

# 学科领域属性抽取及分类

北京邮电大学

2017年7月

# 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

- “学科领域属性抽取及分类”项目，设计将特征词提取技术与中文文本分类技术，与科技情报生产线对接的机制及方案，实现一个学科领域属性抽取及分类系统。

# 考核指标

- 1) 学科领域属性数据处理:

对各学科数据的文章进行切词, 去除停用词, 词性标注等处理

- 2) 学科领域属性分类模型生成:

用大量的多学科文章进行模型的训练, 并用多学科的文章对模型进行预测。

- 3) 学科领域属性分类模型优化:

针对分类过程数据稀疏的特点进行分类模型的优化

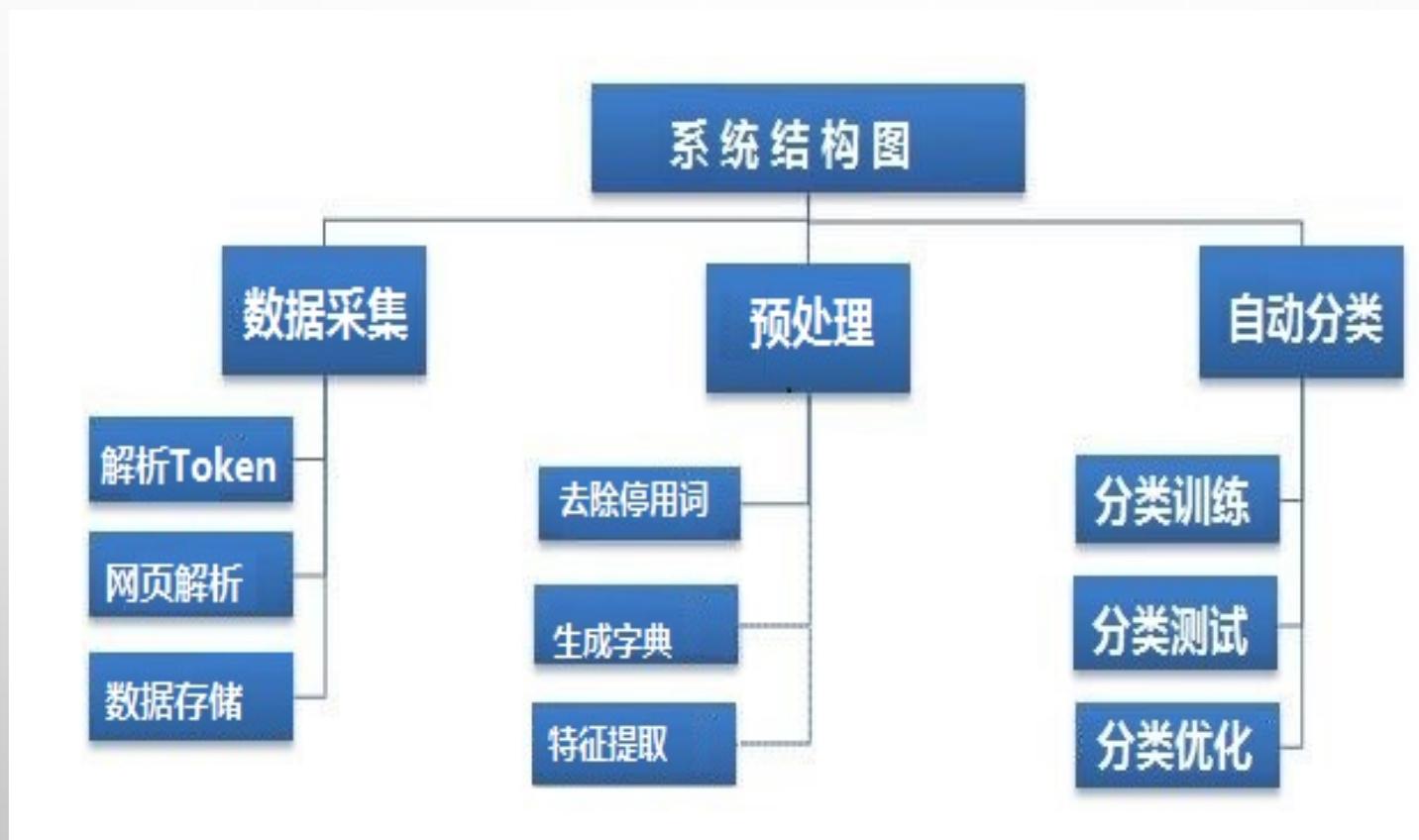
# 开发进度

- **第一阶段：2017年4月-5月**
  - 对各学科数据进行采集和处理，设计实现相应的科技文章规约，清理，抽取和变换；
- **第二阶段：2017年5月-7月**
  - 用大量的多学科文章进行分类模型的训练，并用多学科的文章对模型进行预测
- **第二阶段：2017年7月-10月**
  - 针对分类过程数据稀疏的特点进行分类模型的优化

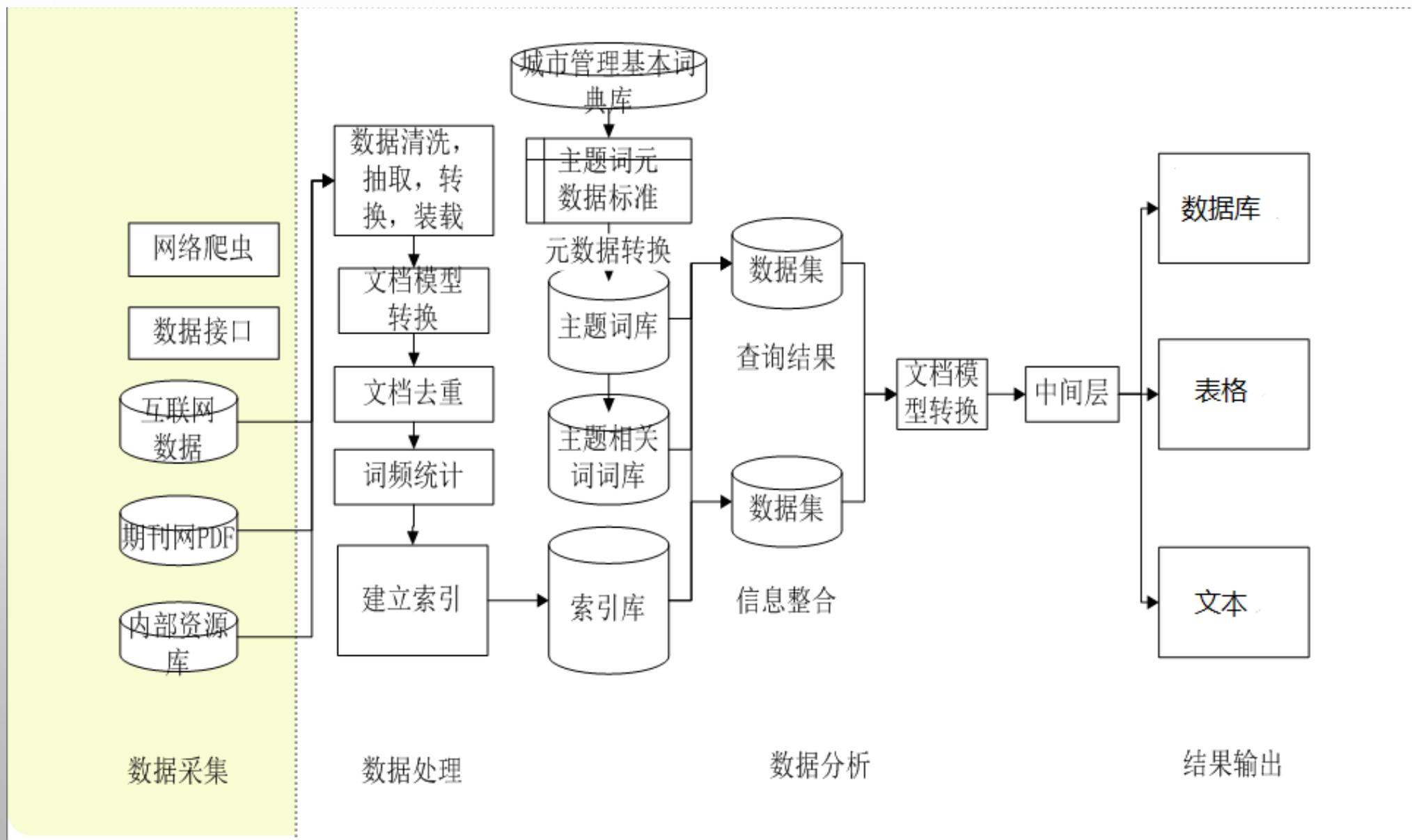
# 成果报告

- 通过引入网络爬虫、文本分类等技术，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” ——学科领域数据处理、学科分类模型生成、学科分类模型优化
- 实现了将学科自动分类，学科领域属性提取等新技术资源，融入到科技情报生产线分析中，提升了情报分析工具的效率。

# 学科领域属性抽取及分类架构

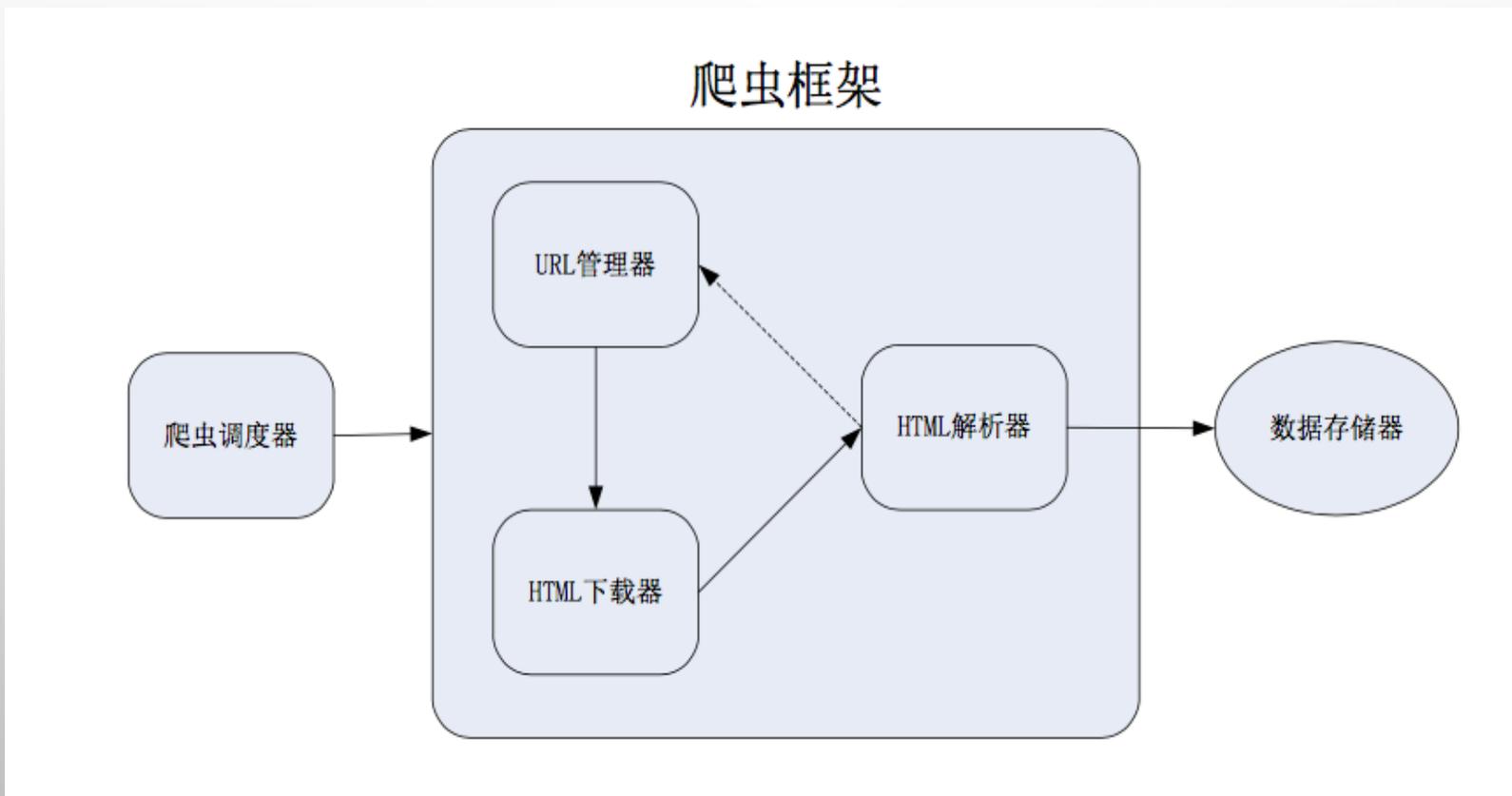


# 系统流程简介



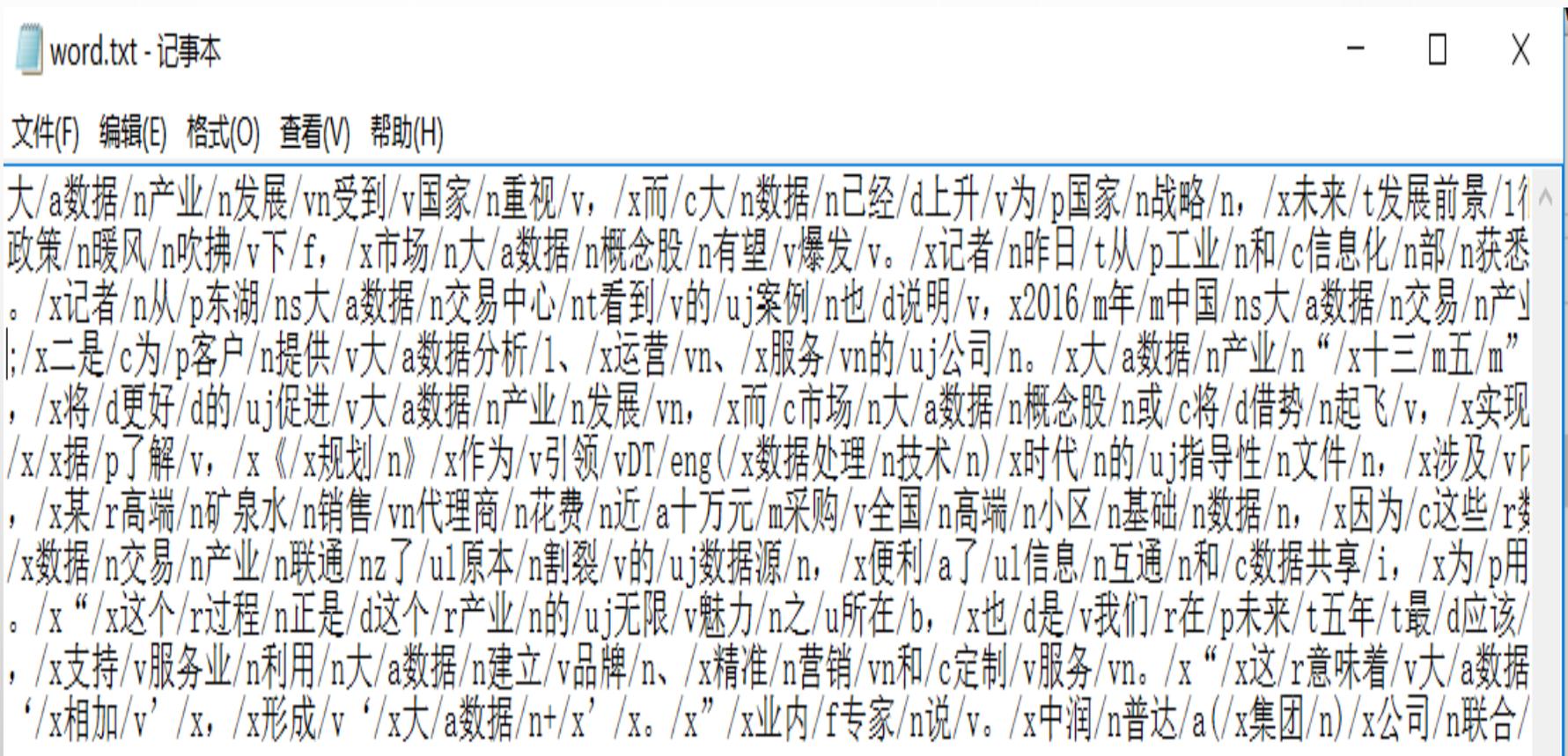
# 科技文章数据采集

爬虫框架主要包括五大模块，分别为爬虫调度器、URL管理器、HTML 下载器、HTML 解析器、数据存储器。



# 科技文章数据预处理

对科技文章进行切词，和词性标注方便后面进行特征提取



# 科技文章数据预处理

中文停用词表(比较全面\_有1208个停用词).txt - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

?  
人 民  
末 ## 末  
啊  
阿  
哎  
哎呀  
哎哟  
唉  
俺  
俺们  
按  
按照  
吧  
吧哒  
把  
罢了  
被  
本  
本着  
比  
比方  
比如  
鄙人  
彼  
彼此  
边  
别的  
别的说  
并且  
不比  
不成

对科技文章切分词后，中间夹杂着大量的无用词，指语气助词，代词等。所以设计了停用词表对这部分词进行剔除

# 科技文章词表生成和存储

科技文章的词表，统计了每个单词在每类中出现的词频，并且对这些单词都打上了标签，用以标注这些单词的学科

{ '语音增强': 1, '行业转型': 1, '软启动器': 2, 'Scholar': 1, '外来物种入侵': 2, '美国能源信息署': 1, '理论价值': 1, '复治肺结核': 1, '作文素材': 2, '性问题': 3, '果园': 16, '反季栽培': 1, '去骨瓣减': 1, 'RBM': 1, '羔羊培育': 1, 'S型柔性立管': 1, '手机新闻摄影': 1, '两腔振荡器': 1, '玉米覆膜': 1, '监狱民警': 1, '绰号': 2, '神经元': 1, '艺术市': 1, '干预对策': 1, '数控代码': 1, '艺术问题': 1, '大型发动机': 1, '产后恶露不绝': 1, '莲荷': 1, '城市生活垃圾': 13, 'Alamar': 1, '锻炼意向': 1, '对外汉语教': 2, '人文气息': 3, '公司总裁': 2, '四层次教学模式': 1, '系统管理总线(SMBu)': 2, '水通道蛋白2': 2, '题型特点': 1, '技术应用': 14, '应急平台': 1, '冻土': 4, '《中国社会科学》': 1, '美利曲辛': 2, '自动清洗系统': 1, '广告翻': 1, '次级轨道效应(SOI)': 1, '阴茎根部': 1, '技术40新': 1, '直播水稻': 3, '复方', '《煤矿安全规程》': 1, '门窗节能': 1, '水银体温计': 1, '夏季林火': 1, '学位论': 3, '思想政治理论课教': 5, '统计与概率': 3, '斜巷运输': 1, '“官二代”': 1, '变', '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', 'rtFoxServer': 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, '熵理论': 2, '泪囊鼻', '商贸物': 1, 'HF0-1234yf': 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, '病理类型': 1, '单体燃烧': 1, '基本公共': 6, '铅锌多金属矿': 1, '重藏少用': 1, '业务核算': 1, '混凝土施工及': 1, '小脑梗死': 1, '移行': 1, '东方航空': 1, '实', '性压缩复合材料': 1, '趋标': 1, '子宫颈病变': 3, '思想政治理论课': 39, '育成': 3, '霍城县': 2, '重油催化裂化装置': 1, '达里湖': 1, '文字录入': 1, '校长培训': 1, '生长结果表现': 1, '战略创新': 1, '不间断电源': 1, '阻碍': 4, '水烟': 1, '获取新知': 1, '护理交班': 1, '氯化氢': 1, '股骨近端骨折': 3, '胞外信号调节激酶': 1, '碍性疾病': 3, '党校图书馆': 5, '平衡带通滤波器': 1, '补肾强骨活血方': 1, '履带板': 1, 'bevacizumab': 1, '重庆电视台': 6, '住房抵押贷款': 3, '区域金融结构': 1, '芯片尺寸封装': 1, '软性隐形眼镜': 1, '设计色彩': 3, '铸造工艺': 7, '图书馆事业': 6, '内侧颞叶癫痫': 1, '超声导引': 1, '飞': 19, '船舶动力装置': 1, '聚光式太阳能', '梯度风': 1, '土壤数值化分类': 1, '管理性违章表现': 1, '园艺设施': 1, '简易诊断': 1, '金属粉末注射成形': 1, 'AVR单片机': 6, '慢性盆腔炎性疼痛': 1, '万托林': 1, 'SNCG': 1, '眼部症状': 1, '会东铅锌矿': 2, '审美联觉': 1, '汽轮机低压缸': 1, '汇流': 2, '二项风险模型': 1, '世界冠军': 7, '叶酸受体- $\alpha$  (FR- $\alpha$ ): 1, '单纯舒张期', '高中生物教学': 8, '国贸': 1, '双鸭山': 2, '航运': 2, '平面铣刀': 1, '风险认知': 2, '经济责任导向审计': 1, '旋转扁壳': 1, '脑萎缩': 1, '红薯': 6, '浅静脉炎', '促进因素': 1, '串口服务器': 1, '既有高速公路': 1, '湍流模型': 4, '过泵': 1, '景气循环': 1, '节目发展': 3, '公共音乐课程, 教学': 1, '种子量': 1, '发动机温度过', '电脑产品': 1, '垄断规': 1, '完备三部曲': 1, '合金组织': 1, '自动化设计': 1, '康普顿散': 1, '猪笼草': 1, '复合型绝热保温涂': 1, '婚检率': 1, 'IT基础设施': 1, '医院保险': 1, '车踵': 1, '三维固定式托槽': 1, '旅游地生命周期': 1, '免疫预防': 2, '西洋参根': 1, '周围性': 2, '操行评语': 1, '虚拟档案': 1, '高中一年级', '血源管理': 1, '侵犯行为': 1, '危机决策': 2, '改进措': 3, '血管球瘤': 2, '电影传播': 2, '约束柱': 1, '系统教学设计': 1, '甲骨文': 2, '胃十二指肠溃疡出': 率: 5, '管理岗位': 2, '多发': 7, '城乡协调': 2, '科学管理方法': 1, '缩微技术': 1, '南京临时参议院': 1, '主要途径': 1, '水上交通': 3, '内烯': 1, '《穆赫兰道》', '衣壳蛋白': 1, '难降解度': 1, '接受者': 1, '贯穿武案': 1, '算法仿': 1, '合金Mg-3.8Zn-2.8': 1, '斜飞模式': 1, '化学振荡反': 1, '新法饲养': 2, '分段卸压崩落法', '曲霉菌病': 1, '女子': 16, '财政教育支出': 2, '农机试验': 2, '尚力': 1, '理论课': 3, '网络模式': 1, '管路封堵器': 1, 'BFC-SIMPLE-VOF算': 1, '保护区划': 1, '花表', '截骨术': 1, '注疏模': 1, '中式卷烟': 2, '心脏电生理': 1, '底部结构': 1, '生态鸡': 1, '优良': 2, '纤溶活性蛋白': 1, '刀具热装机': 1, '云南歌': 1, '福利改革': 1,



# 分类模型的训练

因为训练出的模型为二进制的dat文件，展示出来也都是二进制格式。所以这里我们只放上模型训练的相关代码

```
# 导入训练集
trainpath = "train_word_bag/tfdifspace.dat"
train_set = _readbunchobj(trainpath)

# 导入测试集
testpath = "test_word_bag/testspace.dat"
test_set = _readbunchobj(testpath)

# 训练分类器：输入词袋向量和分类标签，alpha=0.001 alpha越小，迭代次数越多，精度越高
clf = MultinomialNB(alpha=0.001).fit(train_set.tdm, train_set.label)

# 预测分类结果
predicted = clf.predict(test_set.tdm)

for flabel, file_name, expct_cate in zip(test_set.label, test_set filenames, predicted):
    if flabel != expct_cate:
        print(file_name, ": 实际类别:", flabel, " --> 预测类别:", expct_cate)

print("预测完毕!!!")

# 计算分类精度：
from sklearn import metrics
def metrics_result(actual, predict):
    print('精度: {0:.3f}'.format(metrics.precision_score(actual, predict, average='weighted')))
    print('召回: {0:0.3f}'.format(metrics.recall_score(actual, predict, average='weighted')))
    #print('f1-score: {0:.3f}'.format(metrics.f1_score(actual, predict, average='weighted')))

metrics_result(test_set.label, predicted)
```

# 分类模型的测试

精度为 $TP / (FP + TP)$ ,其中TP表示正确分类的文章数, FP表示错分到其他类的文章数。而召回率为 $TP / (TP + FN)$ , FN表示本不属于该类却被分到该类的文章数

```
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
```

# 分类模型的优化

针对稀疏数据的优化处理，我们首先设立了阈值，词频数小于3的都不能加入到字典中。然后进行了稀疏矩阵的乘法优化

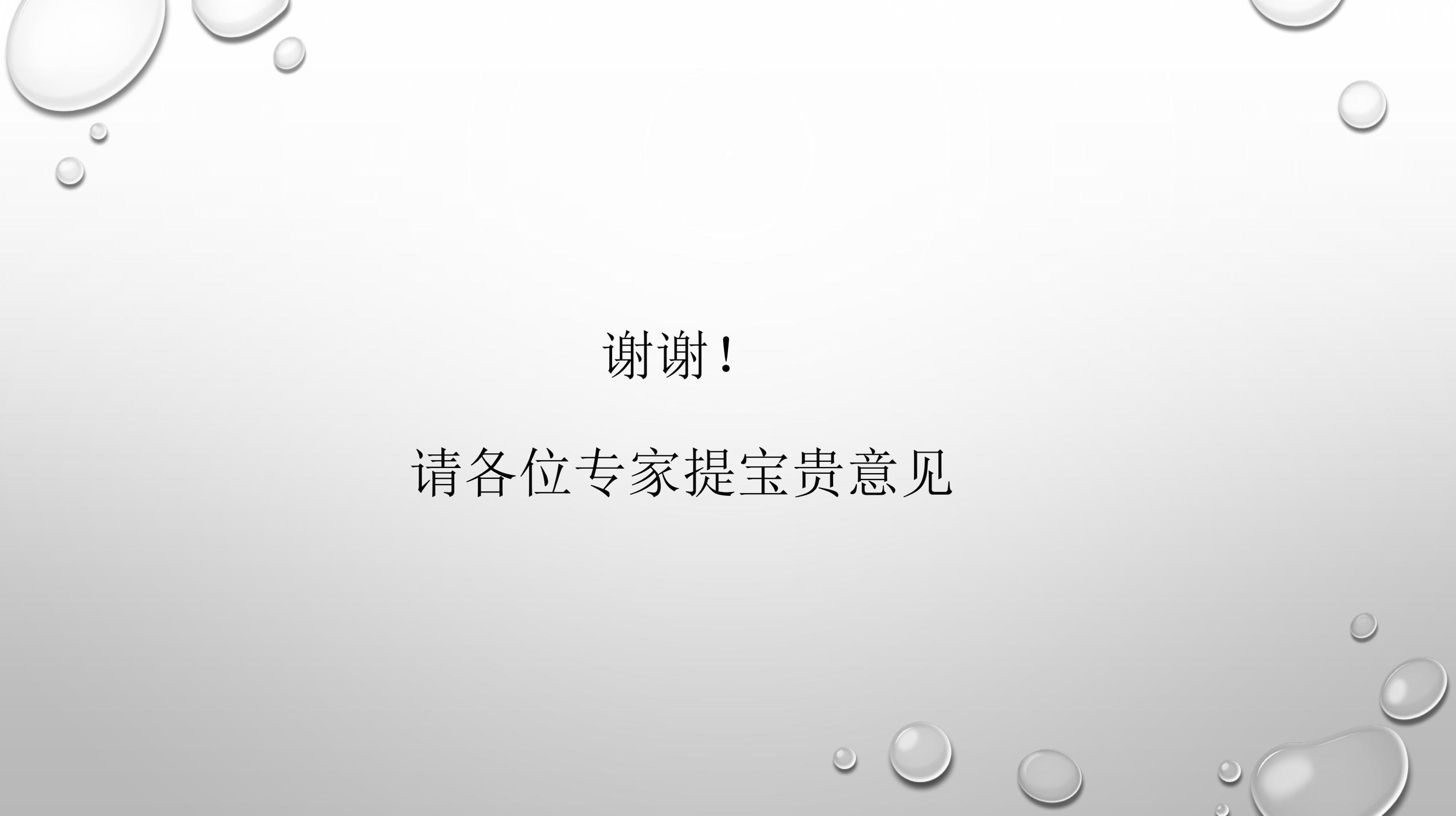
```
matrix multi(matrix a,matrix b)
{
    matrix c;
    memset(c.data,0,sizeof(c.data));
    for(int i=0; i<=n; i++)
    for(int j=0; j<=n; j++)
    {
        //稀疏矩阵的乘法优化
        if(a.data[i][j]) //一个数一个数加进去
        for(int k=0; k<=n; k++)
        //注意这里的ijk已经改变位置
        c.data[i][k]+=a.data[i][j]*b.data[j][k];
    }
    return c;
}

matrix init(matrix *a)
{
    memset((*a).data,0,sizeof((*a).data));
    for(int i=0;i<=n;i++)
    (*a).data[i][i]=1;
    //矩阵乘法的意义:
    //注意这里 (*a).data[n][n]=1; 他的意义是继承上次操作的值
    //(*a).data[i][j]=1;继承的是交换的值 两个值加起来就是新的值
    return *a;
}

matrix pow1(matrix a,ll b)
{
    matrix ans;
    init(&ans);
    while(b)
    {
        if(b&1)
        {
            ans=multi(ans,a);
        }
    }
}
```

# 考核指标完成情况

- 1) 学科领域属性数据处理（完成）：  
完成各学科数据的文章进切词，去除停用词，词性标注等处理，满足学科数据相关处理的要求。
- 2) 学科领域属性分类模型生成（完成）：  
分类模型效果良好，其中召回率达到94.5%，精度达到96.4%。
- 3) 学科领域属性分类模型优化（完成）：  
利用设定阈值和矩阵乘法优化原理完成了对稀疏数据的优化。

The background features a light gray gradient with several realistic water droplets of varying 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