



前沿科技文章评分推荐系统

北京邮电大学

大数据与智能信息处理实验室





大纲

contents

- 1 项目概论及考核指标
- 2 开发过程概述
- 3 成果报告
- 4 项目考核完成情况



一、项目概况及考核指标





项目概况

大数据中蕴含了丰富的价值与巨大的潜力，能够高效的从数据中获取到需要的信息变得越来越重要。在学术领域同样有相似的需求，快速的**根据已有的文献推荐出相似且评分高的文献**刻不容缓。

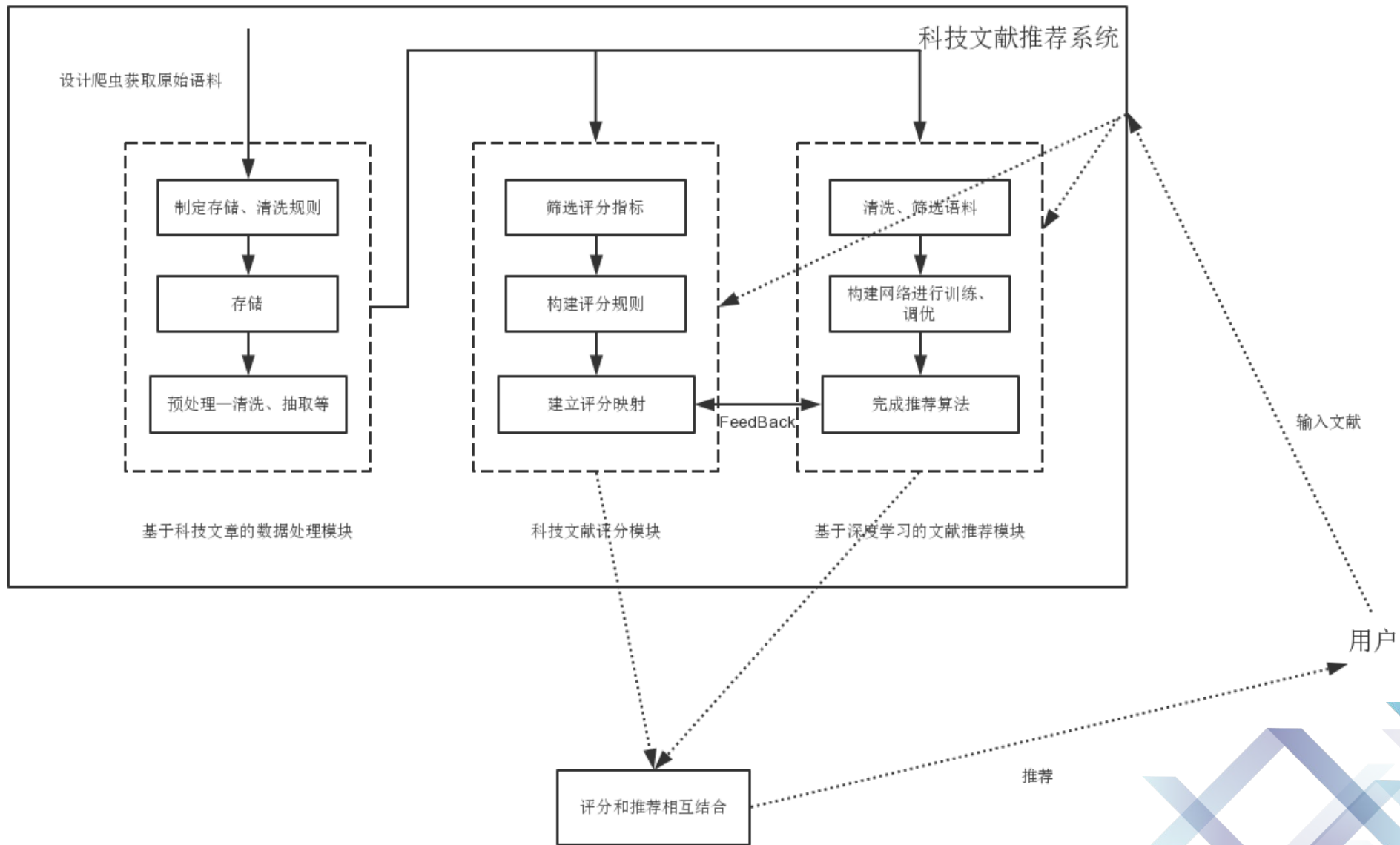
本项目对知网中互联网技术分类的科技文献进行爬取，总计60000篇，每篇文献爬取相关的**标题、摘要、关键词、发表时间、下载量和引用量**等因素，除此之外还提取了发表文章平台的**复合影响和综合影响因子**。

然后基于深度学习的Doc2Vec技术和TF-IDF提取方法对每一篇科技文章进行向量化并使用PCA对向量进行降维操作，最后结合评分系统的结果对科技文章进行评分，通过评分推荐的组合给出高相似度和高质量的科技文章。





项目简介





考核指标

- **科技文章的数据处理模块**
 - 对科技文章相关指标的快速爬取
 - 快速数据清洗与预处理
 - 文本的规约整理与高效存储
- **基于深度学习的科技文献评分模块**
 - 确定科技文章的评分指标
 - 设置科技文章的初始评分指标
 - 建立文献摘要和文献评分的权值矩阵
- **前沿科技文献的评分推荐模块**
 - 新科技文献的抓取与数据清洗
 - 调优网络，评分与推荐的相互反馈





二、开发过程概述





考核指标

- **科技文章的数据处理模块**
 - 对科技文章相关指标的快速爬取
 - 快速数据清洗与预处理
 - 文本的规约整理与高效存储
- **基于深度学习的科技文献评分模块**
 - 确定科技文章的评分指标
 - 设置科技文章的初始评分指标
 - 建立文献摘要和文献评分的权值矩阵
- **前沿科技文献的评分推荐模块**
 - 新科技文献的抓取与数据清洗
 - 调优网络，评分与推荐的相互反馈





开发过程概述

- 信息科技
 - 无线电电子学
 - 电信技术
 - 计算机硬件技术
 - 计算机软件及计算机应用
 - 互联网技术
 - 自动化技术
 - 新闻与传媒
 - 出版
 - 图书情报与数字图书馆
 - 档案及博物馆

```
管理员: C:\Windows\system32\cmd.exe - python myMain35.py 35
程序挂掉了重新运行
117.242.147.181:34281
初始页码 1
程序挂掉了重新运行
123.139.56.238:9999
初始页码 1
正在获取第1条信息, FileName=DZXU200108024
正在获取第2条信息, FileName=DZYX201003041
正在获取第3条信息, FileName=KZYC200505001
正在获取第4条信息, FileName=ZGDC200009005
正在获取第5条信息, FileName=SJCJ199703002
正在获取第6条信息, FileName=RJDK200911011
正在获取第7条信息, FileName=ZGTB200106003
正在获取第8条信息, FileName=HXJZ603.006
正在获取第9条信息, FileName=HWYJ200203019
正在获取第10条信息, FileName=HWYJ200404003
正在获取第11条信息, FileName=JSJX201408004
正在获取第12条信息, FileName=DZYX200701055
正在获取第13条信息, FileName=FIUE200402003
正在获取第14条信息, FileName=XAJT904.001
正在获取第15条信息, FileName=ZGTB200404004
正在获取第16条信息, FileName=DLXT710.012
正在获取第17条信息, FileName=DLZD200404016
正在获取第18条信息, FileName=ZGDC200102003
半:
```

爬取数据过程展示

对知网中信息科技类中数据进行爬取总计60000篇数据，使用分布式爬虫技术，为绕过反爬机制，使用**动态代理功能**，实际测试爬取速度可达每秒十几篇科技文献。





考核指标

- **科技文章的数据处理模块**
 - 对科技文章相关指标的快速爬取
 - **快速数据清洗与预处理**
 - 文本的规约整理与高效存储
- **基于深度学习的科技文献评分模块**
 - 确定科技文章的评分指标
 - 设置科技文章的初始评分指标
 - 建立文献摘要和文献评分的权值矩阵
- **前沿科技文献的评分推荐模块**
 - 新科技文献的抓取与数据清洗
 - 调优网络，评分与推荐的相互反馈





开发过程概述

区块链是随着比特币等数字加密货币的日益普及而逐渐兴起的一种全新的去中心化基础架构与分布式计算范式,目前已经引起政府部门、金融机构、科技企业和资本市场的高度重视与广泛关注.区块链技术具有去中心化、时序数据、集体维护、可编程和安全可信等特点,特别适合构建可编程的货币系统、金融系统乃至宏观社会系统.本文通过解构区块链的核心要素,提出了区块链系统的基础架构模型,详细阐述了区块链及与之相关的比特币的基本原理、技术、方法与应用现状,讨论了智能合约的理念、应用和意义,介绍了基于区块链的平行社会发展趋势,致力于为未来相关研究提供有益的指导与借鉴.

区块链是比特币数字加密货币普及兴起一种全新中心化基础架构分布式计算范式 政府部门 金融机构 科技企业 资本市场 高度重视 关注 区块链 技术 中心化 时序 数据 集体 维护 可编程 可信 特别适合 构建 可编程 货币 系统 金融 系统 宏观 社会 系统 本文 解构 区块链 核心 要素 提出 区块链 系统 基础架构 模型 详细 阐述 区块链 及 相关 比特币 基本原理 技术 方法 现状 讨论 智能 合约 理念 意义 介绍 区块链 平行 社会 发展趋势 致力于 未来 相关 研究 提供 有益 指导 借鉴





考核指标

- **科技文章的数据处理模块**
 - 对科技文章相关指标的快速爬取
 - 快速数据清洗与预处理
 - 文本的规约整理与高效存储
- **基于深度学习的科技文献评分模块**
 - 确定科技文章的评分指标
 - 设置科技文章的初始评分指标
 - 建立文献摘要和文献评分的权值矩阵
- **前沿科技文献的评分推荐模块**
 - 新科技文献的抓取与数据清洗
 - 调优网络，评分与推荐的相互反馈





开发过程概述

“题名”:“Web服务核心支撑技术:研究综述”,
 “作者”:“岳昆, 王晓玲, 周傲英”,
 “来源”:“软件学报”,
 “发表时间”:“2004-03-30”,
 “被引量”:“1827”,
 “下载量”:“7534”,
 “复合影响因子”:“3.828”,
 “综合影响因子”:“2.277”,
 “关键词”:“[

“Web服务”,
 “服务组合”,
 “语义Web”,
 “服务发现”,
 “安全性”,
 “P2P”,
 “网格”

],

“摘要”:“随着电子商务的迅速崛起,基于Web的应用模式迅速发展,Web应用从局部化发展到全球化,从B2C(business-to-customer)发展到B2B(business-to-business),从集中式发展到分布式,Web服务成为电子商务的有效解决方案.Web服务是一个崭新的分布式计算模型,是Web上数据和信息集成的有效机制.Web服务的新型构架,Web服务的高效执行方式,Web服务与其他成熟技术的有机结合以及Web服务的集成是解决现实应用问题的重要技术.从Web服务研究的不同侧面对其进行了综述,阐述了Web服务的基本概念,分析了当前Web服务的主要研究问题及其核心支撑技术,概括了Web服务中的数据集成技术、Web服务的组合、语义Web服务、Web服务发现、Web服务安全、P2P(Peer-to-Peer)新型计算环境下的Web服务解决方案和网格服务等方面的研究内容,并对这些技术进行了总结,结合已有的研究成果,展望了Web服务未来的研究方向及其面临的挑战.”

“相似文献”:“[

[1]基于语义的Webservice组合技术研究[J].李明翠.电脑知识与技术.2009(14),

<https://kns.cnki.net/kcms/detail/detail.aspx?filename=DHZS200914029&dbcode=CJFD&dbname=CJFD2009&v=>,

[2]基于Petri网的Web服务的创建与描述[J].张正明,马炳先,相东明.系统仿真学报.2011(S1),

<https://kns.cnki.net/kcms/detail/detail.aspx?filename=XTFZ2011S1004&dbcode=CJFD&dbname=CJFD2011&v=>,

[3]WebServices组合的容错方法[J].唐渊.湖南工业大学学报.2010(06),

<https://kns.cnki.net/kcms/detail/detail.aspx?filename=ZZGX201006016&dbcode=CJFD&dbname=CJFD2010&v=>,

[4]基于Petri网的语义Web服务组合方法[J].吴敏敏.泰山学院学报.2016(03).”

题名	读者推荐	相似文献	被引量	下载量	复合影响因子	综合影响因子	关键词	摘要	作者	来源	发表时间
领域自适应的Web服务评价模型	基于遗传算法的QoS感知的Web服务选择,Web服务组合的基于文法的消息处理,Web服务描述...	基于模糊评判的Web服务评价模型,一个基于浏览器与组件技术的Web信息发布模型,Web页面获...	111.0	1137	5.567	3.272	Web服务,评价模型,机器学习,先验知识	Web服务质量的评价是指Web服务的选取与组合的主要手段[1],而目前的Web服务评价模...	杨文军,李涓子,王克宏	计算机学报	2005-04-12 00:00:00

数据存储

数据爬取时根据互联网技术下十个小类别使用JSON格式进行存储。所有数据爬取完毕后，使用MySQL数据库对数据进行存储。





考核指标

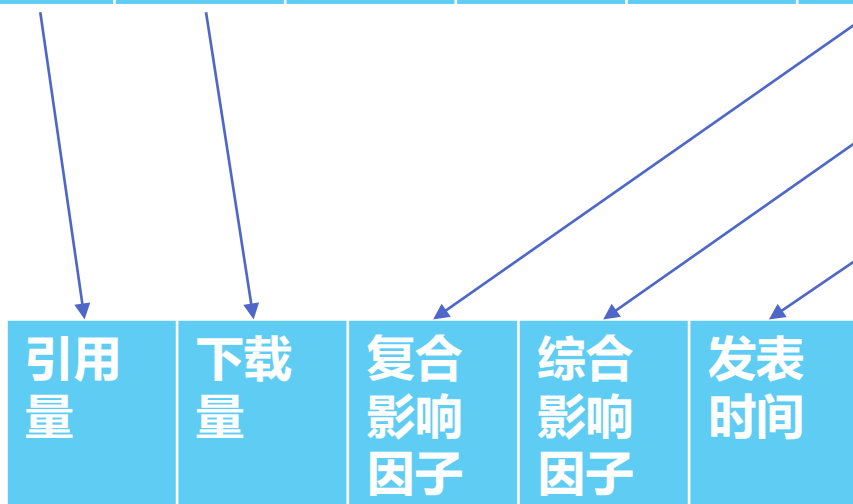
- **科技文章的数据处理模块**
 - 对科技文章相关指标的快速爬取
 - 快速数据清洗与预处理
 - 文本的规约整理与高效存储
- **基于深度学习的科技文献评分模块**
 - **确定科技文章的评分指标**
 - 设置科技文章的初始评分指标
 - 建立文献摘要和文献评分的权值矩阵
- **前沿科技文献的评分推荐模块**
 - 新科技文献的抓取与数据清洗
 - 调优网络，评分与推荐的相互反馈





开发过程概述

题名	读者推荐	引用量	下载量	关键词	摘要	作者	复合影响因子	综合影响因子	发表时间
----	------	-----	-----	-----	----	----	--------	--------	------



科技文献评分指标

根据相关论文和经验，对爬取数据的字段进行筛选，最后选择引用量、下载量、复合影响因子、综合影响因子和发表时间作为评分指标。





考核指标

- **科技文章的数据处理模块**
 - 对科技文章相关指标的快速爬取
 - 快速数据清洗与预处理
 - 文本的规约整理与高效存储

- **基于深度学习的科技文献评分模块**
 - 确定科技文章的评分指标
 - **设置科技文章的初始评分指标**
 - 建立文献摘要和文献评分的权值矩阵

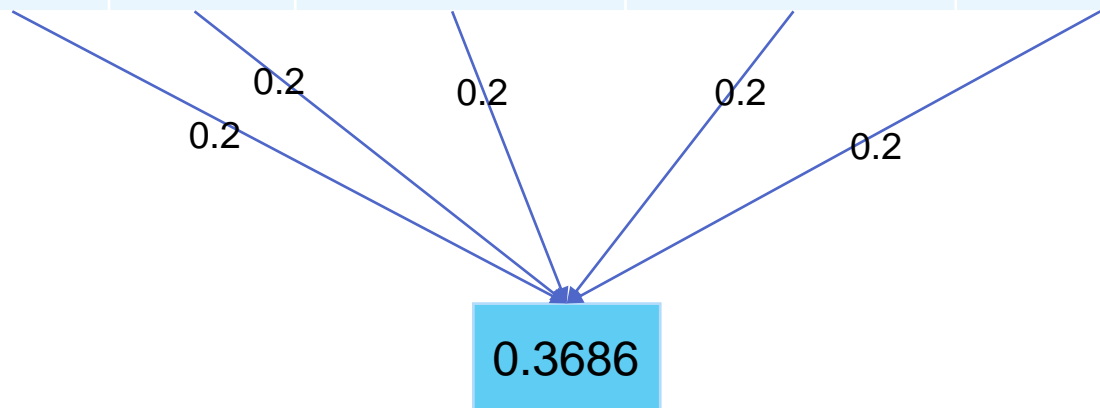
- **前沿科技文献的评分推荐模块**
 - 新科技文献的抓取与数据清洗
 - 调优网络，评分与推荐的相互反馈





开发过程概述

引用量	下载量	复合影响因子	综合影响因子	发表时间
1137	111	5.567	3.272	2005-04-12
0.0121	0.0182	0.5964	0.3878	0.8288



科技文献评分指标

为方便数据处理对五个评分指标进行最大最小归一化处理并设置每个指标的初始权重为0.2。





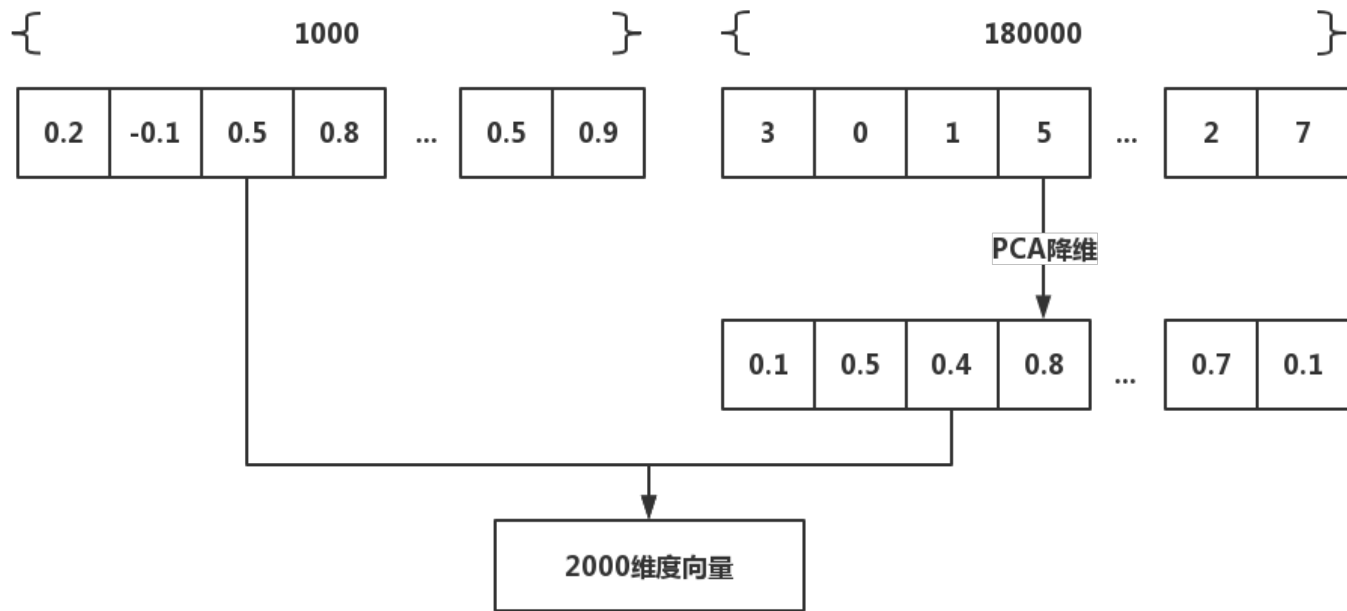
考核指标

- **科技文章的数据处理模块**
 - 对科技文章相关指标的快速爬取
 - 快速数据清洗与预处理
 - 文本的规约整理与高效存储
- **基于深度学习的科技文献评分模块**
 - 确定科技文章的评分指标
 - 设置科技文章的初始评分指标
 - **建立文献摘要和文献评分的权值矩阵**
- **前沿科技文献的评分推荐模块**
 - 新科技文献的抓取与数据清洗
 - 调优网络，评分与推荐的相互反馈





开发过程概述



文本向量	评分指标
2000维度向量	0.35
2000维度向量	0.68
...	...
2000维度向量	0.96

科技文献权值矩阵

对清洗和预处理后的摘要、标题和关键词信息使用Doc2Vec和ITF-IDF进行向量化，使用PCA降低维度，最后将向量化后的结果和评分值组成权值矩阵保存。





考核指标

- **科技文章的数据处理模块**
 - 对科技文章相关指标的快速爬取
 - 快速数据清洗与预处理
 - 文本的规约整理与高效存储
- **基于深度学习的科技文献评分模块**
 - 确定科技文章的评分指标
 - 设置科技文章的初始评分指标
 - 建立文献摘要和文献评分的权值矩阵
- **前沿科技文献的评分推荐模块**
 - 新科技文献的抓取与数据清洗
 - 调优网络，评分与推荐的相互反馈





开发过程概述

$$P(L_u) = \frac{L_u \cap B_u}{L_u}$$

$$P(B_u) = \frac{L_u \cap B_u}{B_u}$$

$$F1 = \frac{2 * P * R}{P + R}$$

F1值由准确率P和召回率R组成，公式中 L_u 为读者推荐和相似文献的组合， B_u 为推荐的文献组合。准确率指推荐列表在读者推荐和相似文献中出现的情况，召回率指读者推荐和相似文献在推荐列表中出现的状况。

爬取新的科技文献并进行相应的清洗操作，除常规爬取数据外，添加知网的读者推荐和相似文献。使用F1值对推荐的结果进行评价，并根据F1值大小使用网格搜索的方式对评分系统指标的权重进行较优设置。





开发过程概述

引用量	下载量	复合影响因子	综合影响因子	发表时间
0.1	0.1	0.1	0.1	0.6
0.1	0.1	0.1	0.2	0.5
0.1	0.1	0.1	0.3	0.3
...
0.1	0.6	0.1	0.1	0.1
...
0.6	0.1	0.1	0.1	0.1

爬取新的科技文献并进行相应的清洗操作，除常规爬取数据外，添加知网的读者推荐和相似文献。使用F1值对推荐的结果进行评价，并根据F1值大小使用网格搜索的方式对评分系统指标的权重进行较优设置。



三、成果汇报



成果报告



北京市科学技术情报研究所

Beijing Institute of Science and Technology Information

The screenshot shows a web application interface for the Beijing Institute of Science and Technology Information (ISTI). The interface includes a dark blue sidebar on the left with navigation options: '前端展示界面', '数据处理', '导入数据', '词分词处理', '推荐列表', and '评分推荐列表'. The main content area is titled '词分词处理' and displays the results of a word segmentation process. The text shown is: '最新提交数据分词结果：区块链是比特币数字加密货币普及兴起一种全新中心化基础架构分布式计算范式政府部门金融机构科技企业资本市场高度重视关注区块链技术中心化时序数据集体维护可编程可信特别适合构建可编程货币系统金融系统宏观社会系统本文解构区块链核心要素提出区块链系统基础架构模型详细阐述区块链及相关比特币基本原理技术方法现状讨论智能合约理念意义介绍区块链平行社会发展趋势致力于未来相关研究提供有益指导借鉴'. The footer of the page contains the copyright information: '© 2019 北京邮电大学'.

数据预处理-清洗、分词

将输入系统的标题、摘要和关键词进行清洗和分词操作。





成果报告



北京市科学技术情报研究所

Beijing Institute of Science and Technology Information

前端展示界面

数据处理

推荐列表

推荐文章展示

评分推荐列表

≡

qingbaosuo

导入数据 × 切词分词处理 × 推荐文章展示 × 推荐结果和评分 ×

最新提交文献的推荐文献：

标题名称	作者	相似度
区块链技术综述	沈鑫; 裴庆棋; 刘雪峰	0.9287659305534308
区块链性能的量化分析研究	王旭; 甘国华; 吴凌云	0.9015141057905964
区块链技术安全威胁分析	孙国梓; 王纪涛; 谷宇	0.900299707505039
区块链技术及其研究进展	朱岩; 王巧石; 秦博涵; 王中豪	0.8897162916965637
区块链技术与应用前瞻综述	何蒲; 于戈; 张岩峰; 鲍玉斌	0.880213001761252
基于纠错码的区块链系统区块文件存储模型的研究与应用	赵国锋; 张明聪; 周继华; 赵涛	0.8769881742197944
区块链重塑网络舆论环境和治理	钟欢	0.8677130535858244
基于区块链的食品溯源系统的存储设计	沈政启	0.866086884499901
区块链技术在数字著作权保护中的运用与法律规制	王清; 陈潇婷	0.86111454893538
基于区块链债转平台的供应链融资决策分析	唐丹; 庄新田	0.8560489695403353

© 2019 北京邮电大学

文章推荐

根据训练好的模型对文章相关信息切词后的数据进行向量化，并将相似度高的五个文献展示。





成果报告



北京市科学技术情报研究所

Beijing Institute of Science and Technology Information

- 前端展示界面
- 数据处理
- 推荐列表
- 评分推荐列表
- 推荐结果和评分**

最新提交文献的推荐和评分列表：

标题名称	作者	下载量	引用量	发表时间	复合影响因子	综合影响因子	相似度	评分值
区块链技术综述	沈鑫; 裴庆祺; 刘雪峰	14681	199.0	2016-11-15 00:00:00	1.362	0.752	0.9287659305534308	0.2176539508631895
区块链技术与应用前瞻综述	何涌; 于戈; 张岩峰; 鲍玉斌	7460	166.0	2017-04-15 00:00:00	1.554	0.919	0.880213001761252	0.17518895009969454
基于纠删码的区块链系统区块文件存储模型的研究与应用	赵国锋; 张明聪; 周继华; 赵涛	106	0.0	2019-02-10 00:00:00	1.925	1.189	0.8769881742197944	0.13450328035324893
区块链性能的量化分析研究	王旭; 甘国华; 吴凌云	404	0.0	2019-06-06 13:42:00	1.4480000000000002	0.8740000000000001	0.9015141057905964	0.12795673706603153
基于区块链债转平台的供应链融资决策分析	唐丹; 庄新田	809	0.0	2019-10-12 11:17:00	1.269	0.657	0.8560489695403353	0.12649024114133997

© 2019 北京邮电大学

文章评分推荐

通过评分系统对上述推荐的结果进行再次排序给出评分高、高相似度的文章。





四、项目考核完成情况

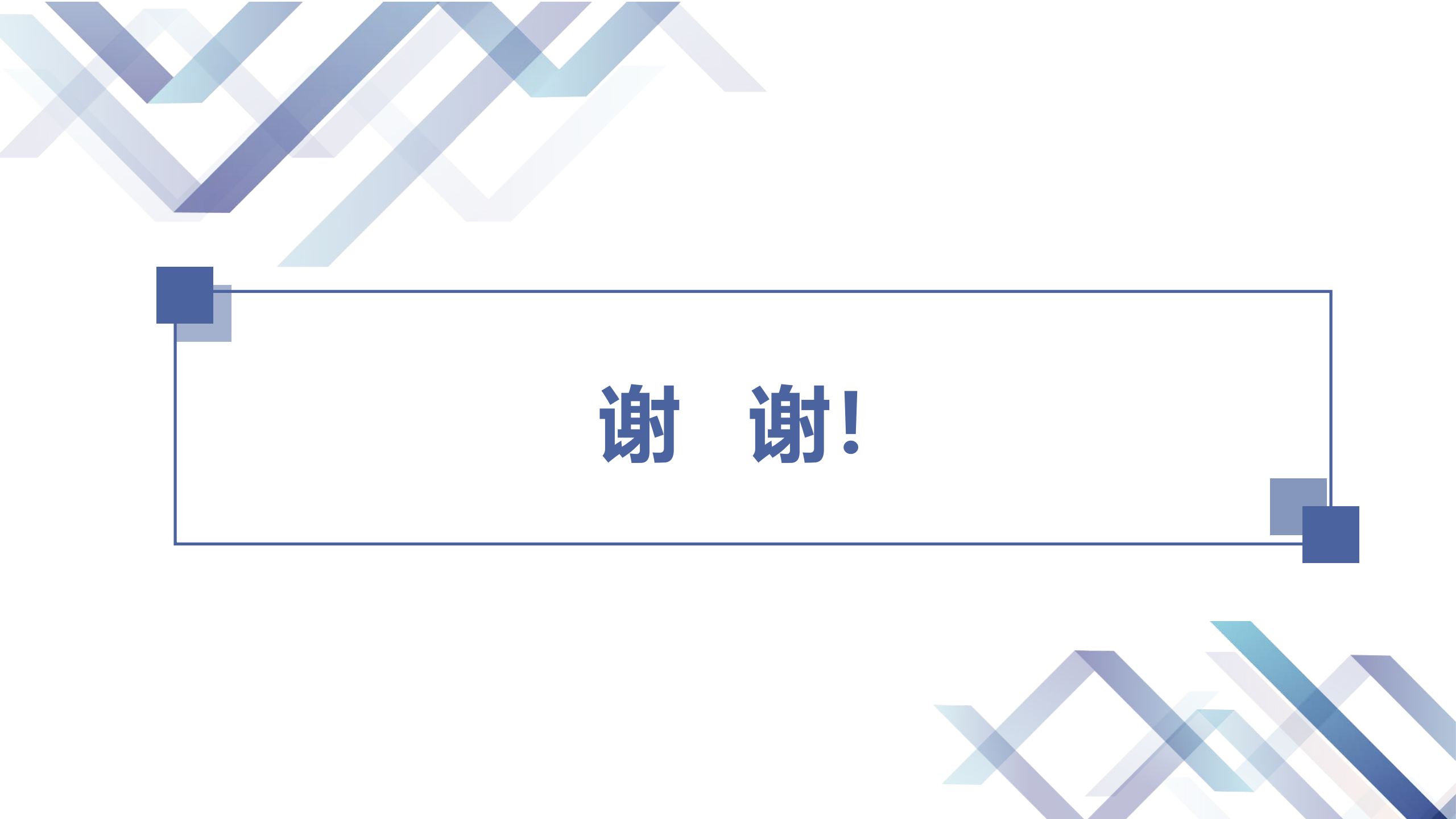




开发过程概述

基于科技文章的数据处理模块	完成
基于深度学习的科技文献评分模块	完成
前沿科技文献的评分推荐模块	完成
用户交互和展示模块	完成





谢 谢!