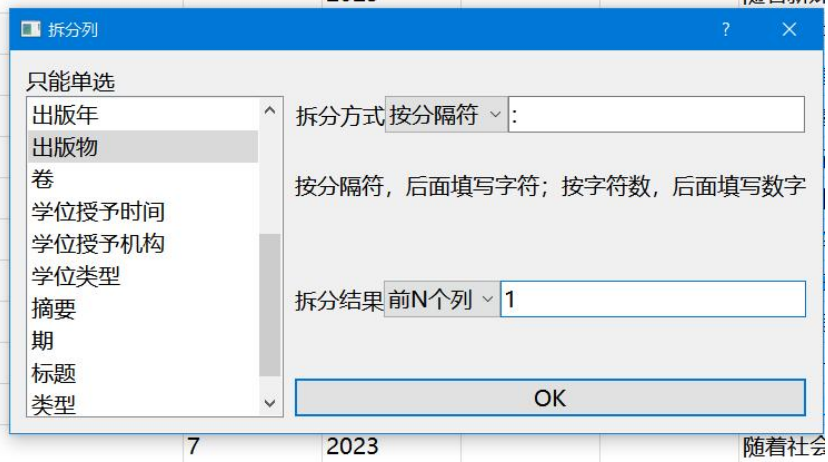
2.7.1.场景 1



2.7.2.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。

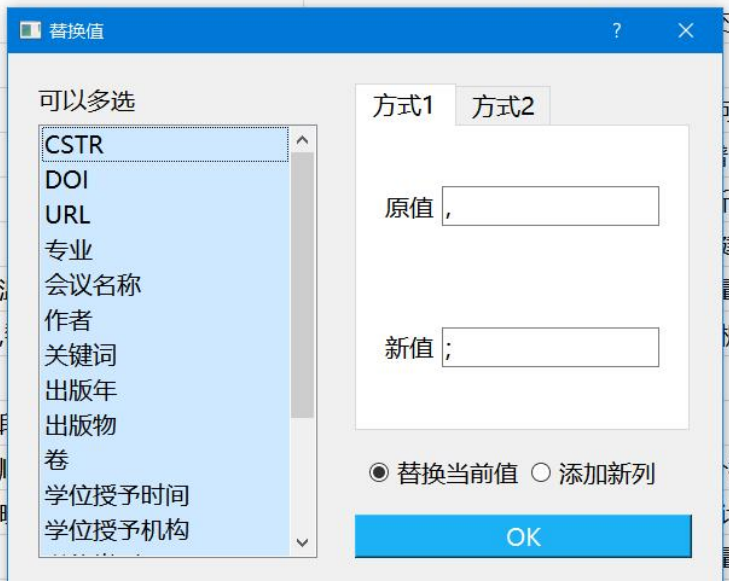
出版物	卷	学位授予时间	学位授予机构	学位类型	摘要
绿色科技	25	2023			股权激励
新丝路:上旬		2023			从党的十
中央民族大学学报:自然科学版	32	2023			为了解我
健康体检与管理	4	2023			目的：采
新闻前哨		2023			随着新媒
柳州职业技术学院学报					
华侨大学学报:自然科学版					
中国合理用药探索					
按摩与康复医学					
山东农业工程学院学报					
预防医学情报杂志					
生态学报					
林业与生态科学					
供应链管理					
甘肃科技					
科技创新发展战略研究	7	2023			随着社会
现代盐化工	50	2023			ZVI/nZV
中华耳科学杂志	21	2023			目的了解
中国医疗设备	38	2023			目的分析
云南地理环境研究	35	2023			采用文献
黑龙江工业学院学报:综合版	23	2023			绿色建筑
医药前沿	13	2023			目的：基



2.8. 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。

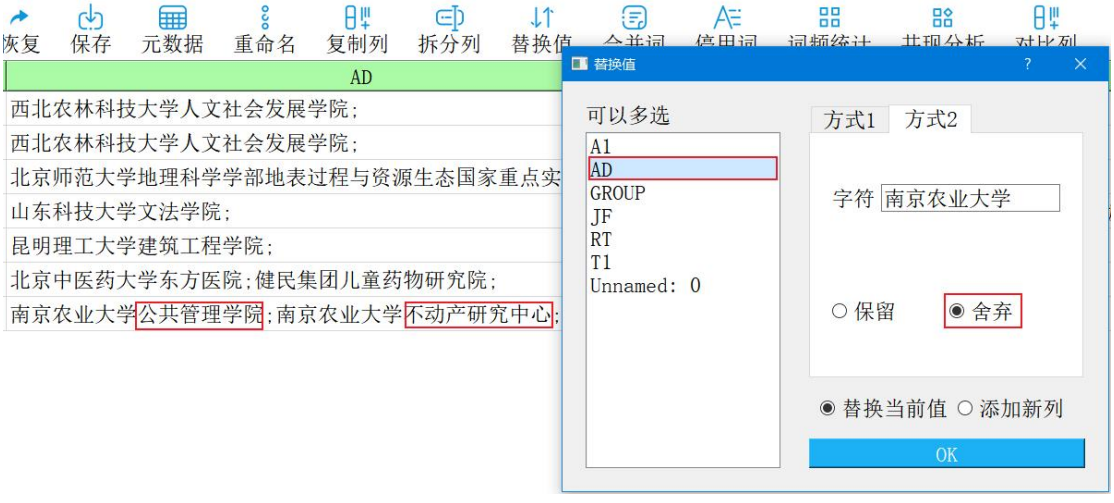
作者	关键词
崔禄海,聂雨菲,唐文文,耿谊宸,杨敏	护理,抗逆力,CiteSpace软件
任静静	思政课,评价,高校,文献计量
王新友,王玉娇	耕地撂荒,CiteSpace,可视化,治理,复耕
何洁,李文昭,颜雄,隋常玲,陈志峰	核心期刊,生物炭,土壤,文献计量学
王蕾,王佳轩,姚允龙,贾佳,翟雅琳,闫海龙	土地管理,黑土监测,综述,文献计量,对比分...
于晓繁,王富荣,张倩,支岭	高校,课程思政,核心作者,文献计量分析
柴梦阳,刘德钦	股权激励,成就,CiteSpace,发展方向
刘妍	
万亭君	
张牧坤,洪海鸥	
陈贤聪,唐卫琳,陈先林	
周文娟	
秦旋,陈康	
孙文平,匡泽民,刘佐军,	
林健,陈倩,吕丽琼,张芸,	
王强	
刘丽莎,王彦丁,贾昕婧,	
阚志毅,陈光程,陈彬,陈	
吴超玉,贾小旭,刘晨,牛	
姜岩,史航,郭连成	



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。

2.8.1.场景 1

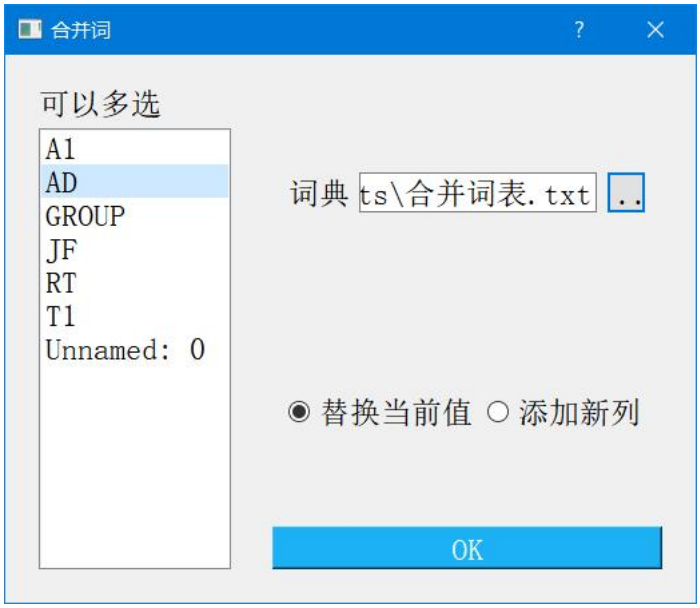
比如打算分离二级机构。我们只统计南京农业大学下属的各个机构，就可以使用替换值。从而对所有行的内容进行处理。



如果有大量的词需要处理，使用下面的合并词功能。

2.9. 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。

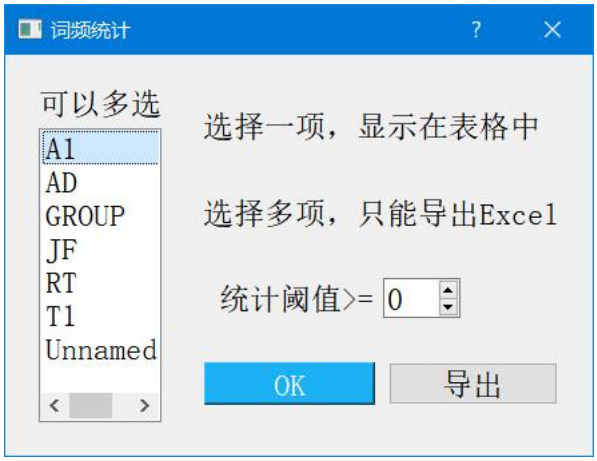
2.10. 停用词

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。

2.11. 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。

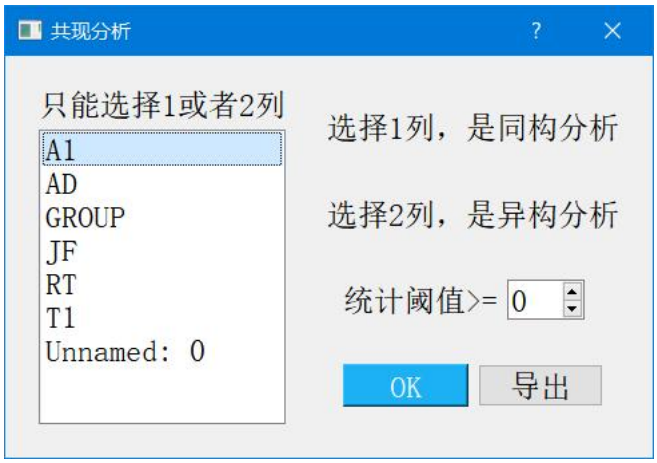


上面选择的是 A1 作者，下面是效果。

	A1	次数
0	崔彩贤	2
1	远佳怡	2
2	伊雅楠	2
3	王忠港	2
4	沈霖	2
5	李元文	1
6	王健	1
7	吴群	1
8	杨一鸣	1
9	肖飞	1
10	邓宇童	1
11	蔡玲玲	1
12	胡博	1

2.12. 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。

	肖飞	叶靓俏	伊雅楠	胡博	李纬	任雪雯	杨
肖飞	0	0	0	1	1	1	0
叶靓俏	0	0	0	0	0	0	0
伊雅楠	0	0	0	0	0	0	0
胡博	1	0	0	0	1	1	0
李纬	1	0	0	1	0	1	0
任雪雯	1	0	0	1	1	0	0
杨一鸣	0	0	0	0	0	0	0
尹彩春	0	1	0	0	0	0	0
崔彩贤	0	0	2	0	0	0	0
王忠港	0	0	2	0	0	0	0
章胜平	0	0	0	0	0	0	0
宋志刚	0	0	0	0	0	0	0

2.13. 对比列

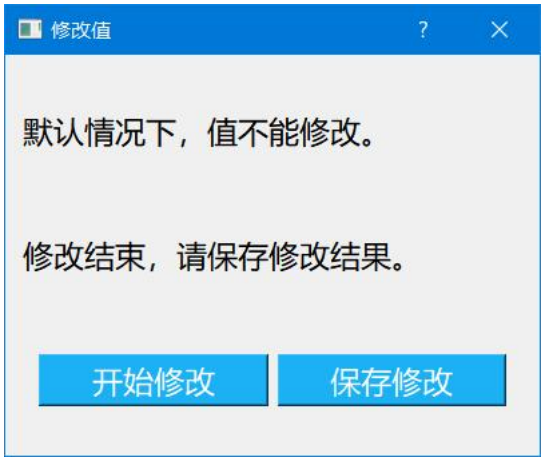
如果有一列是修改之前的值，有一列是修改后的值，可以使用该功能进行对比分析，查看变化。如果有不同的值，使用灰色标出。下面的例子是前面拆分列后，通过对出版物和出版物-1 进行对比，可以查看修改的变化，非常方便。

出版物	出版物-1
卫生职业教育	卫生职业教育
新丝路:下旬	新丝路
干旱区地理	干旱区地理
现代园艺	现代园艺
国土资源科技管理	国土资源科技管理
对外经贸	对外经贸
绿色科技	绿色科技
新丝路:上旬	新丝路
中央民族大学学报:自然科学版	中央民族大学学报
健康体检与管理	健康体检与管理
新闻前哨	新闻前哨
柳州职业技术学院学报	柳州职业技术学院学报
华侨大学学报:自然科学版	华侨大学学报
中国合理用药探索	中国合理用药探索
按摩与康复医学	按摩与康复医学
山东农业工程学院学报	山东农业工程学院学报

如果恢复以前的颜色，去掉灰色，选中该列，点击“恢复颜色”。

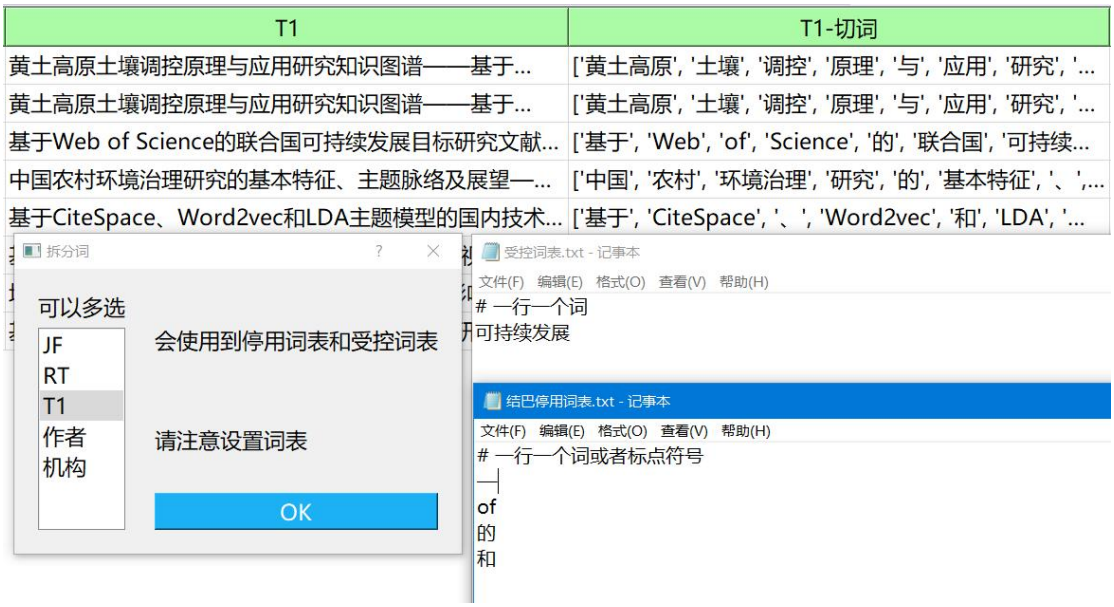
2.14. 修改值

表格内容是不能修改的。如果修改，需要先点击“开始修改”，修改完毕后，再点击“保存修改”。



2.15. 切分词

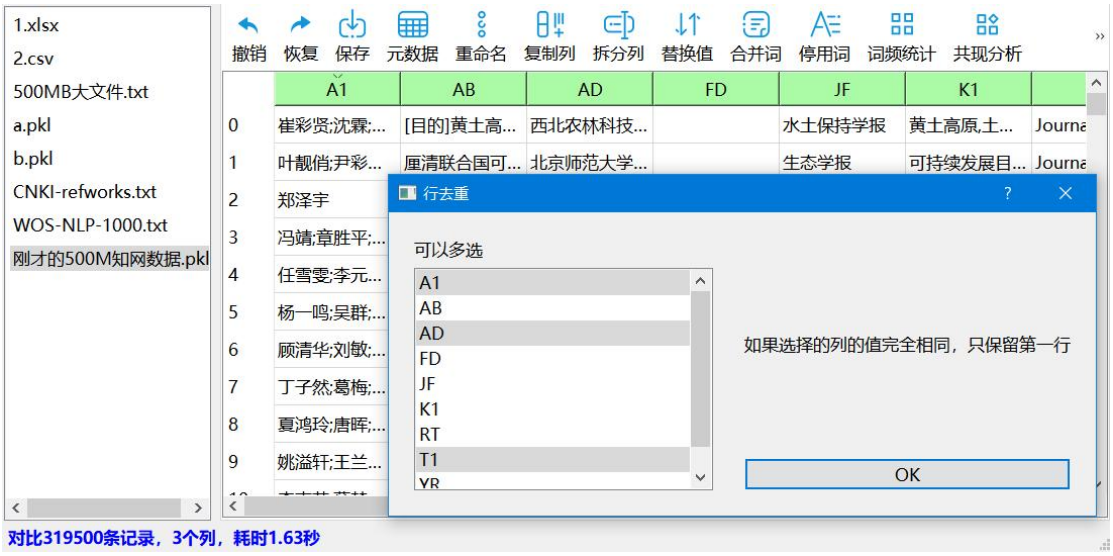
使用 jieba 进行切词，需要用到停用词表对结果进行过滤，需要用到受控词表。这都需要在配置中设置。



2.16. 行去重

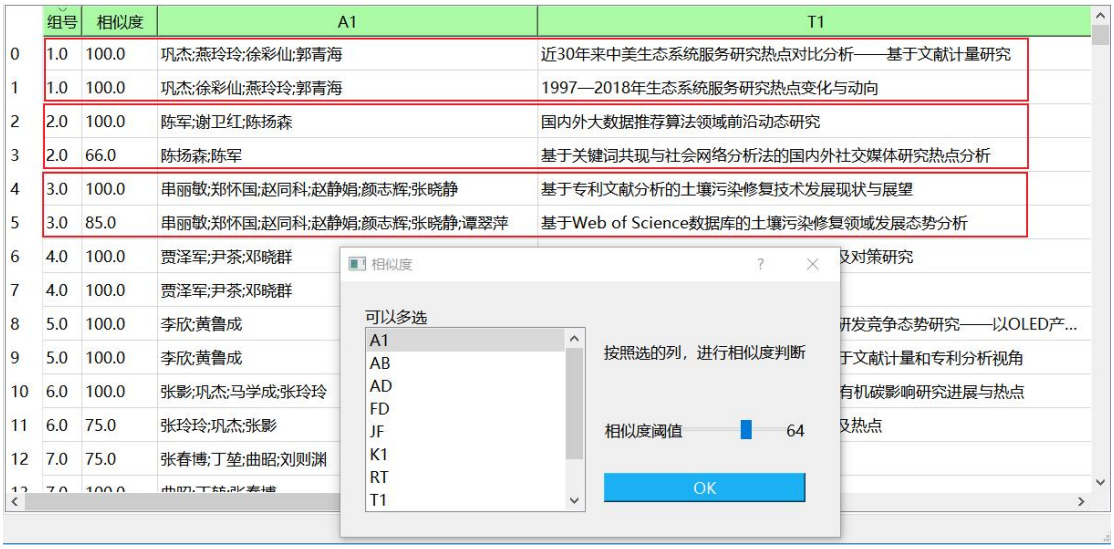
如果有多个数据来源的文件，可能存在内容重复。那么，可以使用该方法进行去重。
需要选择按照哪些列判断重复。

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。

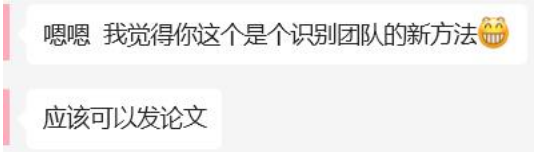


2.17. 相似度

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。



一位老师觉得对于发现合作团体，该方法有帮助。



2.18. 删除行

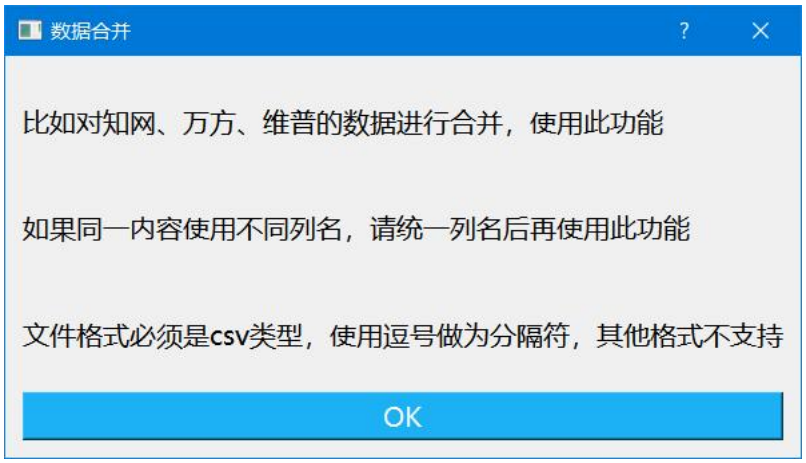
可以删除一行，也可以同时删除多行。

2.19. 删除列

可以删除一列，也可以同时删除多列。

2.20. 数据合并

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。