第八届“泰迪杯”数据挖掘挑战赛——
C 题：“智慧政务”中的文本挖掘应用

题目： 基于文本挖掘的智慧政务研究

目　录

基于文本挖掘的智慧政务研究

摘要： 近年来，随着网络用户人数的迅速增加以及网络问政平台的不断完善，网络问政平台成为公民向政府部门反映问题、提出建议的重要渠道。本文在大数据背景下利用自然语言处理和文本挖掘的方法，结合朴素贝叶斯模型、随机森林模型、逻辑回归模型、LDA 模型、F-Score 模型等对群众留言进行热点挖掘，并针对相关部门的答复意见给出了评价方案。将自然语言处理技术与机器学习运用到网络问政平台中，不仅能够顺应网络问政的发展，而且可以及时处理网络问政留言，减轻人工处理的工作量节约人力物力财力，大规模降低智慧政务系统的工作成本。

关键词： 文本挖掘；LDA 主题模型；分类器；线性支持向量机

Abstract: In recent years, with the rapid increase of the number of network users and the continuous improvement of the network politics platform, the network politics platform has become an important channel for citizens to reflect problems and make suggestions to government departments. In this paper, we use natural language processing and text mining methods under the background of big data, combined with naive Bayes model, random forest model, logical regression model, LDA model, F-score model, etc. to mine the hot spots of the public message, and give the evaluation scheme for the reply of relevant departments. The application of natural language processing technology and machine learning to the network politics platform can not only conform to the development of network politics, but also deal with the network politics message in time, reduce the workload of manual processing, save human resources and financial resources, and reduce the work cost of the intelligent government system on a large scale.

Keywords: Text mining; LDA thematic model; Classifier; Linear support vector machines

一、问题重述

1.1 问题背景

随着经济社会的不断发展和我国改革开放的继续深化,政府问政、听政的改革与治理问题引起了政府相关部门的重视和广泛关注，网络问政逐步出现在了社会公众与政府的视野当中。为获取人民的表达意见,政府部门凭借网络了解民情、体察民意,让人民通过网络参与政府决策并监督政府各项公共事务的办理,这就是网络问政。

伴随着网络问政平台使用率的不断增加，问政留言数目直线上升，数据量级更是呈指数级增长，留言信息人工处理的弊端暴露无遗：无法实时处理问政留言；面对海量数据信息耗费大量人力物力且效率低下；更有人工处理的无意识错误给问政信息的反馈带来隐患。但随着大数据、人工智能等技术的发展，基于自然语言处理技术的网络问政平台已经成为电子政务系统新的发展方向。随着处理技术的进步,目前,自然语言处理技术已经在文本信息检索、机器翻译、智能问答等领域拥有了广泛的应用。将网络问政与自然语言处理技术相结合,实现问政内容与技术处理的顺畅对接，既可以使传统的治理模式向依托大数据的治理模式转变,又能够提高政府机构对虚拟社会的管理水平，为行政决策科学化提供保证。

文本分类是信息检索领域多年来一直在研究的课题，一方面以文本搜索应用为目的来提高搜索的有效性和效率；另一方面，文本分类也是数据挖掘和经典机器学习领域的一个重要研究方向，应用十分广泛。在文本分类的模型方面，基础模型有多项式朴素贝叶斯模型、随机森林模型、逻辑回归模型和线性支持向量机模型等。在 2016 年，贺鸣、成颖[1]等在《基于朴素贝叶斯的文本分类研究综述》中详细地对基于朴素贝叶斯的文本自动分类研究进行了系统的综述。他们介绍了多项式模型和多元伯努利模型等经典的朴素贝叶斯分类方法，并重点分析了经典的特征选择方法等多种方法。对于随即森林算法，在 2014 年，吕红燕和冯倩[2]在《随机森林算法研究综述》中详细介绍了它的定义，并具体表述了随机森林算法的原理和性质,然后综述了近几年来随机森林算法的改

进研究及应用领域,最后对随机森林算法研究做出了总结。2020 年，杨锋[3]在《基于线性支持向量机的文本分类应用研究》中利用公共文本数据集利用 线性支持向量机模型进行了文本分类实验,对线性支持向量机与传统支持向量机在文本分类时的训练时间和分类准确率进行了对比分析,结果表明线性支持向量机具有更快的训练速度和更好的泛化能力。

在搭建 LDA 主题模型方面，2015 年，祖弦与谢飞[4]在《LDA 主题模型研究综述》中系统阐述 LDA(Latent Dirichlet Allocation)主题模型参数估计和 Gibbs 抽样算法,介绍了常见的 LDA 改进和扩展模型,最后分析 LDA 模型在文本挖掘领域的应用情况。为本文模型的搭建奠定了基础。

1.2 问题重述

自然语言处理技术是人工智能领域重要的研究方向之一。Python 作为近几年最火的编程语言，常常被应用于数据分析之中。附件给出了收集自互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见。下面利用自然语言处理和文本挖掘的方法解决下面的问题。

1.群众留言分类。在处理网络问政平台的群众留言时，工作人员首先按照一定的划分体系对留言进行分类，以便后续将群众留言分派至相应的职能部门处理。目前，大部分电子政务系统还是依靠人工根据经验处理，存在工作量大、效率低，且差错率高等问题。请根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型。

2. 热点问题挖掘。某一时段内群众集中反映的某一问题可称为热点问题，如“XXX 小区多位业主多次反映入夏以来小区楼下烧烤店深夜经营导致噪音和油烟扰民”。及时发现热点问题，有助于相关部门进行有针对性地处理，提升服务效率。请根据附件 3 将某一时段内反映特定地点或特定人群问题的留言进行归类，定义合理的热度评价指标，并给出评价结果，按表 1 的格式给出排名前 5 的热点问题。按表 2 的格式给出相应热点问题对应的留言信息。

3.答复意见的评价。针对附件 4 相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案，并尝试实现。

1.3 问题分析

对于问题一：

分类是指计算机自动对给定数据数据进行标注。附件二是用户留言的信息及其一级标签分类情况，要根据用户的留言详情，建立关于一级标签分类模型，本小题的解决方案为，选择朴素贝叶斯分类器进行训练，训练完成之后，就可以对留言进行分类，之后再进行三种模型的对比分析，评估各个模型的平均准确率，再选出准确率最高的模型，查看该模型下各分类的 F1-score（多分类模型一般不使用准确率 accuracy 来评估模型的质量，因为当训练数据不平衡(有的类数据很多,有的类数据很少)时，accuracy 不能反映出模型的实际预测精度，这时候我们就需要借助于 F1-score、ROC 等指标来评估模型。）

对于问题二：

附件三是用户留言详情及其点赞数与反对数的情况，我们的思路是先对附件三进行常规的预处理（去除停用词、分词等），然后就是搭建 LDA 主题模型，通过 LDA 主题模型确定了全部留言内容的五个主题数目。

对于问题三：

附件 4 中主要给出了针对用户留言详情的答复意见。我们想对答复意见的质量给出一套评价方案，那么主要考虑的就是答复意见与留言的相关性是否较强，所以主要从相似度入手。首先分析了留言详情与答复意见的总体相似度：分词后根据词袋模型统计词在每篇文档中出现的次数，形成向量，最后计算余弦相似度。之后具体分析每一行答复对于留言的相似度：利用第二题中 LDA 模型学习到的主题，来比较二者相似度。

二、符号说明

符号	定义
lei_id_df	七个分类的 id 情况
cut_liuyan	数据预处理之后的留言详情（已分词）

三、数据预处理

所给出的原始数据中存在大量的冗余数据，会影响算法的运行判断，所以

需要对原始数据进行预处理。

3.1 附件二：用户留言一级标签分类

在附件 2 的一级分类标签中给出了七个标签（城乡建设、环境保护、交通运输、教育文体、劳动和社会保障、商贸旅游、卫生计生），共计 9210 条留言数据。

我们的目的是要把不同的留言详情分到不同的分类中去，并且每条留言只能对应七个分类标签中的一个。

我们统计了七个类别的各自的数据量，结果如下：

	一级标签	数量
0	城乡建设	2009
1	劳动和社会保障	1969
2	教育文体	1589
3	商贸旅游	1215
4	环境保护	938
5	卫生计生	877
6	交通运输	613

图 3-1 各个分类的数量分布情况

接下来用图形化的方式查看各个类别的分布，如图 3-2 所示：

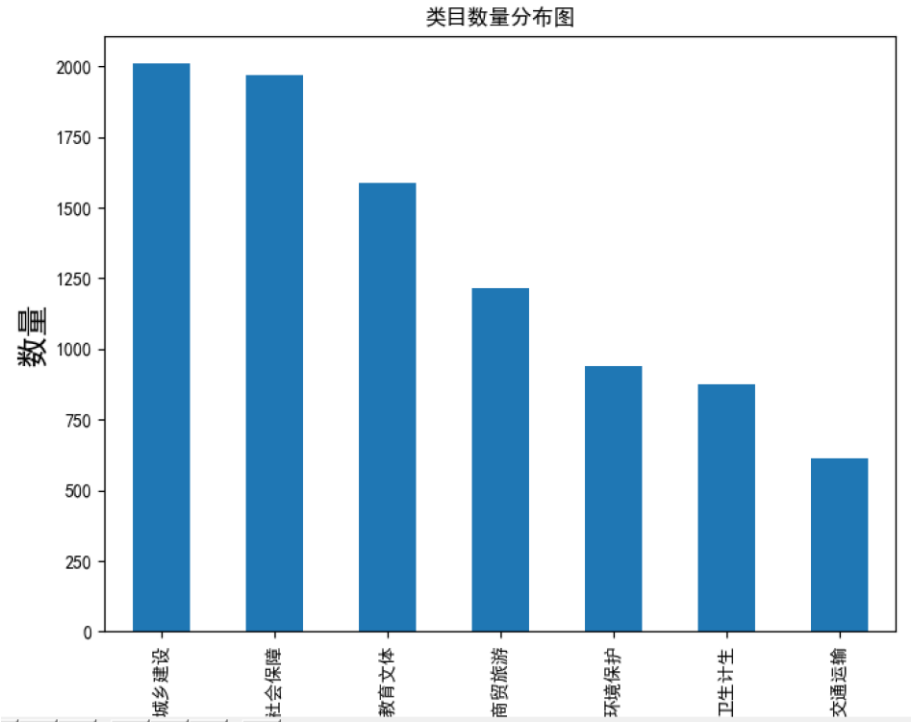


图 3-2 各个分类的数量分布条形图

将七个一级分类转换成 id，这样方便后面的分类模型的训练。

	一级标签	lei_id
0	城乡建设	0
1	环境保护	1
2	交通运输	2
3	教育文体	3
4	劳动和社会保障	4
5	商贸旅游	5
6	卫生计生	6

图 3-3 七个分类转为 ID

由于我们的留言内容都是中文，所以要对中文进行一些预处理工作，这包括删除文本中的标点符号，特殊符号，还要删除一些无意义的常用词（此处所用停用词表为 GitHub 上下载的中文停用词表），因为这些词和符号对系统分析预测文本的内容没有任何帮助，所有在使用这些文本数据之前必须要将它们清除干净。我们过滤掉了留言详情中的标点符号和一些特殊符号，并生成了一个新的字段，接下来我们要在该字段的基础上进行分词，把每个评论内容分成由空格隔开的一个一个单独的词语。

经过分词以后我们生成了又一个新字段。接下来我们要在该字段的基础上生成每个分类的词云图（前 100 个高频词的词云图）。

七个分类各自的词云图如图 3-4 所示：

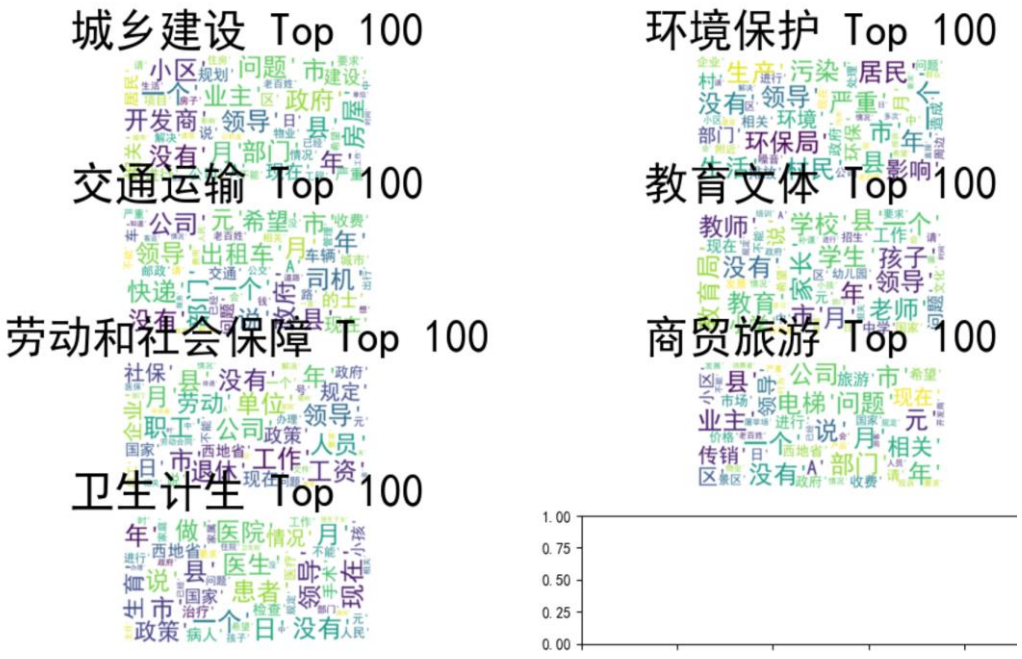


图 3-4 七个分类的并列词云图

3.2 附件三：留言详情以及点赞数与反对数情况

先用正则表达式 re.compile(u"[^a-zA-Z0-9\u4E00-\u9FA5]")匹配出中文，再对处理之后的文本进行去除停用词、分词等操作，还要删除少于两个字的词，由于有些无用的词性对文本的分析与预测没有作用，因此我们需要将每一条留言详情转化为只包含地名、名词、动名词、动词的文本，此处运用了 apply 函数中筛选词性的一个参数 allowPOS，表达为 allowPOS=('ns', 'n', 'vn', 'v')。

3.3 附件四：留言答复意见情况

读取附件 4 的留言详情与答复意见两列数据并分别存为 txt 文档 message 与 response，之后对处理之后的文本去除停用词、分词，后再分别保存至字典中，以便后续计算。

四、问题一研究与分析

4.1 分类器的选择

为了训练监督学习的分类器,我们首先将“留言详情”转变为包含数字的词向量。例如我们前面已经转换好的 tf-idf 的特征值（features）。当我们有了词向量以后我们就可以开始训练我们的分类器。分类器训练完成后,就可以对没有见过的留言详情进行预测。

4.1.1 朴素贝叶斯分类器

朴素贝叶斯分类器最适合用于基于词频的高维数据分类器,最典型的应用如垃圾邮件分类器等,准确率可以高达 95%以上。这里我们使用的是 sklearn 的朴素贝叶斯分类器 MultinomialNB，我们首先将留言详情转换成词频向量,然后将词频向量再转换成 TF-IDF 向量,这里我们还是按照一般的方式将生成 TF-IDF 向量分成两个步骤：1.生成词频向量 2.生成 TF-IDF 向量。

最后我们开始训练我们的 MultinomialNB 分类器。

当模型训练完 预测函数，使用该函数测试一下，该预测效果符合预期，效果如图 4-1：

```
    ....
In[11]: print(myPredict("一家生鲜超市经常从事炒菜作业，油烟污染及其严重，搞的街面到处是油烟，由于灶台移动性方便"))
城乡建设
```

图 4-1 预测效果展示

4.2 模型的选择

接下来我们尝试不同的机器学习模型,并评估它们的准确率，我们将使用如下三种模型: (Multinomial) Naive Bayes(多项式朴素贝叶斯)、Linear Support Vector Machine(线性支持向量机)、Random Forest(随机森林)。从箱体图 4-2 以及准确率结果图 4-3 上可以看出随机森林分类器的准确率是最低的，因为随机森林属于集成分类器(有若干个子分类器组合而成)，一般来说集成分类器不适合处理高维数据(如文本数据)，因为文本数据有太多的特征值，使得集成分类器难以应付。

其中线性支持向量机的准确率最高，其平均准确率达到了 86.98%，其次是朴素贝叶斯和随机森林。

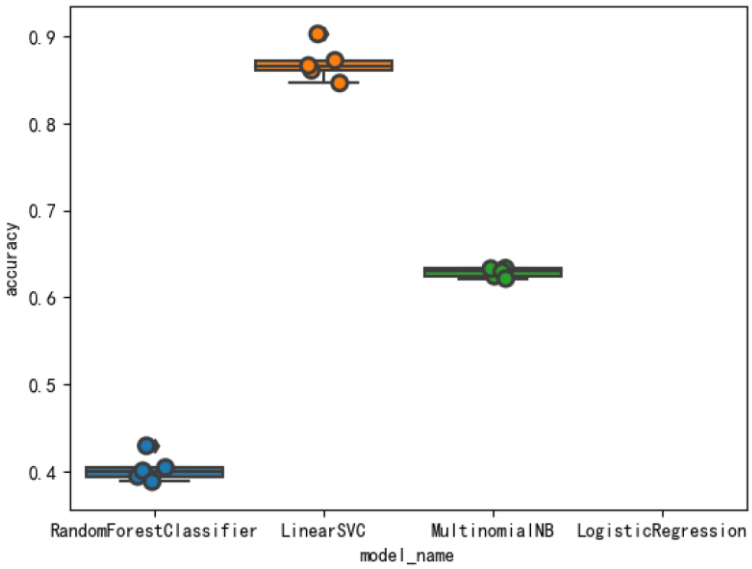


图 4-2 各个模型的准确率的箱线图

```
model_name
LinearSVC                    0.869815
LogisticRegression                NaN
MultinomialNB                0.628447
RandomForestClassifier       0.403366
Name: accuracy, dtype: float64
```

图 4-3 各个模型的准确率展示

4.3 模型评估

下面我们就针对平均准确率最高的 LinearSVC（线性支持向量机）模型，我们将查看混淆矩阵，并显示预测标签和实际标签之间的差异。

我们查看混肴矩阵的结果如图 4-4：

图 4-4 混肴矩阵的结果

混淆矩阵的主对角线表示预测正确的数量，除主对角线外其余都是预测错误的数量。从上面的混淆矩阵可以看出"交通运输"、"环境保护"和"卫生计生"类预测最准确，预测错误较少。“城乡建设”和“劳动和社会保障”预测的错误数量较多。

根据所查资料可知，多分类模型一般不使用准确率(accuracy)来评估模型的质量，因为 accuracy 不能反应出每一个分类的准确性，因为当训练数据不平

衡(有的类数据很多,有的类数据很少)时，accuracy 不能反映出模型的实际预测精度,这时候我们就需要借助于 F1-score、ROC 等指标来评估模型。

图 4-5 为七个分类的 F1-score 情况，

accuracy 0.9006578947368421

	precision	recall	f1-score	support
城乡建设	0.82	0.94	0.88	663
环境保护	0.96	0.94	0.95	310
交通运输	0.96	0.70	0.81	202
教育文体	0.94	0.94	0.94	525
劳动和社会保障	0.90	0.95	0.92	650
商贸旅游	0.90	0.82	0.86	401
卫生计生	0.94	0.85	0.89	289
accuracy			0.90	3040
macro avg	0.92	0.88	0.89	3040
weighted avg	0.90	0.90	0.90	3040

图 4-5 七个分类的 F1-score 分布情况

从以上 F1 分数上看，"环境保护"类以及“教育文体”类的 F1 分数最大，“交通运输”类 F1 分数最低，为 0.81，究其原因可能是因为“交通运输”分类的训练数据最少，为 202 条，使得模型学习的不够充分，导致预测失误较多。

下面我们来观察一些预测失误的例子，如图 4-6~图 4-8 所示。

交通运输 预测为 城乡建设 ： 43 例.

	一级标签	留言详情
3220	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t 沿河路在水一方处到L9县路连接的一段，也...
3301	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t 每天晚间在楚江A市段楚江一桥至二桥之间都...
2947	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t地处月亮岛街道时代倾城小区一二三期之间的公共...
3460	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的谭书记：您好！ 我认为I4县交通令人...
3067	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t 近来，每当夜幕降临，老百姓出门慢步休闲在...
3351	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t 我们因为工作的原因经常要去中国邮政局A6...
3541	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t 1.隧道技术已成熟。建成合武高速那样连续...
3148	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t我是F8市弼时镇湄江村村民彭元中，我家是建档...
3502	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t地铁1号线从南湖路站-马厂站偏离人流量极大的...
3024	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t C市金薮遍远乡镇街道下雨污水严重，路面狭...
2984	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t周县长： 您好！我是朱良桥乡杈梓桥村草坝子...
2950	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t楚府路快速化改造在2016年启动，说是201...
3231	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t L12市岩垅乡江丘村三组没有通水泥路，何时...
3536	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t 粗制滥造 九标施工偷工减料以次充好事实...
3193	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t K6县到K7县路段公交车在节假日旅客多时...
3553	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t关于D2区黄茶路55号渣土随意运输，倒放的问...
3026	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的陈书记： 您好！灰虞公路潭市...

图 4-6 交通运输预测为城乡建设的实例

商贸旅游 预测为 城乡建设 ： 44 例.

	一级标签	留言详情
8009	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t尊重的领导：你好！对于大汉二期翡翠湾房屋质量...
7279	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t市长先生： 你好！ 在B9市一个小区买了...
7229	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的领导：我于2018年5月5日在A市A5...
8287	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t前几天看新闻说天然气会要涨价，所以特地关注了...
7657	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\tL5县交通两大工具：1，火车，以前县城火车很...
7742	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t 各位政府领导好，苍冲村到兰里街上不到20...
8086	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t 西地省A9市华盛彩虹城，至少百分之40的...
8043	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t 华嘉地产公司汇金城二期项目于2018年1...
7313	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t 书院路长征新村和鑫源花苑入口处，有一小摊...
7945	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\tA市安置房存在质量问题电梯，发生多次掉层事故...
7660	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t请问L5县至低庄的公交车为啥又涨价了？请问有...
7599	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t近来得知消息,K市的公交车全部实行一波涨价政...
8101	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t 我购买的A市世茂房产，开发商于0000-...
7420	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t我想知道E市夏季特色旅游线路是什么？想带小孩...
7777	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t一，元月6日上午，我在群丰路公交车站等车回东...

图 4-7 商贸旅游预测为城乡建设的实例

卫生计生 预测为 劳动和社会保障 ： 25 例.

	一级标签			留言详情
8611	卫生计生	\n	\n	我妻子是省医保的缴费者，今天到A市口腔医院补牙（属于省医保的...
9191	卫生计生	\n	\n	您好！向您咨询个问题，请在百忙中回复为盼!根据国家卫生健康委员会...
8576	卫生计生	\n	\n	2014年卫生系统高级职称报名时间：5月5日-5月11日24...
9158	卫生计生	\n	\n	2019.9.12日，西地省卫健委职改办发布了一份“假冒期刊”，...
8660	卫生计生	\n	\n	詹鸣书记： 请问2014年主治医师考试西地省内乡镇基层分数线何...
8780	卫生计生	\n	\n	正在楚雅附一医院感染科52床进行紧急救治的E市E7县六都寨镇...
8498	卫生计生	\n	\n	尊敬的西地省卫生厅长： 您好！我们是F7县2012年参加全...
8914	卫生计生	\n	\n	领导，我工作在D市，户口在A市，现退休住在A市。看病时D市的...
9064	卫生计生	\n	\n	K3县康原肾病专科医院于2018年7月11日上午发生火灾，被K3...
9157	卫生计生	\n	\n	下是西地省F市F9市横铺镇村医反应情况。西地省F市F9市横铺镇2...
9054	卫生计生	\n	\n	西地省农村卫生信息系统频繁出问题：不是登不上去，就是登上了不能翻...
8480	卫生计生	\n	\n	尊敬的厅长： 我是医院的一名职工，我们医院的肛肠科承包给了私人...
9070	卫生计生	\n	\n	一、K2区凭什么不认可全国通用的护士执业资格证书？目前本人已参考...
8850	卫生计生	\n	\n	我是一名公务员，积极响应国家单独二胎政策，于今年3月底即将分...

图 4-8 卫生计生预测为劳动和社会保障的实例

五、问题二研究与分析

5.1 隐狄利克雷分配模型（LDA 主题模型）

LDA 是非监督的机器学习模型，并且使用了词袋模型。一篇文章将会用词袋模型构造成词向量。LDA 需要我们手动确定要划分的主题的个数，LDA 算法的输入与输出如图 5-1 所示：

LDA 算法的输入与输出
算法输入： 分词后的文章集（通常为一篇文章一行） 主题数 K，超参数 α 和 β
算法输出： 1.每篇文章的各个词被指定(assign)的主题编号：tassign-model.txt 2.每篇文章的主题概率分布 θ：theta-model.txt 3.每个主题下的词概率分布 ϕ：phi-model.txt 4.程序中词语 word 的 id 映射表：wordmap.txt 5.每个主题下 ϕ 概率排序从高到低 top n 特征词：twords.txt

图 5-1 LDA 算法的输入与输出

其伪代码如下：

LDA 中，生成文档的过程如下：

1. 按照先验概率 $p(d_i)$ 选择一篇文档 d_i

2. 从 Dirichlet 分布 α 中取样生成文档 d_i 的主题分布 θ_i，主题分布 θ_i 由超参数

为α的Dirichlet分布生成

3. 从主题的多项式分布θ_i中取样生成文档d_i第 j 个词的主题$z_{i,j}$

4. 从Dirichlet分布β中取样生成主题$z_{i,j}$对应的词语分布$\varphi_{z_{i,j}}$，词语分布$\varphi_{z_{i,j}}$由参数为β的Dirichlet分布生成

5. 从词语的多项式分布$\varphi_{z_{i,j}}$中采样最终生成词语$\omega_{i,j}$[5]

5.2 问题研究

先对附件三的数据进行预处理，去除了不常使用的停用词并进行分词，接着删除少于两个字的词，通过条形图检测文本中前三十五的高频词，接着我们将每一条留言详情转化为只包含地名、名词、动名词、动词的文本，因为这些词性对文本的分析具有重要的影响程度。

在对附件三的文本进行完预处理之后，下面我们来搭建LDA主题模型了。我们先对语料创建术语词典，之后，我们要用创建好的词典将评论转化成文件术语矩阵，我们可生成LDA主题模型学习到的主题，如下图5-2所示：

```
In[6]: lda_model.print_topics()
Out[6]:
[(0,
  '0.007*"车位" + 0.006*"希望" + 0.005*"景园" + 0.004*"滨河" + 0.004*"捆绑" + 0.004*"职工" + 0.004*"停车" + 0.004*"请问" + 0.004*"领
 (1,
  '0.015*"居民" + 0.015*"小区" + 0.012*"影响" + 0.012*"部门" + 0.009*"没有" + 0.009*"生活" + 0.008*"领导" + 0.008*"希望" + 0.007*"相
 (2,
  '0.007*"没有" + 0.007*"工作" + 0.006*"请问" + 0.005*"领导" + 0.005*"西地省" + 0.005*"办理" + 0.005*"是否" + 0.005*"公司" + 0.004*"
 (3,
  '0.014*"业主" + 0.013*"没有" + 0.009*"问题" + 0.009*"开发商" + 0.008*"政府" + 0.008*"相关" + 0.008*"领导" + 0.007*"部门" + 0.007*"
 (4,
  '0.010*"学校" + 0.010*"希望" + 0.008*"没有" + 0.006*"学生" + 0.006*"国家" + 0.006*"领导" + 0.006*"政策" + 0.006*"孩子" + 0.005*"时
```

图5-2 LDA分成的各个主题的词频分布情况

最后我们进行主题的可视化，为了在二维空间中对我们的主题进行可视化，我们用的是pyLDAvis库。生成的主题可视化结果展示在网页上，截图进行展示，如图5-3~5-7所示（此处展示了五个主题的各自的分布情况）：

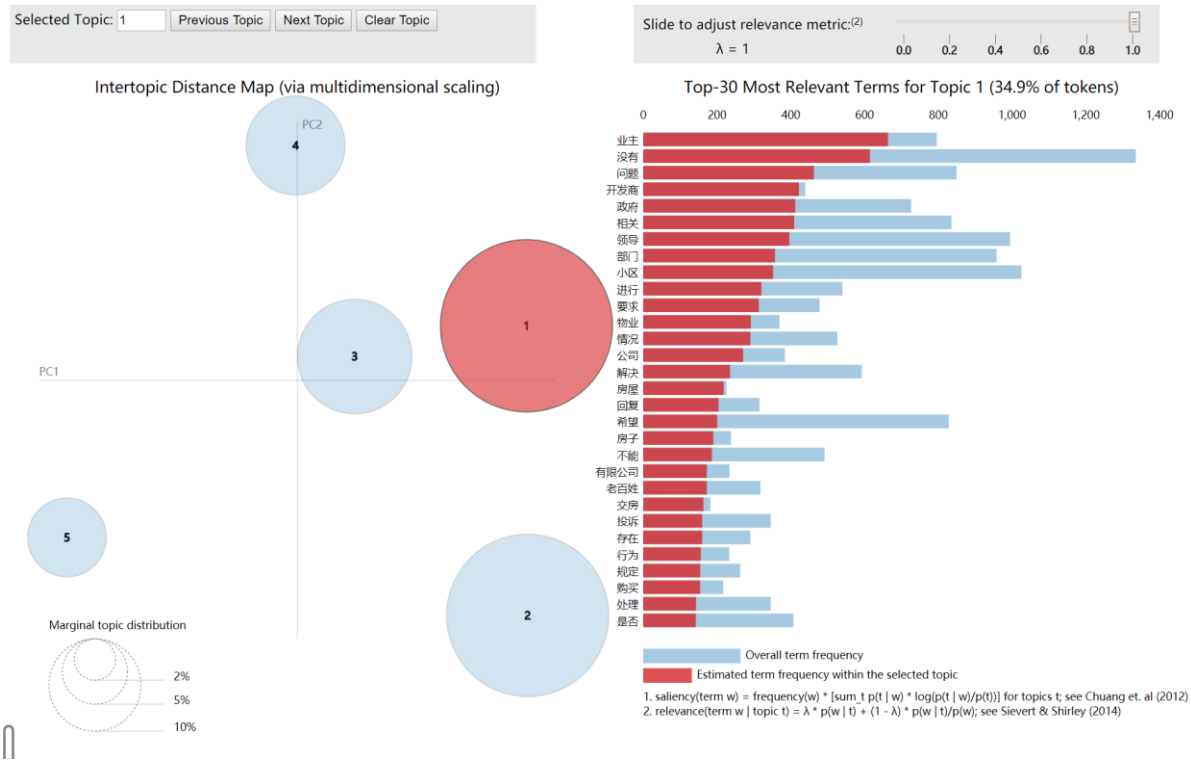


图 5-3 topic1 的分布情况

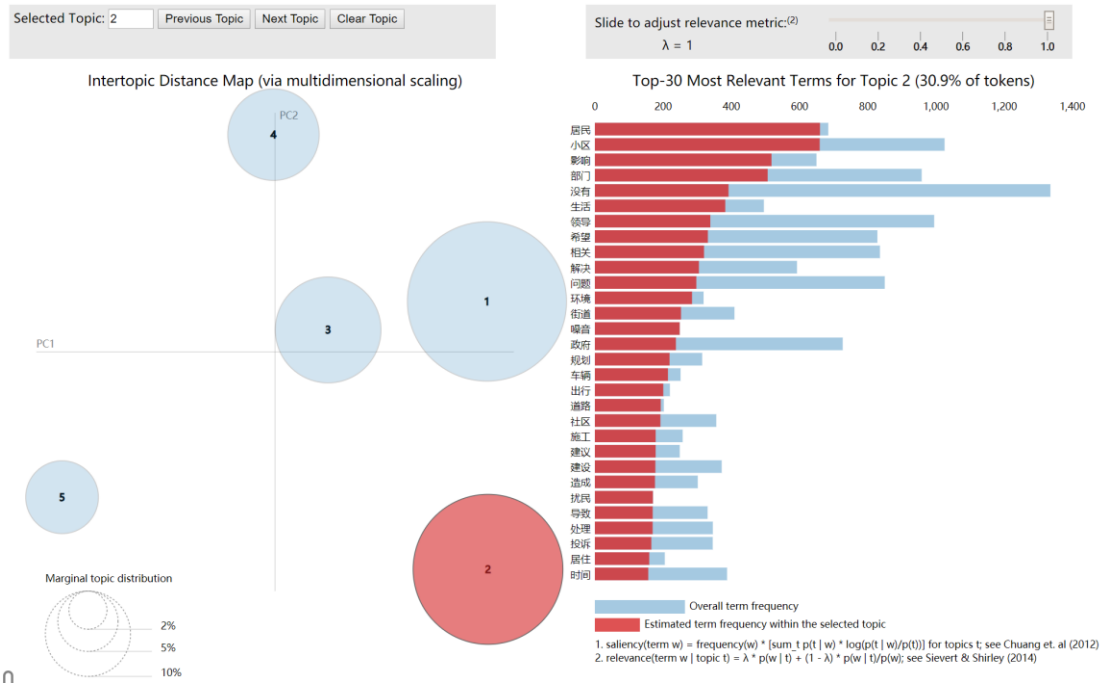


图 5-4 topic2 的分布情况

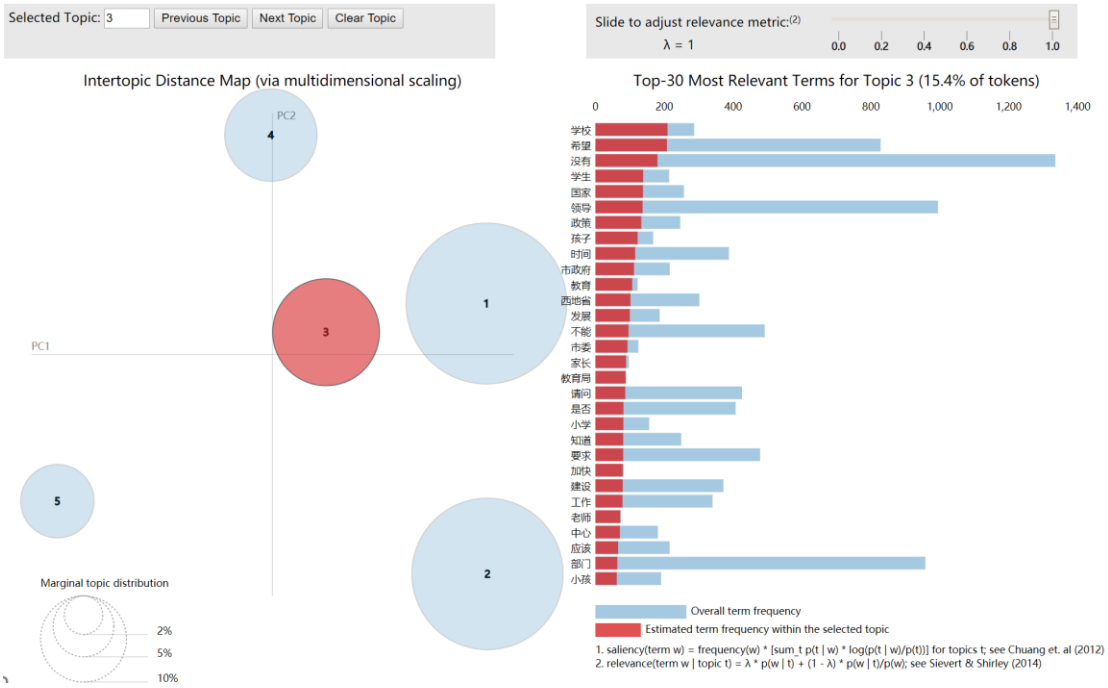


图 5-5 topic3 的分布情况

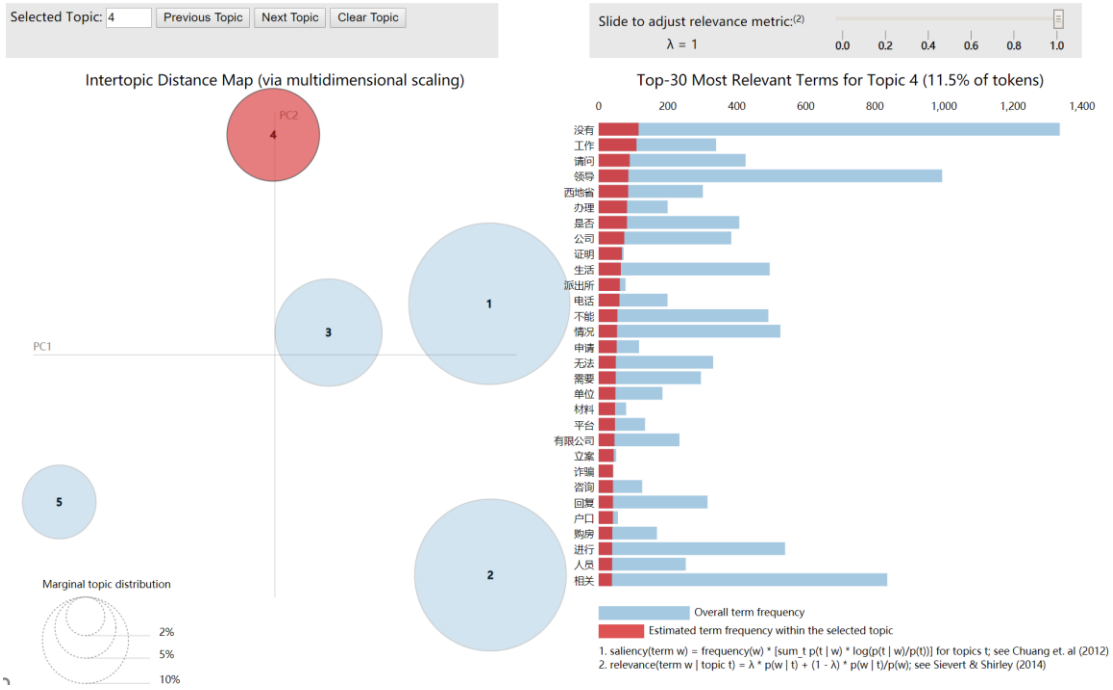


图 5-6 topic4 的分布情况

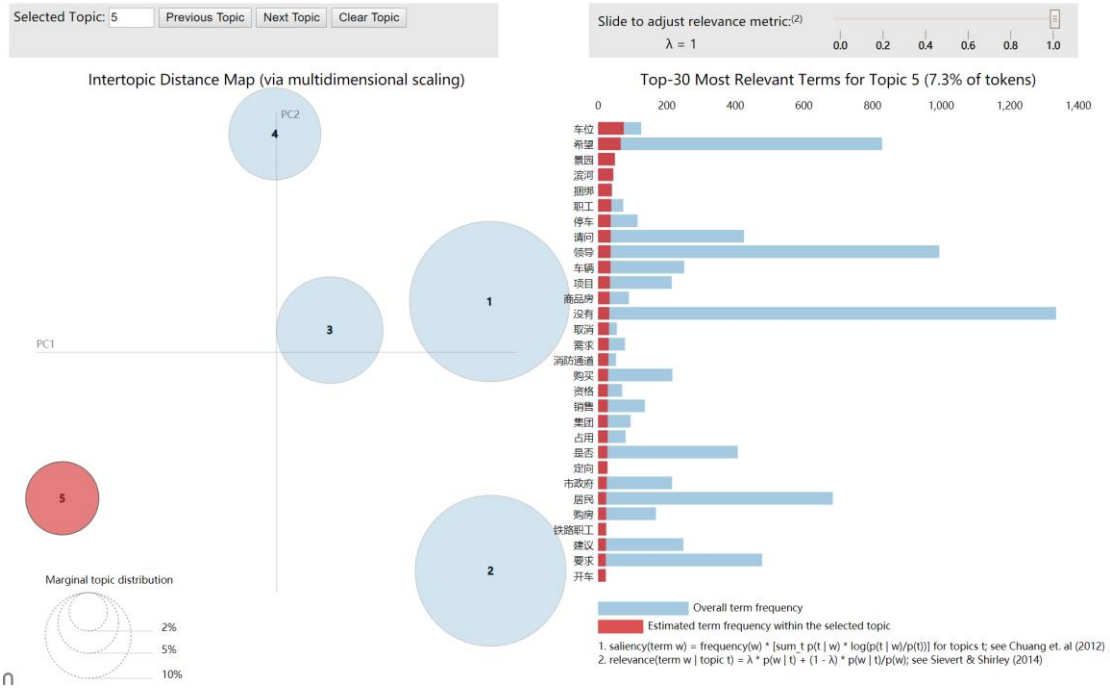


图 5-7 topic5 的分布情况

上图将所有的留言详情分成了五个主题，点击各个圆圈可看到每个主题的词频分布情况，进而可以了解热门的留言主题。

5.3 结论与分析

我们可以看出，通过搭建 LDA 主题模型选择出的五个主题，第一个主题的高频词集中为“业主”，“开发商”，“政府”，“部门”，“物业”等词，可看出它主要反映的是小区物业方面的问题，希望开发商以及政府予以重视；第二个主题的高频词集中为“居民”，“影响”，“生活”，“环境”，“噪音”等词，可看出它主要反映的是居民生活上的问题，比如噪音等；第三个主题的高频词集中为“学校”，“学生”，“政策”，“教育”，“希望”等词，可看出它主要反映的是教育方面的问题；第四个主题的高频词集中为“工作”，“领导”，“办理”，“申请”，“材料”等词，可看出它主要反映的是材料办理的问题；第五个主题的高频词集中为“景园”，“车位”，“停车”，“车辆”，“领导”等词，可看出它主要反映的是业主停车的问题。

六、问题三研究与分析

6.1 理论模型

主要运用了词袋模型理论。

在不考虑停用词的情况下，文档中出现频率越高的词语，越能描述该文档。因此可以统计每个词项在每篇文档中出现的次数，即词项频率，记为$tf_{t,d}$，其中 t 为词项，d 为文档。将文档中的每个词语都赋予 tf 权重，那么这篇文档就变成了词与权重的集合，通常称为词袋模型（bag of words model）。用词袋模型来描述一篇文档，可以省略编程语言中中文转码的额外步骤，而且结果更加易于解释。词袋的含义就是说，像是把一篇文档拆分成一个一个的词条，然后将它们扔进一个袋子里。在袋子里的词与词之间是没有关系的。因此词袋模型中，词项在文档中出现的次序被忽略，出现的次数被统计。语序颠倒但不影响词意将被识别为同一词语。与将词项在每篇文档中出现的次数保存在向量中，这就是这篇文档的文档向量。

6.2 隐狄利克雷分配模型（LDA 主题模型）

此处 LDA 是非监督的机器学习模型，并且使用了词袋模型。并且将会用词袋模型构造成词向量。搭建 LDA 主题模型时，需要对语料创建术语词典。之后，我们要用创建好的词典将评论转化成文件术语矩阵。最后生成 LDA 模型学习到的主题。

以下为对回复意见提取的 LDA 主题（前五个）：

```
[(0,
  '0.039*"反映" + 0.034*"问题" + 0.025*"网友" + 0.018*"调查" + 0.015*"收悉" + 0.015*"相关" + 0.013*"留言" + 0.013*"情况" + 0.012*"部门" + 0.01
(1,
  '0.018*"回复" + 0.018*"工作" + 0.017*"收悉" + 0.016*"支持" + 0.013*"建设" + 0.011*"进行" + 0.011*"情况" + 0.010*"留言" + 0.009*"理解" + 0.00
(2,
  '0.016*"进行" + 0.014*"工作" + 0.013*"情况" + 0.012*"回复" + 0.012*"问题" + 0.011*"收悉" + 0.011*"支持" + 0.010*"要求" + 0.009*"反映" + 0.00
(3,
  '0.013*"工作" + 0.011*"收悉" + 0.011*"问题" + 0.011*"回复" + 0.011*"情况" + 0.011*"规定" + 0.010*"咨询" + 0.010*"相关" + 0.010*"办理" + 0.00
(4,
  '0.008*"车辆" + 0.005*"停车" + 0.004*"道路" + 0.004*"管理" + 0.004*"交通" + 0.003*"加强" + 0.003*"标准" + 0.003*"机动车" + 0.003*"停放" + 0.
```

图 6-1 回复意见提取的 LDA 主题

6.3 问题研究

首先计算留言详情与答复意见的总体相似度。

引入 pycharm 中的 xlrd 模块来读取 excel 表格中数据并存为 txt 文档。

总体模型的中文思路如下：

1.使用 jieba 包分别对两篇中文 txt 文件进行分词，得如[‘宿舍’，‘小区’，‘水电费’]的字符串列表。

2.对得到的分词后的数组通过进行词袋模型统计，得到他们每个词在文中出现的次数向量；

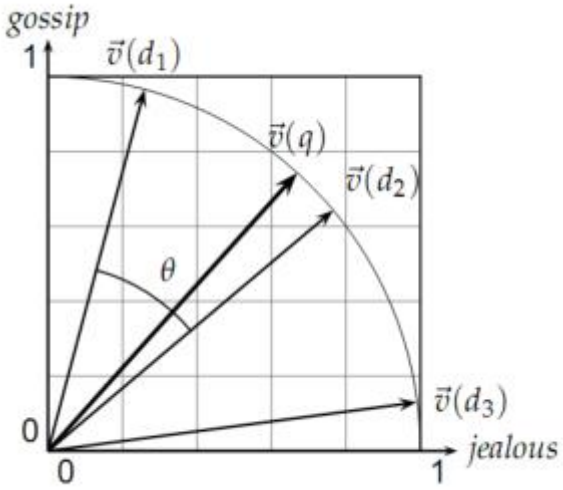
3.对得到的两个次数的矩阵进行余弦相似度计算，得到余弦相似度作为它们的文本相似度。

$$\mathrm{sim}(d_1, d_2) = \frac{\vec{V}(d_1) \cdot \vec{V}(d_2)}{|\vec{V}(d_1)||\vec{V}(d_2)|}$$

图 6-2 余弦相似度

余弦相似度计算需要引入文档向量。其中每个分量代表词项在文档中的相对重要性。一系列文档在同一向量空间的表示称为 VSM（Vector Space Model）。VSM 是词袋模型。向量空间模型是信息检索、文本分析中基本的模型。通过该模型，可以进行有序文档检索、文档聚类、文档分类等。当然，现在的研究有新发展。出现了很多模型代替 VSM。每篇文档在 VSM 中用向量表示，那么计算两篇文档的相似度自然的想到用两个向量的差值。但是，可能存在的情况是。如果两篇相似的文档，由于文档长度不一样。他们的向量的差值会很大。余弦相似度是使用的非常广泛的计算两个向量相似度的公式，它可以去除文档长度的影响。

公式的分母是两个向量的欧几里得长度之积，分子是两个向量的内积。这样得到的 sim 实际上就是两个欧式归一化的向量之间的夹角的余弦。如下图：



最终得到两篇文档的相似度为 0.115000。

```
>>> print("\n余弦相似度为 ：%f"%result)

余弦相似度为 ：0.115000
>>>
```

图 6-3 余弦相似度结果

由于相似度较低，与预先设想不太相同，接下来我们将留言详情与答复意见逐条对比。

将 LDA 模型学习到的主题作对比，然后打印出相似度大于 0.2 的结果。后面我们发现这个结果与直接分词计算的结果相差无几，所以我们给出简单提取的结果。

运行结果表明，答复意见和留言详情相似度在 20%以上的共有 48 条：

0.23529411764705882

0.20408163265306123

0.5161290322580645

0.2926829268292683

0.22

0.22535211267605634

0.3333333333333333

0.25806451612903225

0.3448275862068966

0.2222222222222222

0.2807017543859649

0.2077922077922078

0.4166666666666667

0.20689655172413793

0.2857142857142857

0.2682926829268293

0.20833333333333334

0.25

0.23076923076923078

0.2727272727272727

0.2222222222222222

0.21052631578947367

0.23529411764705882

0.24

0.21052631578947367

0.3076923076923077

0.3333333333333333

0.20689655172413793

0.21052631578947367

0.475

0.25

0.21052631578947367

0.21428571428571427

0.26666666666666666

0.2222222222222222

0.3076923076923077

0.2647058823529412

0.23529411764705882

0.28125

0.2222222222222222

0.20588235294117646

0.25

0.20512820512820512

0.46153846153846156

0.23529411764705882

0.26666666666666666

0.25

0.23529411764705882

附件 4 中的答复意见一共 2816 条，也就是说大部分答复意见较留言详情的相似度是偏低的。

6.4 结论与分析

通过上述结果，由于答复意见与留言详情相似度偏低，2816 个匹配项目中只有 48 个项目的相似度在 0.2 以上，所以我们只好认为相关部门给出的答复不够好，不符合预期要求。

究其原因，当下我国大多数的网络问政平台都采用了单向信息流模式、三极沟通模式和二极直接互动模式来构建模型，而当下我国网络问政中存在政府信息公开不足导致问政信息流割裂、政府与公民的议题互动度仍需提高、常态下的网络问政活动实际效果并不理想等问题。有关部门需要提高信息流通能力，使政府部门可以扩大信息公开度，从而与民众更好交流互动。

七、总结

本文主要研究了网络问政平台的热点问题挖掘以及评价模型，针对题目所给出的留言主题、答复意见等大量数据，分类后进行数据分析，建立模型，提取出热点词，给出评价指标。主要利用 python 进行程序运行并给出结果。

在对公众留言问题进行分类时，首先进行数据的预处理，而后结合自然语言处理技术，采用多项式朴素贝叶斯模型、随机森林模型、逻辑回归模型和线性支持向量机模型等模型对问政留言信息进行分类，最后借助于 F1-score、ROC 等指标来评估模型：从混淆矩阵可以看出"交通运输"、"环境保护"和"卫生计生"类预测最准确，预测错误较少，“城乡建设”和“劳动和社会保障”预测的错误数量较多。从 F1 分数上看，"环境保护"类以及“教育文体”类的 F1 分数最大，“交通运输”类 F1 分数最低，为 0.81。究其原因可能是因为“交通运输”分类的训练数据最少，为 202 条，使得模型学习的不够充分，导致预测失误较多。

对热点问题的关注方面，将自然语言处理技术与政务系统相结合，在对留言信息进行处理之后，采用无监督的聚类算法将留言进行聚类，通过聚类结果将热点留言问题提取出来。通过搭建 LDA 主题模型选择出的五个主题，第一个

主题的高频词集中为“业主”，“开发商”，“政府”，“部门”，“物业”等词，第二个主题的高频词集中为“居民”，“影响”，“生活”，“环境”，“噪音”等词，第三个主题的高频词集中为“学校”，“学生”，“政策”，“教育”，“希望”等词，第四个主题的高频词集中为“工作”，“领导”，“办理”，“申请”，“材料”等词，第五个主题的高频词集中为“景园”，“车位”，“停车”，“车辆”，“领导”等词。

在对回复情况进行评价方面，在自然语言处理方法的支持下，利用文本相似性度量回复与留言问题相关程度，通过综合情况对回复情况做出评价方案。运用自然语言处理技术处理公众网络问政留言，在保证准确率的情况下科学高效地对留言内容进行分组，以便相关部门处理对应权职范围内的群众留言；通过自定义的热度评价指标对群众留言的热点问题进行筛选，以便职能部门可以迅速对热点问题进行了解与解决；针对留言问题的回复情况，通过考察回复的相关性与及时性等情况，制定合理的评价指标，对答复情况的质量给出评价。将自然语言处理技术与机器学习运用到网络问政平台中，不仅能够顺应网络问政的发展，而且可以及时处理网络问政留言，减轻人工处理的工作量节约人力物力财力，大规模降低智慧政务系统的工作成本。

对于本题的模型来讲，我们仅仅是按照题目要求，按照自己的理解给出了评价模型。本文还有许多值得深入思考的地方，譬如针对回复意见的解释性我们能否通过简单地建模解决所有针对性的问题，又比如为了改善普通网络问政中政府发生不足的问题，能否从扩大政府信息公开、提高公众知情权、参与权等权利的法律的保障、网络问政流程制度的完善等几个层面建立新的模型。

总得来说，自然语言处理与文本挖掘的方法在大数据时代的今天拥有极高的价值，结合机器学习与各种计算机语言，可以帮助各行各业的人们处理许多人工难以解决的问题。以热点问题挖掘来说，公司可以通过分析海量数据来针对受众客源制定利润最大化的运行方案，而现实中如日中天的大企业背后一定离不开海量的数据支撑。

最后感谢泰迪杯给我们提供了这样一个锻炼自己的机会，让我们有机会将理论与实践紧密结合，为今后的工作学习提供指导意义。无论成功与否都将受益匪浅。

参考文献

[1] 贺鸣,孙建军,成颖，南京大学，基于朴素贝叶斯的文本分类研究综述

[2] 吕红燕,冯倩，河北经贸大学，随机森林算法研究综述

[3] 杨锋，基于线性支持向量机的文本分类应用研究

[4] 祖弦,谢飞，LDA 主题模型研究综述，合肥师范学院学报

[5] 一文详解LDA主题模型[EB/OL]，https://zhuanlan.zhihu.com/p/31470216