基于LDA算法的多文本主题提取软件

目录

[基于LDA算法的多文本主题提取软件 1](#_Toc127195568)

[1引言 1](#_Toc127195569)

[2软件说明 1](#_Toc127195570)

[2.1功能介绍 2](#_Toc127195571)

[2.1.1单文本分词 2](#_Toc127195572)

[2.1.2多文本主题分析 4](#_Toc127195573)

[2.1.3添加用户词典 7](#_Toc127195574)

[2.2性能介绍 8](#_Toc127195575)

[2.2.1代码健壮性 8](#_Toc127195576)

[2.2.2响应时间 9](#_Toc127195577)

[3参考资料 9](#_Toc127195578)

**表目录**

[图 1软件界面预览 2](#_Toc127195590)

[图 2单文本词频分析 3](#_Toc127195591)

[图 3分词数据导出excel 3](#_Toc127195592)

[图 4分词数据预览 4](#_Toc127195593)

[图 5多文本主题提取 5](#_Toc127195594)

[图 6主题困惑度计算 5](#_Toc127195595)

[图 7操作顺序警告 8](#_Toc127195596)

[图 8文件类型违法警告 9](#_Toc127195597)

# 1引言

随着人工智能和大数据时代的到来,人类日常生活的方方面面都离不开网络和信息技术，现实世界的文本信息更多的呈现为电子版，文本挖掘也成为信息领域的研究热点和学习重点。用计算机实现海量文本的识别和分析成为研究的热门话题。

现如今，文本的载体逐渐从纸面向荧屏转移，文本挖掘与文本分析也逐渐成为研究热点，然而，市面上用于文本挖掘与分析的工具和方法却存在着学习门槛高和步骤繁杂的问题。开发基于LDA算法的多文本主题提取软件的目的可以概括如下：第一，为文本分析提供窗口化工具，降低操作门槛；第二，集成文本分词与主题提取功能，减少用户繁琐劳动；第三，加入可视化分析工具，提高提取进度与分析效率。

# 2软件说明

软件设计所使用的语言为python语言，主要使用的软件包为PyQt5，开发环境为pycharm和qt designer，并使用Pyinstaller将软件打包。

以下将从功能和性能两方面对软件进行介绍，软件窗口界面如图1。

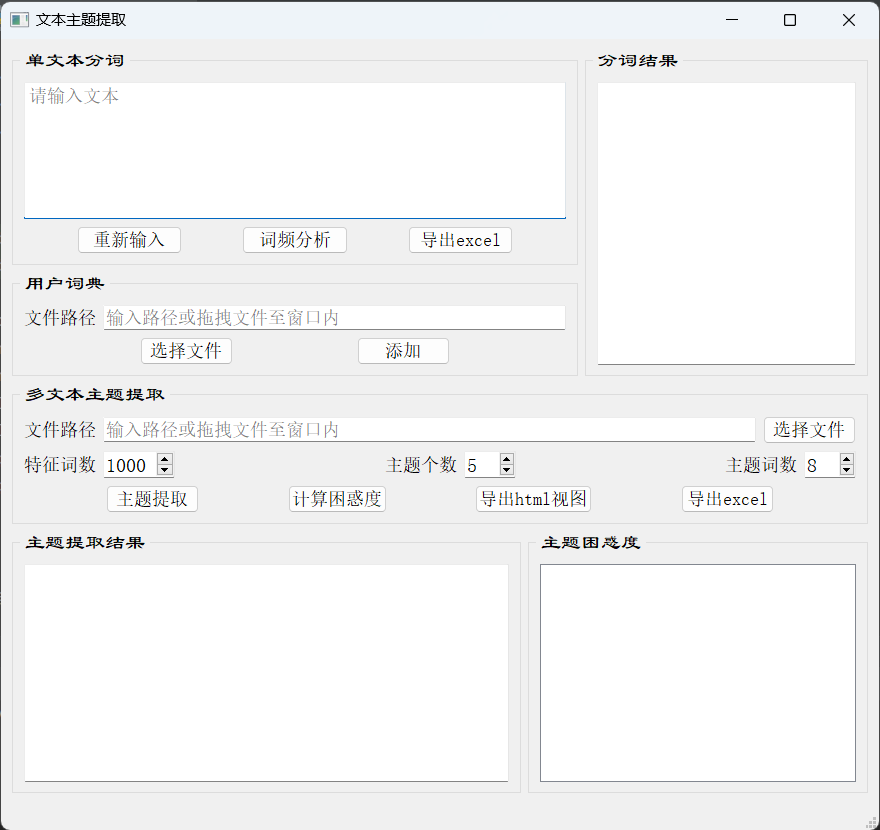


图 1软件界面预览

## 2.1功能介绍

软件包含单文本分词和多文本主题提取两个主要功能，在此基础上，用户可以按照规定文件格式自定义导入用户词典，并导出数据处理结果。

### 2.1.1单文本分词

用户在“单文本分词”编辑框中输入将要分析的文本，若要清除编辑内容，可点击“重新输入”按钮清除内容。文本编辑好后点击“词频分析”按钮，即可在右边“分词结果”显示框内查看词频最高的10个词汇（图2）。若要完整的词频数据，用户可点击“导出excel”按钮导出词频数据（图3），xlsx文件默认保存至当前文件夹下，用户可自行点击查看（图4）。

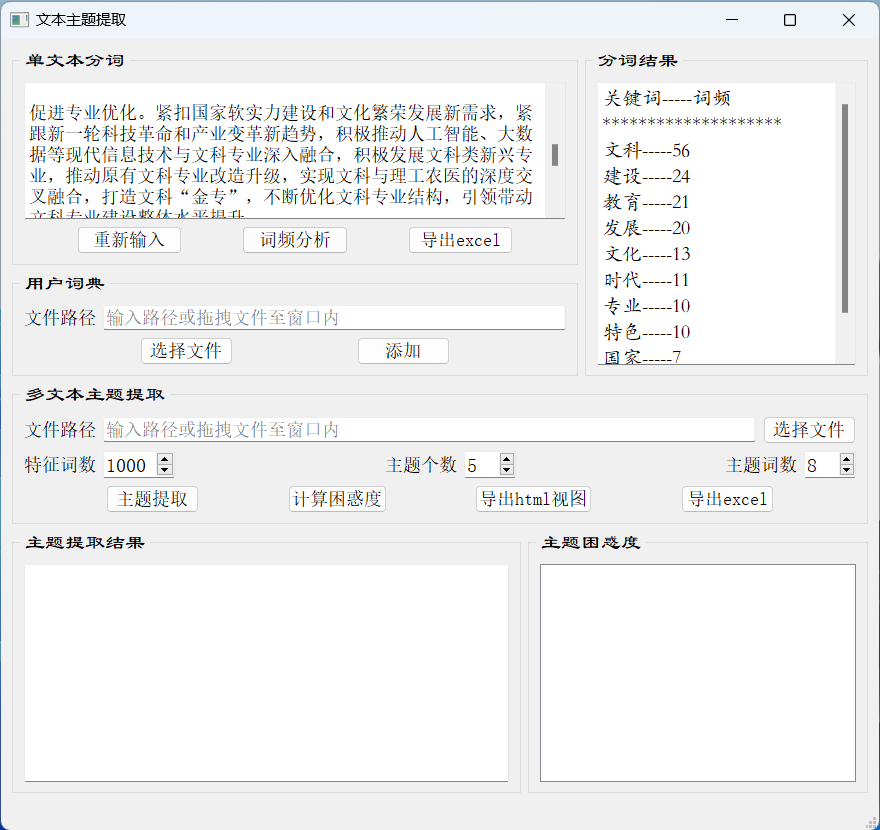


图 2单文本词频分析

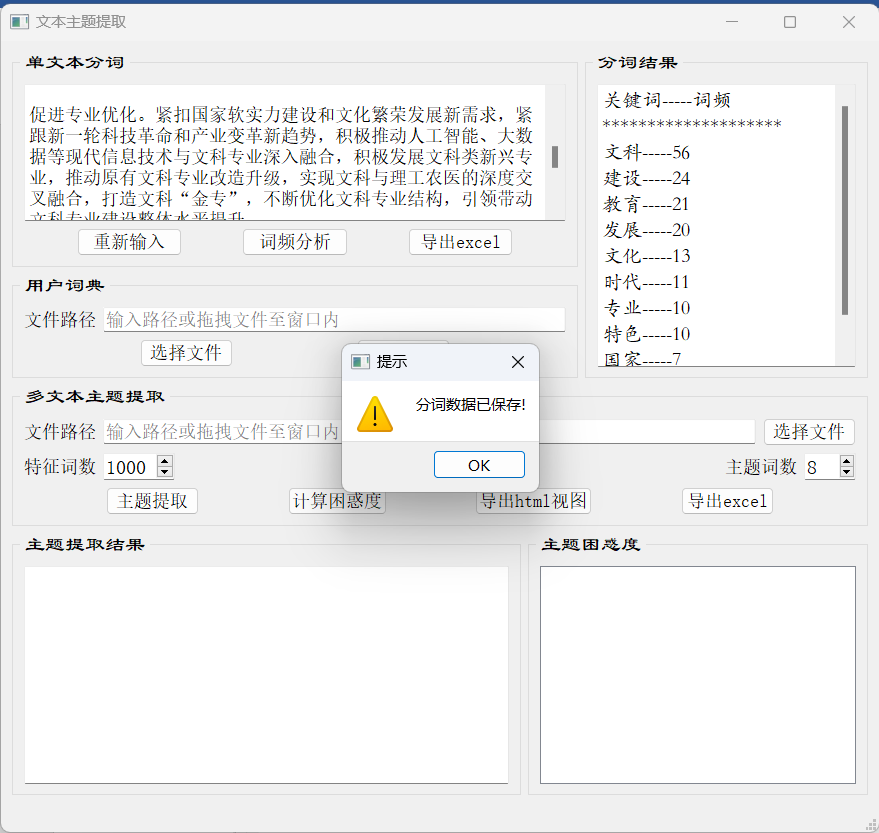


图 3分词数据导出excel

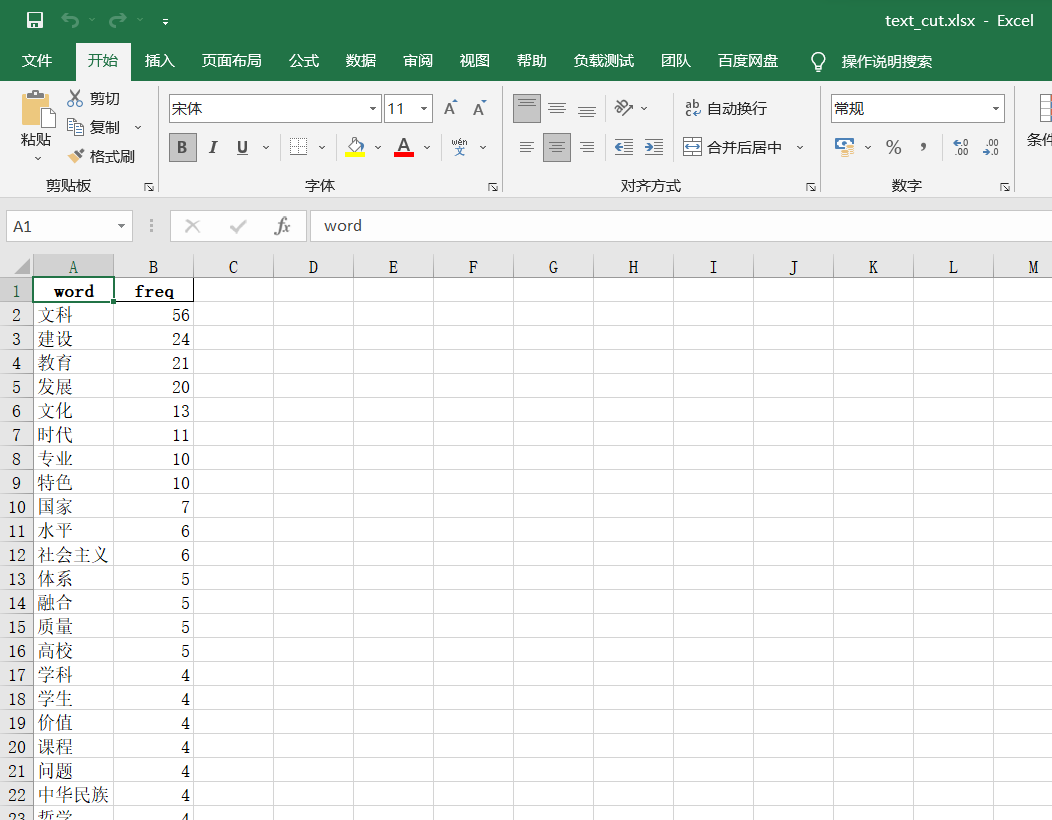


图 4分词数据预览

### 2.1.2多文本主题分析

多文本主题分析是该软件的核心功能，用户可以将所要处理数据的xlsx文件（或xls文件）拖入窗口内以获取文件路径，或者点击“文件选择”浏览文件，也可以在编辑框内直接输入路径。路径设置好之后，用户可根据文本数据大小和特点自行设置提取的特征词数、主题个数和主题词数，用户未输入情况下，软件默认提取1000个特征词，归类为5个主题，并展示8个体现主题的关键词。点击“主题提取”按钮，即可在“主题提取结果”内查看各个主题的关键词（图5）。点击“计算困惑度”按钮，用户可在窗口右下角显示框内查看到困惑度曲线（图6），并根据曲线调整LDA模型的参数，再次进行主题提取。在已进行主题提取操作后，点击“导出html视图”可将主题提取的详细结果以图形的方式保存至目录（图7），打开目录下的html文件可在浏览器查看各主题的关键词及其主题概率（图8）。用户点击“导出excel”按钮即可导出数据（图9）。导出数据如图10。

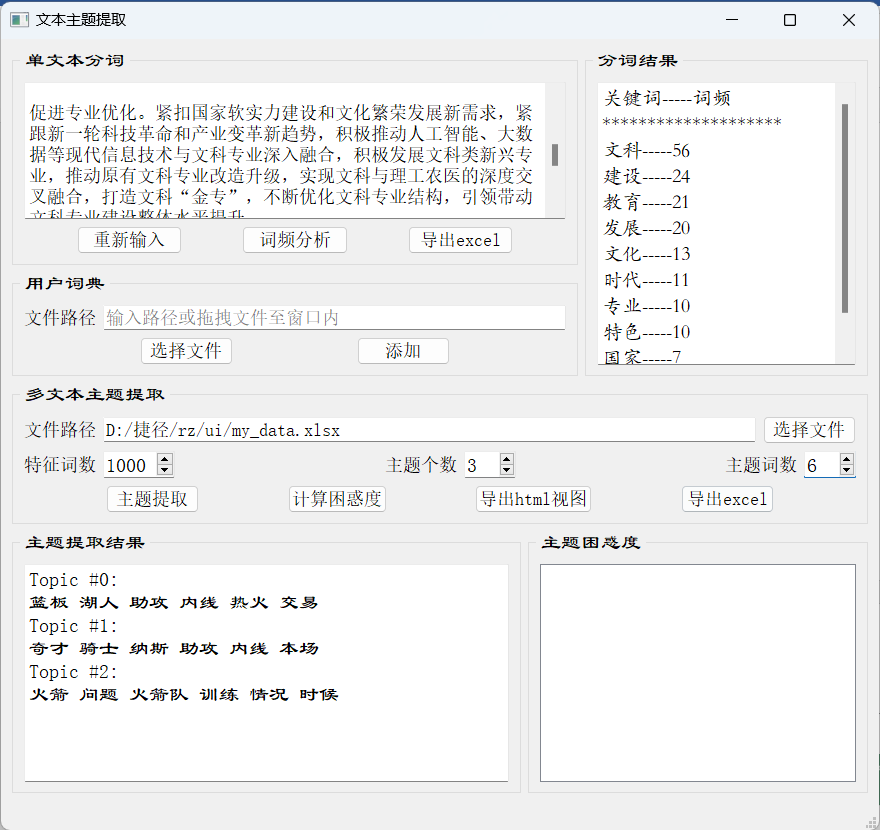


图 5多文本主题提取

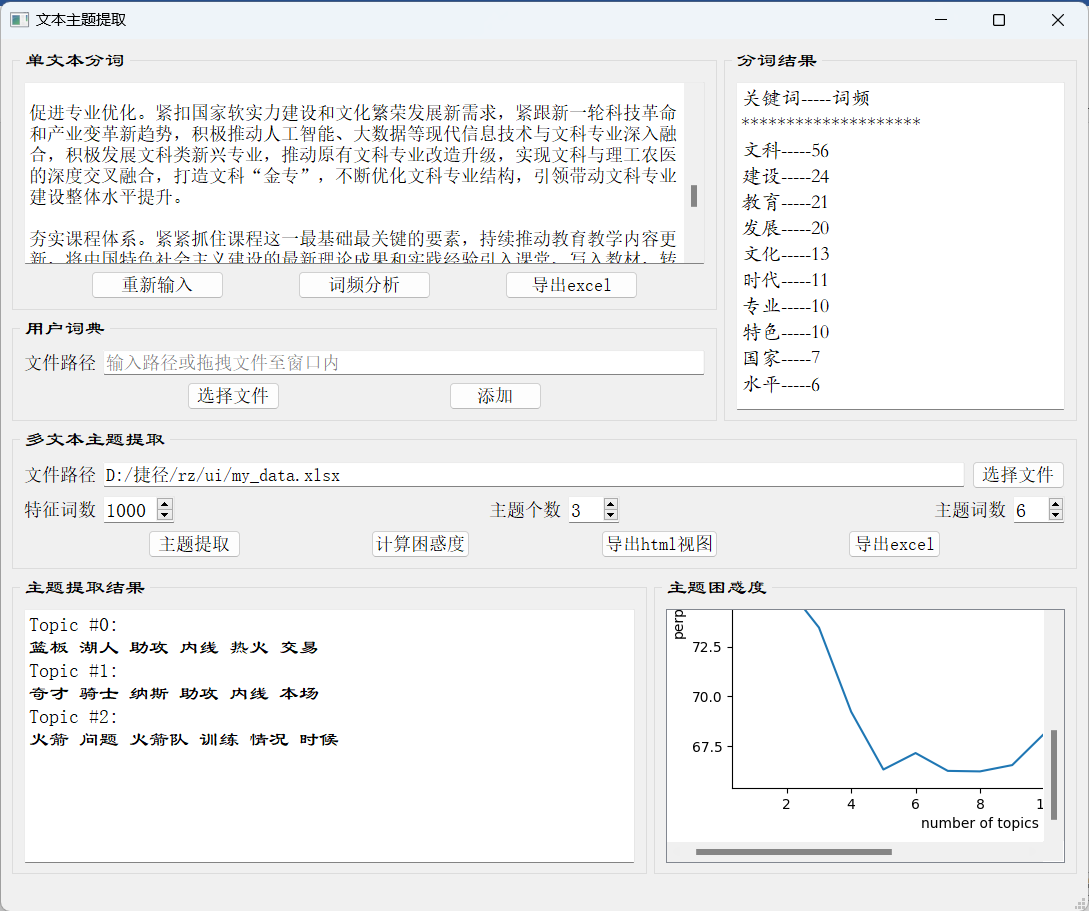


图 6主题困惑度计算

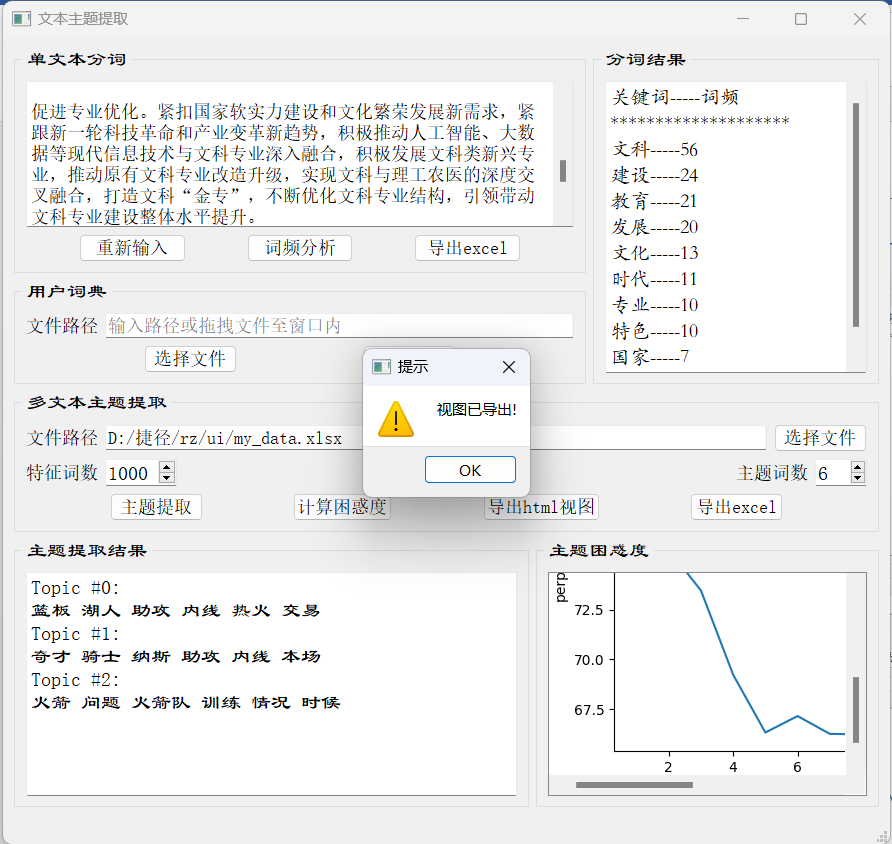


图 7可视化html导出

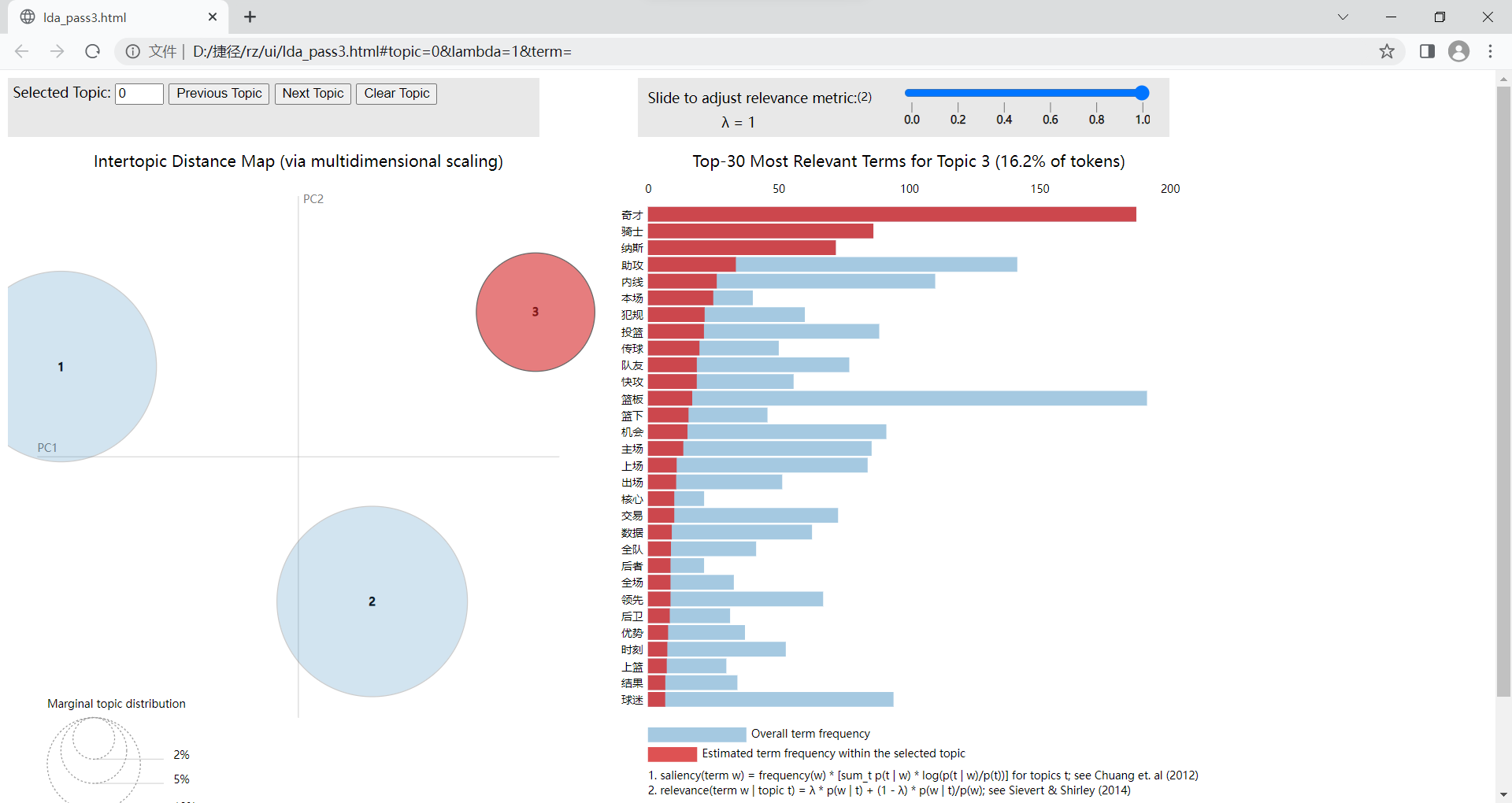


图 8主题提取可视化

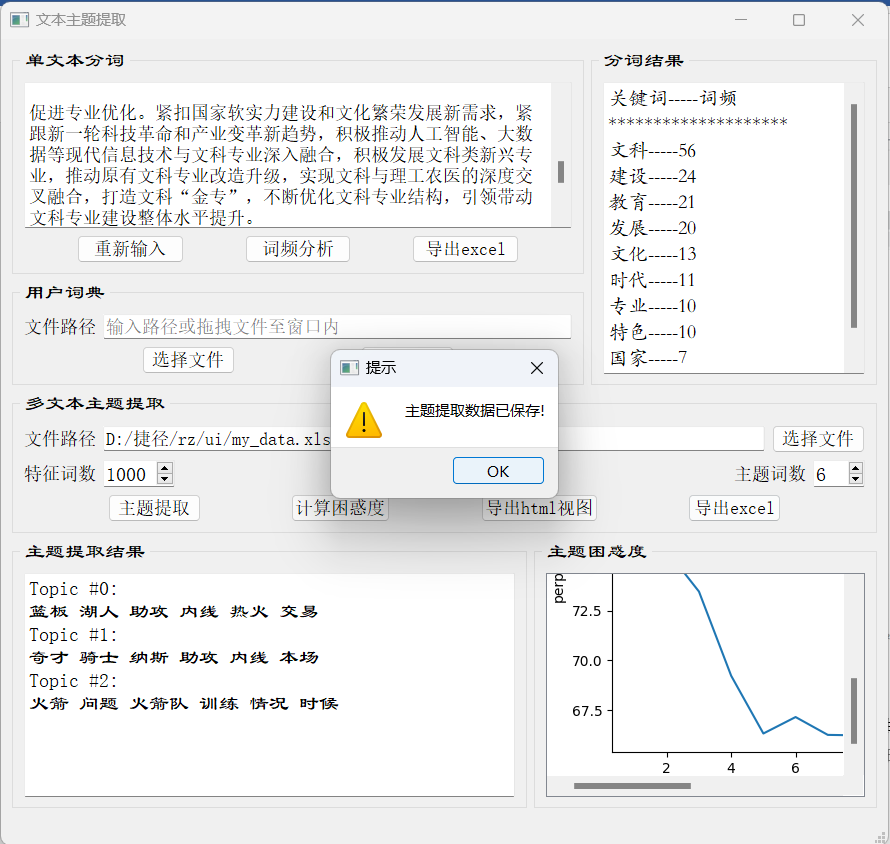


图 9主题词数据导出excel

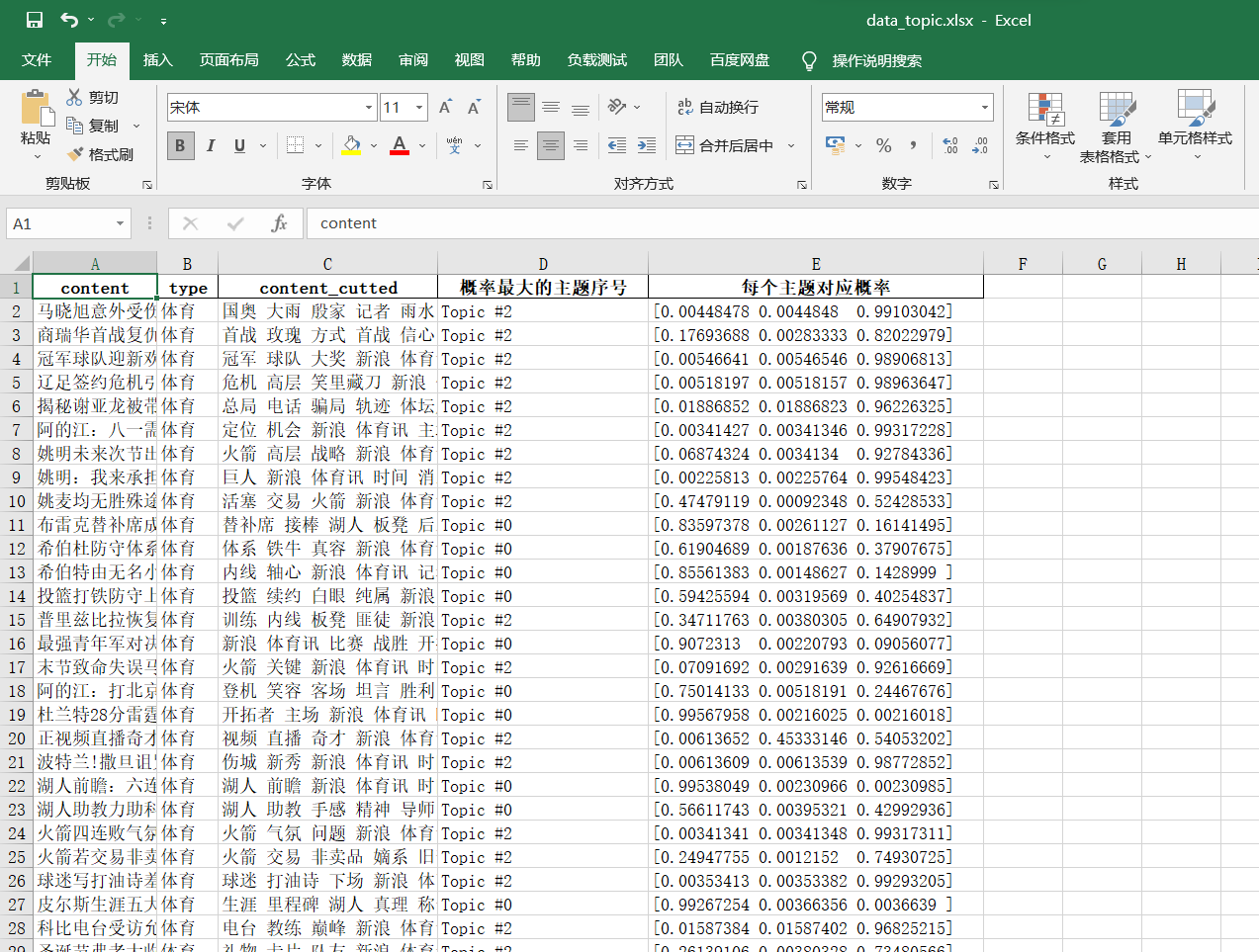


图 10主题提取数据预览

### 2.1.3添加用户词典

用户在进行单文本分词或多文本主题提取时，可通过该功能添加用户词典，以避免在分词处理时关键词丢失，如图6。文件导入方式与多文本分析类似，用户可通过拖拽、点击浏览和直接输入三种方式设置用户词典路径，并且，用户在进行文件拖拽时，只能将“txt”，“xls”，“xlsx”三种类型文件拖入，软件会根据文件类型对号入座：若拖入txt文件，则将路径设置为用户词典路径，若拖入xls文件或xlsx文件，则设置为分析文本数据路径。

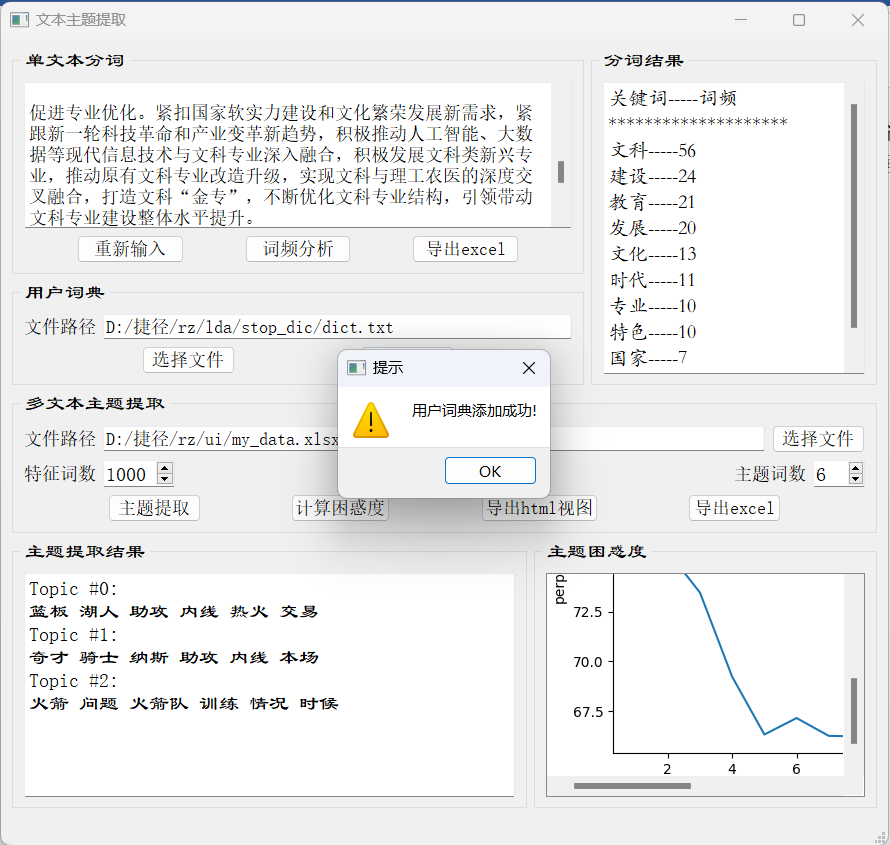


图 11添加用户词典

## 2.2性能介绍

### 2.2.1代码健壮性

软件的各个功能和操作之间有严格的先后顺序，对导入文件也有严格的要求。对于违规操作，软件会进行弹窗警告并中止操作，如：

在“词频分析”之前“导出分词excel或在”主题提取“之前导出html文件和excel文件，软件会有相应的弹窗警告（图12、13）。

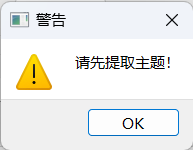
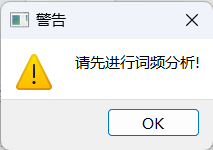


图 7操作顺序警告

当导入用户词典或分析文本时，若所选文件类型不合法，软件会立即弹窗警告，并提示选择正确的文件类型。

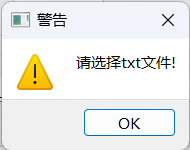
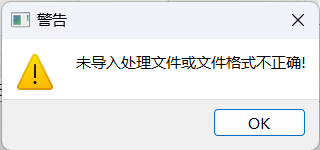


图 8文件类型违法警告

此外，对于某些操作结果无明显窗口展示的情况，软甲页设置了相应的窗口提示，如上图9、11。

### 2.2.2响应时间

当处理数据量较大时，在进行“主题提取“和”计算困惑度“操作时，软件会有明显的”迟钝“现象。该现象属于正常现象，由于软件需要时间对lda模型进行训练拟合，所以以上两个操作会有一定的响应延时。

# 3参考资料

1. <https://www.bilibili.com/video/BV1cJ411R7bP/?spm_id_from=333.337.search-card.all.click&vd_source=da5e0dd60be12895200f4a821f52fb26>
2. https://www.bilibili.com/video/BV1LQ4y1Q7xv/?spm\_id\_from=333.337.search-card.all.click&vd\_source=da5e0dd60be12895200f4a821f52fb26
3. https://mp.weixin.qq.com/s/hMcJtB3Lss1NBalXRTGZlQ
4. https://blog.csdn.net/qq\_39496504/article/details/107125284
5. https://blog.csdn.net/TiffanyRabbit/article/details/76445909
6. https://www.bilibili.com/video/BV1t54y127U8
7. https://www.jianshu.com/p/5c510694c07e
8. https://blog.csdn.net/weixin\_43343486/article/details/109255165
9. https://blog.csdn.net/weixin\_39676021/article/details/112187210