

互联网文本观点抽取

北京邮电大学

2017年

汇报目录



 一、项目概况及考核指标.....

 二、开发过程概述.....

 三、成果报告.....

 四、项目考核指标完成情况.....

项目简介

- “互联网文本观点抽取”项目，设计将网络信息采集技术和信息处理技术，与科技情报生产线对接的机制及方案，实现一个情报采集加工模块。

考核指标

- 原始语料库处理子模块

中文互联网文本数据收集和清理、变换；文本数据存储及观念抽取；文本观念数据的规约；

- 过程语料库生成子模块

中文互联网文本的过程语料库生成和存储；语料模型的特征选择和抽取功能；语料模型的数据训练和预测；

- 模板语料库处理平台子模块

模板语料数据的管理模型；数据处理过程中的问题优化；语料库的性能调优。

开发进度

- **第一阶段：2017年4月-5月**

原始语料库处理子模块：中文互联网文本数据收集和清理、变换；文本数据存储及观念抽取；文本观念数据的规约；

- **第二阶段：2017年6月-9月**

过程语料库生成子模块：中文互联网文本的过程语料库生成和存储；语料模型的特征选择和抽取功能；语料模型的数据训练和预测；

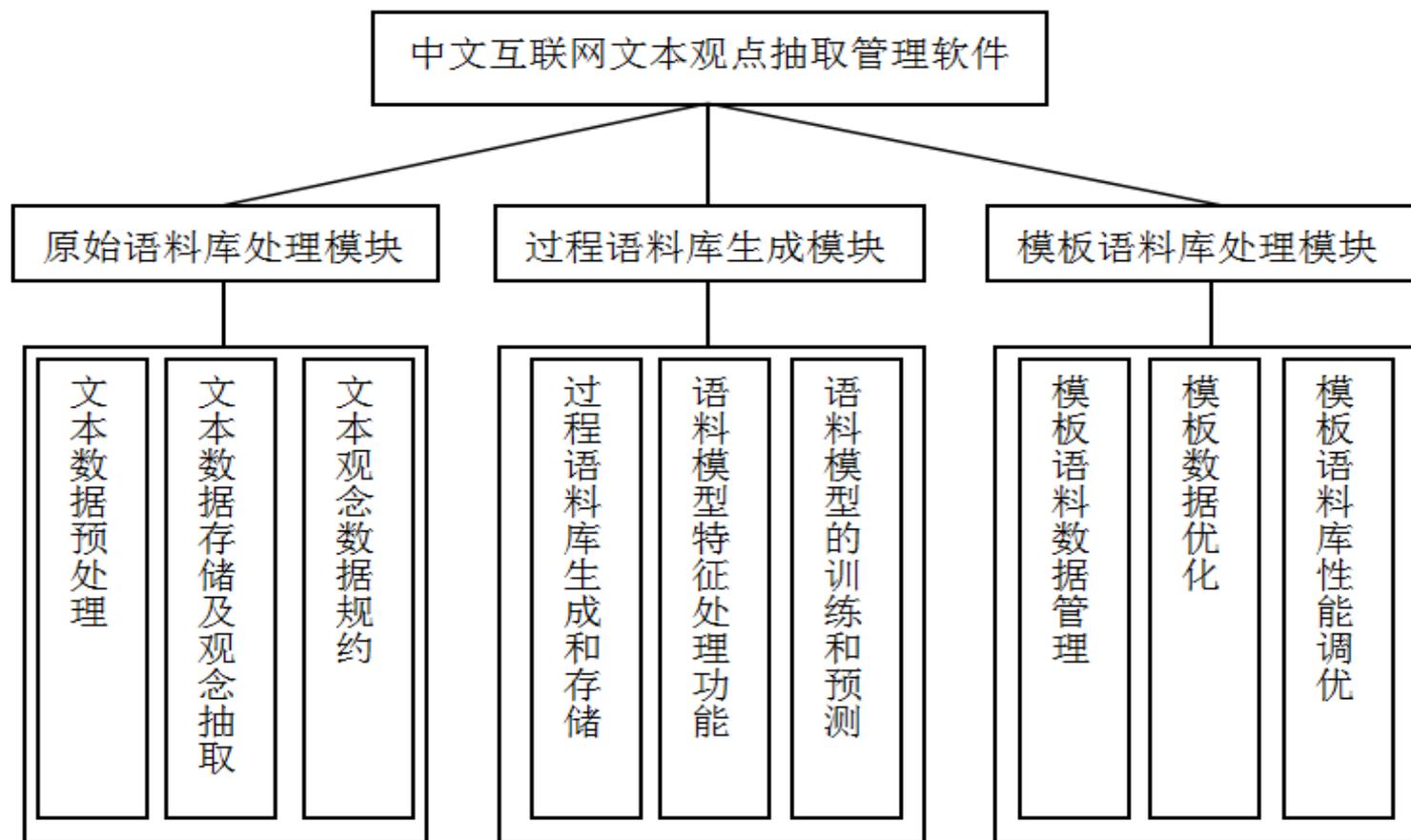
- **第三阶段：2017年9月-11月**

模板语料库处理平台子模块：模板语料数据的管理模型；数据处理过程中的问题优化；语料库的性能调优。

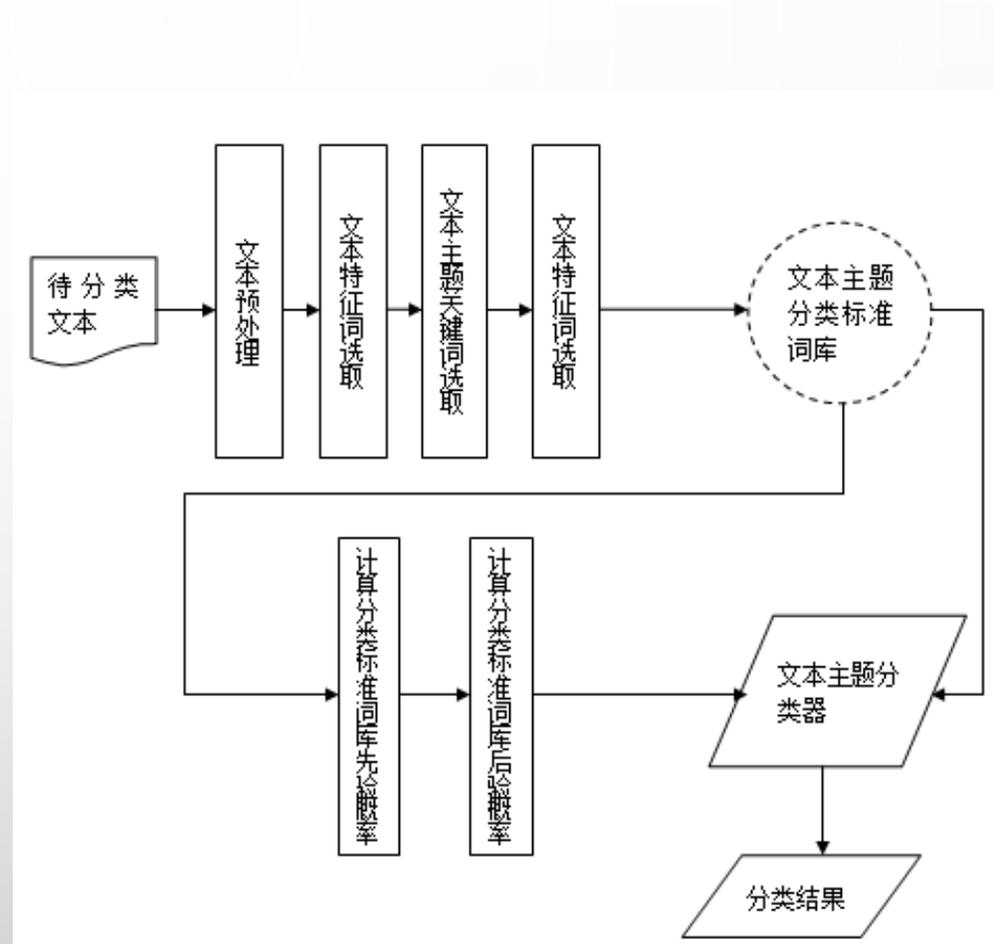
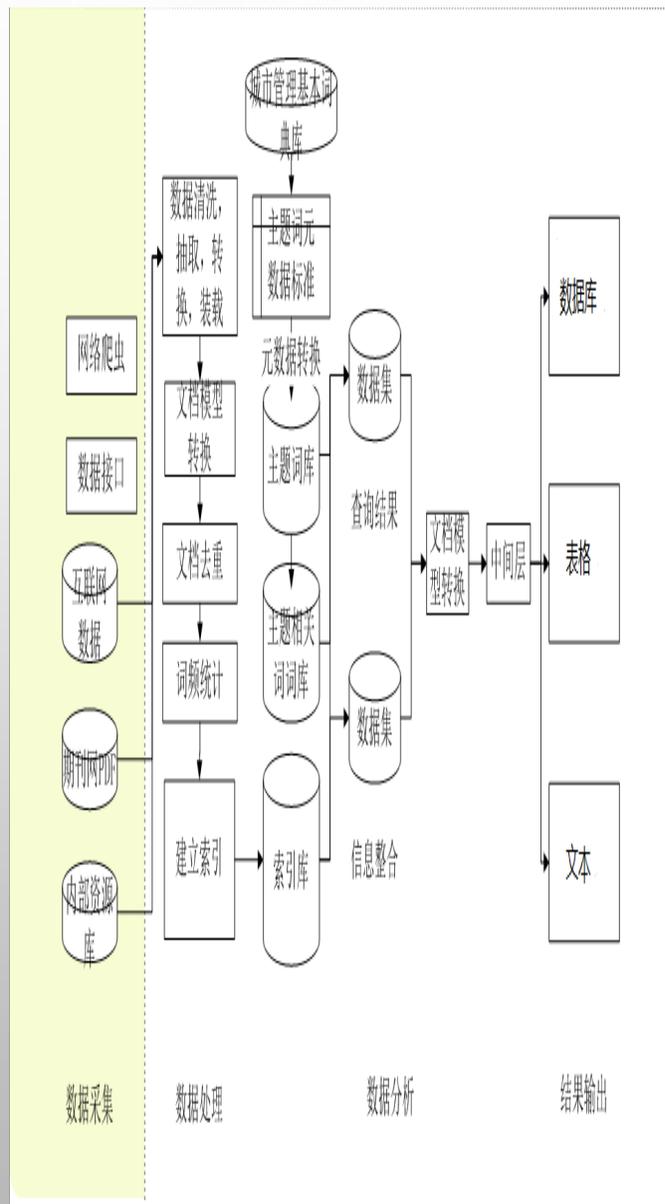
成果报告

- 通过引入TF-IDF、互信息等技术，以大数据技术为指导，设计了本系统，取得了以下成果：
 - “集成不同平台” ——LINUX平台、WINDOW平台
 - “三个子功能模块” ——原始语料库处理、过程语料库生成、模板语料库处理平台子模块
 - “两种不同用户” ——管理人员、普通用户
- 实现了将网络采集和信息查询等新技术资源，融入到科技情报生产线采集与分析中，提升了情报采集工具的效率。

互联网文本观点抽取架构

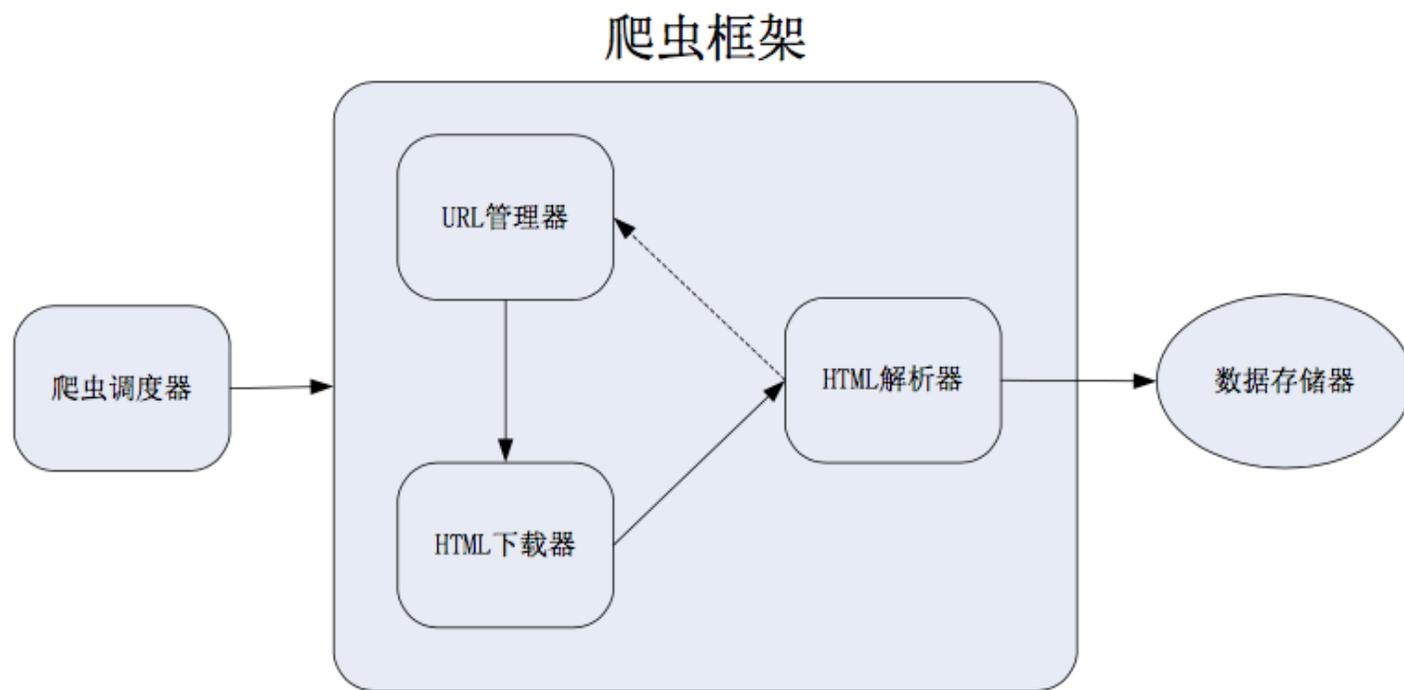


系统流程简介

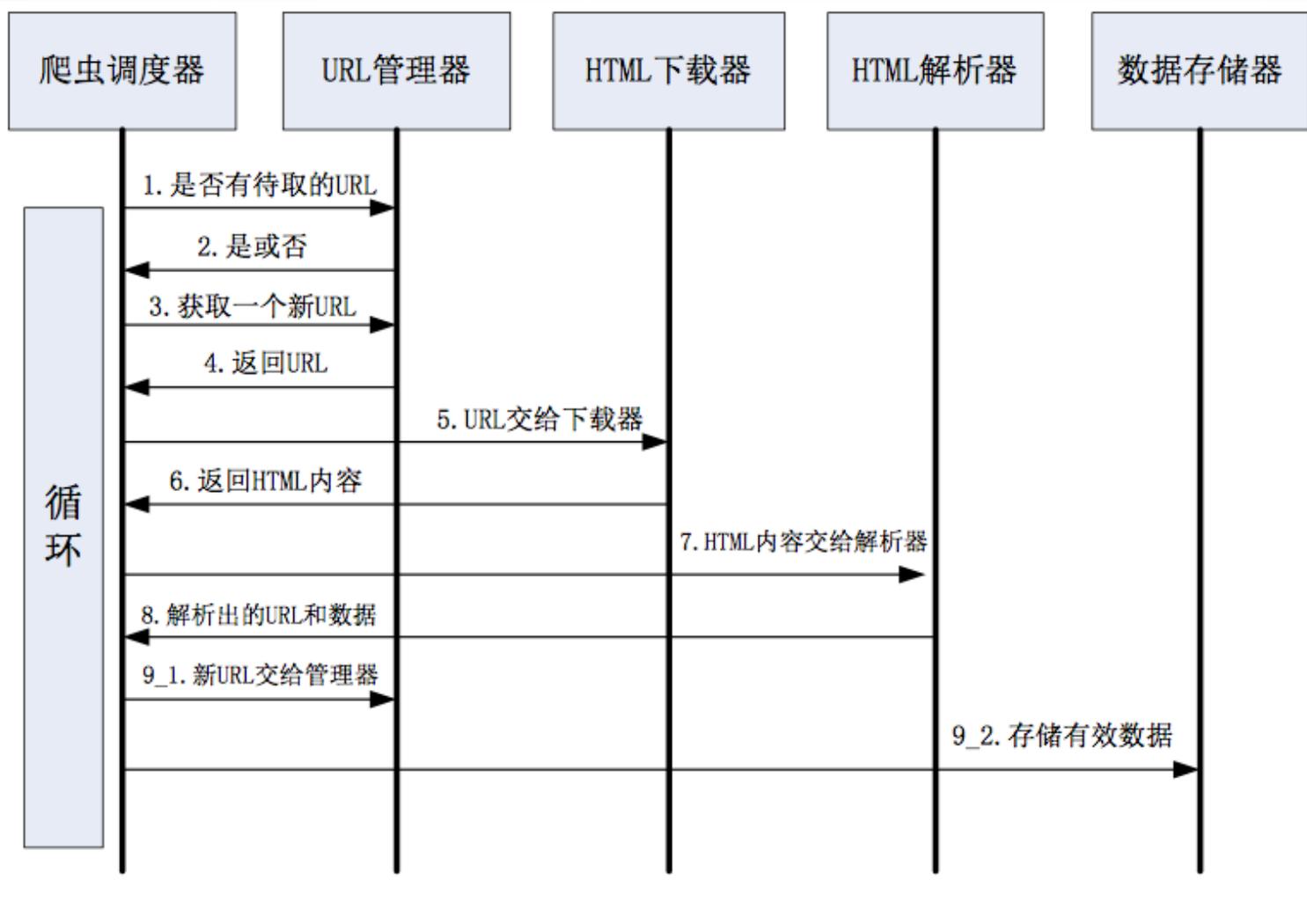


文本数据收集和清理、变换

原始语料处理模块首先要针对一系列的目标网页制定具体的网页解析策略，然后将站点的URL加入待爬取URL队列，初始化爬虫。



数据爬取流程



文本数据存储及观念抽取

名称	修改日期	类型	大小
C29-Transport002.txt	1997/9/28 15:19	文本文档	1 KB
C29-Transport003.txt	1997/9/28 15:19	文本文档	5 KB
C29-Transport004.txt	1997/9/28 15:20	文本文档	2 KB
C29-Transport006.txt	1997/9/28 15:23	文本文档	5 KB
C29-Transport008.txt	1997/9/28 15:24	文本文档	2 KB
C29-Transport010.txt	1997/9/28 15:25	文本文档	2 KB
C29-Transport012.txt	1997/9/28 15:25	文本文档	1 KB
C29-Transport014.txt	1997/9/28 15:26	文本文档	2 KB
C29-Transport016.txt	1997/9/28 15:26	文本文档	1 KB
C29-Transport018.txt	1997/9/28 15:27	文本文档	4 KB
C29-Transport020.txt	1997/9/28 15:29	文本文档	1 KB
C29-Transport022.txt	1997/9/28 15:17	文本文档	2 KB
C29-Transport024.txt	1997/9/28 15:30	文本文档	1 KB
C29-Transport026.txt	1997/9/28 15:31	文本文档	2 KB
C29-Transport028.txt	1997/9/28 15:31	文本文档	3 KB

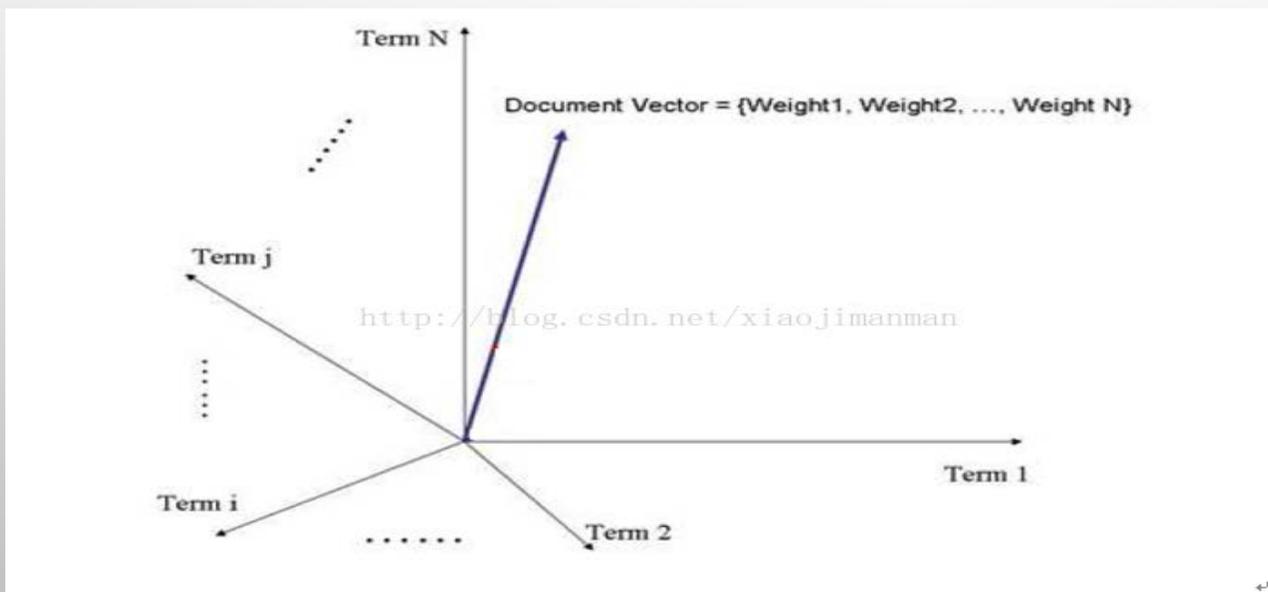
文本观念数据的规约

我们可以把一篇文档看作是一系列词元的集合，每个词元都有一个权重，如下：

Document = {term1, term2, term3 termN}

DocumentVector = {weight1, weight2, weight3 weightN}

我们把所有的文档放到一个N维的空间向量中（所有文档分成N个词元，就N维向；其中文档D在m坐标上的映射为文档D中的m词元的权重），如下图所示。



过程语料库生成和存储

名称	修改日期	类型	大
电子	2017/11/1 17:48	文件夹	
国际	2017/11/1 17:48	文件夹	
环境	2017/11/1 17:48	文件夹	
计算机	2017/11/1 17:49	文件夹	
交通	2017/11/1 17:48	文件夹	
教育	2017/11/1 17:48	文件夹	
矿产	2017/11/1 17:48	文件夹	
历史	2017/11/1 17:48	文件夹	
美术	2017/11/1 17:49	文件夹	
能源	2017/11/1 17:49	文件夹	
太空	2017/11/1 17:48	文件夹	
文学	2017/11/1 17:48	文件夹	
哲学	2017/11/1 17:48	文件夹	

通过分词、切分、去停用词后结果

切除修复:1, '节约资源':2, '芯棒接手':1, '肌肉注射':2, '双向防震铁路桥梁':1, '义务教育物理课程':1, '萃':8, '临界点':2, '血流量':2, '地缘经济关系:欧氏':1, '层
次聚':1, '出口市场':5, '知识交流':3, '堆肥技术':1, '新古典':1, '公路护坡':1, '急性扁桃腺炎':2, '代表议案':3, '落':2, '现代物流体系':1, '无人机系':1, '突出屋
面':1, '图形库':1, '高中历史教材':2, '建筑施':5, '房地产预警':1, '冯惟敏':1, '复杂结构':1, '企业并购重组':5, '折叠尾翼':1, '考试座位表':1, '液量检测':1, '重
钙粉体':1, '微型种植体支抗':2, '口头表达':2, '驱动轮':1, '带状疱疹性角膜炎':1, '网络暴力游戏':1, '文化艺术':11, 'Lotka-Volterra混':1, '多人多出-正交频分':1,
'盈通':2, '《当代经济科学》':1, '加工产业':2, '耗散粒子动力学':1, '模式转换':2, '骨折术':1, '选项课':2, '法学角':1, '组合导航':7, '现状管理':1, 'PC104':1,
网上订购':2, '激素耐药型':1, '计算机体层':1, '2,4-二氯苯氧乙酸':1, '甘油转化产品':1, '护理要':1, '地锚桩':1, '白色':7, '人力资源价值':1, '并行运算':1, '激光
平地机':2, '马蹄形隧洞过水断':1, '侗寨':1, '冷思考':7, '表达':24, '初中数学教材':2, '中华草龟':1, '灵感':8, '版面编排':3, '大学适应期':1, '大渡河':3, '消火
栓系':1, '输尿管小结石':1, '土壤有机碳含量':1, '网上答疑':1, '衰老标记蛋白30':1, '径流分析':1, '试点县':1, '家族企业家':2, '民间化':1, '票据电子化':1, '国际
专利':1, '现役混凝土坝':1, '高频热水消融':1, '大学文化':6, '神经影':1, '轮胎压路机':2, '社会福利院':9, '饱和度':1, '创作误区':1, '巨尾校人工林':1, '工程验
收':1, '碧岩录':1, '以学':2, '粮食补贴':2, '主要死因':1, '发声':5, '合成金刚石':1, '肿瘤/中医药疗法':1, '负压创面治疗技术':1, '货币金融':1, '居住时间':1, '居
孩子的内心':1, '都市型现':1, '农村金融支持':1, '厌氧活性污泥':1, '风电成本':1, '分闸孪子':1, '李贺诗歌':1, '颈动脉疾病':6, '吉安地区':1, '周口':1, '典型大用
户':1, '团队情商':1, '课室申请系统':1, '蓄电池箱':1, '近场聚焦光束形成':1, '旅游市':1, '量子线':2, '聋哑患者':2, '泪囊炎':4, '非法入侵探测':1, '鸡传染性法氏
囊病':17, '感性消费':1, '公共图书馆服务':1, '尿调蛋白':1, '保证金':2, '内部审':3, '注册会计师':7, '安全贮藏':1, '矿渣微粉':2, '改革关键':1, '学术思想':10,
多巴胺':9, '万有引力定律':1, '乙氧亚甲基丙二腈':1, '冻融循':1, '政治等级':1, 'SmartFoxServer':1, '数值信息':1, '区(县)公共图书馆':1, '湍冲吸收塔':1, '长谷川':
3, '村域':1, '提升服务':3, '短信服务':2, '果报观':1, '焯理论':2, '泪囊鼻腔吻合':1, '四联用':1, '自然之':2, '涡轮钻法':1, '血液透析低血压':1, '配合力':3, '卓
越工程师':9, '有杆泵':1, '达芬奇平':1, '数学理解力':1, '混凝土拱桥':1, '艾灵顿':1, '局部屈曲':1, '直线段检测':1, '以数据为中心':1, '吡啶甲酸铂(III)配':1, '达
来诺尔湖泊':1, '多环境':1, '巷道稳定':2, '慢性肾功':1, '南北民歌':1, 'ASICs3':1, '免疫接种':1, '党员':19, '民法实现':1, '电力信息':1, '中国科学院高能物':2,
'生长率':4, '内存均衡':1, '失业保险金期间':1, '重庆市质量技术监督':1, '拔节期':1, '数额':1, '鼻咽肿瘤,老年/放':1, '充压水封':1, '持续发':1, '春耕':2, '人才培养
工作':1, '莎士比':1, '设计艺术教育':1, '断层裂隙水':1, '断层分析':1, '酶解法':2, '解读':86, '温岭市':3, '热休克蛋':1, '视频镜头分类':1, '木前栓塞':2, '影响
性':1, '简支T型梁桥':1, '抗寒性':8, '根系':15, '动态特':1, '农业科技信息':1, '五味消毒饮':2, '难处理金矿':2, '腰椎骨':1, '现代城市':10, 'MAPK通路':1, '关键
性指':1, '西安市档案馆':1, '甲状腺肿瘤/病理学':4, '高校技术转移':1, '科学技术委员会':4, '押扣':1, '关键词:化学灌浆':1, '异常行为':1, 'Snodgras':1, '中小型家族
企业':1, '布拉氏酵母菌':1, '崇明':2, '资源节约型农业':1, '价值诉求':1, '异读':1, '永靖县':2, '水稻氮素':1, '《风趣的他》':1, '结露':1, '商贸物':1, 'HFO-
1234yf':1, '煤岩':2, '髌关节置换':6, '产业技术进步':1, '海岸沙丘表面':1, 'CP':3, '草酸钙结石':1, '测厚':3, '管理和服务':1, '尾孢菌':1, '埋地金属':1, '控制回
路':1, '无创':2, '资源建':4, '尼尔雌':1, '德国《劳务派遣法》':1, '动机理':1, '活起来':3, '脾实质内不':1, '第三方在线支付':1, '经济开发区建设':1, '古城保护':1,
'间作套':2, '精准定位':1, 'Structures':1, '克拉克瓷':1, '起诉裁量':2, '水泥矿山':2, '稳定期COPD':1, '区域水资源':1, '建构知识':1, '牛奶蛋白':1, '亚胺培南耐药
鲍曼':1, '人地关系地域系统':1, '牢笼':1, '空间结构优化':1, '土地集约化经':1, '介入化疗':3, '投资对象':1, '农村宅基地':11, '地下集装箱运输系':1, '无线移动':1,
'化学药物':2, '增值税政策':1, 'DCS节能管理法':1, '开架借':1, '重组组织型纤溶酶':1, '白芨多糖':1, '企业博客营销':1, '红外线灯':1, '顶空气相色谱法':1, '分子标记
辅助育种':1, '华':11, '高维稀疏数据':1, '薄膜反射镜':1, '声控技术':1, '超声波清洗技术':1, '舰炮':3, '人民医院':8, '皮肤创伤':1, '甲骨卜':2, '农业技术推广体系

语料模型的特征选择和抽取功能

我们通过文本特征词选择算法选择对文本分类影响较大的词汇，通过文本特征词提取算法对词汇进行权重计算。通过TF-IDF得出文本词汇分值，对分值进行排序，提取较高分值词汇，结果如图所示。

政府竞争': 2.4591776509935076e-05, '游客转化': 2.4591776509935076e-05, '小花南芥': 2.4591776509935076e-05, '学前教育': 2.4591776509935076e-05, '面部表情': 2.4591776509935076e-05, '培养模': 2.4591776509935076e-05, '英语歧义': 2.4591776509935076e-05, '碱性蛋白酶': 4.918355301987015e-05, '排球选': 2.4591776509935076e-05, '000m障碍': 2.4591776509935076e-05, '眼神语言': 2.4591776509935076e-05, '超声检查,多普勒': 2.4591776509935076e-05, '体育教育': 2.4591776509935076e-05, '跨文化交际': 2.4591776509935076e-05, '封存点': 2.4591776509935076e-05, '历史情结': 2.4591776509935076e-05, '排水管道工程': 2.4591776509935076e-05, '资源共': 2.4591776509935076e-05, '小学语文教学': 2.4591776509935076e-05, '甘肃农村': 2.4591776509935076e-05, '信息服务质量': 2.4591776509935076e-05, '收入分配改革': 2.4591776509935076e-05, '典型模型': 2.4591776509935076e-05, '闸北区': 2.4591776509935076e-05, '实施障碍': 2.4591776509935076e-05, '实验教': 2.4591776509935076e-05, '创业富': 2.4591776509935076e-05, '路面下封层': 2.4591776509935076e-05, '西方教育': 2.4591776509935076e-05, '三位一体': 2.4591776509935076e-05, '中船集团公司': 2.4591776509935076e-05, '一体化管理模式': 2.4591776509935076e-05, '口传艺术': 2.4591776509935076e-05, '病因病': 2.4591776509935076e-05, '深度处理': 2.4591776509935076e-05, '直接支配': 2.4591776509935076e-05, '土壤培肥': 2.4591776509935076e-05, '纳什均': 2.4591776509935076e-05, '距离': 2.4591776509935076e-05, '信息泛滥': 2.4591776509935076e-05, 'K-燃料(R)技术': 2.4591776509935076e-05, '紫外透射光谱': 2.4591776509935076e-05, '人物形': 2.4591776509935076e-05, '茶银尺罐': 2.4591776509935076e-05, '拔': 2.4591776509935076e-05, '心血管病': 2.4591776509935076e-05, '行星地球': 2.4591776509935076e-05, '国外优秀射箭运动': 2.4591776509935076e-05, '说话发声': 2.4591776509935076e-05, '割制首缩': 2.4591776509935076e-05, '军人': 2.4591776509935076e-05, '政策实践': 2.4591776509935076e-05, '习字': 2.4591776509935076e-05, '连续采煤机': 2.4591776509935076e-05, '设计及传': 2.4591776509935076e-05, '资源化技术': 2.4591776509935076e-05, '地图网站系统': 2.4591776509935076e-05, '篮球选项课': 2.4591776509935076e-05, '荨麻疹/中医药疗法': 2.4591776509935076e-05, '会计本质': 2.4591776509935076e-05, '马克思恩格斯': 2.4591776509935076e-05, '敏感场分析': 2.4591776509935076e-05, '道德风险': 2.4591776509935076e-05, '腺嘌呤': 2.4591776509935076e-05, '折射体绘制': 2.4591776509935076e-05, '可信密码模块': 2.4591776509935076e-05, '传统手工艺': 2.4591776509935076e-05, '战略绩效': 2.4591776509935076e-05, '致辞': 2.4591776509935076e-05, '待人接物': 2.4591776509935076e-05, '水产品产量': 2.4591776509935076e-05, '时代转型': 2.4591776509935076e-05, 'C/SiC-ZrC': 2.4591776509935076e-05, '斯奇康': 2.4591776509935076e-05, '肺与大肠相表里': 2.4591776509935076e-05, '内容与': 2.4591776509935076e-05, '综合货运': 2.4591776509935076e-05, '提起': 2.4591776509935076e-05, '环境保护与可持': 2.4591776509935076e-05, '同种异体': 2.4591776509935076e-05, '具': 2.4591776509935076e-05, '医学论': 2.4591776509935076e-05, '资源开发': 4.918355301987015e-05, '迁移与创伤': 2.4591776509935076e-05, '无人机图像': 2.4591776509935076e-05, '菲利普斯曲线': 2.4591776509935076e-05, 'SiC': 2.4591776509935076e-05, '异': 2.4591776509935076e-05, '血样': 2.4591776509935076e-05, '批判理论': 2.4591776509935076e-05, '不同雾化吸入方案': 2.4591776509935076e-05, '科学教师': 2.4591776509935076e-05, '第一次反'围': 2.4591776509935076e-05, '适应性记忆': 2.4591776509935076e-05, '教材建设': 2.4591776509935076e-05, '风电发展': 2.4591776509935076e-05, '被害人保': 2.4591776509935076e-05, '艺术的角度': 2.4591776509935076e-05, '凸二次规划': 2.4591776509935076e-05, '中国石油天然气集': 2.4591776509935076e-05, '广西人民': 2.4591776509935076e-05, '服务企业': 2.4591776509935076e-05, '轻型战术轮式车辆': 2.4591776509935076e-05, '案例教学法': 2.4591776509935076e-05, '时间同步': 2.4591776509935076e-05, '利用外资': 2.4591776509935076e-05, '寺院文': 2.4591776509935076e-05, '讲课比赛': 2.4591776509935076e-05, '断': 2.4591776509935076e-05, '双曲抛物': 2.4591776509935076e-05, '甲醇制烯烃': 2.4591776509935076e-05, '伐昔洛韦': 2.4591776509935076e-05, '生殖激素': 2.4591776509935076e-05, '医院网络': 2.4591776509935076e-05, '估': 2.4591776509935076e-05, '小学体育课': 2.4591776509935076e-05, '模糊子坡(理想)': 2.4591776509935076e-05, '高考数学试题': 2.4591776509935076e-05, '营销渠': 2.4591776509935076e-05, '乡村小说': 2.4591776509935076e-05, '学习数学': 2.4591776509935076e-05, 'α-萜烯': 4.918355301987015e-05, '电网运行': 2.4591776509935076e-05, '风险感知': 2.4591776509935076e-05, '摩崖': 2.4591776509935076e-05, '机械专业学生': 2.4591776509935076e-05, '农业清洁生产': 2.4591776509935076e-05, '牙拔': 2.4591776509935076e-05, '赤贫指数': 2.4591776509935076e-05, '法律现象': 2.4591776509935076e-05, '篮球教学': 2.4591776509935076e-05, '相': 4.918355301987015e-05, '络合': 2.4591776509935076e-05, '椎间盘突出症': 2.4591776509935076e-05, '中国信': 2.4591776509935076e-05, '正常流动': 2.4591776509935076e-05, '氟化物中毒': 2.4591776509935076e-05, '襄阳': 2.4591776509935076e-05, '角色 爱因斯坦场': 2.4591776509935076e-05, '电声演': 2.4591776509935076e-05, '信息类理': 2.4591776509935076e-05, '特型理': 2.4591776509935076e-05

语料模型的数据训练和预测

通过前面抽取的特征，构建文本的向量模型，运用合适的算法将需要分类的文本数据和经过特征抽取的语料库模型进行比对，并对需要分类的文本数据进行类别预测。分类模型效果图如下。

```
D:\Anaconda\python.exe E:/PyWorkSpace/chinese_text_classification-master/NBayes_Predict.py
C:/Users/FYM/Desktop/test_corpus_seg/C3-Art/C3-Art1027.txt : 实际类别: C3-Art -->预测类别: C7-History
C:/Users/FYM/Desktop/test_corpus_seg/C3-Art/C3-Art1035.txt : 实际类别: C3-Art -->预测类别: C7-History
C:/Users/FYM/Desktop/test_corpus_seg/C3-Art/C3-Art1057.txt : 实际类别: C3-Art -->预测类别: C7-History
C:/Users/FYM/Desktop/test_corpus_seg/C3-Art/C3-Art1067.txt : 实际类别: C3-Art -->预测类别: C7-History
C:/Users/FYM/Desktop/test_corpus_seg/C3-Art/C3-Art1069.txt : 实际类别: C3-Art -->预测类别: C7-History
C:/Users/FYM/Desktop/test_corpus_seg/C3-Art/C3-Art1071.txt : 实际类别: C3-Art -->预测类别: C7-History
C:/Users/FYM/Desktop/test_corpus_seg/C3-Art/C3-Art1075.txt : 实际类别: C3-Art -->预测类别: C7-History
C:/Users/FYM/Desktop/test_corpus_seg/C3-Art/C3-Art1086.txt : 实际类别: C3-Art -->预测类别: C7-History
C:/Users/FYM/Desktop/test_corpus_seg/C3-Art/C3-Art1090.txt : 实际类别: C3-Art -->预测类别: C7-History
预测完毕!!!
精度:0.964
召回:0.945

Process finished with exit code 0
```

模板语料数据的管理模型

通过对数据库的操作可以实现上传语料信息、删除语料信息、模板语料库检索等功能。

id	title	department_name	lead_research_organ	abstract	funded_value	funded_period	funder	project_status	project_category	project_reference
286	Visible I	Chemistry	University of Cambri	Titanium dic	644,075	十月 07 - 二月 11	EPSRC	Closed	Research Grant	EP/F00253X/1
287	Photorr	Physics	University of Cambri	The submilli	969,457	四月 15 - 三月 18	STFC	Active	Research Grant	ST/M000818/1
288	Hydrog	Chemistry	University of Bath	The most cc	47,977	九月 12 - 十一月 13	EPSRC	Closed	Research Grant	EP/G068038/2
289	Structu	Biology and Biochen	University of Bath	Angiotensir	595,238	二月 16 - 一月 20	MRC	Active	Research Grant	MR/M026647/1
290	Long-ra	Sch of Geography	University of Nottin	The Arctic is	26,291	二月 10 - 八月 13	NERC	Closed	Research Grant	NE/G019193/1
291	Treache	Humanities Arts Lan	London Metropolita	My propose	10,139	四月 10 - 六月 10	AHRC	Closed	Research Grant	AH/H003649/1
292	Deuteri	Sch of Mathematical	University of Sussex	One of the	564,400	十月 11 - 九月 15	EPSRC	Closed	Research Grant	EP/I029230/1
293	Learnin	Moray House Schoc	University of Edinbu	Much is alre	77,704	五月 07 - 十月 09	ESRC	Closed	Research Grant	ES/E025099/1
294	Learnin	Moray House Schoc	University of Edinbu	Much is alre	77,704	五月 07 - 十月 09	ESRC	Closed	Research Grant	ES/E025099/1
295	Structu	Biology and Biochen	University of Bath	Angiotensir	595,238	二月 16 - 一月 20	MRC	Active	Research Grant	MR/M026647/1
296	Quantu	Oxford Physics	University of Oxford	Electrons in	545,702	十月 10 - 三月 15	EPSRC	Closed	Research Grant	EP/H014934/1
297	Deuteri	Sch of Mathematical	University of Sussex	One of the	564,400	十月 11 - 九月 15	EPSRC	Closed	Research Grant	EP/I029230/1
298	Quantu	Oxford Physics	University of Oxford	Electrons in	545,702	十月 10 - 三月 15	EPSRC	Closed	Research Grant	EP/H014934/1
299	Learnin	Moray House Schoc	University of Edinbu	Much is alre	77,704	五月 07 - 十月 09	ESRC	Closed	Research Grant	ES/E025099/1
300	Photorr	Physics	University of Cambri	The submilli	969,457	四月 15 - 三月 18	STFC	Active	Research Grant	ST/M000818/1
301	The Ma	Dept of Mathematic	Imperial College Lor	One of the	461,979	二月 14 - 七月 17	EPSRC	Active	Research Grant	EP/K041134/1
302	Visible I	Chemistry	University of Cambri	Titanium dic	644,075	十月 07 - 二月 11	EPSRC	Closed	Research Grant	EP/F00253X/1
303	Treache	Humanities Arts Lan	London Metropolita	My propose	10,139	四月 10 - 六月 10	AHRC	Closed	Research Grant	AH/H003649/1
304	Visible I	Chemistry	University of Cambri	Titanium dic	644,075	十月 07 - 二月 11	EPSRC	Closed	Research Grant	EP/F00253X/1
305	Photorr	Physics	University of Cambri	The submilli	969,457	四月 15 - 三月 18	STFC	Active	Research Grant	ST/M000818/1
306	Learnin	Moray House Schoc	University of Edinbu	Much is alre	77,704	五月 07 - 十月 09	ESRC	Closed	Research Grant	ES/E025099/1

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	研究領域	研究領域(略)	研究課題/領域	研究種目	配分区分	審査区分	研究分野	研究機関	研究代表者	研究期間(年度)	研究課題ステータス	配分額	注記	キーワード	
2	次世代水素ステーションに	16K21672	若手研究(B)	基金			構造・機能材料 エネルギー学	国立研究開発		2016-04-01 - 2018-03-31	交付(2016年度)	20	水素/水素昇圧/水素ステーション		
3	基礎物理現象に基づく先端	16K21673	若手研究(B)	基金			計測工学 分析化学	国立研究開発		2016-04-01 - 2019-03-31	交付(2016年度)	20	計測工学/計測機器		
4	化学物質の有害性評価手法	16K21674	若手研究(B)	基金			環境政策・環境社会システム 環境影響評価	国立研究開発		2016-04-01 - 2020-03-31	交付(2016年度)	20	応用数学/統計数学/有害化学物質/動物実験代替法		
5	ビッグデータ解析環境への	16K21675	若手研究(B)	基金			マルチメディア・データベース 高性能計算	国立研究開発		2016-04-01 - 2018-03-31	交付(2016年度)	20	データストレージ/ビッグデータ解析/データステージング		
6	プラズマ中の微粒子が及ぼ	16K21676	若手研究(B)	基金			プラズマエレクトロニクス プラズマ科学	国立研究開発		2016-04-01 - 2018-03-31	交付(2016年度)	20	プラズマ加工/プラズマ・核融合		
7	信頼性の高い3次元地質情報	16K21677	若手研究(B)	基金			自然災害科学・防災学 地質学	国立研究開発		2016-04-01 - 2019-03-31	交付(2016年度)	20	地質学		
8	移動体への高効率ワイヤレ	16K21678	若手研究(B)	基金			計測工学 航空宇宙工学	国立研究開発		2016-04-01 - 2018-03-31	交付(2016年度)	20	計測工学		
9	加速器BNCT用中性子エネ	16K21679	若手研究(B)	基金			量子ビーム科学 医学物理学・放射線技術学	国立研究開発		2016-04-01 - 2019-03-31	交付(2016年度)	20	加速器/放射線/放射線計測/中性子/放射線治療		
10	ミトコンドリア内膜プロテ	16K21680	若手研究(B)	基金			生命・健康・医療情報学 生物物理学	国立研究開発		2016-04-01 - 2018-03-31	交付(2016年度)	20	ミトコンドリア/タンパク質分解/ストロマ/呼吸連鎖/ATP合成/Inhibiting		
11	新規重希土類資源としての	16K21681	若手研究(B)	基金			地球・資源システム工学 岩石・鉱物・鉱床	国立研究開発		2016-04-01 - 2019-03-31	交付(2016年度)	20	重希土類資源/アパタイト/リン酸/45%		
12	マルチ同位体分析による次	16K21682	若手研究(B)	基金			岩石・鉱物・鉱床学 地球・資源システム工	国立研究開発		2016-04-01 - 2019-03-31	交付(2016年度)	20	鉱床学/地球化学/リチウム鉱床/同位体		
13	レプリン酸を活用した新規	16K21683	若手研究(B)	基金			持続可能システム グリーン・環境化学	国立研究開発		2016-04-01 - 2018-03-31	交付(2016年度)	20	バイオマス/有機化学/触媒/化学プロセス/高分子合成/		
14	(社会的養護の家庭化)に	16K21684	若手研究(B)	基金			社会学 子ども学(子ども環境学)	国立社会保障		2016-04-01 - 2019-03-31	交付(2016年度)	20	社会学/福祉社会学/家族社会学/社会的養護/質的研究		
15	体重階級別競技選手の減量	16K21685	若手研究(B)	基金			スポーツ科学 身体教育学	独立行政法人		2016-04-01 - 2019-03-31	交付(2016年度)	20	減量/脱水/酸化ストレス/抗酸化能/減量期間		
16	二重課題遂行サルを用いた	16K21686	若手研究(B)	基金			実験心理学 認知科学	国立研究開発		2016-04-01 - 2018-03-31	交付(2016年度)	20	認知神経科学/前頭葉機能/動物行動科学		
17	視聴覚モダリティにまたが	16K21687	若手研究(B)	基金			認知科学 実験心理学	国立研究開発		2016-04-01 - 2018-03-31	交付(2016年度)	20	認知科学/実験心理学		
18	ウェアラブル脳波計を用い	16K21688	若手研究(B)	基金			生命・健康・医療情報学 脳計測科学	国立研究開発		2016-04-01 - 2019-03-31	交付(2016年度)	20	脳/神経/生体情報/脳波/聴覚定常状態応答		
19	脳の機能的構造の観点から	16K21689	若手研究(B)	基金			認知科学 実験心理学	国立研究開発		2016-04-01 - 2019-03-31	交付(2016年度)	20	実験心理学/認知科学		
20	大気汚染物質沈着量に対す	16K21690	若手研究(B)	基金			環境動態解析 気象・海洋物理・陸水学	一般財団法人		2016-04-01 - 2019-03-31	交付(2016年度)	20	環境変動/環境政策/地球化学/気象学/大気汚染防止・浄		
21	ホスホリパーゼDに着目し	16K21691	若手研究(B)	基金			疼痛学 生理学一般	大学共同利用		2016-04-01 - 2019-03-31	交付(2016年度)	20	生理学/細胞生物学/体性感覚/かゆみ/TRPチャネル		
22	列車振動を用いた不整形地	16K21692	若手研究(B)	基金			構造工学・地震工学・維持管理工学 地球	公益財団法人		2016-04-01 - 2018-03-31	交付(2016年度)	20	列車振動/反射法探査/地震波干渉法/不整形地盤		
23	肢位の違いが鍵振動刺激に	16K21693	若手研究(B)	基金			身体教育学 実験心理学	地方独立行政		2016-04-01 - 2019-03-31	交付(2016年度)	20	運動障害/鍵振動刺激/知覚運動協応/知覚運動学習/感覚		
24	圧力を用いた生体鉱物模倣	16K21694	若手研究(B)	基金			岩石・鉱物・鉱床学 有機・ハイブリッド材	地方独立行政		2016-04-01 - 2019-03-31	交付(2016年度)	20	複合材料/物性/結晶成長/生体鉱物/炭酸カルシウム		
25	常同行動の神経基盤-線条体	16K21695	若手研究(B)	基金			基盤・社会脳科学 神経化学・神経薬理学	地方独立行政		2016-04-01 - 2019-03-31	交付(2016年度)	20	ドーパミン/常同行動/パッチクランプ/大脳基底核/線条		
26	要介護高齢者を対象とした	16K21696	若手研究(B)	基金			高齢看護学 社会福祉学	地方独立行政		2016-04-01 - 2019-03-31	交付(2016年度)	20	高齢者/介護福祉/QOL		
27	医療情報データベースを用	16K21697	若手研究(B)	基金			リハビリテーション科学・福祉工学 医療社	独立行政法人		2016-04-01 - 2019-03-31	交付(2016年度)	20	虚血性心疾患/心臓リハビリテーション/データベース/診		
28	高速自動解析手法を用いた	16K21698	若手研究(B)	基金			生物多様性・分類 環境動態解析	国立研究開発		2016-04-01 - 2019-03-31	交付(2016年度)	20	海洋生態学/環境影響評価/メタオファウナ		
29	地震波形を再現する三次元	16K21699	若手研究(B)	基金			固体地球惑星物理学 自然災害科学・防災学	東京大学		2016-04-01 - 2019-03-31	交付(2016年度)	20	内部構造/地震動		
30	北極海の海洋乱流エネルギー	16K21700	若手研究(B)	基金			気象・海洋物理・陸水学 環境動態解析	国立研究開発		2016-04-01 - 2019-03-31	交付(2016年度)	20	海洋物理/北極海/海洋乱流/気候変動		
31	北大洋における微小動物	16K21701	若手研究(B)	基金			環境動態解析 水圏生産科学	国立研究開発		2016-04-01 - 2019-03-31	交付(2016年度)	20	Microzooplankton / Functional response / Dilution technique		
32	堅果類の豊凶同期パターン	16K21702	若手研究(B)	基金			森林科学 文化財科学・博物館学	国立研究開発		2016-04-01 - 2019-03-31	交付(2016年度)	20	豊凶同期/堅果/ツキノワグマ/成長動態/博物館資料		
33	電力自由化及び炭素税導入	16K21703	若手研究(B)	基金			環境政策・環境社会システム 持続可能シ	公益財団法人		2016-04-01 - 2018-03-31	交付(2016年度)	20	地球温暖化ガス排出削減/再生可能エネルギー/省エネルギー		
34	客の意思決定と混雑現象の	16K21704	若手研究(B)	基金			社会システム工学・安全システム 数理情報	防衛大学校		2016-04-01 - 2018-03-31	交付(2016年度)	20	OR/待ち行列/意思決定/ゲーム理論/Nash均衡		
35	ホール係数を用いた金属表	16K21705	若手研究(B)	基金			計測工学 機械材料・材料力学	独立行政法人		2016-04-01 - 2019-03-31	交付(2016年度)	20	計測工学/金属物性/物性実験		
36	博物館が所蔵する写真の資	16K21706	若手研究(B)	基金			文化財科学・博物館学 文化人類学・民俗学	山梨県立博物		2016-04-01 - 2020-03-31	交付(2016年度)	20	博物館資料/民俗学/写真		
37	皮膚ガスに含まれる活性酸	16K21707	若手研究(B)	基金			生体工医学・生体材料学 内科学一般(内科)	中部大学		2016-04-01 - 2018-03-31	交付(2016年度)	20	ヒドロキシルラジカル/生体ガス/レーザー誘起蛍光法/質		
38	小動物による補助人工心臓	16K21708	若手研究(B)	基金			生理学一般 心臓血管外科学	国立研究開発		2016-04-01 - 2018-03-31	交付(2016年度)	20	生理学/循環力学/循環調/心筋疾患外科学		
39	ヒトIPS細胞を用いたパレ	16K21709	若手研究(B)	基金			感染症内科学 衛生学・公衆衛生学	大阪市立環境		2016-04-01 - 2018-03-31	交付(2016年度)	20	パレコウイルス/筋炎/筋痛症		
40	ベンソトリアゾール系紫外	16K21710	若手研究(B)	基金			衛生学・公衆衛生学 放射線・化学物質影響	大阪府立公衆		2016-04-01 - 2018-03-31	交付(2016年度)	20	紫外線吸収剤/ダイオキシン/代謝/DNA損傷		
41	水辺の習俗行事にかかわる	16K21711	若手研究(B)	基金			環境農学(含ランドスケープ科学) 文化財科	兵庫県立人と		2016-04-01 - 2019-03-31	交付(2016年度)	20	文化的景観/水辺空間/習俗行事/空間変容/社会変容		
42	放射光X線を用いた微生物	16K21712	若手研究(B)	基金			高分子・繊維材料 持続可能システム	(公財)高輝度		2016-04-01 - 2019-03-31	交付(2016年度)	20	バイオマス由来プラスチック/型放射光/高次構造形成メカ		
43	アミロソフェロイドによる	16K21713	若手研究(B)	基金			薬理系薬学 薬理学一般	公益財団法人		2016-04-01 - 2018-03-31	交付(2016年度)	20	薬理学/生理学		
44	β-Klotho/FGF15システムの	16K21714	若手研究(B)	基金			食生活学 食品科学	公益財団法人		2016-04-01 - 2019-03-31	交付(2016年度)	20	栄養学/β-Klotho/FGF15/子宮内発育遅延/ノックアウト		
45	地理情報システムを用いた	16K21715	若手研究(B)	基金			考古学 人文地理学	独立行政法人		2016-04-01 - 2020-03-31	交付(2016年度)	20	考古学/日本史/地理情報システム(GIS)/古代交通		
46	次世代シーケンサーによる	16K21716	若手研究(B)	基金			土木環境システム 地球・資源システム工学	福岡県保健環		2016-04-01 - 2019-03-31	交付(2016年度)	20	廃棄物処理/最終処分場/微生物群集解析/次世代シーケン		
47	大型板鯨類の生殖状態を反	16K21717	若手研究(B)	基金			水圏生命科学 動物生産科学	一般財団法人		2016-04-01 - 2018-03-31	交付(2016年度)	20	大型板鯨類/飼育下繁殖/性ステロイドホルモン/ピテロジ		
48	実世界センシングを用いた	16K21718	若手研究(B)	基金			知能情報学 ウェブ情報学・サービス情報学	株式会社国際		2016-04-01 - 2019-03-31	交付(2016年度)	20	人工知能/情報検索		
49	Deep Learning Based Phys	16K21719	若手研究(B)	基金			知能情報学 ヒューマンインタフェース・イ	株式会社国際		2016-04-01 - 2019-03-31	交付(2016年度)	20	知能情報処理/知能ロボット		
50	思春期主体脳・生活・人	16K21720	若手研究(B)	基金			新学術領域研究(研究領	東京大学		2016-06-30 - 2021-03-31	交付(2016年度)	20	認知科学		
51	パレオアジアパレオアジア	16K21721	若手研究(B)	基金			新学術領域研究(研究領	東京大学		2016-06-30 - 2021-03-31	交付(2016年度)	20	考古学/旧石器時代/ホモ・サピエンス/人類進化/新人旧		
52	国際活動支援 特異構造の結	16K21722	若手研究(B)	基金			新学術領域研究(研究領	東京大学		2016-06-30 - 2021-03-31	交付(2016年度)	20	結晶工学/電子・電気材料/半導体物性/格子欠陥/先端機		
53	「ネオウイルス」ネオウイルス	16K21723	若手研究(B)	基金			新学術領域研究(研究領	東京大学		2016-06-30 - 2021-03-31	交付(2016年度)	20	ウイルス/宿主との共生/生態		

项目考核指标完成情况

- 原始语料库处理子模块（完成）

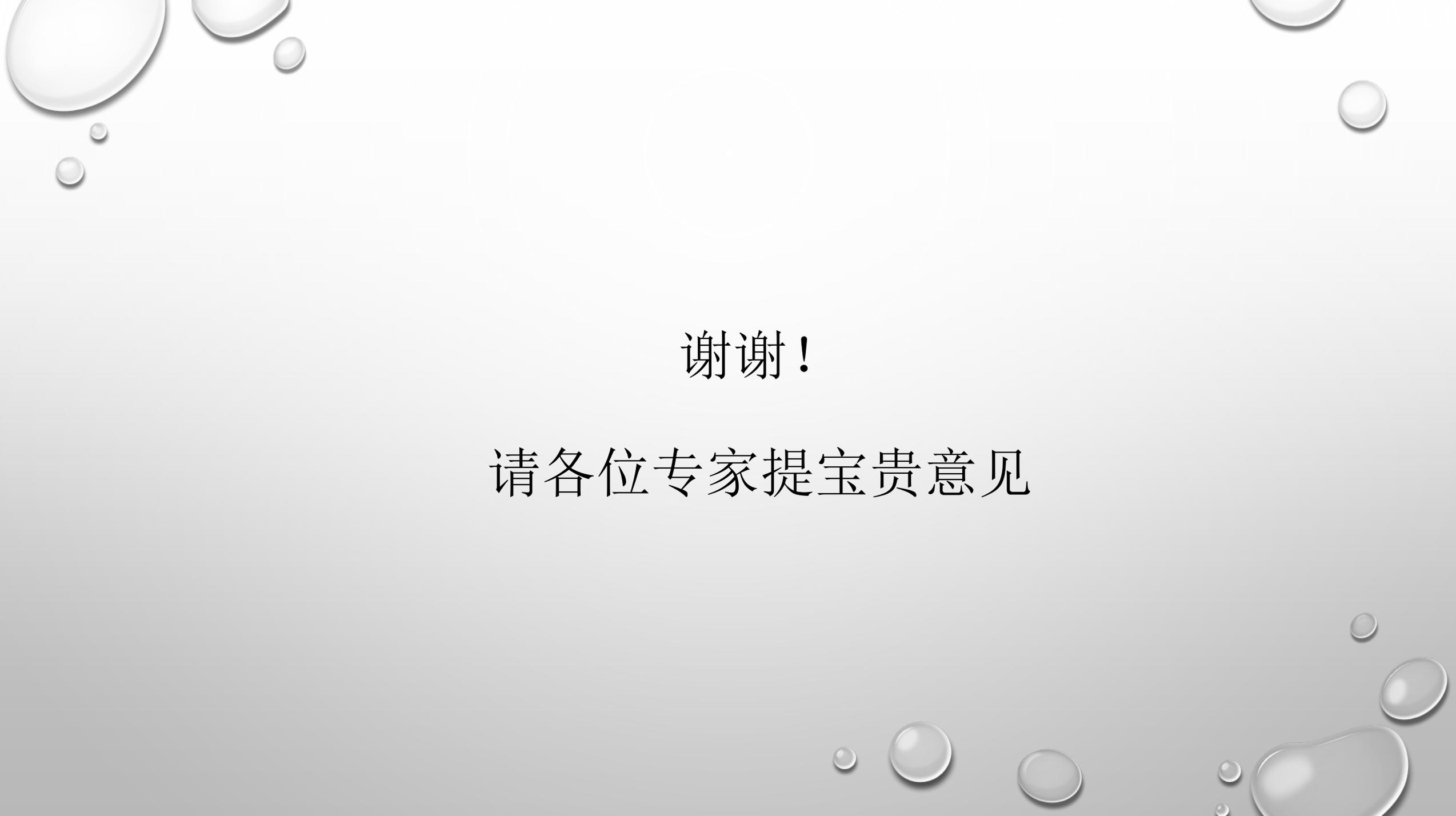
中文互联网文本数据收集和清理、变换；文本数据存储及观念抽取；文本观念数据的规约的制定；

- 过程语料库生成子模块（完成）

中文互联网文本的过程语料库生成和存储；语料模型的特征选择和抽取功能；语料模型的数据训练和预测；

- 模板语料库处理平台子模块（完成）

模板语料数据的管理模型；数据处理过程中的问题优化；语料库的性能调优。

The background features a light gray gradient with several realistic water droplets of various sizes scattered in the corners. The droplets have highlights and shadows, giving them a three-dimensional appearance.

谢谢！

请各位专家提宝贵意见