
# 爱文本(AItexts)智能文本分析系统

# 操作手册

目录

“爱文本(AItexts)智能文本分析系统”是由知名人工智能团队开发，经禾公司负责运维和市场推广的新一代智能文本分析产品。系统致力于为财经领域研究者提供精确、高效和可视化的文本分析工具，在文本数据、自由搜索、文本分析和文本挖掘四大领域打造零代码文本分析平台。

系统包括网页版和客户端版两个版本，其中网页版主要提供了文本数据筛选和下载和个人账户管理功能，客户端主要提供了自由搜索、文本分析和文本挖掘等功能(www.aitexts.com)。

# 1. 网页版

## 1.1. 登录

登录爱文本网址(www.aitexts.com)，输入个人账户的用户名和密码，点击登录，可以勾选“记住密码”方便二次登录(图 1-1)。


**图 1-1 网页版登录**

## 1.2. 首页

登录后，进入爱文本网页版，如图 1-2 所示。

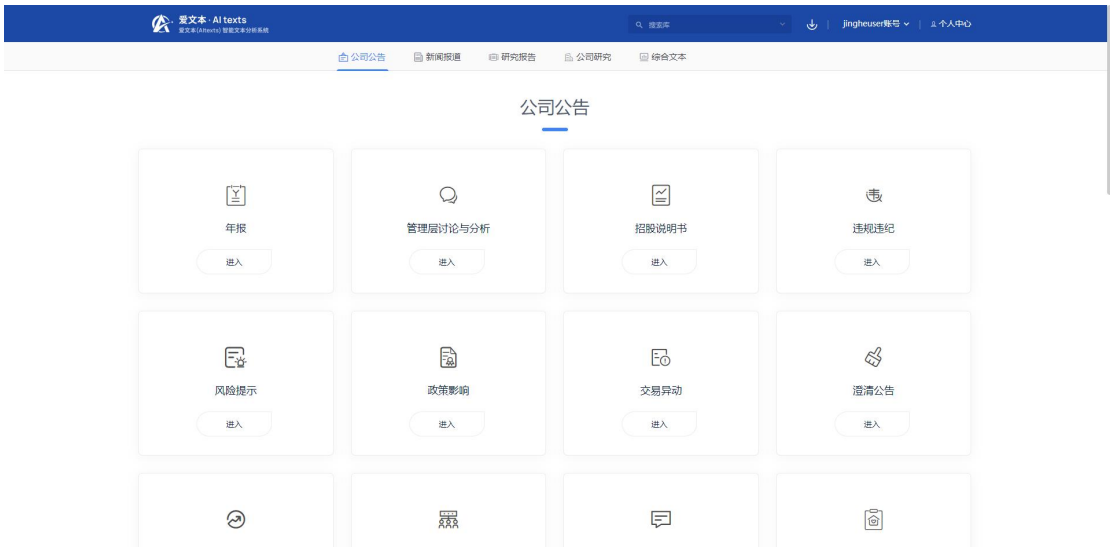



**图 1-2 网页版**

爱文本将数据库按公司公告，新闻报道，研究报告，公司研究，综合文本分门别类，用户可点击以上五个模块了解每个模块包含的数据库；也可根据自己所需数据库直接在搜索栏输入数据库名称快速精准找到所需数据(1-3)。


**图 1-3 网页版模块**

## 1.3. 搜索功能

爱文本上方搜索框可搜索数据库名称，帮助用户快速找到所需数据库以及相关数据库，点击数据库名称即可进入对应的数据库。以年报为例；搜索框中输入年报，查出先关数据库名称，点击银行年报，进入银行年报数据库


**图 1-4 关键词搜索**

# 1.4. 数据获取

## 1.4.1. 数据库列表

进入模块下面的具体数据库，如公司公告模块年报数据库。点击公司公告，弹出以下页面，找到年报数据库，点击进入。

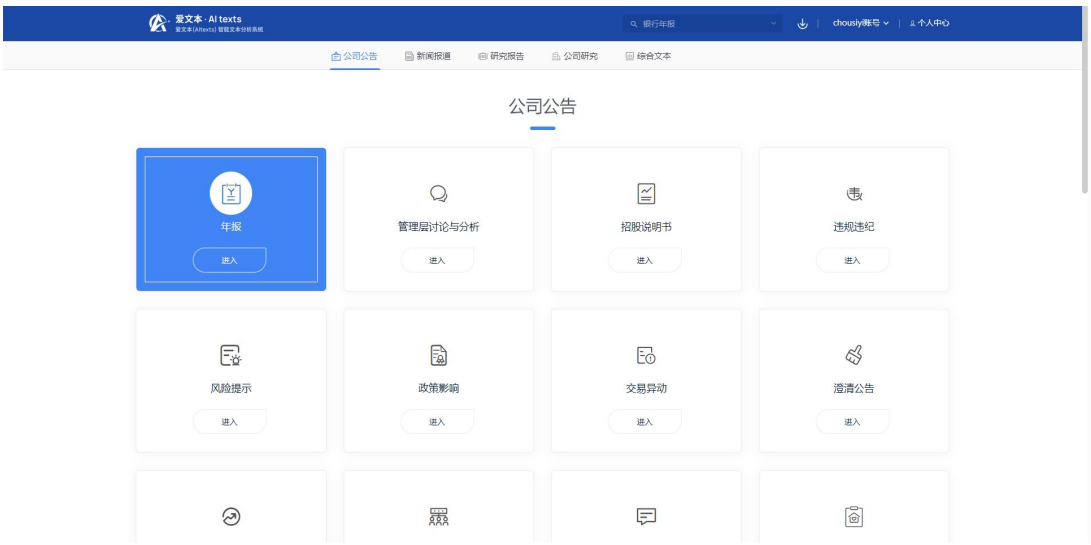



**图 1-5 数据库列表**

点击样本与字段说明，可了解表格相关的样本数据以及字段说明，如下图所示(图 1-6)。

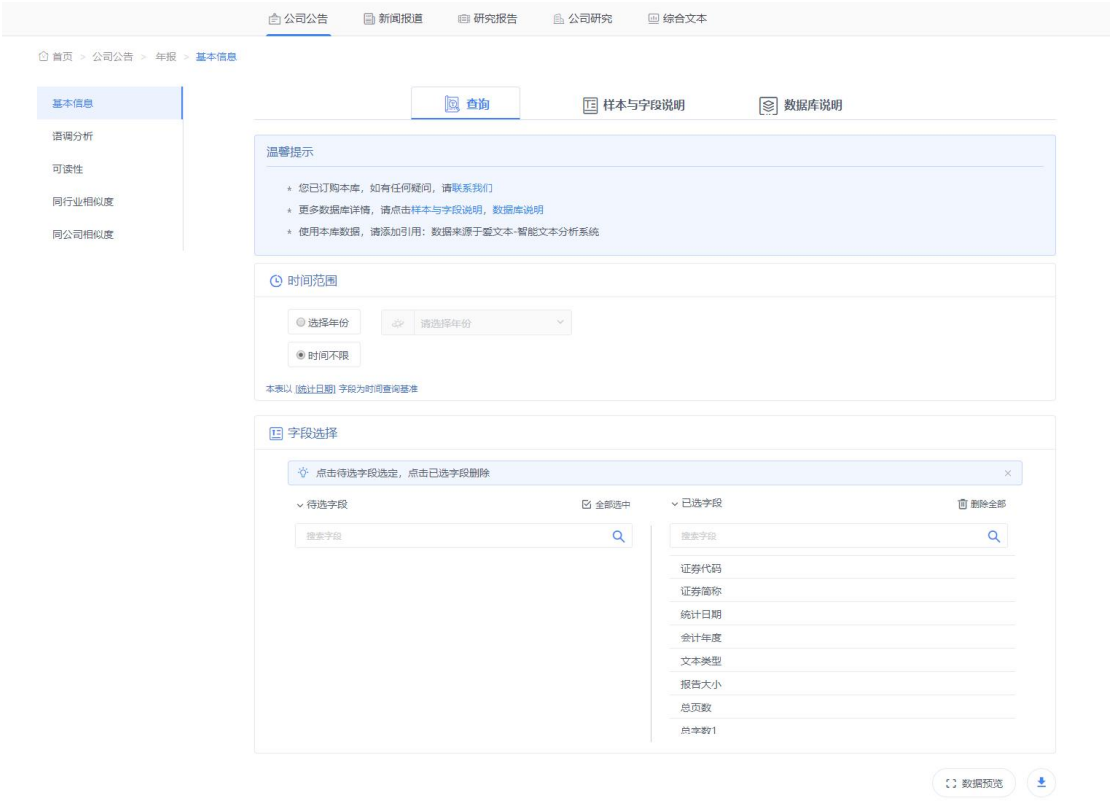



**图 1-6 数据预览和字段说明**

## 1.4.2. 数据下载

进入具体数据库后，可根据左上角选择所需下载的数据表格。可按需求在时间范围区域选择所需年份，字段选择区域选择所需字段，然后点击下载数据即可。如需下载所有年份全部数据，时间范围默认时间不限，字段选择默认全部选中，点击下载即可(图 1-7、1-8)。

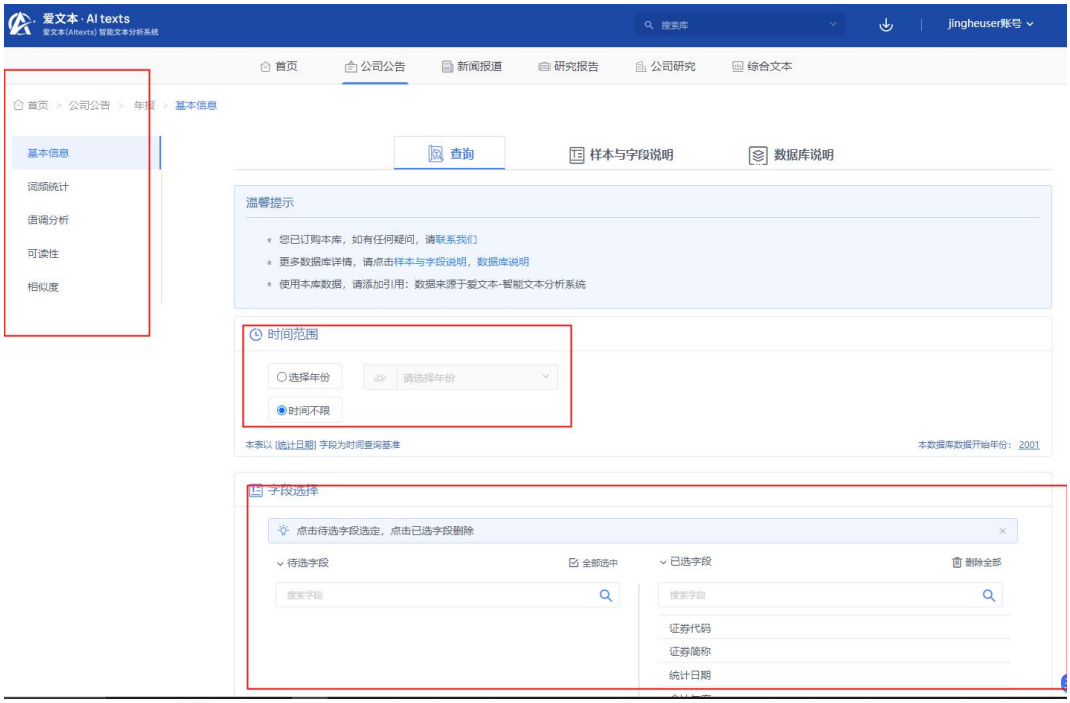



**图 1-7 下载字段选择**


**图 1-8 数据下载**

## 1.4.3. 个人中心

点击系统右上角“个人中心”，可以查看最近的下载记录，并且可以修改密码和改绑手机号(图 1-9)。

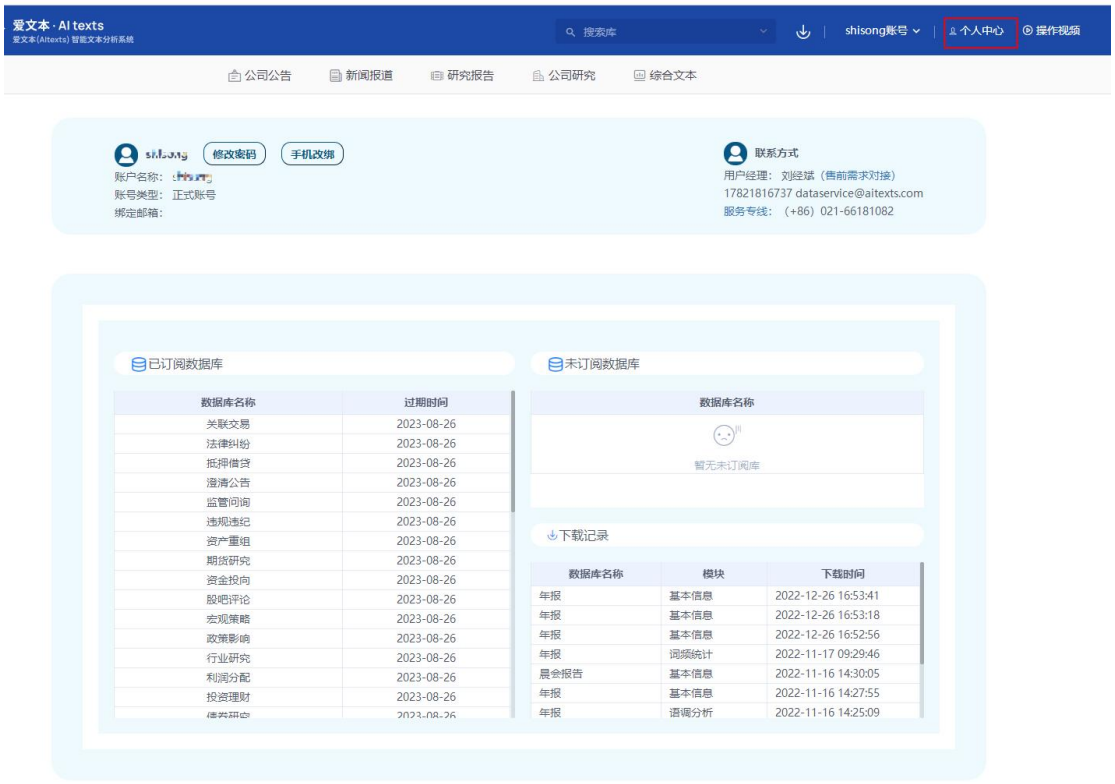



**图 1-9 个人中心**

## 1.4.4. 操作视频

用户登录后，点击系统右上角“操作视频”，可以查看网页版和客户端具体操作的相关多媒体视频，该多媒体录像包括音频和字幕(图 1-10)。

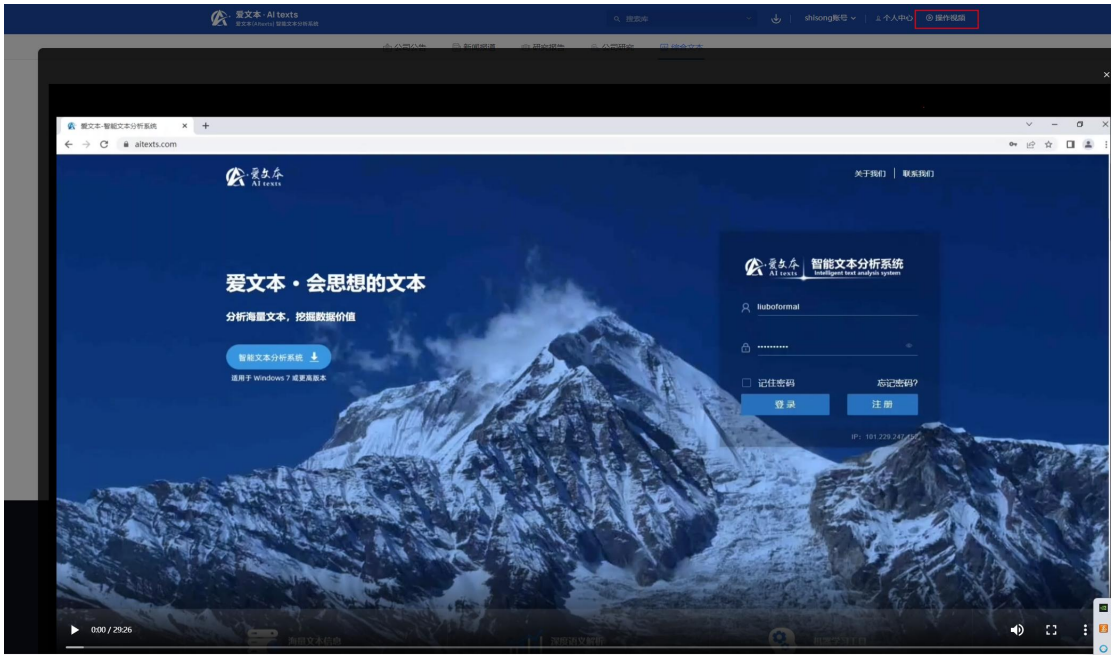



**图 1-10 操作视频**

# 2. 客户端

爱文本客户端致力于为财经领域研究者提供便利的可视化文本分析工具。系统主要包括自由搜索、文本分析、文本挖掘等三个模块。

## 2.1. 下载安装

下载客户端安装包，在网页版首页(www.aitexts.com)下载客户端安装包，下载完成后点击安装(图 2-1)。


**图 2-1 安装包下载**

## 2.2. 登录

登录客户端，输入个人账户的用户名和密码，点击登录，可以勾选“记住密码”方便二次登录(图 2-2)。

**图 2-2 客户端登录**

登录后，进入爱文本客户端首页，如下图所示。


**图 2-3 客户端首页**

## 2.3. 自由搜索

自由搜索模块提供了系统已有财经文本分析结果的一键下载导入搜索，用户也可以导入自定义文档实现一键自由搜索，用户可以在输入框中输入多个关键词（需要用英文逗号分开），也可以通过上传本地关键词文件（需要上传 TXT 文件，每个关键词占一行）(图 2-4)。

**图 2-4 自由搜索首页**

## 2.3.1. 自定义文本

在这一模块，用户可以导入自定义数据，统计用户输入关键词在各个文件中出现频率以及关键词相关词在各个文件中的出现频率等一些数据，点击右侧导出表格，可以导出 excel 格式的表格文件。

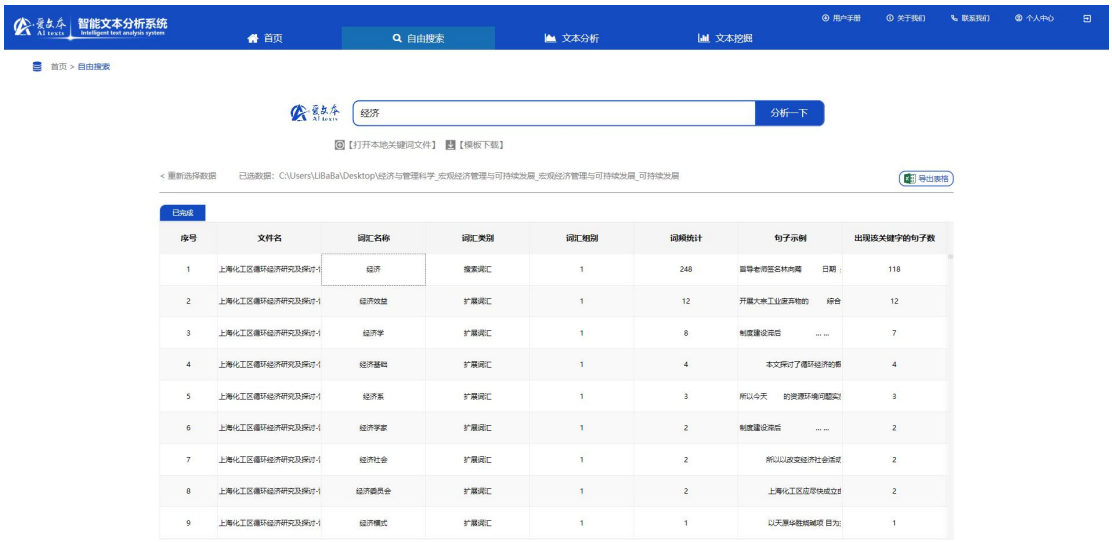



**图 2-5 自定义文本搜索结果**

## 2.3.2. 系统文本

在这一模块，用户查看系统已有文本，点击对应文本类型名称，查看该类型已有文本以及下载情况，可点击下载，删除，打开本地文件夹。


**图 2-6 系统已有文本类型列表**


**图 2-7 系统已有文本详细列表**

在这一模块，用户下载已有文本后导入，统计用户输入关键词在各个文件中出现频率以及关键词相关词在各个文件中的出现频率等一些数据，点击右侧导出表格，可以导出 excel 格式的表格文件。

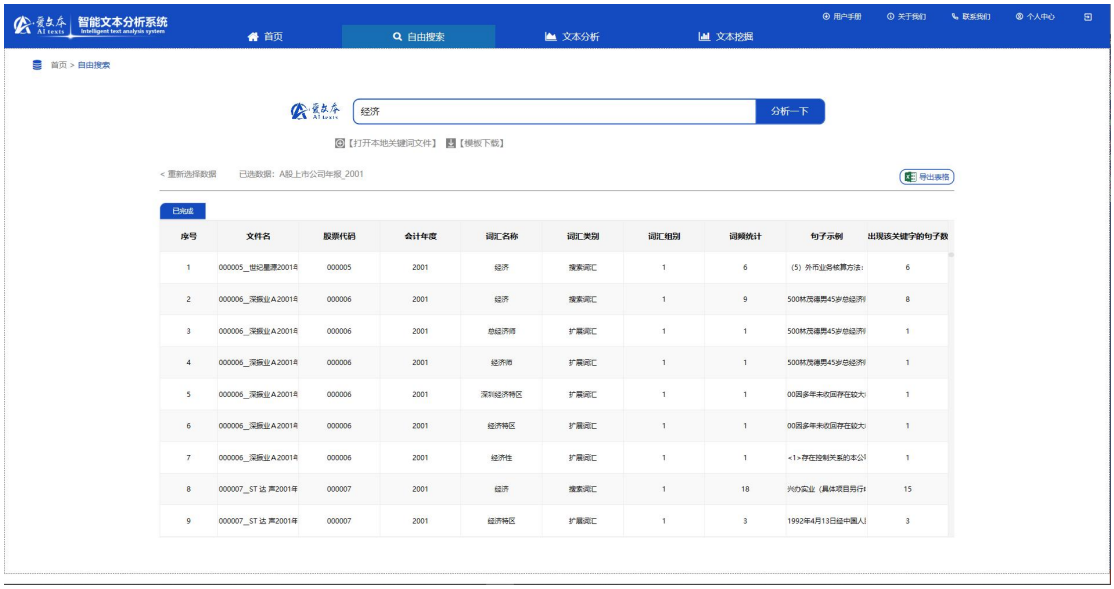



**图 2-8 文本搜索结果**

## 2.4. 文本分析

文本分析模块提供了文本研究领域常见指标的自动统计，包括基本信息统计、自定义情感分析、可读性统计等。

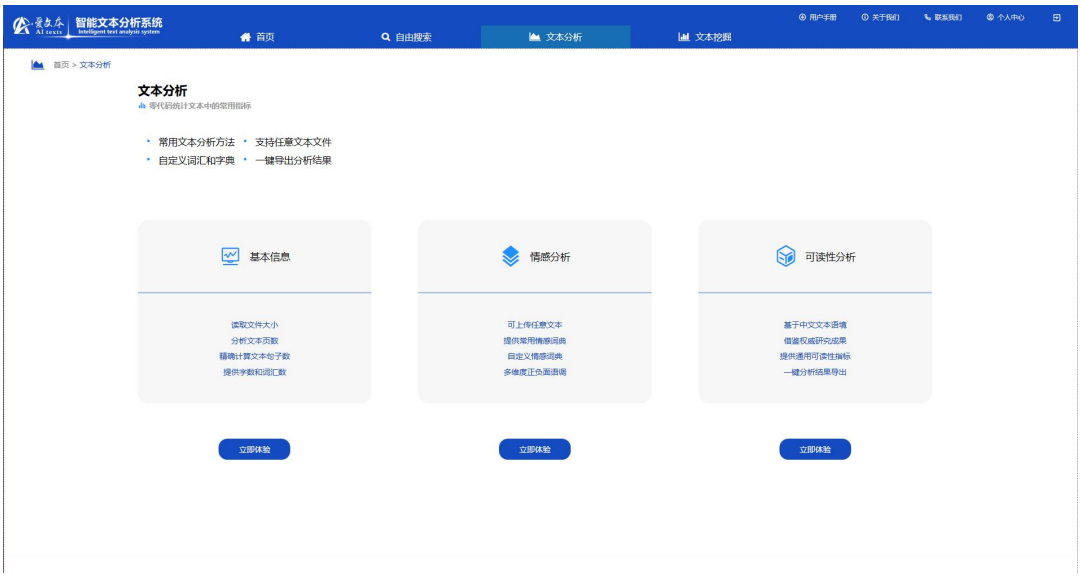



**图 2-9 文本分析模块**

### 2.4.1. 基本信息

该模块自动计算了用户自定义文本的一些基本信息，例如，文本 PDF 大小、

页数和字数等(图 2-10)。

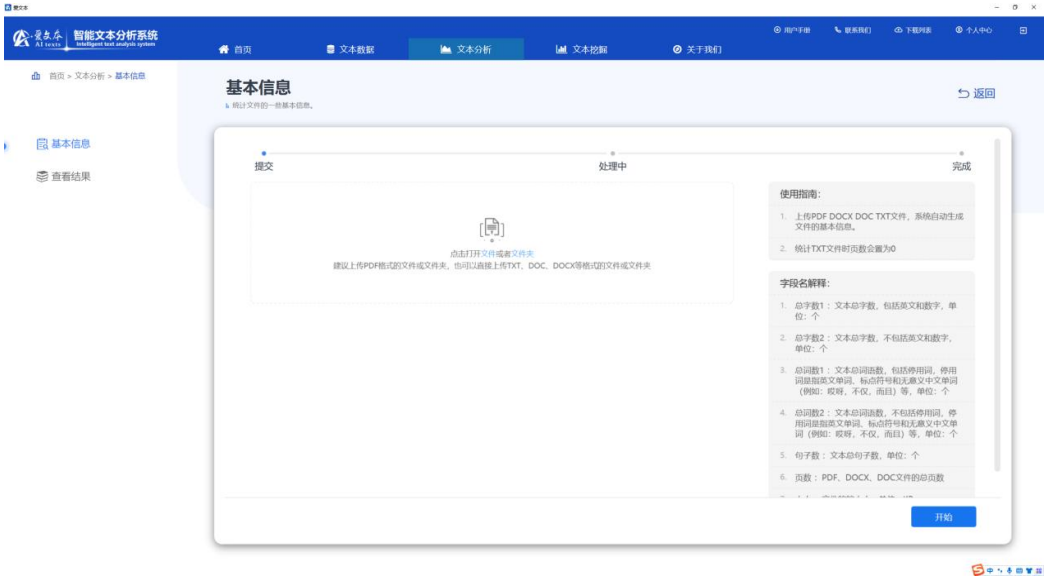



**图 2-10 基本信息统计**

## 2.4.2. 情感分析

用户可以单独或批量上传用户自定义的文件，根据爱文本系统提供的情感词典或者用户自定义情感词典，计算积极/消极词汇数、TONE 等情感分析结果。


**图 2-11 情感分析**

### 2.4.3. 可读性

用过可以选择自定义的文本或文本集合，系统自动统计一些较常用的可读性指标，并提供统计结果的 Excel 文档(图 2-12)。


**图 2-12 可读性分析**

## 2.5. 文本挖掘

这一模块首先将文本转化为在计算机中可以处理的中间形式，接下来集成了文本挖掘中的常见模型，包括相似度统计、文本分类、文本聚类和主题模型。用户无需编程，即可在这一模块挖掘得到文本中的模式和知识，并且可以进行模式和知识的评价、展示等。

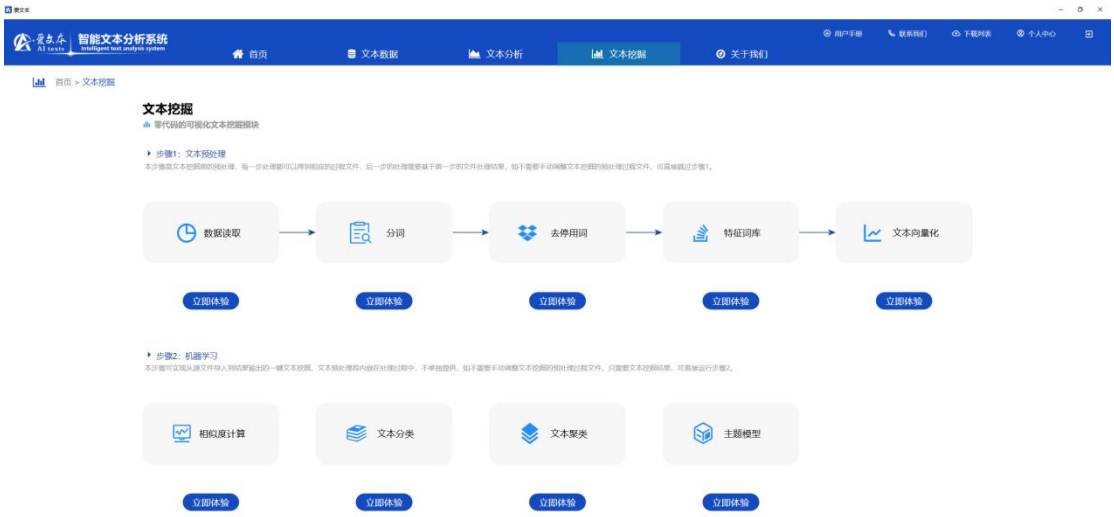



**图 2-13 文本挖掘主界面**

## 2.5.1. 数据读取

在财经领域，常见的文本包括年报、分析师报告、招股说明书、银行年报等，这些报告通常存储为 PDF、Doc、Docx 等格式。爱文本智能文本分析系统提供了将 PDF、Doc、Docx 直接转换为 TXT 的功能，用户可以通过输入文件或文件夹，直接读取 PDF、Doc、Docx 的内容，转换成方便计算机处理的 TXT 格式。

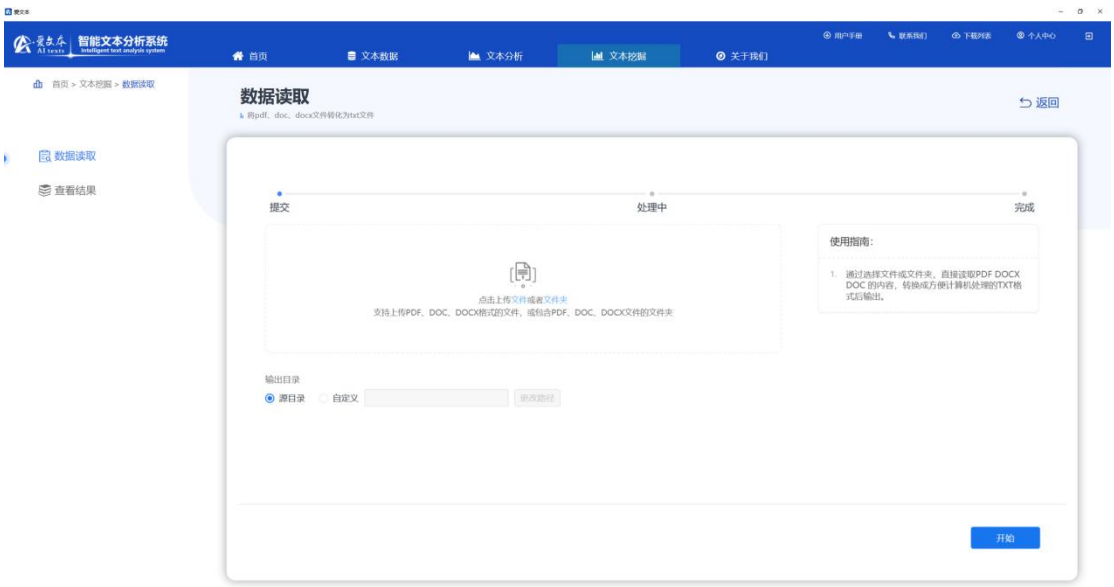



**图 2-14 数据转换**

## 2.5.2. 分词

中文分词是中文文本处理的一个基础步骤，也是爱文本智能文本分析系统的基础模块。不同于英文的是，中文句子中没有词的界限，因此在进行中文自然语言处理时，通常需要先进行分词，分词效果将直接影响文本分类、文本聚类等模块的效果。在系统中，单独或批量上传用户自定义的文件，可以得到全模式、精准模式和搜索引擎模式等多种分词结果。

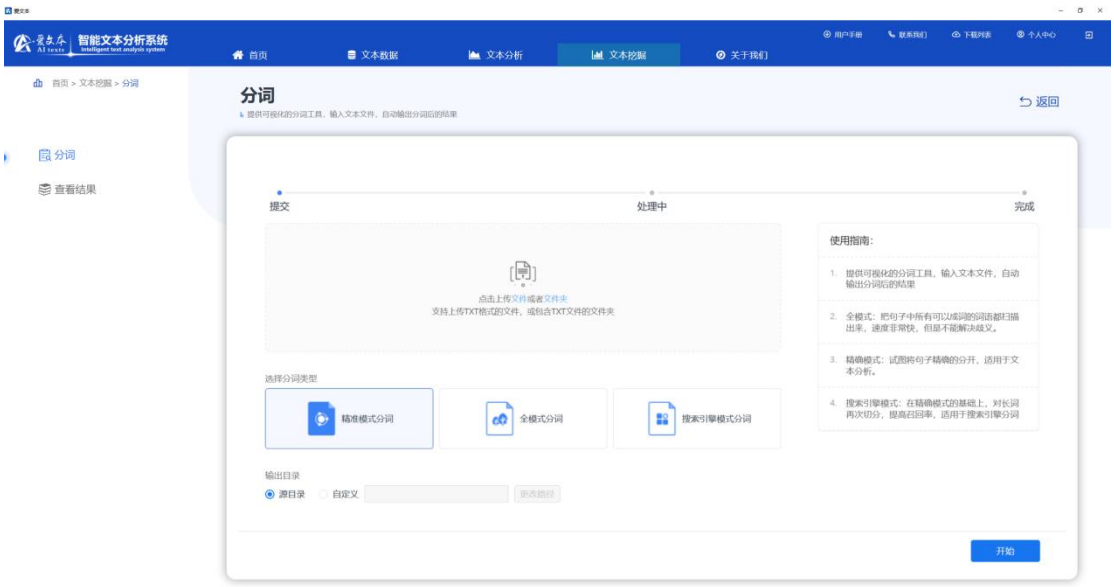



**图 2-15 分词界面**

## 2.5.3. 去停用词

自然语言处理过程中，原始语料库存在许许多多的停用词。停用词(Stopwords)是指在文 中出现频率较高但自身并无明确意义的词，比如，中文常见的“的、得、了、例如、吗、 大约、这、地”等词，这些词主要是副词、介词、连接词、助词等。

去停用词不但可以节省计算机的存储空间、提高执行效率，而且能在一定程度上使文本关键词更加集中、突出，文本语义表达更加明确，进一步提高正在执行任务的效果。在爱文本智能文本分析系统中，用户可以使用系统提供的停用词表去停用词，也可以通过自定义词表的方式去停用词。

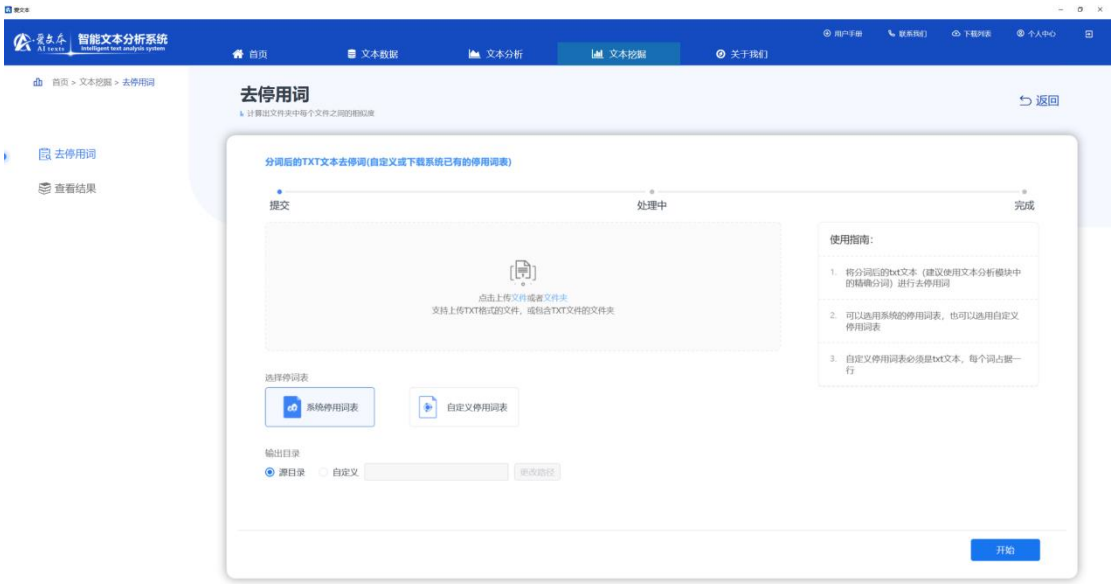



**图 2-16 去停用词**

## 2.5.4. 特征词库

在分词、去停用词文本集合中建立特征词库是文本向量化的基础工作。这部分工作内容包括从文本中提取关键词列表和统计每个词的文档频率两个部分(图 2-17)。


**图 2-17 构建特征词库**

图 2-18 为爱文本构建的特征词库示例，包括 5 个字段，分别是：①序号，特征词的编号；②文档数，包括特征词的文档数目；③词语，特征词；④IDF 值，特征词的逆向文档频(公式 3-1)；⑤文本总数，文本集合中的文档总数。

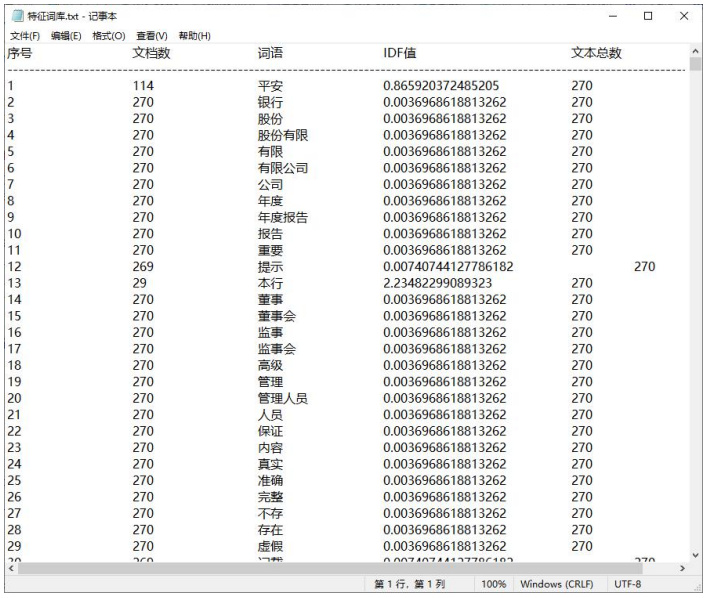



**图 2-18 特征词库示例**

## 2.5.5. 文本向量化

文本在计算机中存储和表示的方式，也会对文本挖掘任务产生较大影响。爱文本客户端提供了文本挖掘中常见的几种特征权重计算模块，分别是 TF-IDF 权重(TF-IDF Weighting)、TFC 权重(TFC Weighting)、LTC 权重(LTC Weighting)和熵权重(Entropy Weighting)。系统提供了预处理向量化(图 2-19)和一键向量化两种方式(图 2-20)。其中预处理向量化在分词、去停用词后的文本集合中进行向量化，一键向量化是在原始数据上直接进行向量化。

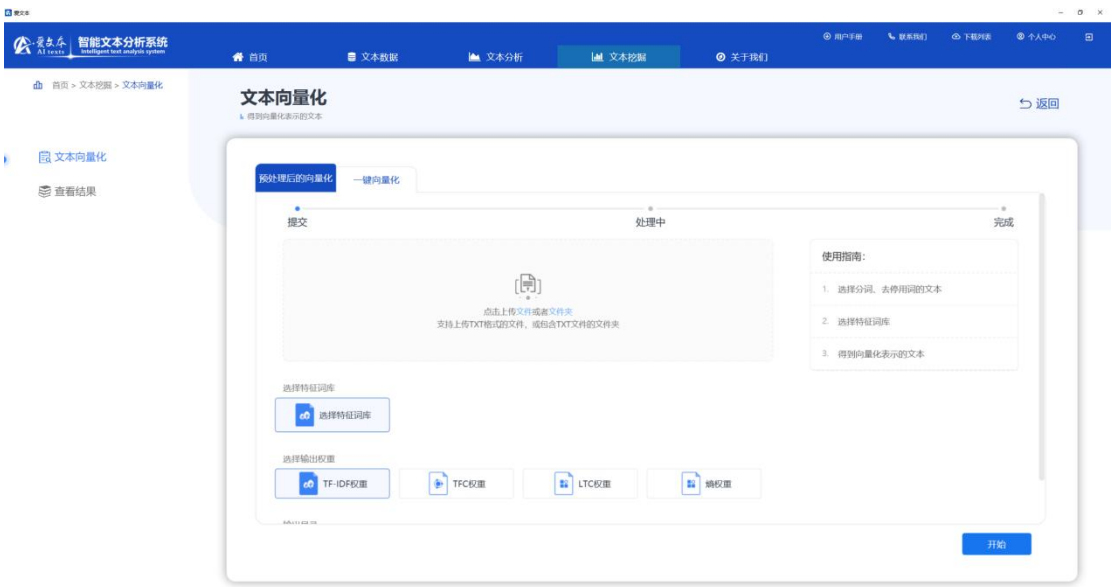


图 2-19 预处理后的向量化

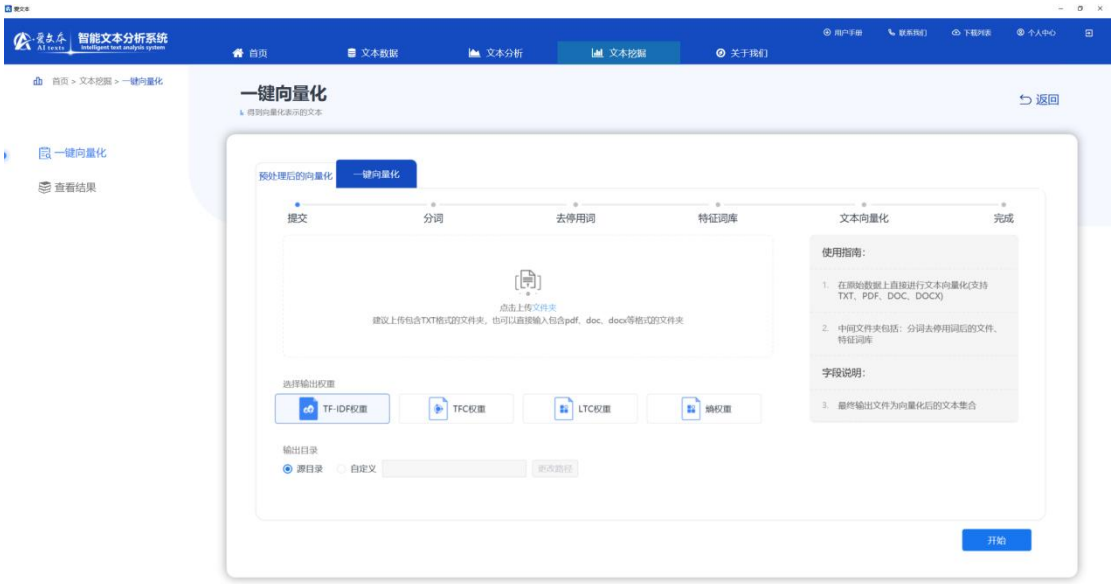



图 2-20 一键向量化

## 2.5.6. 相似度

在自然语言处理中，相似度是一种非常有用的工具，可以帮助研究者解决很多问题。爱文本智能文本分析系统提供了文本相似度计算模块，这一模块实现了三种常见的文本相似度计算方式，分别是：欧式距离、余弦距离和 Jacard 距离。系统统提供了预处理相似度分析(图 2-21)和一键相似度分析(图 2-22)。其中预处理相似度计算在向量化的文本集合中进行向量化，一键相似度分析是在原始数据上直接进行向量化。

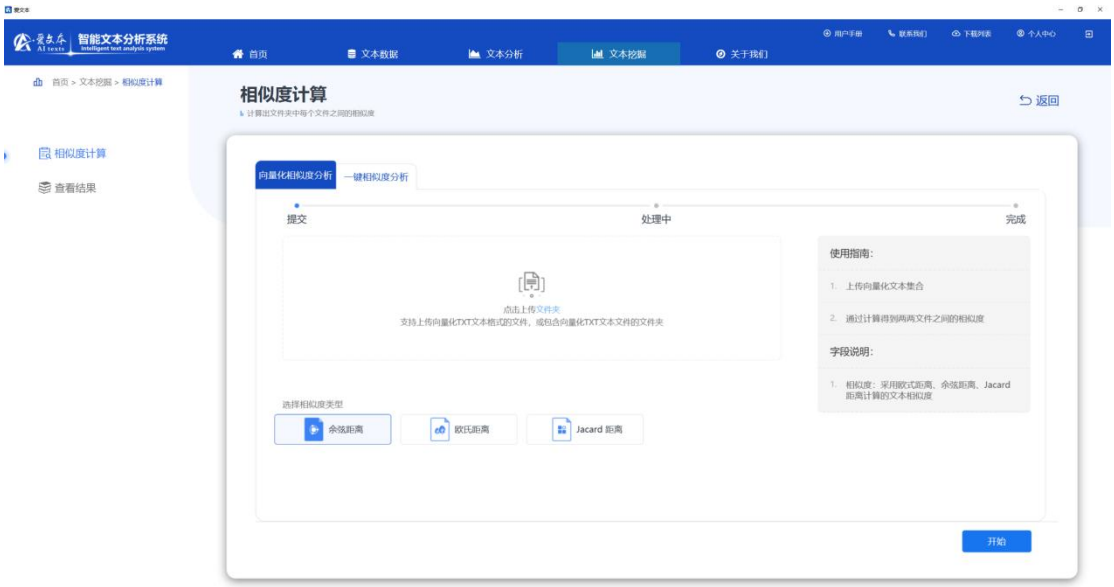



**图 2-21 预处理相似度分析**

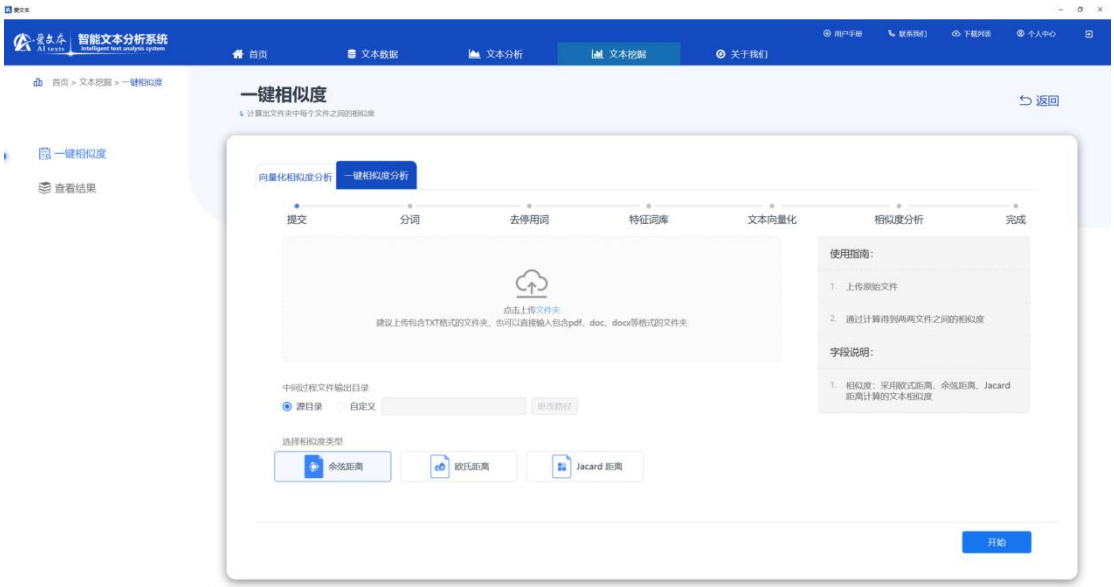



**图 2-22 一键相似度分析**

## 2.5.7. 文本分类

文本分类是爱文本智能文本分析系统的核心模块，在这一模块，需要将数据分为三大部分，分别是训练集、测试集、预测集。其中，训练集用于模型构建；测试集用于评估模型的准确率；构建好的模型在预测集中预测每个文本的类别。系统提供了预处理文本分类(图 2-23)和一键文本分类(图 2-24)。其中预处理相似度计算在向量化的文本集合中进行文本分类，一键相似度分析是在原始数据上直接进行文本分类。

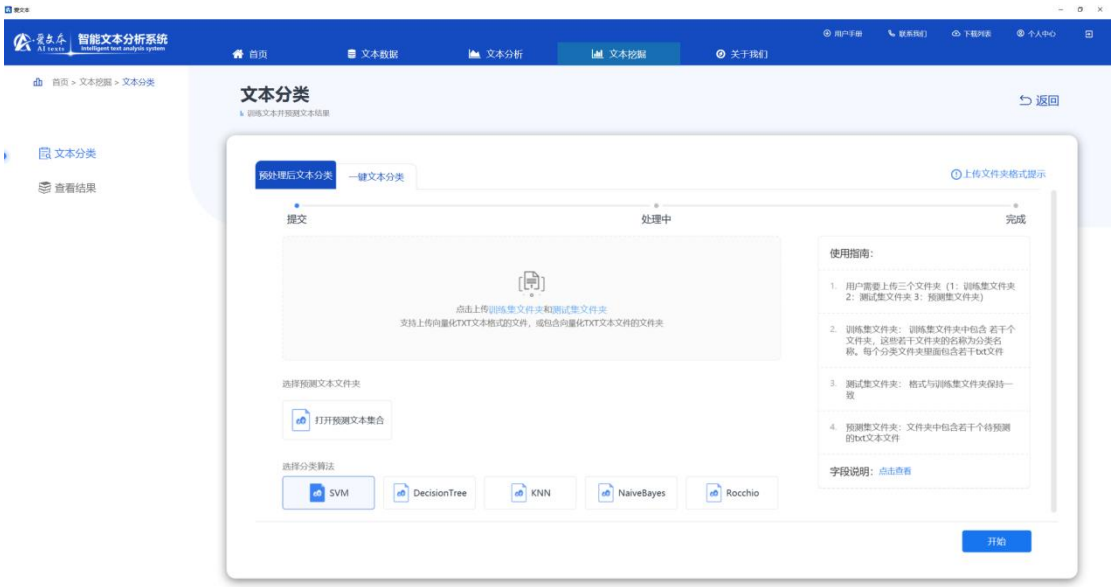



**图 2-23 预处理后的文本分类**

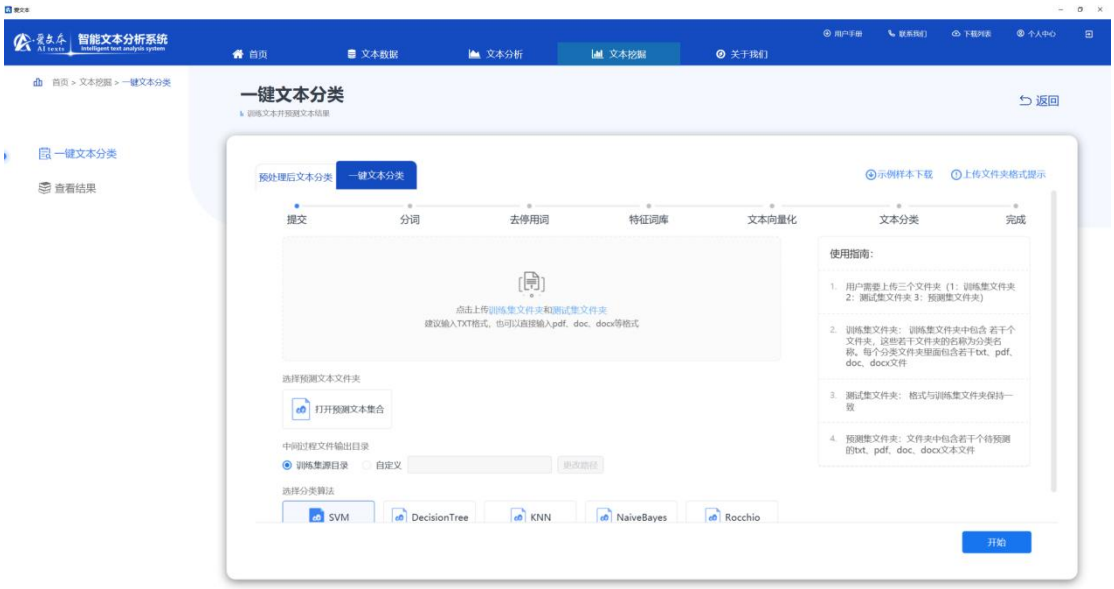



**图 2-24 一键文本分类**

## 2.5.8. 文本聚类

文本聚类就是根据在文档数据集中文档表示形式及文档间关系的信息，将文档对象分组的过程。其目标就是使聚类后组内的文档之间是相似的，组间的文档是相异的。组内的相似性(同质性)越大，组间差别越大，聚类效果就越好。爱文本智能文本分析系统提供了两种常见的文本聚类算法，分别是 K-Means 聚类和 DBSCAN 聚类算法。

系统提供了预处理文本分类(图 2-25)和一键文本分类(图 2-26)。其中预处理相似度计算在向量化的文本集合中进行文本分类，一键相似度分析是在原始数据上直接进行文本分类。

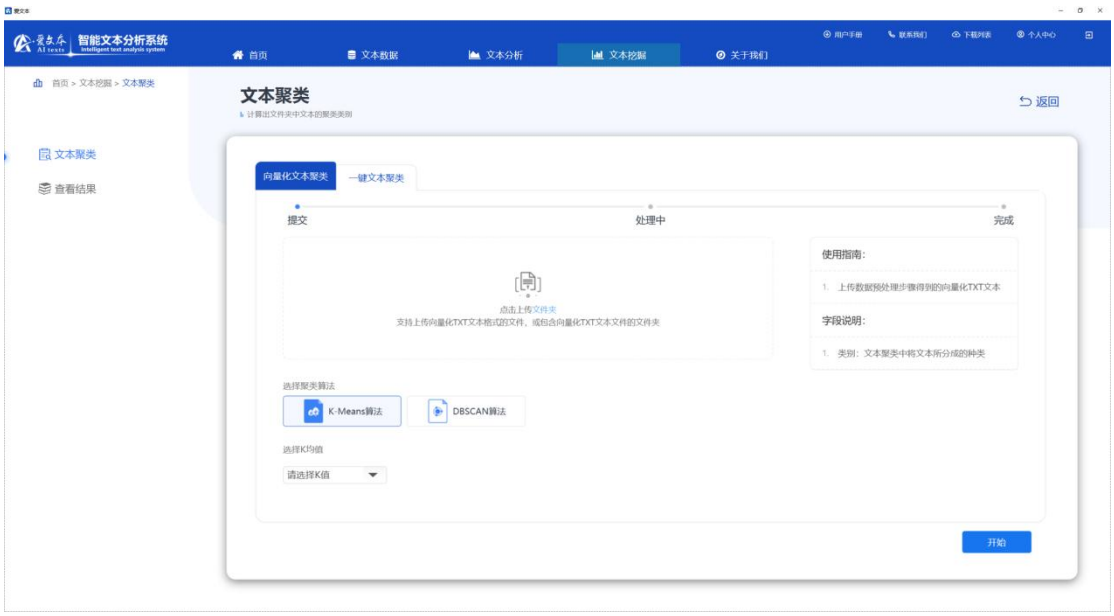



**图 2-25 预处理后的文本聚类**

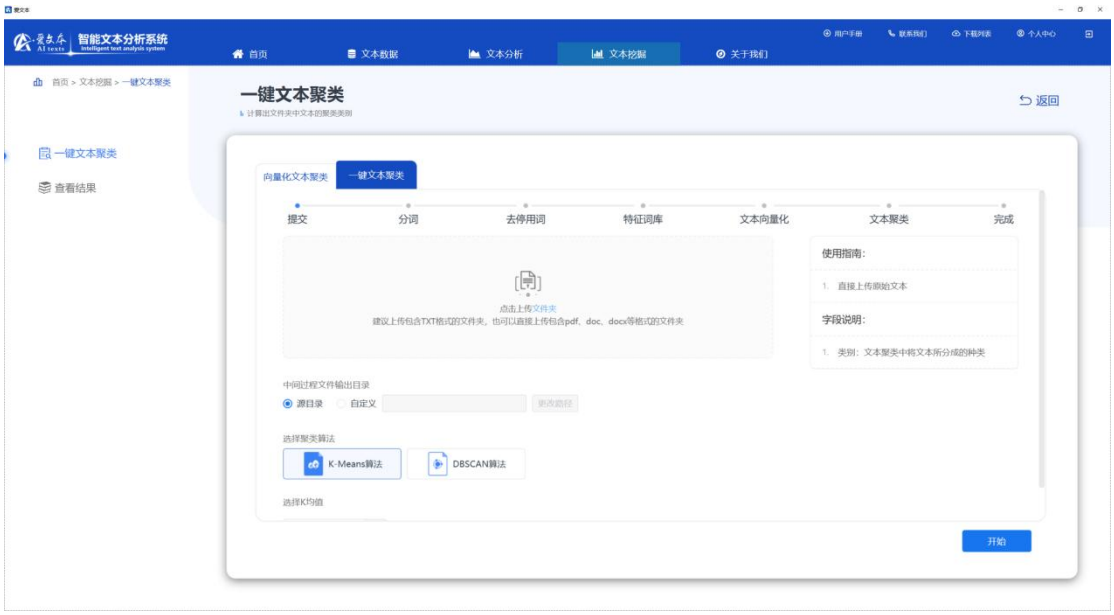



**图 2-26 一键文本聚类**

## 2.5.9. 主题模型

爱文本智能文本分析系统实现了 LDA 主题模型(图 2-27)，输入为文本集合，输出为不同主题的主题词(图 2-28)和不同文本的主题分布(图 2-29)。

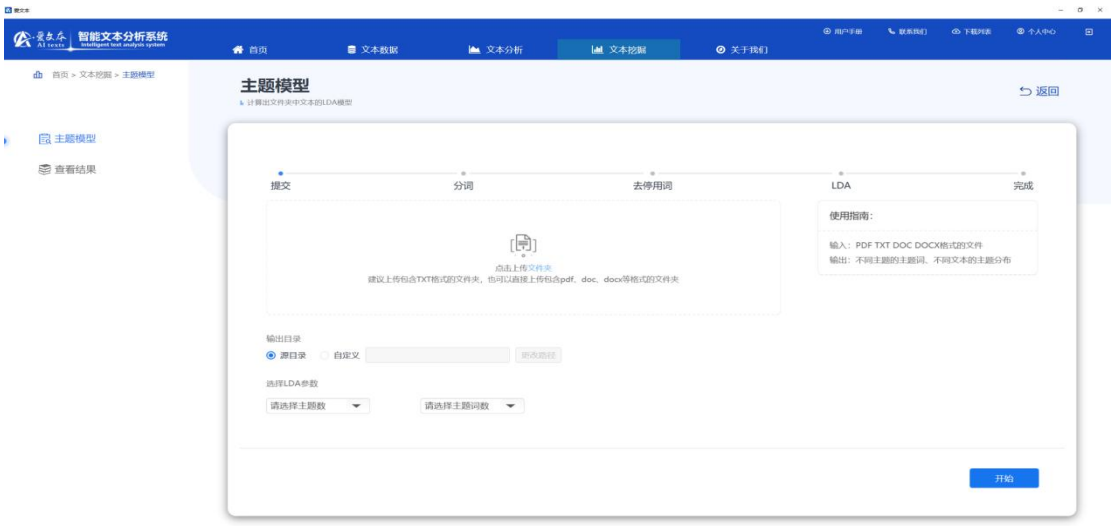
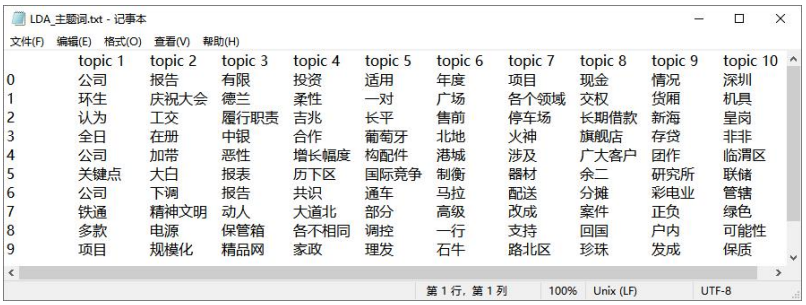



**图 2-27 LDA 主题模型**



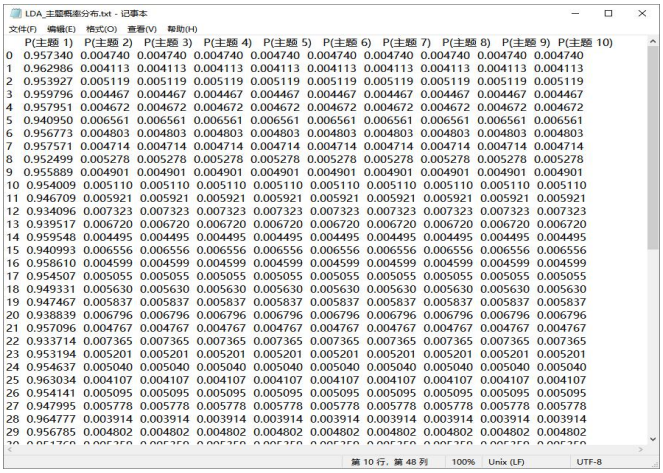
LDA_主题词.txt - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

| | topic 1 | topic 2 | topic 3 | topic 4 | topic 5 | topic 6 | topic 7 | topic 8 | topic 9 | topic 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 公司 | 报告 | 有限 | 投资 | 适用 | 年度 | 项目 | 现金 | 情况 | 深圳 |
| 1 | 环生 | 庆祝大会 | 德兰 | 柔性 | 一对 | 广场 | 各个领域 | 交权 | 货用 | 机具 |
| 2 | 认为 | 工交 | 履行职责 | 吉兆 | 长平 | 售前 | 停车场 | 长期借款 | 新海 | 皇岗 |
| 3 | 全日 | 在册 | 中银 | 合作 | 葡萄牙 | 北地 | 火神 | 旗舰店 | 存贷 | 非非 |
| 4 | 公司 | 加带 | 恶性 | 增长幅度 | 构配件 | 港城 | 涉及 | 广大客户 | 团作 | 临渭区 |
| 5 | 关键点 | 大白 | 报表 | 历下区 | 国际竞争 | 制衡 | 器材 | 余二 | 研究所 | 联储 |
| 6 | 公司 | 下调 | 报告 | 共识 | 通车 | 马拉 | 配送 | 分摊 | 彩电业 | 管辖 |
| 7 | 铁通 | 精神文明 | 动人 | 大道北 | 部分 | 高级 | 改成 | 案件 | 正负 | 绿色 |
| 8 | 多款 | 电源 | 保管箱 | 各不相同 | 调控 | 一行 | 支持 | 回国 | 户内 | 可能性 |
| 9 | 项目 | 规模化 | 精品网 | 家政 | 理发 | 石牛 | 路北区 | 珍珠 | 发成 | 保质 |

第 1 行，第 1 列 100% Unix (LF) UTF-8

**图 2-28 主题关键词**



LDA_主题概率分布.txt - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

| | P(主题 1) | P(主题 2) | P(主题 3) | P(主题 4) | P(主题 5) | P(主题 6) | P(主题 7) | P(主题 8) | P(主题 9) | P(主题 10) |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.957340 | 0.004740 | 0.004740 | 0.004740 | 0.004740 | 0.004740 | 0.004740 | 0.004740 | 0.004740 | 0.004740 |
| 1 | 0.962986 | 0.004113 | 0.004113 | 0.004113 | 0.004113 | 0.004113 | 0.004113 | 0.004113 | 0.004113 | 0.004113 |
| 2 | 0.953927 | 0.005119 | 0.005119 | 0.005119 | 0.005119 | 0.005119 | 0.005119 | 0.005119 | 0.005119 | 0.005119 |
| 3 | 0.959796 | 0.004467 | 0.004467 | 0.004467 | 0.004467 | 0.004467 | 0.004467 | 0.004467 | 0.004467 | 0.004467 |
| 4 | 0.957951 | 0.004672 | 0.004672 | 0.004672 | 0.004672 | 0.004672 | 0.004672 | 0.004672 | 0.004672 | 0.004672 |
| 5 | 0.940950 | 0.006561 | 0.006561 | 0.006561 | 0.006561 | 0.006561 | 0.006561 | 0.006561 | 0.006561 | 0.006561 |
| 6 | 0.956773 | 0.004803 | 0.004803 | 0.004803 | 0.004803 | 0.004803 | 0.004803 | 0.004803 | 0.004803 | 0.004803 |
| 7 | 0.957571 | 0.004714 | 0.004714 | 0.004714 | 0.004714 | 0.004714 | 0.004714 | 0.004714 | 0.004714 | 0.004714 |
| 8 | 0.952499 | 0.005278 | 0.005278 | 0.005278 | 0.005278 | 0.005278 | 0.005278 | 0.005278 | 0.005278 | 0.005278 |
| 9 | 0.955889 | 0.004901 | 0.004901 | 0.004901 | 0.004901 | 0.004901 | 0.004901 | 0.004901 | 0.004901 | 0.004901 |
| 10 | 0.954009 | 0.005110 | 0.005110 | 0.005110 | 0.005110 | 0.005110 | 0.005110 | 0.005110 | 0.005110 | 0.005110 |
| 11 | 0.946709 | 0.005921 | 0.005921 | 0.005921 | 0.005921 | 0.005921 | 0.005921 | 0.005921 | 0.005921 | 0.005921 |
| 12 | 0.934096 | 0.007323 | 0.007323 | 0.007323 | 0.007323 | 0.007323 | 0.007323 | 0.007323 | 0.007323 | 0.007323 |
| 13 | 0.939517 | 0.006720 | 0.006720 | 0.006720 | 0.006720 | 0.006720 | 0.006720 | 0.006720 | 0.006720 | 0.006720 |
| 14 | 0.959548 | 0.004495 | 0.004495 | 0.004495 | 0.004495 | 0.004495 | 0.004495 | 0.004495 | 0.004495 | 0.004495 |
| 15 | 0.940993 | 0.006556 | 0.006556 | 0.006556 | 0.006556 | 0.006556 | 0.006556 | 0.006556 | 0.006556 | 0.006556 |
| 16 | 0.958610 | 0.004599 | 0.004599 | 0.004599 | 0.004599 | 0.004599 | 0.004599 | 0.004599 | 0.004599 | 0.004599 |
| 17 | 0.954507 | 0.005055 | 0.005055 | 0.005055 | 0.005055 | 0.005055 | 0.005055 | 0.005055 | 0.005055 | 0.005055 |
| 18 | 0.949331 | 0.005630 | 0.005630 | 0.005630 | 0.005630 | 0.005630 | 0.005630 | 0.005630 | 0.005630 | 0.005630 |
| 19 | 0.947467 | 0.005837 | 0.005837 | 0.005837 | 0.005837 | 0.005837 | 0.005837 | 0.005837 | 0.005837 | 0.005837 |
| 20 | 0.938839 | 0.006796 | 0.006796 | 0.006796 | 0.006796 | 0.006796 | 0.006796 | 0.006796 | 0.006796 | 0.006796 |
| 21 | 0.957096 | 0.004767 | 0.004767 | 0.004767 | 0.004767 | 0.004767 | 0.004767 | 0.004767 | 0.004767 | 0.004767 |
| 22 | 0.933714 | 0.007365 | 0.007365 | 0.007365 | 0.007365 | 0.007365 | 0.007365 | 0.007365 | 0.007365 | 0.007365 |
| 23 | 0.953194 | 0.005201 | 0.005201 | 0.005201 | 0.005201 | 0.005201 | 0.005201 | 0.005201 | 0.005201 | 0.005201 |
| 24 | 0.954637 | 0.005040 | 0.005040 | 0.005040 | 0.005040 | 0.005040 | 0.005040 | 0.005040 | 0.005040 | 0.005040 |
| 25 | 0.963034 | 0.004107 | 0.004107 | 0.004107 | 0.004107 | 0.004107 | 0.004107 | 0.004107 | 0.004107 | 0.004107 |
| 26 | 0.954141 | 0.005095 | 0.005095 | 0.005095 | 0.005095 | 0.005095 | 0.005095 | 0.005095 | 0.005095 | 0.005095 |
| 27 | 0.947995 | 0.005778 | 0.005778 | 0.005778 | 0.005778 | 0.005778 | 0.005778 | 0.005778 | 0.005778 | 0.005778 |
| 28 | 0.964777 | 0.003914 | 0.003914 | 0.003914 | 0.003914 | 0.003914 | 0.003914 | 0.003914 | 0.003914 | 0.003914 |
| 29 | 0.956785 | 0.004802 | 0.004802 | 0.004802 | 0.004802 | 0.004802 | 0.004802 | 0.004802 | 0.004802 | 0.004802 |

第 10 行，第 48 列 100% Unix (LF) UTF-8

**图 2-29 文档主题分布**

# 全文结束

**感谢您的关注！**
**如有任何问题和需求，请联系我们！**
**电话：021-66181082**
**邮箱：liujingbin@aitexts.com**