

传承文脉 致敬经典

中国近代文献数据库

—— 1840-1949 ——

浙江越生文化传媒集团
浙江 · 绍兴



浙江越生文化传媒集团

浙江越生文化传媒集团成立于1992年，经过三十年发展，实现了从实业生产到文化出版再到数字化建设的跨越式发展，目前已经成为一家集策划、编纂、设计、印刷、出版、发行、展陈、艺术品拍卖、互联网出版物平台交易于一体的全产业链现代文化企业。下属六家实体公司，在绍兴和北京均设有办公地点。集团多次荣获中国包装印刷产品质量评比金奖和多项技术专利，出版图书先后入选国家“十三五”“十四五”重点出版项目、浙江省重大文化研究工程，相关数据库入选“国家文化产业发展项目库”，先后被评为“浙江省第四批成长型企业”“国家文化类高新企业”“浙江省数字文化示范企业（30强）”。

近年来，浙江越生全力建设“中国近代文献保护工程”和“数字人文建设平台”两大工程，加强优秀文化创造性转化与创新性利用，推动文化基因数字解码；全力打造“全国百强名社数字出版物交易阅读平台”，助力“文化共富”，推动校企合作的同时取得了良好的社会效益、经济效益和学术影响力。

中国近代文献保护工程简介

“中国近代文献保护工程”与“中国近代文献数据总库”

“中国近代文献保护工程”抓住信息化时代机遇，从中华优秀传统文化根基中寻找重点题材，树立开放性、系统性、科学性理念，旨在借势再造民族文化凝聚力与创造力。集团已通过全球范征集**1840年--1949年**图书十几万种，为“工程”的成果--《中国近代文献史料丛刊》的出版提供了内容的丰富性、版本的优质性。在此基础上，由国内各领域内顶尖专家组成编委会，对内容进行甄选整理，并与国内著名出版社合作出。

“中国近代文献数据总库”在《中国近代文献史料丛刊》基础上，运用图书情报、数字人文、中文信息化、互联网、大数据、区块链等相关标准与技术，构建 B/S 模式的，在云端为广大读者、学者、决策者等使用的资源检索与知识服务平台，及大方便了广大读者，研究人员等阅读、利用这些珍惜资源。该数据库分为七个主要分库，分别为文学、经济、教育、艺术、史地、科技和政治。

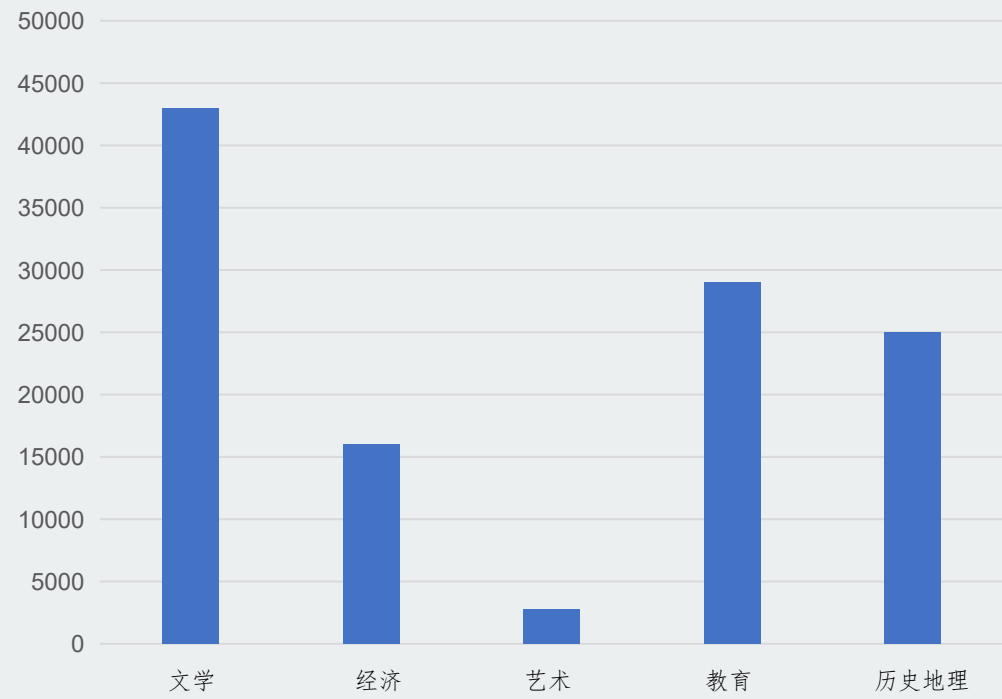
目前，已建成发布三个子库《中国近代文学文献数据库》、《中国近代教育文献数据库》和《中国近代艺术文献数据库》。

中国近代文献保护工程

- 文学 已采集 43000种
- 经济 已采集 16000种
- 艺术 已采集 2800种
- 教育 已采集 29000种
- 历史地理 已采集 25000种

注：合计十余万种，各种资源仍在全球范围内持续采集中

数据表



通过全球性征集1840年 — 1949年期间的图书

中国近代文献数据总库--特点

- **高质量、专业的资源:**由越生集团联合国内各领域顶尖专家组成编委会共同策划开发，旨在为学术界和研究领域提高质量、专业的资源。拒绝“捡到篮子里就是菜”。
- **历史和文献收藏价值:**库内收集的文献约80%都是从20世纪70年代起就再没印刷出版过，具有珍贵的历史价值和文献收藏价值。
- **权威性和准确性:**数据库的内容经过了专业人员的审核和整理，并由国内著名出版社出版，纸电同步，保证了其内容的权威性和准确性。
- **阅读更方便:**所有数据经过OCR深度处理，支持图文对照（横排，竖排，简繁）阅读，让用户能够在检索结果中同时查看文本和相关图片，从而更加直观地了解文献内容。

中国近代文献数据总库--特点

- **使用更灵活：**多种标签组合浏览功能，如“分类导航”、“文献出版者”导航、“作者信息”导航、“文献出版年代”导航；书目各项单一及组配检索，文献内全文检索功能--实现“**字字可检索，句句可查询**”。
- **方便深度应用：**系统提供个人空间，对于专业研究用户，可实现下载、收藏、上传个人文献等功能，构建真正的个人书架，方便研究。
- **前沿技术应用：**数据库于2023年开发完成，推向市场，较前些年应用在数据库领域技术得到了巨大发展，尤其是人工智能技术的广泛应用，使本库具有实时语意搜索、知识图谱、热点分析等独有功能，极大的方便了读者。
- **文献著录标准规范：**不同于其它文库资源堆砌或简单分类，而是在文献元数据标引等遵循《中国文献编目规则》《中国机读目录格式》等标准规范。
- **丰富文献资源：**目前已征集十几万种资源，并仍在持续征集，可以为学术研究持续提供丰富、权威的文献资源。

标准和核心技术

RDF+Schema三元组描述标准

三元组知识本体标准描述文内知识体

异体字认同标准与流程

基于国际标准ISO/IEC 10646:2017中的异体字认同标准

字符集标准

ISO/IEC 10646 : 2017共收录87875汉字

文献著录标准规范

即文献元数据标引，不仅遵循《中国文献编目规则》
《中国机读目录格式》，还要根据建库制定的实施细则，要求建库人员共同遵守。

04

03

02

01

05

06

07

08

自动版面分析技术及OCR识别技术

运用神经网络技术，训练海量的近代文献版面及字迹，使得识别率达到百分之九十以上,并配有在线校对工具。

自动分词技术

利用自然语言理解，对文献文本内容进行自动分词标引，应用于热词词云的形成，经济类语言发展演变等功能中。

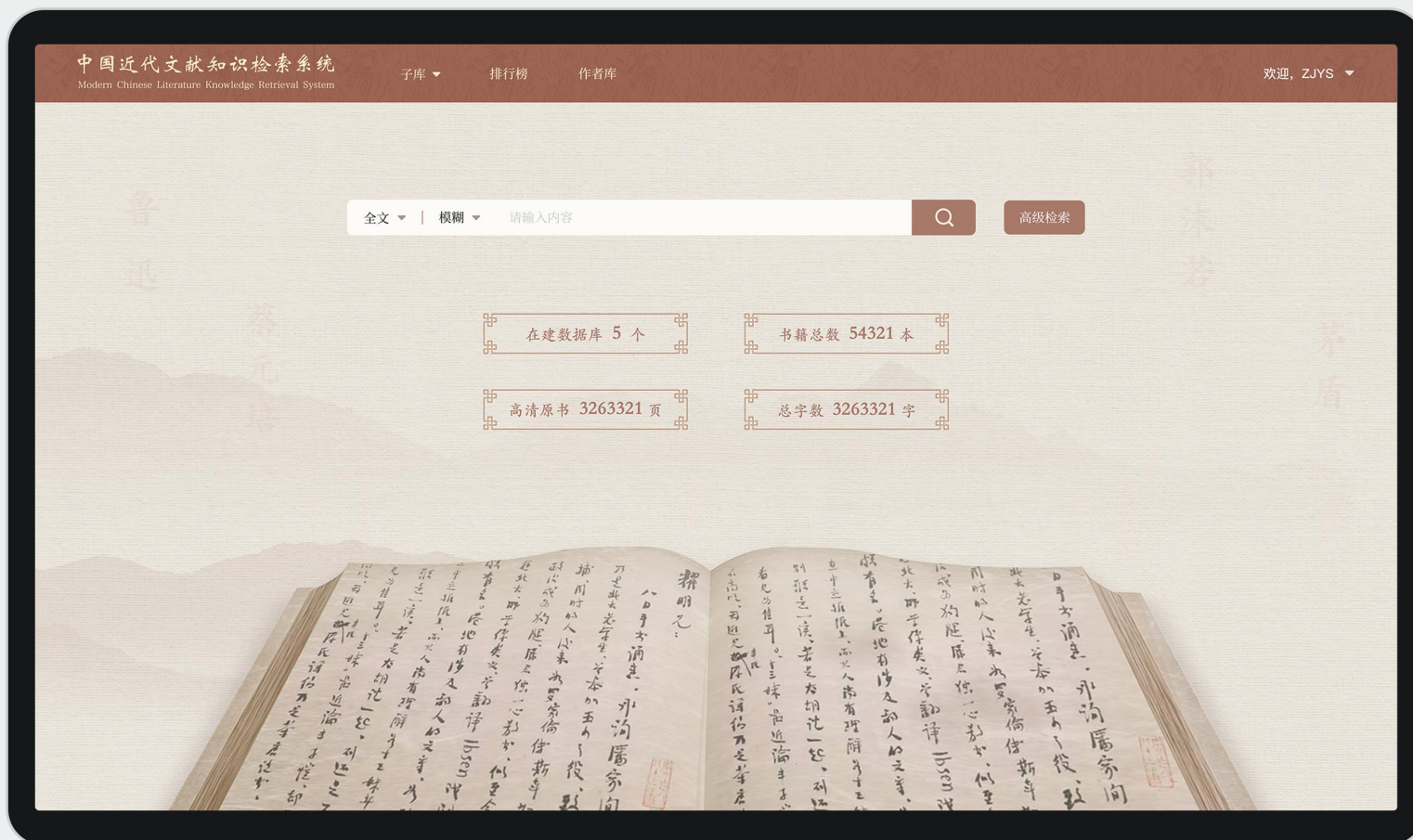
图数据库、知识图谱技术

Neo4j图数据库，自动生成文献内知识点之间的关联

知识性全文检索

字字可查，句句可检，并进行语义性关联检索。

中国近代文献数据总库--功能



- 1. 内容检索功能
- 2. 分类导航功能
- 3. 文献信息展示
- 4. 文献预览功能
- 5. 文献阅读功能
- 6. 文献图册展示
- 7. 文献相关内容展示
- 8. 作者知识库
- 9. 用户中心

功能特色

(1) 多种标签组合浏览功能

- 分类导航功能
- “作者信息” 导航功能
- “文献出版者” 导航功能
- “文献出版年代” 导航功能

(2) 文献内全文检索功能--实现“字字可检索，句句可查询”

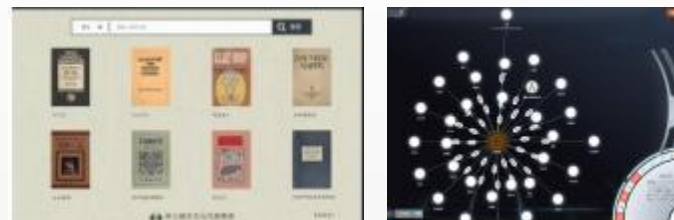
- 书目各项单一及组配检索
- 内文全文检索

(3) 检索结果重构重组及补充

(4) 文献知识服务功能（知识荟萃）

- 知识图谱建设
- 内文热词分析
- 某知识体的解释

(5) 热点发现功能



功能特色

(6) 文献浏览功能

- 图文对照

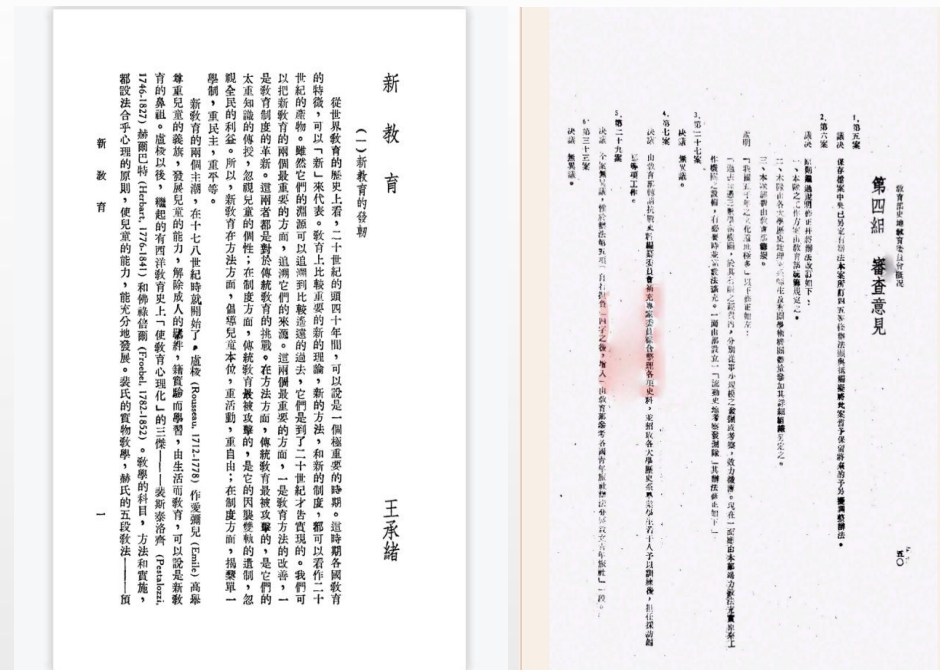
本检索系统支持图文对照（横排，竖排，简繁）阅读，让用户能够在检索结果中同时查看文本和相关图片，从而更加直观地了解文献内容。

- 高清清晰度

本系统的文献均为高清清晰度，进行了去污、排版等处理，方便读者阅读和研究文献。

- “繁”“简”转换

本系统的文献均为即将繁体字转换为简化字，方便读者阅读和研究文献。



功能特色

(7) 个性化服务功能

- **个人收藏**（“个人书架”）
- **轨迹分析**：通过阅读历史可自动定位到最后阅读的位置。

(8) 后台用户权限管理功能

(9) 用户可自建资源上传，与该库本身内容合并共享数据库功能



子库一：《中国近代文学文献数据库》

“中国近代文学文献数据库（1840-1949）”通过全球性征集1840年—1949年期间的图书，内容涵盖从鸦片战争、洋务运动到北洋时期、新文化运动时期、左翼文学时期、延安文学时期各个阶段的著作。入库图书以文学经典的初版本和稀见作品为遴选标准。目前已甄选5000种文献入库，收入了大量以往不受重视的晚清文人的作品，以及民国旧派文人的作品，包括武侠、言情、侦探和小报连载作品等，以真实展现中国近现代文坛的复杂面貌。该数据库以结构化后的全文数据资源为主，图文并茂、应收尽收，为广大科研工作者、读者和国内外馆藏机构提供一个资源共知、共享、共建的信息化平台。

学术价值

中国近代文学文献数据库分为小说、散文、戏剧、诗歌、文学研究、汉译文学六大类。甄选入库的5000种文献，其中约80%从20世纪70年代起就再没印刷出版过。本数据库以纸电同步的传播方式，提供全面而系统的第一手文献，是人文学科开展学术研究的基础性资料工程。

该数据库是由国内文学领域的顶尖专家团队，对浙江越生多年来在海内外征集的数万种近代文学文献整理而成。编委会主任陈平原、副主任陈子善，编委会委员包括孙郁、关爱和、陈思和、张伟、李今、李冬木、胡晓明、高远东、夏晓虹和解志熙等，都是我国文化名家、大家，具有极高的学术权威性。

检索展示

分类导航



新书上架



教育大家



文献详情

文献信息



《心理教育学》

责任者: (日) 九保田清 著 分类: 教育丛书 其他
出版年份: 1962.11 页数: 154
出版社: 广文书局

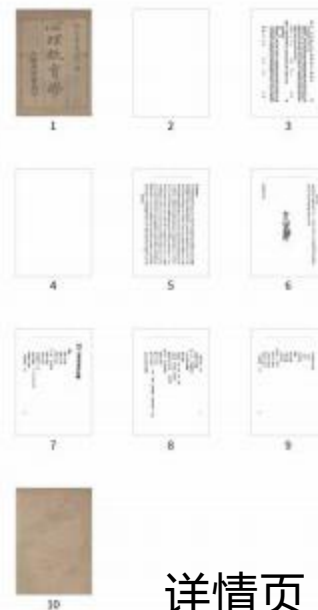
☆ 收藏

阅读

目录

- 封面
- 目录
- 总义
- 第一章 心身总论
- 第二章 心身关系论
- 第三章 心身发育论
- 第四章 注意力
- 第五章 感觉力
- 第六章 知覚力
- 第七章 记忆力
- 第八章 想像力
- 第九章 概念力
- 第十章 断定力

文献预览



详情页

目录

Q 目录预览

- 封面
- 目录
- 第一章 绪论
- 第二章 教育史
- 第三章 教育行政
- 第四章 教育理论
- 第五章 教育实践
- 第六章 教育心理学
- 第七章 教育会议
- 第八章 教育论文
- 第九章 教育丛书
- 第十章 各类教育
- 第十一章 教育实践
- 第十二章 教育心理学



数据库特色

- 全景反应中国文学史发展脉络

本数据库分期建设，全部完成后图书品种达到5万种，成为世界范围内数量最多、品种最全、包含最多珍稀版本的文学图书数据库。

- 进一步拓展和丰富研究视角

本数据库搜罗内容更全面、提供知识服务的全文检索数据库为更丰富的研究视角提供可能性。

- 持续更新

本数据库分期建设完成后，仍将持续进行更新。发现的新文献，将持续增加，逐步完善。

- 技术先进

利用最新技术，有效整合不同图书馆的近代文学文献资料信息资源，及时更新、快速处理，为读者提供方便快捷的现代阅读模式，真正实现信息资源的保护与共享。

子库二：《中国近代教育文献数据库》

《中国近代教育文献全文检索知识服务数据库》，依托于浙江越生文化传媒集团实施的“中国近代文献保护工程”，通过海内外全球征集，目前已累计采集近代教育文献资料书种约2.9万种，运用现代信息科技，建成一个基于网络的知识服务系统，为广大读者提供了便利的全文资源检索和知识服务功能，既可再现这些教育类文献的原貌，又符合图书馆馆藏结构调整的大趋势。

“中国近代文献”是建设教育强国学理知识的重要组成部分，是哲学社会科学学科体系建设的重要内容。而《中国近代教育文献数据库》的建设将对此类文献进行系统性、创新性、权威性的保护，具有十分重要的时代意义。2022年7月14日，该库成功入选中央文改办“国家文化产业发展项目库”第二批。

子库二：《中国近代教育文献数据库》

数据库定位

中国近代教育文献全文索引知识服务数据库是通过全球性征集1840年~1949年期间涵盖从鸦片战争、洋务运动到民国时期、新中国成立之前的有关全面反映教育领域的各类文献资料，运用知识图谱技术，实现教育文献内容可视化，建成一个以结构化后的全文数据资源为主，图文并茂，形成一定规模，提供社会服务的、能持续发展的、质量较高的教育类特色文献全文数据库，为广大教育科研工作者、读者和国内外馆藏机构提供一个资源共知、共享、共建的信息化平台。

数据库规模

浙江越生从2018年开始整理《中国近代教育文献数据库》所存的中国近代教育文献，通过整合权威的综合力量进行甄选、整理，并完成了基础性的收集和扫描整理工作。拟建数据库采取分批遴选、分步实施、分期建设，根据对近代教育文献数据库调研，首期规模精选5000种。

专家团队

田正平：浙江大学文科资深教授、浙江大学 中外教育现代化研究所所长、中国教育学会教育史分会名誉理事长、博士Th导师。长期从事中国教育史和中外教育文化交流史的教学与研究工作。代表作有《留学日本热潮与清末教育》《留学Th与中国教育近代化》《世纪之理想》等。

刘海峰：厦门大学中国教育考试中心主任、国家教育咨询委员会委员、国家考试委员会委员、中国教育学会教育史分会副理事长、厦门大学博士Th导师。主要研究方向为高考改革与科举学、高等教育历史与理论。代表作有《科举学 导论》《高考改革论》《高等教育历史与理论研究》等。

杜成宪：中国教育学会教育史分会副理事长、华东师范大学教育系学术委员会主任、博士Th导师。研究领域为中国教育史、中国考试史、中国传统文化等。代表作有《教育史学》《中国教育史九十年》等。

肖 朗：中国教育学会教育史分会副理事长、浙江大学教育学院学术

石 鸥：全国教育文物研究会理事长、首都师范大学中国基础教育教材研究院院长、博士Th导师。主要

中共中央宣传部

文改办发函〔2022〕0063号

关于国家文化产业发展项目库 第二批入库项目的通知

浙江省委宣传部：

你单位推荐的“中国近代教育文献全文检索知识服务数据库”等6个项目确定为国家文化产业发展项目库第二批入库项目，详情见附件。

请认真组织好项目实施，并做好项目结项情况报告。

附件：入库项目名单

中央文化体制改革和发展工作领导小组办公室

2022年7月5日

附件

入库项目名单

序号	项目名称	推荐单位	申报单位	备注
1	中国近代教育文献全文检索知识服务数据库	浙江省委宣传部	浙江越生文化传媒有限公司	入库项目
2	剧场演艺装备运维管理服务管理平台	浙江省委宣传部	浙江大丰实业股份有限公司	入库项目
3	电子竞技网络游戏——梦三国2亚运版	浙江省委宣传部	杭州电魂网络科技有限公司	入库项目
4	盐官音乐小镇	浙江省委宣传部	海宁盐官古城旅游股份有限公司	入库项目
5	上海滩影视旅游拍摄基地建设	浙江省委宣传部	浙江横店影视城有限公司	入库项目
6	基于全媒体智能中台的天目云3.0平台	浙江省委宣传部	浙报融媒体科技(浙江)股份有限公司	入库项目

子库三：《中国近代艺术文献数据库》

数据库定位

“中国近代艺术文献数据库(1840-1949)”是依托于浙江越生文化传媒集团所实施的“中国近代文献保护工程”而建立的。通过全球范围内的广泛征集，目前已成功采集近代艺术文献达2800种。本数据库全面且系统地整合并精选了中国1840年至1949年间的艺术文献资源，内容丰富，涵盖了艺术论著、艺术家自述、作品评论、展览资料以及艺术教育相关文档等多个方面，横跨绘画、雕塑、建筑、书法、摄影及民间艺术等多个艺术领域。

数据库所收录的文献，不仅包含了经典著作的珍贵初版，还囊括了一些市面上罕见且具有重要历史意义的文本。著作者群体汇聚了中国近现代美术史上的众多热点人物，名家辈出，如徐悲鸿、林风眠、刘海粟、潘玉良、王远勃、吴恒勤、钱鼎、丁光燮、朱配瞻、李可染、江小鹤、刘开渠等。通过呈现和解读这些珍稀文献，本数据库为相关美术史研究提供了宝贵的补正资料，并有望改变长期以来因“重作品、轻文献”而导致的艺术资源保护中的偏颇和缺失现象。

专家团队

本数据库第一辑的入库内容，是基于邵大箴、卢辅圣担任顾问，李超担任主编的《中国近代艺术文献丛刊·美术卷》而编撰的。为了确保每一份文献的真实性和权威性，我们将继续邀请更多艺术界的泰斗级人物加入我们的团队，共同守护这份珍贵的艺术遗产。

服务群体

高等院校、各类图书馆

搭建学科建设 孵化培养
人才 大数据创新研发

出版机构

洞悉业界动态 研究领域
趋势 宣传学术发展



学术机构、团体

共享权威资源 交流传承
学术 占领学术制高点

国家、地方智库

纵览前瞻课题 充实智库
精华 提供决策化依据



服务方式

服务方式灵活

方式一：云服务，按年付费（网络版）

方式二：连续续费五年，可转为买断（镜像版）

产品地址：<https://www.jdwxk.com/>

万页文本分析系统

----解开文字里的学术密码

数字人文全新视角做研究、看历史、观世界

1、研发背景

自20世纪90年代以来，随着数字技术和信息技术的不断发展，尤其是大数据、人工智能等先进技术正逐渐深入人文学科领域，改变了人文学科的传统研究范式。人文学科的研究方法，正在由传统意义上的纸质媒体为主要对象向电子文本转变，由文本细读向“远距离可视化阅读”为主要方法转变。

在建设新文科背景下，以学科交叉为特征的“数字人文”理念已被广大专家学者所认同，大家普遍认为数字人文将成为人文学科研究的下一个突破点！利用AI技术将文本数据化、可视化，实现定性研究和定量研究的完美结合，为提高学习效率，为教学和科研提供全新的角度。

教育部在建设新文科发展意见中明确提出“融合性”，要与多个学科的交叉、融合、渗透或拓展的特点；《社会科学司2020年工作要点》中明确提出“重点支持建设一批文科实验室，促进研究方法创新和学科交叉融合，引领学术发展”。

2、基本介绍

万页文本分析系统，是在大数据、人工智能大时代背景下，集结了国内高水平的算法工程师、专业作家与编辑、行业资深经理人，以机器学习为基础，开发人工智能模型处理文本，为海量文本的学习、研究提供专业化、技术化服务。

系统运用**NLP**自然语言处理、句法分析、语义依赖分析、无监督学习算等技术，开发了国内首个基于数字人文理念，针对文科专业的人工智能文本分析系统。该系统包含三大文本处理模块『特征分析』、『比较研究』和『摘要提取』。

本系统旨在从源头解决文科院系师生教学、科研、学习、阅读等质量与效率问题，为人文学科研究、论文写作提供更便捷的工具，更多元的研究角度。

3、系统介绍

国内首个
基于数字人文理念针对人文学科研究而开发的文本分析系统

分秒之间分析出高频词、关键词、TOP人物、TOP地点、情节发展曲线、人物地点之间的关系

浓缩数百万字长篇秒读，分秒内掌握情节走向，浓缩文本精华，1年读完1800本书



4、功能展示



5、适用对象

一切文本皆可分析！

➤ 人文社科

分析词频、关键词、TOP人物、TOP地点、情节发展曲线、人物地点之间的关系、提取大事年表、地理元素、人物事件关系等等。

➤ 英语专业

系统支持英文文本的分析，可以将英文文本象分析中文一样进行深度分析等等。

➤ 其他专业

提取图书、论文摘要精华，重要概念、定义、分析词频、情感起伏曲线分析等等。

➤

一切以文本为研究对象的专业均可适用！

6、技术相关

- NLP自然语言处理、句法分析、语义依赖分析、无监督学习算法……
- 通过对模型输入层和输出层的优化，秒速、按需提取全书摘要精华。
- 支持中文、英语文本的分析。
- 支持单个文本和批量文本分析。
- 支持纯文本、PDF（可编辑）文本格式。
- 支持对分析参数的设定和人工干预。
- 支持百万级文字分析。
- 支持文言文、繁体字文本分析。
- 分析结果可与原文对应查看。

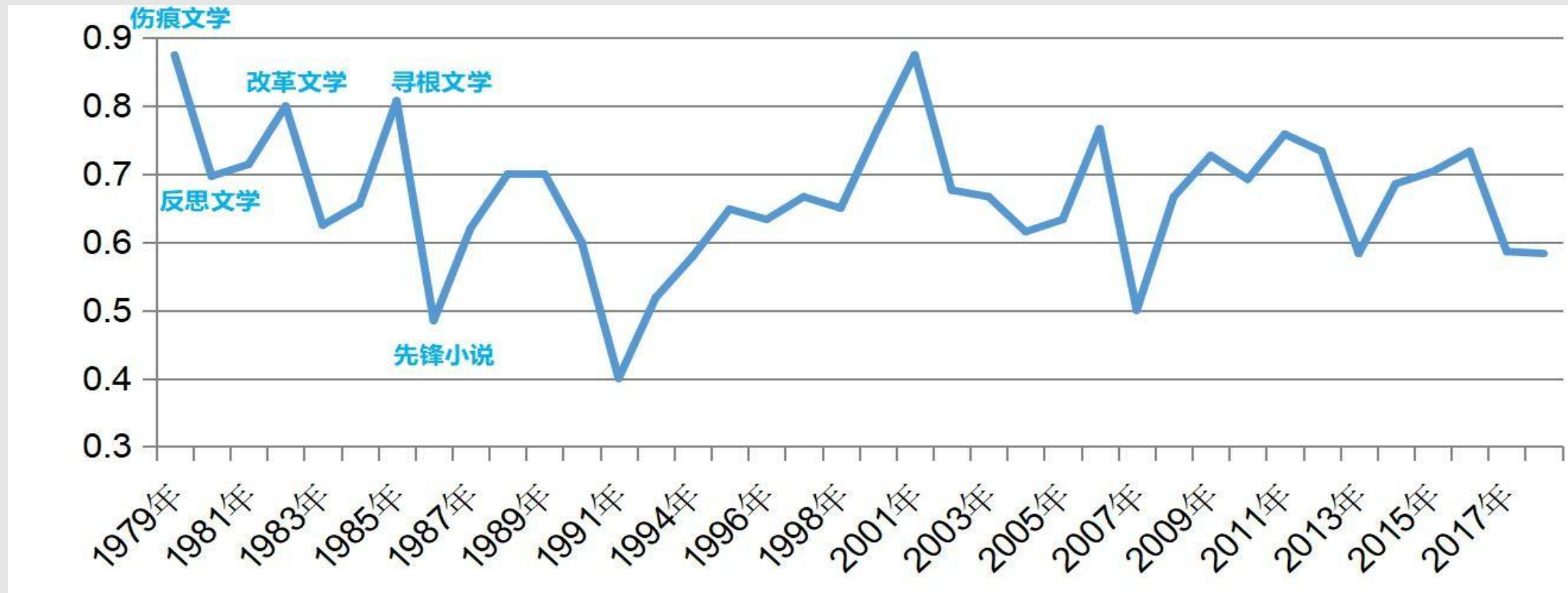
7、应用示例

示例一 文学领域分析研究

以分析《收获》杂志40年趋势为例
《收获》1979年-2018年 680位作者长中短篇小说1618篇

（《〈收获〉刊载小说的数据性考察（1979--2018）》
---- 《中国现代文学研究丛刊》）

8、应用示例



《收获》年度 明快-沉郁 作品比例变化图

7、应用示例



《收获》（1979-1989）

小说高频词统计结果



《收获》（1990-1999）

小说高频词统计结果



《收获》（2000-2009）

小说高频词统计结果



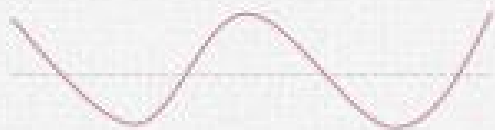
《收获》（2010-2018）

小说高频词统计结果

9、系统介绍

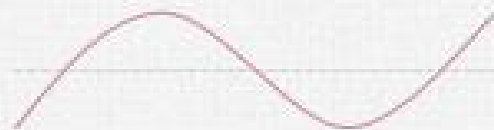
所有故事都可以用一条线来展现

W 型



《琅琊榜》

N 型



《那些年，我们一起追过的女孩》

V 型



《白夜追凶》

M 型



《霸王别姬》

倒 N 型



《长安十二时辰》

倒 V 型



《鬼吹灯之精绝古城》

10、应用示例

分秒间获取更多的知识

瞬间轻松提取论文摘要

对篇幅较长的著作进行精华提取，迅速阅读、掌握

提取著作中要点精华，形成导读版图书

在提取全书精华的基础上形成课程、思维导图

无限应用，尽在研究者任意发挥

11、成果展示

利用本系统分析成结果，部分已发表论文成果

- 《〈收获〉刊载小说的数据性考察（1979--2018）》 ---- 《中国现代文学研究丛刊》
- 《网络小说的数据法与类型论--以2018年的749部中国网络小说为考察对象》 ---- 《扬子江评》
- 《数据分析视角下的茅盾文学奖研究》 ---- 《中国比较文学》
- 《探索未来，一半靠人工，一半靠AI》 ---- 《南方文坛》
- 李强 | 《作为数字人文思维的“网文算法”》 ——以“明穿”小说为例 -- 《中国现代文学研究丛刊》
- 吉云飞 | 《作为“计算批评”的“远读”》 ——以网络小说“升级文”中的节奏与情绪为例 -- 《中国现代文学研究丛刊》

13、与其他AI工具的区别

目前出现了大量以chat gpt为代表的人工智能工具，如“KIMI”、“文心一言”等，这些工具在文本语言的支持上、底层模型、算法和功能设计上有着各自的侧重点。

- 1、应用领域不同，其它工具涵盖广泛，万页只针对科研，垂直性更强。
- 2、功能上，其它工具多是问答，根据要生成文本的形式体现，没有更细致的分析功能。
- 3、核心内容分析提取方式不同，语义分析可读性强，但有可能不够准确。
- 4、多语言支持上，多数工具仅仅是翻译。
- 5、大多数工具都是以自动生成文档为核心卖点，但是这一行为已被认定为学术不端行为，很多学术期刊包括各领域的顶级期刊也明确拒绝AI生成的论文。

12、数字人文实验室

我们的目标

建成一个集资源建设，研究工具，人才培养，学术研究和交流为一体平台



更多精彩即将推出！

欢迎试用

目前系统主要的服务方式是针对机构用户开放服务，只需提供校园网IP段，开通后校园网内师生均可免登录使用。

THANKS



杨 文

13501287661 (微信)

产品地址: <https://www.jdwxk.com>