

基于网页文本的分类

[实现新的网页的分类。支持交互式 **URL** 输入，或者输入一个文本，文本每行都是一个 **URL**，系统输出结果文本，每行对应输入文本的类别号。]

成员： 冯云 苏俊玮 苏莉娅

尹捷 张晓欧

指导老师：王斌

2016 年 12 月 26 日

说明文档

- 1. 项目概述及人员：3
 - 1.1 项目概述.....3
 - 1.2 人员.....3
- 2. 需求分析.....3
- 3. 系统设计.....4
 - 3.1 总体设计.....4
 - 3.2 系统功能模块.....5
- 4. 测试及分析.....17
 - 4.1 交叉验证的方法.....17
 - 4.2 交叉验证结果.....18
 - 4.3 结果分析.....18
 - 4.4 可优化地方.....19
- 5 结束语.....19

1. 项目概述及人员：

1.1 项目概述

实现一个分类系统。

分类体系为：财经(类别号：1)、科技(类别号：2)、汽车(类别号：3)、房产(类别号：4)、体育(类别号：5)、娱乐(类别号：6)、其它类(类别号：7)，利用网站的新闻主页(可以下载 Sogou 语料)，训练一个分类器(训练集合不能少于 5000 篇文档)。

结果：实现新的网页的分类。支持交互式 URL 输入，或者输入一个文本，文本每行都是一个 URL，系统输出结果文本，每行对应输入文本的类别号。

1.2 人员

姓名	学号	分工
冯云	201618018629135	爬虫编写
苏俊玮	2016E8018661180	界面编写
苏莉娅	201628018629065	分类器编写文档整理
尹捷	201628018627124	爬虫编写 ppt 整理
张晓欧	2016E8018661198	界面设计文档整理

2. 需求分析

(1) 功能需求：新闻或资讯展示平台需要将得到的最新资讯进行自动分类处理，以快速的发布更新资讯。

支持交互式提交新的网站链接，得到分类器分析结果。

(2) 性能需求：爬取数据要保证速度的同时不被封号；分类器分类结果的正确率要在 60%以上为优。

3. 系统设计

3.1 总体设计

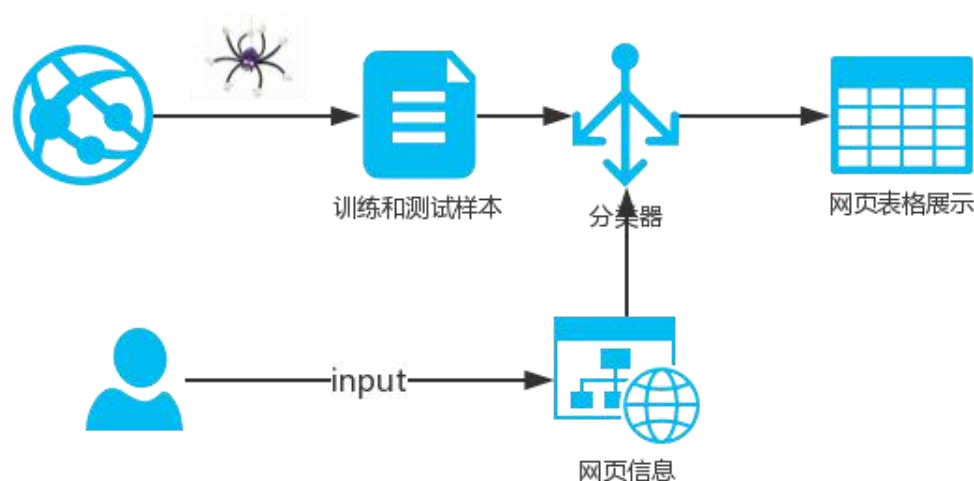


图 1：系统总设计图

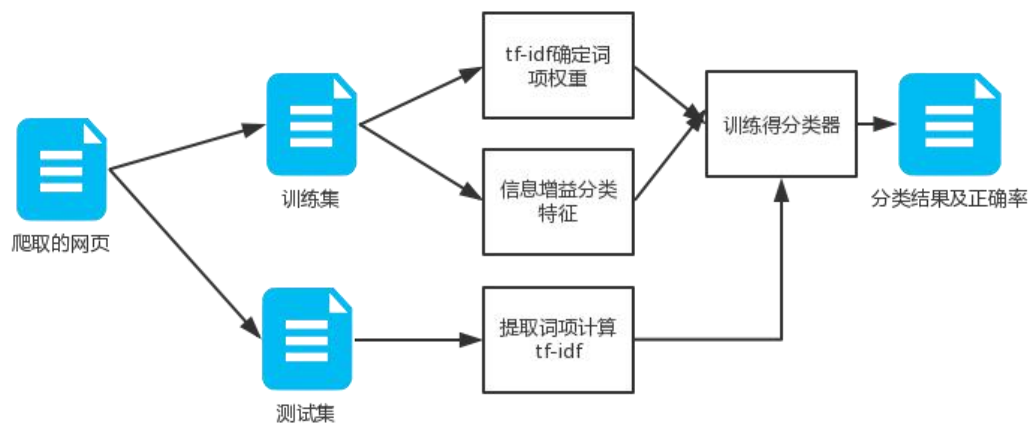


图 2：分类器设计图

系统分为生成分类器和结果测试两个部分。

生成分类器包括：训练和测试两个部分。训练部分采用的主要方法为 KNN 和贝叶斯两个方法，使用信息增益为特征，根据 tf-idf 设置权重。

结果测试部分通过网页与用户交互，用户在网页上提交网页链接，系统自动调用分类器，通过条形图可视化的展示分类的效果。

3.2 系统功能模块

3.2.1 爬取部分

1、爬虫原理：

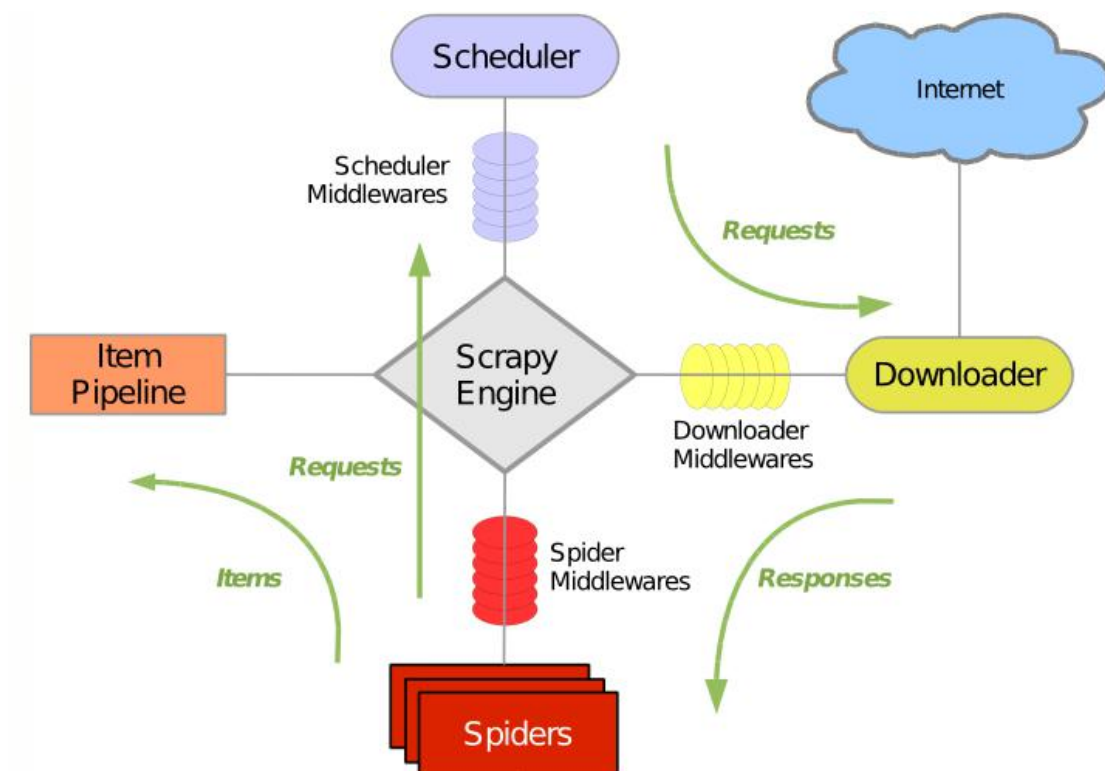


图 3：Scrapy 整体架构

首先获取第一个 URL 的初始请求，当请求返回后调取一个回调函数。第一个请求是通过调用 `start_requests()` 方法。该方法默认从 `start_urls` 中的 Url 中生成请求，并执行解析来调用回调函数。

在回调函数中，你可以解析网页响应并返回项目对象和请求对象或两者的迭代。这些请求也将包含一个回调，然后被 Scrapy 下载，然后有指定的回调处理。

在回调函数中，你解析网站的内容，同程使用的是 Xpath 选择器（但是你也可以使用 BeautifulSoup, lxml 或其他任何你喜欢的程序），并生成解析的数据项。

最后，从蜘蛛返回的项目通常会进驻到项目管道。

2、具体实现：

在新闻爬取部分，由于不同类别的新闻来源有所不同，网页数据提取方式不同，因此对每种类别的新闻分别进行爬取。

利用 Scrapy 框架，首先在程序中定义目标新闻列表页面，用 XPath 进行网页数据提取，从初始页面爬取单条新闻子链接，然后对每个新闻子链接进行爬取，获得新闻标题及内容，按照{类别号，id 号，标题，URL，内容}的格式储存为 JSON 文件。

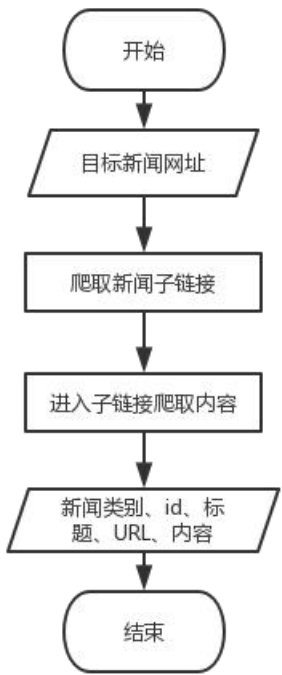


图 4：爬取的流程如下图

```
财经 "http://203.192.8.57/was5/web/search?channelid=214510&prepage=1000&searchword=extend5%3D%27%2511105289%25%27"
科技 "http://203.192.8.57/was5/web/search?channelid=234968&searchword=extend5%3D%27%2511109303%25%27&prepage=1000&list=&page=1"
汽车 "http://203.192.8.57/was5/web/search?channelid=276589&prepage=1000&searchword=extend5%3D%27%2511109357%25%27"
房产 "http://www.chinanews.com/house/gd.shtml" "http://fdc.fang.com/news/more/11806/1.html"
体育 "http://www.bjnews.com.cn/sport/list-28-page-1.html"
娱乐 "http://www.shcaon.com/wy/"
其他 "http://203.192.8.57/was5/web/search?channelid=276589&searchword=extend5%3D%27%2511109449%25%27&prepage=1000&list=&page=1"
```

图 5：新闻爬取地址

```
C:\Windows\system32\cmd.exe
age-4.html>
2016-12-25 09:59:54 [scrapy] DEBUG: Crawled (200) <GET http://www.bjnews.com.cn/sport/2016/12/08/426383.html> (referer: http://www.bjnews.com.cn/sport/list-28-page-4.html)
2016-12-25 09:59:54 [scrapy] DEBUG: Crawled (200) <GET http://www.bjnews.com.cn/sport/2016/12/09/426533.html> (referer: http://www.bjnews.com.cn/sport/list-28-page-4.html)
2016-12-25 09:59:54 [scrapy] DEBUG: Crawled (200) <GET http://www.bjnews.com.cn/sport/2016/12/09/426544.html> (referer: http://www.bjnews.com.cn/sport/list-28-page-4.html)
http://www.bjnews.com.cn/sport/2016/11/02/421899.html
http://www.bjnews.com.cn/sport/2016/11/01/421835.html
http://www.bjnews.com.cn/sport/2016/10/25/420975.html
http://www.bjnews.com.cn/sport/2016/09/21/417544.html
http://www.bjnews.com.cn/sport/2016/09/20/417478.html
http://www.bjnews.com.cn/sport/2016/09/20/417408.html
http://www.bjnews.com.cn/sport/2016/09/20/417401.html
http://www.bjnews.com.cn/sport/2016/09/18/417155.html
http://www.bjnews.com.cn/sport/2016/09/18/417154.html
http://www.bjnews.com.cn/sport/2016/09/30/418662.html
http://www.bjnews.com.cn/sport/2016/09/29/418611.html
http://www.bjnews.com.cn/sport/2016/09/29/418572.html
http://www.bjnews.com.cn/sport/2016/09/05/415910.html
http://www.bjnews.com.cn/sport/2016/09/05/415817.html
http://www.bjnews.com.cn/sport/2016/09/05/415816.html
http://www.bjnews.com.cn/sport/2016/09/04/415771.html
http://www.bjnews.com.cn/sport/2016/08/26/414808.html
半:
```

图 6：运行示意图



图 7：得到的 JSON 文件

共爬取财经类新闻 992 条，科技类新闻 999 条，汽车类新闻 330 条，房产类新闻 584 条，体育类新闻 1000 条，娱乐类新闻 485 条，其他类新闻 1000 条。

然后，为方便后续分类器的训练，编写 j2e.py 脚本将 JSON 文件内容储存在 Excel 表格中，每个类别的新闻分别放在一个标签页中，得到 news.xlsx，如下图所示：

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	
1	aid	title	link	desc														
2		1 胡润全球华人富豪榜：王健林	http://news.xinhuanet.com/foi	胡润研究院19日发布《2015新资本 胡润全球华人富豪榜》，李嘉诚20年来首次被大陆富豪超越，王健林以2600亿的财富成为全球华人新首富。紧随其后，马														
3		2 华彬信托拟打 信托证券	http://news.xinhuanet.com/foi	最近，信托公司证券投资部门的员工闲了许多。 证券时报记者采访的多位信托公司相关业务人士表示，在监管层态度核查华彬信托及场外配资之后，目前														
4		3 论牌商得股权转让竞拍结束	http://news.xinhuanet.com/foi	论牌商得19日上午在西南联合产权交易所结束网络竞价，最终成交价定格为23.51元/股，已有买家成功摘牌。至此，这件好事多磨的股权转让案终于尘埃落定，														
5		4 美国成亚洲海外房地产投资	http://news.xinhuanet.com/foi	据海外媒体报道，世邦魏理仕表示，美国取代英国成为亚洲房地产投资者的首选国家，今年上半年美国吸引到的房地产投资规模为61亿美元，英国为44亿美元。														
6		5 90后网络创业先有“钱”不	http://news.xinhuanet.com/foi	在国家政策层面鼓励万众创新、大众创业的背景下，如今在中国创业的门槛已经非常低了。即便是对于众多的90后的“小鲜肉”而言，只要有一个好点子，也很容														
7		6 偏于在广州火车站售票	http://news.xinhuanet.com/foi	火车站售票员有防骗告示，其中就是提醒乘客提防工作人员。可人多时，乘客很谨小慎微到这些告示。 火车站地站站务员团队提醒：卖高价“黄牛”其														
8		7 二手车买卖不过户 推利	http://news.xinhuanet.com/foi	记者近日在深圳、上海、北京、厦门等地采访发现，二季度以来，国家和地方一系列政策效应显现，楼市回暖，车价亦升，但一些地方出现房价涨幅过快，														
9		8 增加直接税 财税改革成	http://news.xinhuanet.com/foi	《经济参考报》记者从多位权威专家处了解到，财税改革有望成为“十三五”规划的重点之一。“十三五”期间将基本完成十八届三中全会制定的财税改革目标，														
10		9 工企证金“双金”会联手	http://news.xinhuanet.com/foi	上市公司日前密集披露了“国家队”成员中国证券金融股份有限公司（以下简称“证金”）和中央汇金投资有限公司（以下简称“汇金”）的特权情况，其中汇金														
11		10 新浪微博同比增长广告	http://news.xinhuanet.com/foi	8月19日，新浪微博公布了截至2015年6月30日的第二季度未经审计的财务报告。报告显示，归属于微博普通股股东的净收益为420万美元，去年同期为净亏														
12		11 海南椰岛涉嫌违法产品启动	http://news.xinhuanet.com/foi	昨日，海南椰岛集团股份有限公司发布澄清事项公告，针对国家食药监总局日前通告，51家企业在69种保健食品、配制酒中违法添加了西地那非（俗称“伟哥”）等														
13		12 同花顺、安硕信息面临暂停	http://news.xinhuanet.com/foi	同花顺昨日晚间公告称，公司因涉嫌违反证券期货法律法规，目前正在被中国证监会立案调查。根据有关规定，如公司存在或涉嫌存在重大违反证券法律法规														
14		13 央行向14家金融机构注入	http://news.xinhuanet.com/foi	昨天傍晚，央行官方微博发布信息证实，当日对14家金融机构开展中期借贷便利操作（MLF）1100亿元，期限6个月，利率3.35%。 事实上，近期市场利率持														
15		14 新三板监管升级28家公司	http://news.xinhuanet.com/foi	持续火爆的新三板监管再度升级。昨天，针对2014年年报逾期财务报表附注、违规披露2015年一季度财务报告、股票发行程序违规等行为，全国股转公司对														
16		15 以房养老“蛋糕”为何碎	http://news.xinhuanet.com/foi	将房产抵押给保险公司，每月领取“养老金”，将老时房产归保险公司处置，你愿不愿意？ 2014年7月1日起，中国平安会在北京、上海、广州、武														
17		16 大宗交易溢价折价“冰火两重天”	http://news.xinhuanet.com/foi	在近期成交量骤减的大宗交易市场上，一些比较热门的股票被溢价交易，一些相对冷门的股票反而被贱卖交易。 统计显示，6月大宗交易平台发生208笔交易														
18		17 人民币汇率四个交易日连续	http://news.xinhuanet.com/foi	在维持连续三个交易日的贬值之后，人民币汇率近日持续回调，决策层安抚市场的策略暂时取得成功。 中国人民银行外汇交易中心公布，8月19日拆														
19		18 百度基金入驻客CEO否认	http://news.xinhuanet.com/foi	百度集团日前宣布，任命60岁的Micky Pant出任百胜中国事业部CEO，接替原中国事业部主师苏敬斌的位置。这也是百胜中国成立以来首任换帅。 63岁														
20		19 同花顺、恒生电子涉嫌违法	http://news.xinhuanet.com/foi	8月18日晚间，同花顺、恒生电子均发布公告称，近日收到证监会《调查通知书》。《通知书》称，因涉嫌违反证券、期货法律法规，根据《中华人民共和国														
21		20 携程独大并非终局 OTA是	http://news.xinhuanet.com/foi	互联网行业向来就是巨头的游戏，实物类电商风起云涌，服务类电商也战战兢兢。在旅游产业链内，携程收购艺龙后，去哪儿指其垄断市场，去哪儿也公然叫														
22		21 新三板市场分层方案不会	http://news.xinhuanet.com/foi	8月18日，《证券日报》记者从全国中小企业股份转让系统（俗称新三板）内部权威人士处独家获悉，新三板市场分层方案正在推进中，待成熟后向市场公														
23		22 名牌一时奢侈品品牌PRADA	http://news.xinhuanet.com/foi	在整个奢侈品行业都在为了对抗颓势而各自招徕之时，PRADA显得格外低调，业绩改善之迟缓。近日公布的上半年财报中，收益净额仅增4%。一系列指标														
24		23 内部电话使用过热 苹果	http://news.xinhuanet.com/foi	昨日，记者从国家食药监总局获悉，苹果公司已向国家食药总局提交了召回计划，并将于8月19日起召回2014年11月生产的iPhone6s和iPhone6s Plus两款手机														
25		24 工行回应“降薪”：上半	http://news.xinhuanet.com/foi	针对日前有媒体报道的四大行“降薪”、员工工资砍近半等，工商银行相关负责人昨天回应表示，上半年工资总额同上年基本持平，并														
26		25 资本外流压力增大 央行	http://news.xinhuanet.com/foi	在资本市场近期流动性开始趋紧之时，央行昨日紧急进行了1200亿元逆回购操作缓解市场情绪，创下自2014年1月以来单日最大操作量。分析人士表示，央行通														
27		26 小贷行业“没得选” 要	http://news.xinhuanet.com/foi	“宽带提速降费”，今年以来一直都是坊间热门话题。无论是国家层面方兴未艾的“互联网+”战略实施，还是百姓需求日益强烈的信息消费，都离不开宽带网络														
28		27 网贷“单双号限行”航	http://news.xinhuanet.com/foi	明天起，北京市实施单双号限行临时措施，交管部门将最大限度将警力投入路面，严查违反单双号限行。同时，全市道路上的高清摄像头也将调整记录违法														
29		28 网贷模式 京东巨头	http://news.xinhuanet.com/foi	京东巨头电商爆发至今，依然处于“半耕半读”，自给自足“状态”。从诞生开始，京东商城“黑金”业务如影随形，但这并不影响巨头们一轮又一轮的扩张市场。														

图 8： excel 表格存储结果

3.2.2 分类器部分

1、分类原理：

kNN 的分类原理：

k 近邻法(k-nearest neighbor. k-NNs)是一种基本分类与回归方法，k 近邻法的输入为实例的特征向量，对应于特征空间的点；输出为实例的类别，可以取多类。k 近邻法假设给定一个训练数据集，其中的实例类别已定。分类时，对新的实例，根据其 k 个最近邻的训练实例的类别通过多数表决等方式进行预测。因此，k 近邻法不具有显式的学习过程。k 近邻法实际上利用训练数据集对特征向量空间进行划分，并作为其分类的“模型”。k 值的选择、距离度量及分类决策规则是 k 近邻法的三个基本要素。

k 值的选择会对 k 近邻法的结果产生重大影响。k 值的减小就意味着整体模型变得复杂，容易发生过拟合；如果选择较大的 k 值，就相当于用较大邻域中的训练实例进行预测，其优点是减少学习的估计误差。但缺点是学习的近似误差会增大。在应用中，k 值一般取一个比较小的数值。通常采用交叉验证法来选取最优的 k 值。

特征空间中两个实例点的距离是两个实例点相似程度的反映。k 近邻模型的特征空间一般是 n 维实数向量空间 R^n 。本次我们的程序中使用的是欧氏距离。

k 近邻法中的分类决策规则往往是多数表决，即由输入实例的 k 个邻近的训练实例中的多数类决定输入实例的类。

贝叶斯分类器原理：

朴素贝叶斯(native Bayes)法是基于贝叶斯定理与特征条件独立假设的分类方法。对于给定的训练数据集，首先基于特征条件独立假设学习输入/输出的联合概率分布；然后基于此模型，对给定的输入 x ，利用贝叶斯定理求出后验概率最大的输出 y 。朴素贝叶斯法 实现简单，学习与预测的效率都很高，是一种常用的方法。

利用训练数据学习 $P(X|Y)$ 和 $P(Y)$ 的估计，得到联合概率分布：

$$P(X, Y) = P(Y)P(X|Y)$$

概率估计方法可以是极大似然估计或贝叶斯估计。

朴素贝叶斯法的基本假设是条件独立性，这是一个较强的假设。由于这一假设，模型包含的条件概率的数量大为减少，朴素贝叶斯法的学习与预测大为简化。因而朴素贝叶斯法高效，且易于实现。其缺点是分类的性能不一定很高。

2、具体实现：

网页分类主要分为两部分实现：分类器构建和网页分类。在分类器构建中，首先将爬取到的数据进行分词，然后计算词的信息增益选择特征词，将这些特征词用两种分类方式进行训练（kNN 和贝叶斯），得到两种分类方法的分类器；在网页分类中，将网页的文本首先进行分词，计算 tf-idf 值，选择文本的“关键词”，使用这些词在分类器中的信息增益和词在待查文档中的权重来测试判定网页类别。

1) 分类器构建



图 9：分类器的构建流程图

2) 文本 jieba 分词

首先将得到的新闻文本字符串进行分词，使用 **jieba** 分词工具。本次用到了 **jieba** 分词中适合文本分析的精确模式。在分词的过程中，我们删除了单字词语和标点符号。因为单字词语通常是无意义的“的”、“地”等词语，标点符号并不能做为文本分类的根据。通过分词，我们得到分好词的文本列表 **wordlist**。

3) 计算词项信息增益

计算上步得出的分词的信息增益，首先计算局部类别文档频率 **mdf**（分词后的词项在该类别中的文档频率）和全局文档频率 **ndf**（分词后的词项在全局文档中的概率），词项 *i* 的信息增益计算如下：

$$\text{Gain}(i) = P_1(pi1t1 * \log pi1t1 + pi0t1 * \log pi0t1) \\ + P_0(pi1t0 * \log pi1t0 + pi0t0 * \log pi0t0)$$

其中，参数分别表示： P_1 ：词项在所有文档中出现的概率。

P_0 ：词项在所有文档中未出现的概率。

$pi1t1$ ：词项在该类别文档中出现的频率。

$pi0t1$ ：词项在其他类别文档中出现的频率。

$pi1t0$ ：词项未在该类别文档中出现的频率。

$pi0t0$ ：词项未在其他类别文档中出现的频率。

筛选特征词

将计算得到的词项的信息增益进行比较，选择信息增益较高的词作为特征词。

4) 网页分类



图 10: 网页分类的流程图

- (1) 将网页内容进行分词
- (2) 确定文本的特征词得其权重
- (3) 计算 **tf-idf**

首先将得到分词之后的文本字符串的向量化表示，进行归一化处理，得到词项频率 **tf** 的字典形式（一篇文档对应一个字典）；

然后根据类别文档频率 **mdf** 和全局文档频率 **ndf** 计算每一类文档中各个词条的改进的 **idf** 权重：

$$\text{idf} = \log \left(N * \frac{m}{n} * \frac{m}{n - m + 1} \right),$$

得到每个元素为该类别中的词条的权重列表，按权重由大到小排列，元素为元组，元组的第一个元素单词，第二个元素表示该单词在该类别中的权重；

最后根据上两步得到的 **tf**、**idf** 计算每篇文档的 **tf-idf** 权重向量。

$$\text{tf-idf} = \text{tf} * \text{idf}$$

(4) 计算特征词的信息增益

从上一步可以看出词的权重，从而提取特征词，计算这些特征词在文本中的信息增益。

(5) 用分类器判断网页类别

将上一步的特征词的信息增益带入 kNN 或贝叶斯的分类器中，得到网页类别。

3、实现中遇到的问题

1) 分类器的分类方法选择

开始试图尝试了决策树和神经网络反馈方法进行分类，但是觉得效率不高而且分类效果并不好，最后选择了 kNN 和贝叶斯两种分类方法。

2) 特征词提取的数目

特征词提取的数目是直接影响正确率的，如果特征词提取过少，可能分类的特征就不会很清晰，如果特征词提取过多，可能会出现分类结果与测试集过分拟合的问题，不符合实际情况。最终我们选取 25 个特征词来代表待分类文本。

4、优化方法

1) 使用多种分类方法交叉对比

可以使用多种分类方法交叉对比构建分类器，将分类器构建的分类结果进行比对，分类结果的正确性会提高。

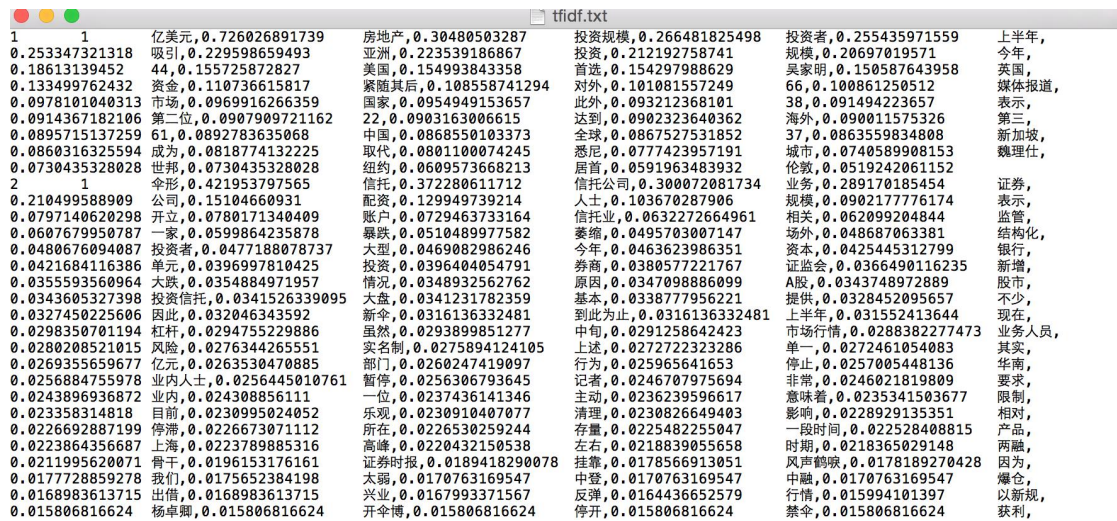
2) 实现 k 近邻法时，主要考虑的问题是如何对训练数据进行快速 k 近邻搜索。这点在特征空间的维数大及训练数据容量大时尤其必要。当训练集很大时，计算非常耗时，这种方法是不可行的。为了提高 k 近邻搜索的效率，可以考虑使用特殊的结构存储训练数据，以减少计算距离的次数。

3.2.3 功能集成

1) 通过数据爬取模块获取训练集，包括新闻的链接、标题和内容，详见爬取模块。

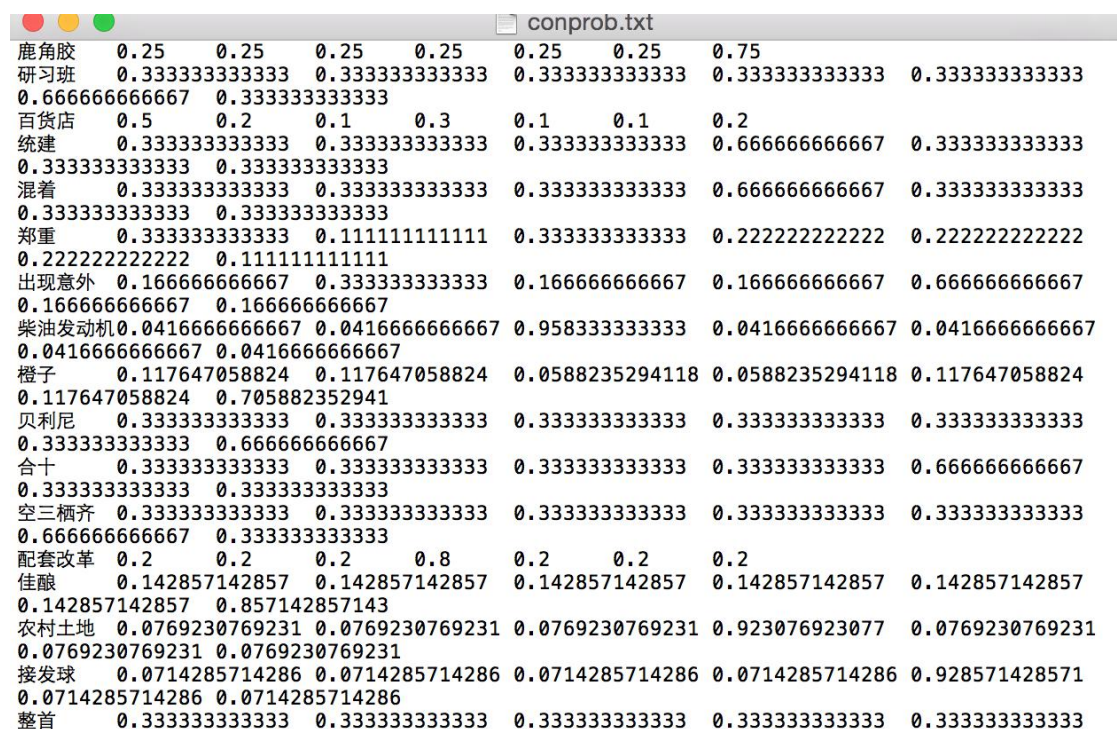
2) 对爬取的数据进行分析。训练集使用结巴分词 (jieba) 插件对文本分词，去除符号和长度为 1 的词。(featureextract.py.tokenext)；然后计算每个文档中每个单词的词项出现的次数 tf；计算文档集合里的 df、idf，这里使用改进算法，

$idf = \log\{N * (m/n) * (m/(n-m+1))\}$; 计算 $td-idf$, 并存储成为 $td-idf.txt$; 计算信息增益 ig 和条件概率, 并存储为 $conprob.txt$



1	1	亿美元, 0.726026891739	房地产, 0.30480503287	投资规模, 0.266481825498	投资者, 0.255435971559	上半年,
0.253347321318	吸引, 0.229598659493	亚洲, 0.223539186867	投资, 0.212192758741	规模, 0.20697019571	今年,	今年,
0.18613139452	44, 0.155725872827	美国, 0.154993843358	首违, 0.154297988629	吴家明, 0.150587643958	英国,	英国,
0.133499762432	资金, 0.110736615817	紧随其后, 0.108558741294	对外, 0.101081557249	66, 0.100861250512	媒体报道,	媒体报道,
0.0978101040313	市场, 0.0969916266359	国家, 0.0954949153657	此外, 0.093212368101	38, 0.091494223657	表示,	表示,
0.0914367182106	第二位, 0.0907909721162	22, 0.0903163006615	达到, 0.0902323640362	海外, 0.090011575326	第三,	第三,
0.0895715137259	61, 0.0892783635068	中国, 0.0868550103373	全球, 0.0867527531852	37, 0.0863559834808	新加坡,	新加坡,
0.0860316325594	成为, 0.0818774132225	取代, 0.0801100074245	悉尼, 0.0777423957191	城市, 0.0740589908153	魏理仕,	魏理仕,
0.0730435328028	世邦, 0.0730435328028	纽约, 0.0609573668213	居首, 0.0591963483932	伦敦, 0.0519242061152		
2	1	信托, 0.421953797565	信托公司, 0.300072081734	业务, 0.289170185454	证券,	证券,
0.210499588909	公司, 0.15104660931	配资, 0.129949739214	人士, 0.103670287906	规模, 0.0902177776174	表示,	表示,
0.0797140620298	开立, 0.0780171340409	账户, 0.0729463733164	信托业, 0.0632272664961	相关, 0.062099204844	监管,	监管,
0.0607679950787	一家, 0.0599864235878	暴跌, 0.0510489977582	萎缩, 0.0495703007147	场外, 0.048687063381	结构化,	结构化,
0.0480676094087	投资者, 0.0477188078737	大型, 0.0469082986246	今年, 0.0463623986351	资本, 0.042545312799	银行,	银行,
0.0421684116386	单元, 0.0396997810425	投资, 0.0396404054791	券商, 0.0380577221767	证监会, 0.0366490116235	新增,	新增,
0.0355593560964	大跌, 0.0354884971957	情况, 0.0348932562762	原因, 0.0347098886099	A股, 0.0343748972889	股市,	股市,
0.0343605327398	投资信托, 0.0341526339095	大盘, 0.0341231782359	基本, 0.0338777956221	提供, 0.0328452095657	不少,	不少,
0.0327450225606	目, 0.032046343592	新伞, 0.0316136332481	到此为止, 0.0316136332481	上半年, 0.031552413644	现在,	现在,
0.0298350701194	杠杆, 0.0294755229886	虽然, 0.0293899851277	中旬, 0.0291258642423	市场行情, 0.0288382277473	业务人员,	业务人员,
0.0280208521015	风险, 0.0276344265551	实名制, 0.0275894124105	上述, 0.027272323286	单一, 0.0272461054083	其实,	其实,
0.0269355659677	亿元, 0.0263530470885	部门, 0.0260247419097	行为, 0.025965641653	停止, 0.0257005448136	华南,	华南,
0.0256884755978	业内人士, 0.0256445010761	暂停, 0.0256306793645	记者, 0.0246707975694	非常, 0.0246021819809	要求,	要求,
0.0243896936872	业内, 0.0243088856111	一位, 0.0237436141346	主动, 0.0236239596617	意味着, 0.0235341503677	限制,	限制,
0.023358314818	目前, 0.0230995024052	乐观, 0.0230910407077	清理, 0.0230826649403	影响, 0.0228929135351	相对,	相对,
0.0226692887199	停滞, 0.0226673071112	所在, 0.0226530259244	存量, 0.0225482255047	一段时间, 0.022528408815	产品,	产品,
0.0223864356687	上海, 0.0223789885316	高峰, 0.0220432158538	左右, 0.0218839055658	时期, 0.0218365029148	两融,	两融,
0.0211995620071	骨干, 0.0196153176161	证券时报, 0.0189412890078	挂帐, 0.0178866913051	风声鹤唳, 0.0178189270428	因为,	因为,
0.0177728859278	我们, 0.0175652384198	太弱, 0.0170763169547	中登, 0.0170763169547	中融, 0.0170763169547	爆仓,	爆仓,
0.0168983613715	出借, 0.0168983613715	兴业, 0.0167993371567	反弹, 0.0164436652579	行情, 0.015994101397	以新规,	以新规,
0.015806816624	杨丰脚, 0.015806816624	开伞博, 0.015806816624	停开, 0.015806816624	禁单, 0.015806816624	获利,	获利,

图 11: $td-idf.txt$



鹿角胶	0.25	0.25	0.25	0.25	0.25	0.75
研习班	0.333333333333	0.333333333333	0.333333333333	0.333333333333	0.333333333333	0.333333333333
0.666666666667	0.333333333333	0.333333333333	0.333333333333	0.333333333333	0.333333333333	0.333333333333
百货店	0.5	0.2	0.1	0.3	0.1	0.2
统建	0.333333333333	0.333333333333	0.333333333333	0.333333333333	0.666666666667	0.333333333333
0.333333333333	0.333333333333	0.333333333333	0.333333333333	0.333333333333	0.666666666667	0.333333333333
混着	0.333333333333	0.333333333333	0.333333333333	0.333333333333	0.666666666667	0.333333333333
0.333333333333	0.333333333333	0.333333333333	0.333333333333	0.333333333333	0.666666666667	0.333333333333
郑重	0.333333333333	0.111111111111	0.111111111111	0.333333333333	0.222222222222	0.222222222222
0.222222222222	0.111111111111	0.111111111111	0.111111111111	0.333333333333	0.222222222222	0.222222222222
出现意外	0.166666666667	0.333333333333	0.333333333333	0.166666666667	0.166666666667	0.666666666667
0.166666666667	0.166666666667	0.166666666667	0.166666666667	0.166666666667	0.166666666667	0.666666666667
柴油发动机	0.0416666666667	0.0416666666667	0.0416666666667	0.958333333333	0.0416666666667	0.0416666666667
0.0416666666667	0.0416666666667	0.0416666666667	0.0416666666667	0.958333333333	0.0416666666667	0.0416666666667
橙子	0.117647058824	0.117647058824	0.117647058824	0.0588235294118	0.0588235294118	0.117647058824
0.117647058824	0.0588235294118	0.0588235294118	0.0588235294118	0.0588235294118	0.0588235294118	0.117647058824
贝利尼	0.333333333333	0.333333333333	0.333333333333	0.333333333333	0.333333333333	0.333333333333
0.333333333333	0.666666666667	0.666666666667	0.666666666667	0.333333333333	0.333333333333	0.333333333333
合十	0.333333333333	0.333333333333	0.333333333333	0.333333333333	0.333333333333	0.666666666667
0.333333333333	0.333333333333	0.333333333333	0.333333333333	0.333333333333	0.333333333333	0.666666666667
空三栖齐	0.333333333333	0.333333333333	0.333333333333	0.333333333333	0.333333333333	0.333333333333
0.666666666667	0.333333333333	0.333333333333	0.333333333333	0.333333333333	0.333333333333	0.333333333333
配套改革	0.2	0.2	0.2	0.8	0.2	0.2
0.142857142857	0.142857142857	0.142857142857	0.142857142857	0.142857142857	0.142857142857	0.142857142857
0.142857142857	0.857142857143	0.857142857143	0.857142857143	0.857142857143	0.857142857143	0.857142857143
农村土地	0.0769230769231	0.0769230769231	0.0769230769231	0.0769230769231	0.923076923077	0.0769230769231
0.0769230769231	0.0769230769231	0.0769230769231	0.0769230769231	0.0769230769231	0.923076923077	0.0769230769231
接发球	0.0714285714286	0.0714285714286	0.0714285714286	0.0714285714286	0.0714285714286	0.928571428571
0.0714285714286	0.0714285714286	0.0714285714286	0.0714285714286	0.0714285714286	0.0714285714286	0.928571428571
整首	0.333333333333	0.333333333333	0.333333333333	0.333333333333	0.333333333333	0.333333333333

图 12: $conprob.txt$

3) 我们选用 `django` 框架来处理 `web` 数据并提供分类视图显示。访问 `127.0.0.1:8000` 显示 `index` 页面。输入待测 `URL`, 使用 `urllib2.urlopen(链接).read()` 读取网页的内容。并使用 `bs4` 插件中的 `BeautifulSoup` 类读取网页中的文本内容。

4) 调用分类器模块中 `kNNtestonefile` 和 `bayestestonefile` 函数对文本内容进行分类处理, 并将选取最终排名结果赋予变量 `d_result_knn` 和 `d_result_bayes`

5) 将分类后得到的各类别的得分即 `prekind_knn` 和 `prekind_bayes` 二维数组中的值取出来存成字典, 并排序。最终, 将给类别的隶属度得分以及分类结果显示到 `result.html` 页面中。

```
d.url_request.GET['url']
d.content = urllib2.urlopen(d.url).read()
soup = BeautifulSoup(d.content)
d.text = soup.get_text()
answer_list = {1: '财经', 2: '科技', 3: '汽车', 4: '房产', 5: '体育', 6: '娱乐', 7: '其他'}
# prekind=ie_knn.answer(d_text)
prekind_knn = textclassify.d_knnTeststonefile(d_text)
prekind_bayes = textclassify.d_bayesteststonefile(d_text)
d_result_knn = answer_list[prekind_knn[0][0]]
d_result_bayes = answer_list[prekind_bayes[0][0]]

d_dict = {}
for i in xrange(0, 7):
    d_dict[prekind_knn[i][0]] = prekind_knn[i][1]
d_list_knn = d_dict.values()[::-1]

d_dict = {}
for i in xrange(0, 7):
    d_dict[prekind_bayes[i][0]] = prekind_bayes[i][1]
d_list_bayes = d_dict.values()[::-1]
# return HttpResponse(rep)
return render(request, 'result.html', {'d_result_knn': d_result_knn, 'd_result_bayes': d_result_bayes, 'd_list_knn': d_list_knn, 'd_list_bayes': d_list_bayes})
```

图 13: 功能集成部分代码展示

3.2.4 展示模块:

1、输入输出流程:

1) 输入待测的 URL

2) 爬取网站中的主体文本。即接收用户输入的链接, 使用 `urllib2.urlopen(链接).read()` 读取网页的内容。并使用 `bs4` 插件中的 `BeautifulSoup` 类读取网页中的文本内容。

3) 对网站文本进行分词处理。

4) 分别使用贝叶斯和 KNN 算法对文本分词数据进行分析, 分别计算其权重和余弦相似度。最终得到排名最高的类别。

5) 输出分类结果

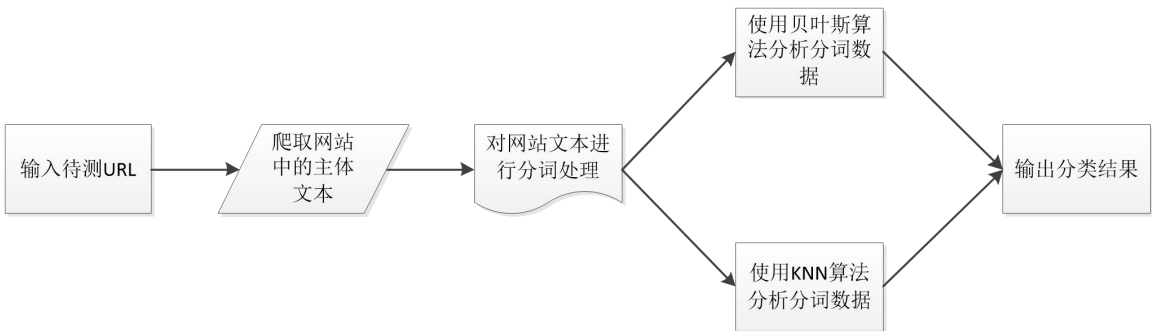


图 14: 展示模块流程图

2、使用过程展示

首页展示：



图 15：首页图

找到一篇与汽车相关的文档：

北京发布网约车细则：京籍京牌限制继续保留

2016年12月21日 16:07 新浪科技 微博



新浪科技讯 12月21日下午消息，北京市交通委员会发布《北京市网络预约出租汽车经营服务管理实施细则》与《北京市私人小客车合乘出行指导意见》，促进出租汽车行业和互联网融合发展，规范网络预约出租汽车经营服务行为。

创事记

谷歌不可能

其中细则规定，车辆申请网约车资质时，须满足本市最新公布实施的机动车排放标准，没有未处理完毕的交通事故和交通违法记录，且符合本市网约车车辆技术规范等条件。



网约车驾驶员须为本市户籍，取得本市核发的驾驶证件，未达到法定退休年龄，身体健康，申请之日前1年内驾驶机动车未发生5次以上道路交通安全违法行为，未被列入出租汽车严重违法信息库。

科技前

网约车平台公司要加强网络安全和信息安全防护，建立健全数据安全管理制度，依法依规采集、使用和保护个人信息，不得泄露涉及国家安全的敏感信息，所采集的

图 16：新闻截图

将其 URL 输入搜索框，点击分类，得到结果。

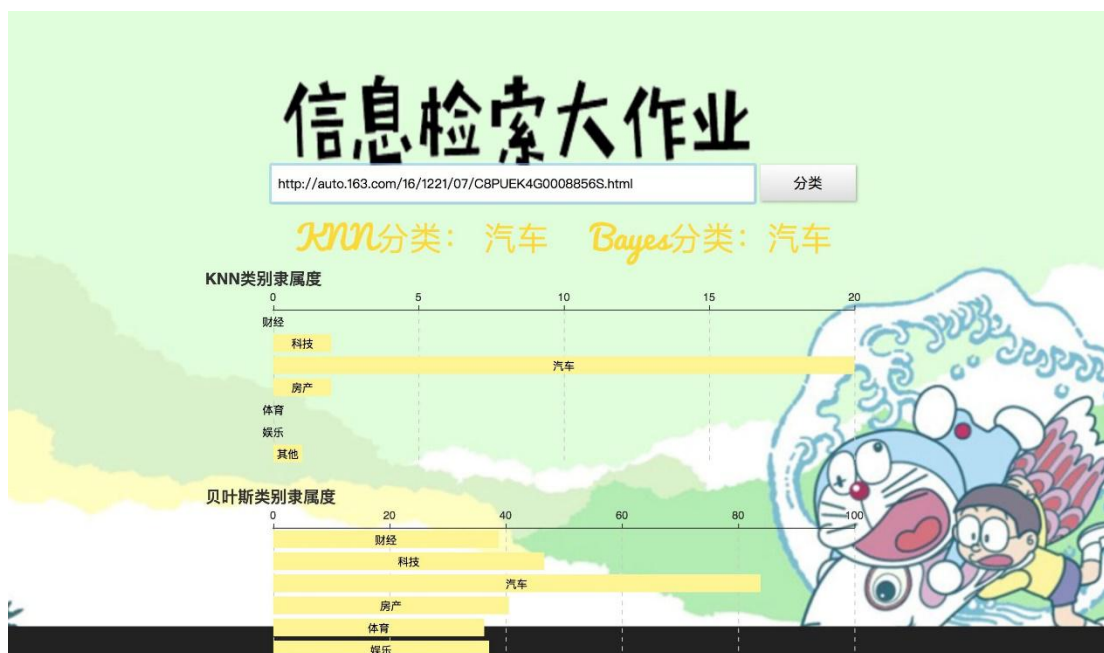


图 17：结果展示图

结果显示：利用 KNN 和贝叶斯算法均能将该文档准确分类为汽车类别。

4. 测试及分析

4.1 交叉验证的方法

交叉验证（Cross Validation）是用来验证分类器的性能一种统计分析方法，基本思想是把在某种意义下将原始数据（dataset）进行分组，一部分做为训练集（training set），另一部分做为验证集（validation set），首先用训练集对分类器进行训练，在利用验证集来测试训练得到的模型（model），以此来做为评价分类器的性能指标。

其中 **K-fold Cross Validation**（**K-折交叉验证**，记为 **K-CV**）：将原始数据分成 K 组（一般是均分），将每个子集数据分别做一次验证集，其余的 K-1 组子集数据作为训练集，这样会得到 K 个模型，用这 K 个模型最终的验证集的分类准确率的平均数作为此 K-CV 下分类器的性能指标。K 一般大于等于 2，实际操作时一般从 3 开始取，只有在原始数据集合数据量小的时候才会尝试取 2。K-CV 可以有效的避免过学习以及欠学习状态的发生，最后得到的结果也比较具有说服力。

4.2 交叉验证结果

在实际的训练中使用了 7 次交叉验证，分别测试 7 次实验的正确率，最终得到平均效果的正确率，如图所示：

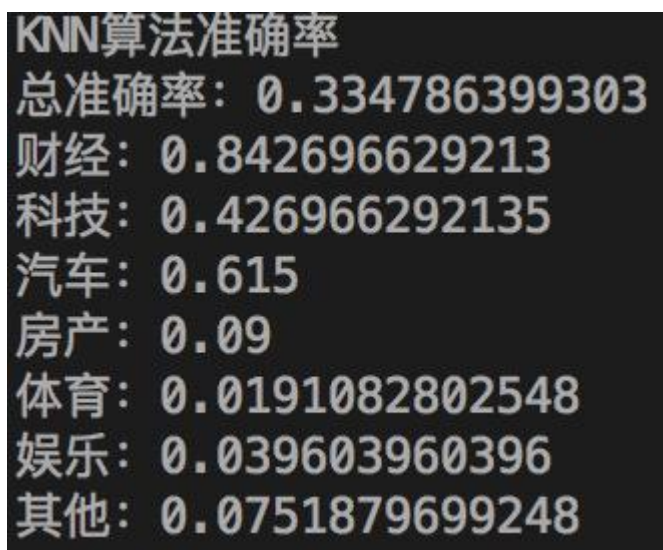


图 18: KNN 算法结果正确率



图 19: 贝叶斯算法结果正确率

4.3 结果分析

从结果可以看到，KNN 算法对于财经科技类的分类效果非常好，房产类体育类娱乐类分类效果不好。分析原因可能是由于训练集相关数据的相似性不够导致向量特征很难表达类别特性。

但是贝叶斯方法对于这七个类别的分类效果都非常好，尤其对于 **KNN** 分类效果不够好的五类，两种算法表现了很好的交叉性。

4.4 可优化地方

1、爬虫爬取数据的类型需要进一步丰富，以支持多类别的数据分类。爬虫在爬取大量数据的时候容易出现被封现象，针对这一点可以进一步从爬虫频率，使用代理等角度进行优化。

2、如果处理更多的数据，需要对贝叶斯和 **KNN** 实现算法实现过程进行优化，可考虑使用特殊的结构存储训练数据，以减少计算的次数。

3、两种算法在类别划分上有很好的交叉性，可以考虑结合两种方法结果达到更好的分类结果。

5 结束语

本项目通过对网页数据的多种分类方法尝试，发现贝叶斯和 **KNN** 分类方法对于此类数据有较好的分类效果，符合实验需求的要求。通过算法实现的过程，对于信息检索的实现过程有了深刻的理解，特别是针对本实验中的数据分类结果针对不同类别分类效果差别很大做了深入的探讨和分析。对于优化算法、提高所有类别分类效果做了很多尝试，深入体会了困难所在，以及一些初步的优化想法，例如通过改善爬虫爬取数据的方法增加训练集的丰富性，通过采用更多的数据特征的方式提高分类效果等。

项目已在 **github** 上公开（https://github.com/serea/IR_work.git），后续将结合相关工作做进一步完善。