

基于“智慧政务”的文本挖掘

摘要

随着微信、微博、市长信箱等网络交流工具成长，网络问政平台逐渐成为政府与相关部门与群众沟通重要渠道。与此同时，各类社情民意的网络文本数量呈现爆炸性增长，内容繁多且格式复杂，给以往用人工进行留言分类带来极大挑战。近年来自然语言处理(NLP)在人工智能领域飞速发展。因此，本文通过自然语言处理与深度学习技术，构建一套基于自然语言处理技术的文本挖掘应用，以此解决这类问题。

整个解题过程分为以下几个步骤：

第一步，对于文本发掘的模型中的留言数据以及答复意见进行数据预处理，先清晰明确训练的数据特征，再对数据进行去重处理、去停用词处理、去特殊字符的处理，保证训练的有效性。

第二步，选取附件 2 的 7 个一级标题中的各 30 个数据组成训练集，通过朴素贝叶斯算法建立模型。应用模型进行对于整个数据集进行主题匹配，匹配达 75.66%。

第三步，构建 TF-IDF 与 LDA 模型。对于留言中的热点问题进行挖掘，同时运用根据改良后的 BosonNLP_sentiment_score 评分模型进行和基于 Jaccard 的文本契合度计算方法，得出对于政府部门的答复详情的得分。

实验最后分析且评估了该文本系统的准确率与召回率，并简要说明了提升准确率与契合度的方法。

【关键词】 自然语言处理 文本分类 TF-IDF LDA 模型

目录

0 引言	3
1 问题分析.....	5
1.1 问题一的分析.....	5
1.2 问题二的分析.....	5
1.3 问题三的分析.....	5
2 数据预处理.....	6
2.1 数据描述	6
3 问题一求解.....	8
3.1 留言分类的主要思路.....	8
3.2 训练集和测试集.....	8
3.3 词频-逆向文件频率模型	9
3.3.1 TF-IDF 算法步骤	9
3.4 朴素贝叶斯分类	10
3.4.1 朴素贝叶斯算法的步骤.....	11
3.4.2 朴素贝叶斯的优缺点	11
4 问题二求解	12
4.1 热点问题挖掘主要思路	12
4.1.1 LDA 模型.....	12
4.2 点赞数占比.....	13
4.3 赋予权重计算得分.....	13
5 问题三求解	14
5.1 关键词得分主要思路:	14
5.1.1 数据预处理	14
5.1.2 留言意见完整性和可解释性: 引入关键词评分文档.....	14
5.2 留言意见相关性文本相似性: Jaccard 相似度.....	15
5.2.1 传统的 Jaccard 系数研究	16
6 评价指标和实验结果	17
6.1 评价指标	17

0 引言

当前，随着社会的不断发展，公众生活水平得到提高，社会诉求也发生了一系列的改变，包括公共服务、社会治理等方面。如微信、微博、市长信箱、阳光热线等网络问政平台逐步成为人民群众诉求的重要渠道。在大数据时代，各类社情民意相关的文本数据量不断攀升，这无疑对政府的社会治理和公共服务提出了更高的要求。反观长期以来，我国政府管理运作都处于人工处理的形态。在面对如此庞大的文本数据时，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。

智能政务系统在社会治理的应用为上述问题提供了解决方案。随着大数据、云计算、人工智能的飞速发展，建立基于自然语言处理技术的智慧政务系统已经是社会治理创新发展的新趋势，对提升政府的管理水平和施政效率具有极大的推动作用。

构建基于自然语言处理技术的文本分类模型，快速对文本数据进行分类，直接对群众留言进行处理，准确定位群众留言对应的社会板块，从而为群众提供有效的答复信息。构建基于自然语言处理技术的文本挖掘模型，从群众留言挖掘出热点问题，快速获知和答复群众的最关心的社会问题，形成高效的回复机制。构建基于自然语言处理技术的文本评价模型，不仅关注和回复群众的留言，还对政府部门的回复建立反馈评价模型，从回复的完整性、相关性、可解释性、合理性等评价对留言的回复，反馈回复存在的问题，希望能提高政府部门回复的有效性和有针对性，最后提高政务的处理效率。

基于对智能政务系统的理解和认识, 本文将立足于以上背景和问题, 构建基于 TF-IDF、LDA 主题、朴素贝叶斯分类器文本分类模型, 完成文本数据的分类。对题目所给的文本数据进行数据分析, 有针对性地对数据预处理, 在准确率和得分模型上都表现出良好的结果。

本文包括引言、问题分析、问题解决、实验结果、总结和展望五个部分。

1 问题分析

1.1 问题一的分析

根据附件 1 提供的内容分类三级标签和附件 2 的留言数据，首先运用 `sklearn` 进行特征处理，模型评估。接着使用 `TF-IDF` 算法处理训练集获得特征词矩阵，调用朴素贝叶斯分类器对附件 2 的留言数据进行一级标签分类，根据 `F-score` 对模型进行评估。

1.2 问题二的分析

根据附件 3 将某一时间段内反映特定地点或特定人群问题的留言进行归类和相似性处理，最后对热点问题的热度进行排名。主要利用 `TD-IDF`、`LDA` 模型得到文本关键词的词频，再结合点赞数占比，通过赋予这两部分合理的权重，得出热度较高的留言。

1.3 问题三的分析

针对附件 4 相关部门对留言的答复意见，建立一套留言评价体系。主要对答复意见语句的相关性、完整性、可解释性的角度进行评价。通过构建引入关键词得分的模型评价语句的完整性和可解释性和利用 `Jaccard` 系数评价答复意见和群众留言的相关性解决问题。

2 数据预处理

2.1 数据描述

通过观察所给数据，可以发现数据量比较大（共 15 万多条文本数据），给出的数据大多为文本格式，需要将其向量化才能对其进行分析。数据中有大量的空格以及重复的情况，如果不做处理会对后续分析造成影响。而且文本信息存在大量的噪声特征，如果把这些数据也引入分词、词频统计乃至文本相似度计算等，则必然会影响到模型的精度，所以本文首先要对数据进行初步的预处理。

步骤一：分词

由于词与词之间没有明显的界限是中文文本的特点，从文本中提取词语时需要分词，本文采用 `jieba` 分词[1]对文本数据进行中文分词。`jieba` 分词系统能提供分词、添加自定义词典、关键词提取、词性标注等功能。

部分分词结果示例：

```
In [1]: 小区楼下的烧烤摊无证经营，导致低楼层租户长期被油烟熏。
```

```
Out[1]: 小区, 楼下, 烧烤, 摊, 无证, 经营, 导致, 低楼层, 租户, 长期, 油烟,
```

步骤二 去除换行、空格符和标点符号等特定字符

```
In [1]: 患者用药后会出现剧烈寒颤、呕吐、心跳、呼吸停止现象 ..... ?
```

```
Out[1]: 患者 ,用药 ,出现, 剧烈, 寒颤, 呕吐, 心跳, 呼吸停止, 现象
```

步骤三 去除停用词

去除换行和空格符后，其中仍旧有大量表达无意义的字词。这些

字词会对后续分析造成极大的影响。在文本处理中，最常用功能性词语，如“的”、“一个”、“这”、“那”等，这些词语的使用较大的作用仅仅是协助一些文本的名词描述和概念表达，没有太大的实际意义。一般在预处理阶段就将停用词删除，以避免对文本，尤其是短文本，造成负面影响。

停用词表部分数据如图 1.3:

《	》	》),	」	『	』	【	】	{	})	{	(\)	一	一.	一
一下	一个	一些	一何	一切	一则	一天	一定	一方面							

停用词表的数据是按行存储的。如此一来，我们就可以定义一个统一的方法来读取留言和停用词表。

3 问题一求解

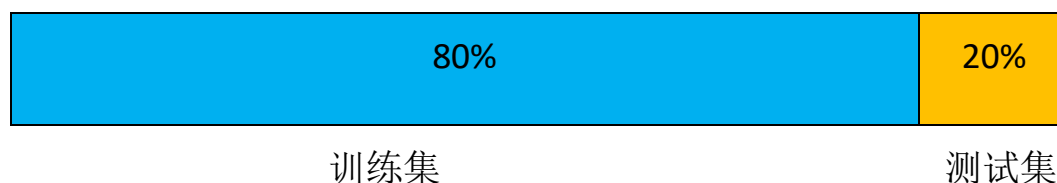
本文对问题一的求解主要将标签列为不同的数字，提供一定的训练集，并将对应标签的内容进行训练，最后得出模型对所有留言的分类。

3.1 留言分类的主要思路



3.2 训练集和测试集

训练集和测试集没有交集，都是从一堆的数据集中直接划分出两部分，一部分是训练集，另一部分就是测试集。如果只有一个数据集，可将数据集分开。训练集是用来建立模型的，而测试集是用来检验最终选择最优的模型的性能如何的。如下图 1.4



3.3 词频-逆向文件频率模型

TF-IDF 算法是一种针对关键词提取的统计分析方法，可以计算出一个词在一篇文章中的重要程度。这个表示重要程度的数值是由两部分相乘（ $TF \times IDF$ ）得到的。其中，**TF(term frequency)**：词频，表示一个词语与其在一篇文章出现次数的正相关性，可以表示为：某词在文章中出现的次数/文章的总词数。**IDF(inverse document frequency)**：逆向文件频率，表示一个词语与其在语料库出现的次数成反比，可以表示为 $\log(\text{语料库的文档总数}/\text{包含该词的文档数}+1)$ 。TF 与 IDF 的乘积越大，表示该特征词对这篇文章的重要性越大。例如"中国"这个词，在某一篇文章中多次出现，但是在所有文章中出现次数不大，因此"中国"这个词就更能体现出这篇文章的主题。

3.3.1 TF-IDF 算法步骤

1) 计算词频

$$\text{词频 (TF)} = \frac{\text{某个词再留言中的出现次数}}{\text{留言的总词数}}$$

2) 计算逆文档频率

$$\text{逆文档频率 (IDF)} = \log\left(\frac{\text{语料库的文档总数}}{\text{包含该词的文档数}+1}\right)$$

3) 计算 TF-IDF

$$\text{TF-IDF} = \text{词频 (TF)} \times \text{逆文档频率 (IDF)}$$

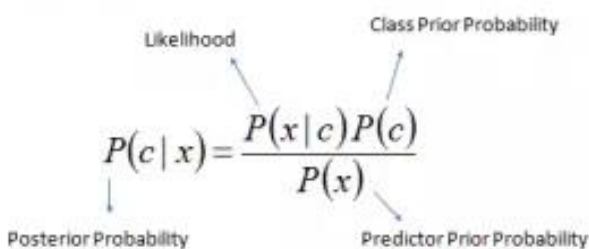
可以看到，TF-IDF 与一个词在文档中的出现次数成正比，与该词在整个语言中的出现次数成反比。所以，自动提取关键词的算法就很清楚了，就是计算出文档的每个词的 TF-IDF 值，然后按降序排列，取

排在最前面的几个词。

3.4 朴素贝叶斯分类

朴素贝叶斯分类器是一种基于贝叶斯定理的概率分类器，是应用最为广泛的分类算法之一。所有朴素贝叶斯分类器都假定样本每个特征与其他特征都不相关，也就是说特征条件是独立假设的。比如一种水果其具有红色、圆形和直径大约 3 英寸的特征，那么我们就可以判定其为苹果。尽管这些特征相互依赖或者依赖于其他特征决定，然而我们认为判定水果是否为苹果的这些特征是独立的。从中可以看出，朴素贝叶斯把问题简单化了，所以它有朴素两个字。朴素贝叶斯分类器很容易建立，且对大型数据库非常有用，这是一种胜过许多复杂算法的高效分类方法。

3.4.1 贝叶斯公式提供了计算后验概率 $P(X|Y)$ 的方式：



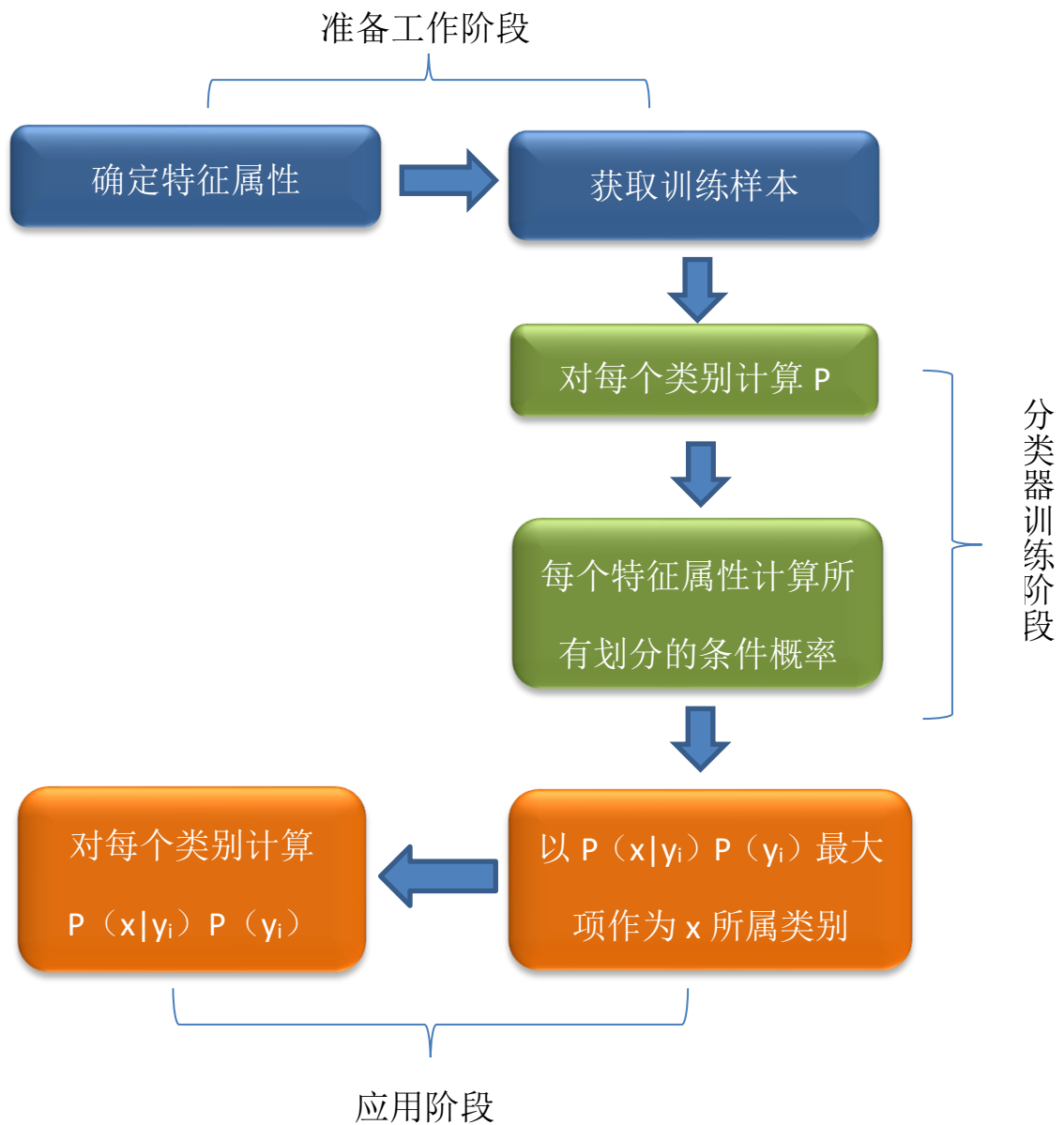
The diagram shows the formula $P(c|x) = \frac{P(x|c)P(c)}{P(x)}$ with arrows pointing from labels to the terms: 'Likelihood' points to $P(x|c)$, 'Class Prior Probability' points to $P(c)$, 'Posterior Probability' points to $P(c|x)$, and 'Predictor Prior Probability' points to $P(x)$.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$
$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

其中，

- 1) $P(c)$ 是样本 c 的概率，即先验概率，用极大似然法可得。
- 2) $P(x)$ 是样本 x 的概率。
- 3) $P(x|c)$ 是样本 c 的条件下求样本 x 的概率。
- 4) $P(c|x)$ 是已知某样本(c ，目标)，(x ，属性)的概率，即后验概率。

3.4.1 朴素贝叶斯算法的步骤



3.4.2 朴素贝叶斯的优缺点

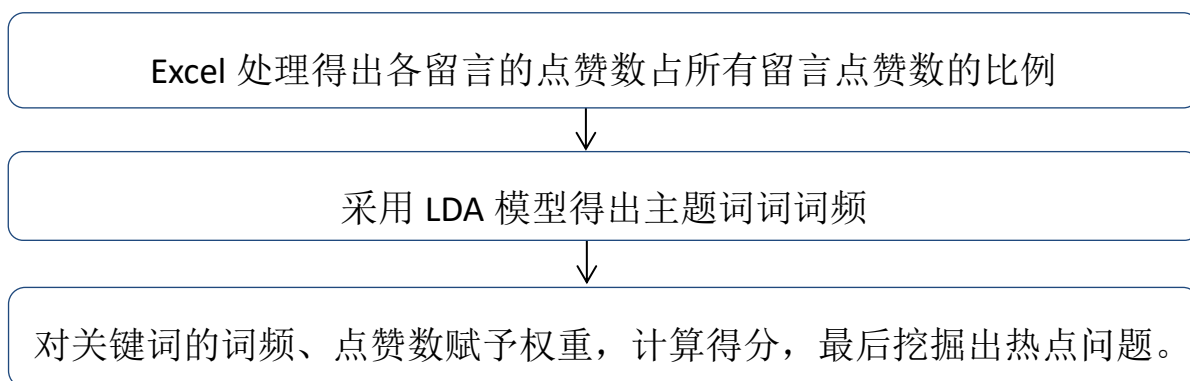
优点：①简单快速，预测表现良好；②对确实数据不敏感，算法简单
③能够处理多分类任务。

缺点：①需要特征条件相互独立，这在实际应用中往往是不成立的；
②特征条件较多和其相关性较大时会影响分类效果。

4 问题二求解

本文对问题二的求解主要采用 LDA 主题模型和 TF-IDF 得出关键词频,对关键词词频和点赞数,建立权重得分模型进行热点问题挖掘,运用关键词提取对留言进行分类。

4.1 热点问题挖掘主要思路



4.1.1 LDA 模型

LDA (Latent Dirichlet Allocation) 是一种文档主题生成模型, 也称为一个三层贝叶斯概率模型, 包含词、主题和文档三层结构。我们认为每段留言的每个词都是通过“以一定概率选择了某个主题, 并从这个主题中以一定概率选择某个词语”这样一个过程得到。文档到主题服从多项式分布, 主题到词服从多项式分布。

它是一种非监督机器学习技术, 可以用来识别大规模文档集 (document collection) 或语料库 (corpus) 中潜藏的主题信息。它采用了词袋 (bag of words) 的方法, 这种方法将每一段留言视为一个词频向量, 从而将文本信息转化为了易于建模的数字信息。每一段留言代表了一些主题所构成的一个概率分布, 而每一个主题又代表了很多单词所构成的一个概率分布。

4.2 点赞数占比

表格为反对数和点赞数总占比和点赞数占比（排名前 20）

留言主题	总占比	点赞数占比
反映 A 市金毛湾配套入学的问题	16.17%	84.02%
请书记关注 A 市 A4 区 58 车贷案	7.51%	39.15%
严惩 A 市 58 车贷特大集资诈骗案保护伞	7.23%	37.67%
承办 A 市 58 车贷案警官应跟进关注留言	6.71%	34.95%
A4 区绿地海外滩小区距长赣高铁最近只有 30 米不到,合理吗?	6.12%	31.90%
A 市富绿物业丽发新城强行断业主家水	2.21%	11.54%
建议西地省尽快外迁京港澳高速城区段至远郊	0.73%	3.81%
请问 A 市为什么要把和包支付作为任务而不让市场正当竞争?	0.71%	3.72%
关于 A6 区月亮岛路沿线架设 110kv 高压线杆的投诉	0.71%	3.72%
A 市三一大道全线快速化改造何时启动?	0.60%	3.15%
A 市长房云时代多栋房子现裂缝,质量堪忧	0.55%	2.62%
关于 A6 区月亮岛路 110kv 高压线的建议	0.52%	2.62%
建议 A 市经开区收回东六路恒天九五工厂地块,打造商业综合体	0.40%	2.10%
建议加大 A7 县东六线榔梨段拆迁力度	0.39%	2.00%
反映 A 市地铁 3 号线松雅湖站点附近地下通道问题	0.38%	2.00%
A3 区郝家坪小学什么时候能改扩建?	0.38%	1.96%
问 A 市经开区东六线以西泉塘昌和商业中心以南的有关规划	0.38%	1.96%
建议 A7 县尽快出让星沙滨湖路以南,特立路以北的土地	0.41%	1.91%
关于加快修建 A 市南横线的建议	0.36%	1.86%
希望 A 市地铁四号线北延线“同心路站”设在雷峰大道上	0.35%	1.81%

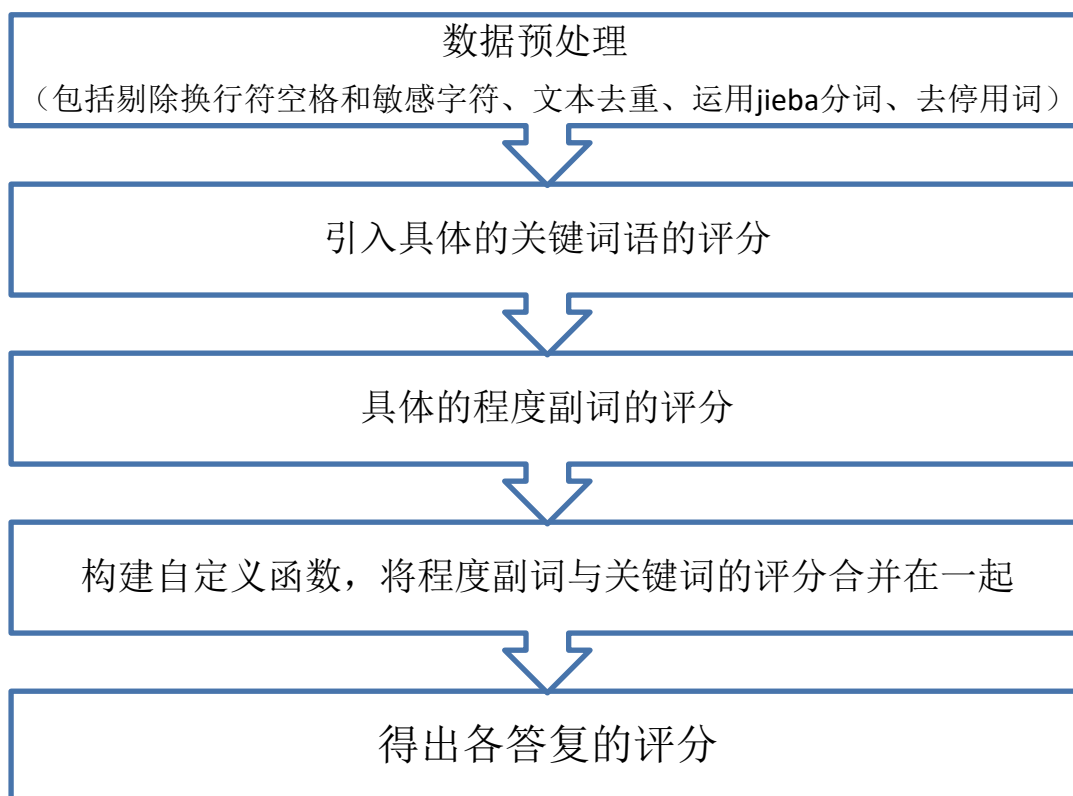
4.3 赋予权重计算得分

权重采用关键词词频得分占 40%，总站比得分占 60%，计算地出前五的热点问题。

5 问题三求解

本文对问题三的求解主要采用关键词得分和文本相似得分

5.1 关键词得分主要思路：



5.1.1 数据预处理

第一步 导入 pandas re jieba 这三个库

第二步 规定元素的范围

第三步 引入附件 4 引入国标 GB18030 作为对于附件 4 的编译

第四步 去除空格符、换行符、去重、结巴分词、去除停用词和去除空白值

5.1.2 留言意见完整性和可解释性：引入关键词评分文档

留言意见完整性即语句语义的完整性和可解释性。先通过语句语义完整性可从是否包括礼貌用语，公文常用语，列出相关的法律法规

等正向词语，是否包括含糊不清和指代不明的负向词语（如相关部门）等内容设置得分标准。可解释性主要通过留言答复中是否涉及列明法律法规、指明相关负责单位、给出咨询方式（如致电、邮箱、网站等）。

关键词获取是通过 excel 筛选出现超过 10 次以上的词语，再把涉及到的法律法规和一般的公文用语比较语义得出分数高低。

部分词语得分如下：	得分
《C 市教育局关于做好 2020 年全市中小学幼儿园寒假工作的通知》/《土地纠纷调解仲裁法》等法律法规	4
您好/谢谢/依法/省厅文件精神	4
你好/咨询电话/致电/加强//推进/监管/核实/办复/限期整改	3
城管局/转交/关心/监督/支持/收悉/高度/重视/反映/调查/了解	2
登记/下达/争取/涉及/预计/审批/关注	1
随意/相关部门	-2

观察到附件 4 的答复以正向词语为主，故采取得分为主的计算方法，统计答复语句的得分情况，得到留言意见的完整性和可解释性的评分。

5.2 留言意见相关性文本相似性：Jaccard 相似度

文本相似性计算有很多方法，通常采用计算样本间的“距离”，如编辑距离、TF-IDF、Jaccard 系数（杰卡德距离）、余弦相似性。编辑距离只考虑句子之间转换的操作次数，忽视了语义差异；TF-IDF 要建立一个大规模的语料库基础上；在实际应用中数据稀疏度过高，通过余弦相似度计算会产生误导性结果。又考虑到留言和答复在语义上

有对应性和相似性，故采用字重叠法即 Jaccard 系数。字重叠法基本思想是通过两个句子词汇交集和词汇并集的比值作为句子相似度

5.2.1 传统的 Jaccard 系数研究

Jaccard index, 又称为 Jaccard 相似系数(Jaccard similarity coefficient) 用于比较有限样本集之间的相似性与差异性。Jaccard 系数值越大，样本相似度越高。

系数定义：给定两个集合 A, B, Jaccard 系数定义为 A 与 B 交集的大小与 A 与 B 并集的大小的比值，定义如下：

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

当集合 A, B 都为空时，J(A,B)定义为 1。

与 Jaccard 系数相关的指标叫做 Jaccard 距离，用于描述集合之间的不相似度。Jaccard 距离越大，样本相似度越低。公式定义如下：

$$d_j(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} = \frac{A \Delta B}{|A \cup B|}$$

6 评价指标和实验结果

6.1 评价指标

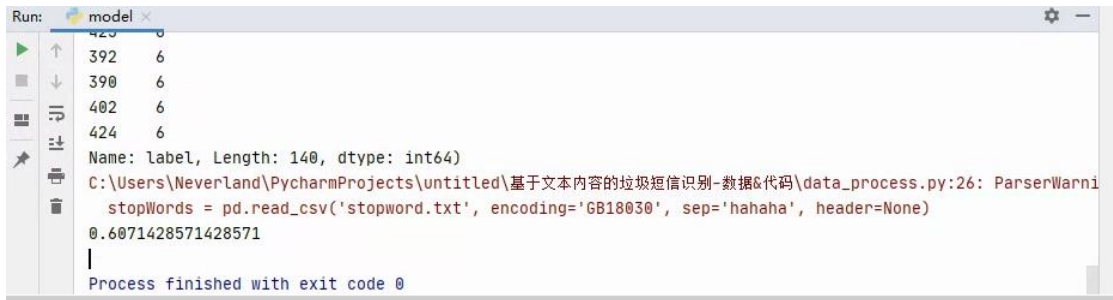
本文采用的评价指标主要包括准确率和 F1-score。准确率又称精度，可表示为：检索出的文档总数/检索出的文档总数，其衡量的的是一个模型的好坏。而单独使用准确率是不够的，于是我们就引入了 F1-score 作为判别模型好坏的另一个指标，其常用于文本分类项目的分类结果评估。F1-score 的公式为：

$$F_1 = \frac{1}{n} \sum_{i=1}^n \frac{2P_i R_i}{P_i + R_i}$$

其中 P_i 为第 i 类的查准率， R_i 为第 i 类的查全率。

6.2 问题一实验结果

分类器精度：0.6071428571428571



```
Run: model x
392 6
390 6
402 6
424 6
Name: label, Length: 140, dtype: int64)
C:\Users\Neverland\PycharmProjects\untitled\基于文本内容的垃圾短信识别-数据&代码\data_process.py:26: ParserWarning:
stopWords = pd.read_csv('stopword.txt', encoding='GB18030', sep='hahaha', header=None)
0.6071428571428571
|
Process finished with exit code 0
```

6.3 问题二实验结果

热点问题前五

留言主题
反映 A 市金毛湾配套入学的问题
请书记关注 A 市 A4 区 58 车贷案
严惩 A 市 58 车贷特大集资诈骗案保护伞
承办 A 市 58 车贷案警官应跟进关注留言
A4 区绿地海外滩小区距长赣高铁最近只有 30 米不到，合理吗？

6.3 问题三实验结果

	A	B	C	D	E	F	G	H	I	J	K
1	留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	分数	文本关联	总分数	
2	2549	A0004558(A2区景容		2019/4/25 9:32	2019年4月25日	现将网友在平台	2019/5/10 14:56	4	12.56	7.424	
3	2554	A0002358(A3区蒲楚		2019/4/24 16:03		网友“A00023583	2019/5/9 9:49	1	2.96	1.784	
4	2555	A0003161(请加快提		2019/4/24 15:40		市民同志：你好！	2019/5/9 9:49	9	11.92	10.168	
5	2557	A0001107(在A市买公		2019/4/24 15:07		网友“A00011073	2019/5/9 9:49	2	9	4.8	
6	2574	A0009233 关于A市公		2019/4/23 17:03		网友“A0009233”	2019/5/9 9:51	0	15.69	6.276	
7	2759	A0007753(A3区含浦		2019/4/8 8:37		网友“A00077538	2019/5/9 10:02	14	0.95	8.78	
8	2849	A0001008(A3区教师		2019/3/29 11:53		网友“A00010080	2019/5/9 10:18	1	7.36	3.544	
9	33970	A0001002(B7县二中		2018/8/12 10:56		网友：您好！现	2018/8/20 9:25	0	5.97	2.388	
10	33978	A0004458(B市601小		2018/8/8 13:15		网民您好：您反	2018/8/17 9:49	5	10.31	7.124	
11	33984	A0009105(咨询在B		2018/8/3 21:26		网民您好：您反	2018/8/7 9:13	8	3.08	6.032	
12											

7 总结和展望

7.1 总结

在面对信息爆炸的政务留言信息，传统的人工分类出错率高，误差大，计算机处理大数据显得十分迫切。近年来人工智能的飞速发展，使得自然语言处理技术也得到很大的提高。各政府部门纷纷建立智慧政务平台，极大地解决了问题。智慧政务构建基于自然语言处理技术的文本分类模型，力求寻求到文本数据快速高效分类，准确定位群众留言对应的社会板块，从而为群众提供有效的答复信息。

为了更好地处理留言和答复意见，我们首先进行数据预处理，接着基于 TF-DIF 权重法提取特征词，然后通过朴素贝叶斯建立模型。通过使用基本的模型，运用关键词、计算词频、生成主题词、朴素贝叶斯分类器、Jaccard 相似度解决文本分类问题、文本相似度问题。

我们根据研究思路撰写论文，通过实验验证了本模型的可行性，基本实现本赛题设立的目标。

7.2 展望

1.在进行 Jaccard 相关系数对比文本相似性时，不仅仅单一对比

词语的交集得出相似性分数。

2.在问题二的分类问题上，希望能得到更好的关键词。

参考文献

[1] <https://github.com/fxsjy/jieba>

[2] 刘志. 基于用户兴趣的协同过滤算法的广告推荐研究[D]. 昆明理工大学, 2014.