基于自然语言处理和文本挖掘的网络问政平台文本分析

摘要

近几年，由于网络的高速发展以及政府对于民众的关注度增强，微信、微博、市长信箱、阳光热线等网络问政平台相继出现，成为了民众向政府反映问题的重要渠道，但与此同时，也带来了各类社情民意相关的文本数据量不断攀升的问题，单纯依靠人工已经远远无法满足文本处理的要求，而在大数据蓬勃发展的今天，搭建智慧政务系统成为了解决问题的有效途径。本文将基于数据挖掘技术对群众留言进行挖掘和分析。

在本次数据挖掘的过程中，首先对所给的评论数据利用 python 进行数据预处理，分词即停用词过滤等操作，实现对居民意见数据的优化，提升其可建模性。

接着我们利用 IF-IDF 进行词向量转化，划分训练集和测试集，之后用高斯朴素贝叶斯模型进行训练，评价分类方法的好坏采用 F-Score 方法。对于第二个问题，先采用 k 均值算法对于数据进行分类，划分为 5 组后，再采用 LDA 的模型构造将每个组的主题按照概率分布的形式给出，通过筛选这些关键词可以找出居民所反映的热点问题，分别是“学校周围交通”“小区垃圾站”“车位和住房捆绑销售”“工地夜间施工”“居民用水困难”。

最后，针对政府的答复意见，从相关性、完整性、可解释性等角度给出评价模型。利用 IF-IDF 进行词向量的转化后，即可进行统计，比较，计算等多种操作

关键词 文本分析；信息提取；语义网络；LDA

Abstract

In recent years, due to the rapid development of network and the government for public awareness enhancement, WeChat, weibo, mayor mailbox, sunshine hotline ask ZhengPing appeared, such as network become an important channel for people to reflect the problem to the government, but at the same time, also brought all kinds of public opinion related to the amount of text data rising problems, rely on artificial far cannot satisfy the requirement of text processing, and in the vigorous development of big data today, building e-government system become the wisdom the effective way to solve the problem.This paper will be based on the data mining technology to the mass message mining and analysis.

In the process of data mining, python was used to preprocess the given comment data, and word segmentation was used to filter the stop words, so as to optimize the residents' opinion data and improve its modelability.

Then, we use if-idf to transform the word vector, divide the training set and test set, and then use gauss naive bayesian model for training, and use f-score method to evaluate the classification method.For your second question, first using k-means algorithm for data classification, divided into 5 groups, and then USES the LDA model construction will be the theme of each group according to the given in the form of probability distribution, by sifting through these keywords can find residents reflects the hot issues, respectively is "traffic around the school" dump "village" "parking and housing tying" night "site construction" "residential water difficult".

Finally, according to the government's response, an evaluation model is proposed from the perspectives of relevance, completeness and interpretability.IF-IDF is used for word vector transformation, statistics, comparison, calculation and other operations can be performed

Keywords text analysis;Information extraction;Semantic web;LDA

1. 问题背景与目标

1.1.问题背景

网络的日益普及和迅速发展使得互联网在中国民众的政治、经济和社会生活中扮演着日益重要的角色，中国公民以网民的身份通过互联网行使知情权、参与权、表达权和监督权，而微信、微博、市长信箱、阳光热线等网络问政平台逐步成为政府了解民意、汇聚民智、凝聚民气的重要渠道。

但是随着各类社情民意相关的文本数据量不断积累和攀升，如果继续按照过去主要依靠人工来进行留言划分和热点整理的相关部门的工作，其效率和准确性显然已经无法满足当下的政务需要。因此利用大数据、云计算、人工智能等技术建立智慧政务系统从而提升政府的管理水平和施政效率，已成为当今社会治理创新发展的新趋势。

因此，预训练的模型表示可以通过一个额外的输出层进行微调，适用于广泛任务的最先进模型的构建，比如问答任务和语言推理，无需针对具体任务做大幅架构修改。针对网络问政平台的群众留言分类、热点问题挖掘和答复意见的评价等问题，如何对群众留言数据进行优化处理，如何建立高效准确的文本分类器模型，是实现智慧政务系统的技术关键。

为此，恰当合理的提取文本数据特征，结合自然语言处理和文本挖掘的技术创新政务发展方式，实现资源整合和信息共享，对电子政务的发展和打造未来智慧型政府具有重要意义。

1.2.问题目标

(1) 针对网络政务平台的群众留言数据，在对文本进行数据清洗预处理、中文分词、停用词过滤后，提取文本特征实现文本的语义表示，通过建立朴素模型和对其进行预训练后，实现对留言内容的一级标签分类和对分类方法的评估。

(2) 在第一步建立的文本分类器的基础上，实现对留言内容的主题摘要和聚类，并定义合理的热度评价指标，建立起热点问题的挖掘模型。

(3) 从文本内容的相关性、完整性、可解释性等角度出发，建立对留言的答复意见的质量评价模型。

2. 分析方法与过程

2.1.问题一的分析方法与过程

2.1.1. 文本留言预处理

文本留言数据里面存在大量信息价值很低甚至没有信息价值的条目，如果将这些评论数据也

引入进行分词、词频统计乃至关键词和主题提取等，则必然会对分析造成很大的影响，得到的结果的准确性也必然是存在问题的。那么在利用到这些文本评论数据之前就必须要先进行文本预处理，把大量的这些无信息价值的评论去除。

2.1.2. 文本评论分词

在中文中，只有字、句和段落能够通过明显的分界符进行简单的划界，而“词”和“词组”的边界模糊，没有一个明确的形式上的分界。所以在进行中文文本挖掘时，首先应对文本分词，即将连续的字序列按照一定的规范重新组合成词序列的过程。分词结果的准确性对后续文本挖掘算法有着不可忽视的影响，如果分词效果不佳，即使后续算法优秀也无法实现理想的效果。例如在特征选择的过程中，不同的分词效果，将直接影响词语在文本中的重要性，从而影响特征的选择。

2.1.3. 停用词过滤

经过中文分词这一步骤，将初始的文本处理成为词的集合，但是文本中，会存在大量的虚词、代词或者没有特定含义的动词、名词，这些词语对文本分析起不到任何的帮助，我们往往希望能去掉这些“停用词”。
停用词的两个特征为：一是极其普遍、出现频率高；二是包含信息量低，对文本标识无意义。例如中文中的“的”、“了”、“地”、“啊”等但是在停用词的去除中，应注意要保留其中的否定词，可以对停用词表进行人工筛选相结合的方式，对停用词进行处理我们首先训练以得到词向量。

2.1.4. 训练生成词向量

对每条留言抽取特征 TF-IDF 值等构成特征向量。

```
Out[9]: array([[0., 0., 0., ..., 0., 0., 0.],
               [0., 0., 0., ..., 0., 0., 0.],
               [0., 0., 0., ..., 0., 0., 0.],
               ...,
               [0., 0., 0., ..., 0., 0., 0.],
               [0., 0., 0., ..., 0., 0., 0.],
               [0., 0., 0., ..., 0., 0., 0.]])
```

然后采用朴素贝叶斯法进行分类。

2.1.5. 过程和结果展示

1. 分词结果

	zhuti	分词
0	西湖建筑集团占道施工有安全隐患	西湖/ 建筑/ 集团/ 占/ 道/ 施工/ 有/ 安全隐患
1	在水一方大厦人为烂尾多年安全隐患严重	在水一方/ 大厦/ 人为/ 烂尾/ 多年/ 安全隐患/ 严重
2	投诉苑物业违规收停车费	投诉/ 苑/ 物业/ 违规/ 收/ 停车费
3	蔡锷南路华庭楼顶水箱长年不洗	蔡锷/ 南路/ 华庭/ 楼顶/ 水箱/ 长年/ 不洗
4	华庭自来水好大一股霉味	华庭/ 自来水/ 好大/ 一股/ 霉味
5	投诉盛世耀凯小物业无故停水	投诉/ 盛世/ 耀凯/ 小/ 物业/ 无故/ 停水
6	咨询楼盘集中供暖一事	咨询/ 楼盘/ 集中/ 供暖/ 一事
7	桐梓坡西路可可小城长期停水得不到解决	桐梓/ 坡/ 西路/ 可可/ 小城/ 长期/ 停水/ 得不到/ 解决
8	反映收取城垃圾处理费不平等的问题	反映/ 收取/ 城/ 垃圾处理/ 费/ 不/ 平等/ 的/ 问题
9	魏家坡小脏乱差	魏家坡/ 小/ 脏乱差
10	魏家坡小脏乱差	魏家坡/ 小/ 脏乱差
11	泰华一村小第四届非法业委会涉嫌侵占小业主公共资金	泰华/ 一村/ 小/ 第四届/ 非法/ 业委会/ 涉嫌/ 侵占/ 小业主/ 公共/ 资金
12	梅溪湖壹号御湾业主用水难	梅/ 溪湖/ 壹号/ 御湾/ 业主/ 用水/ 难
13	鸿涛翡翠湾强行对入住的业主关水限电	鸿涛/ 翡翠/ 湾/ 强行/ 对/ 入住/ 的/ 业主/ 关水/ 限电
14	地铁号线施工导致锦楚国际星城小三期一个月停电来次	地铁/ 号线/ 施工/ 导致/ 锦楚/ 国际/ 星城/ 小/ 三期/ 一个月/ 停电/ 来次
15	润和紫郡用电的问题能不能解决	润/ 和/ 紫/ 郡/ 用电/ 的/ 问题/ 能/ 不能/ 解决
16	锦楚国际新城从月份开始停电好多次	锦楚/ 国际/ 新城/ 从/ 月份/ 开始/ 停电/ 好/ 多次
17	给城南西片城铁站设立的建议	给/ 城南/ 西片/ 城铁/ 站/ 设立/ 的/ 建议
18	请政府加大对滨水新城的绿化建设	请/ 政府/ 加大/ 对滨水/ 新城/ 的/ 绿化/ 建设
19	楚府线几个小经常停电	楚府/ 线/ 几个/ 小/ 经常/ 停电
20	请调查西地省建望集团及西地省辉东安建工程有限公司的违法行为	请/ 调查/ 西地省/ 建望/ 集团/ 及西地/ 省辉/ 东安/ 建/ 工程/ 有限公司/ 的/ 违法行为
21	山水嘉园栋三单元群租房扰民	山水/ 嘉园栋/ 三/ 单元/ 群/ 租房/ 扰民

2. 去停用词结果

	zhuti	分词
0	西湖建筑集团占道施工有安全隐患	西湖 建筑 集团 占 道 施工 安全隐患
1	在水一方大厦人为烂尾多年安全隐患	在水一方 大厦 人为 烂尾 多年 安全隐患 严重
2	投诉苑物业违规收停车费	投诉 苑 物业 违规 收 停车费
3	蔡锷南路华庭楼顶水箱长年不洗	蔡锷 南路 华庭 楼顶 水箱 长年 不洗
4	华庭自来水好大一股霉味	华庭 自来水 好大 一股 霉味
5	投诉盛世耀凯小物业无故停水	投诉 盛世 耀凯 物业 无故 停水
6	咨询楼盘集中供暖一事	咨询 楼盘 集中 供暖 一事
7	桐梓坡西路可可小城长期停水得不到	桐梓 坡 西路 可可 小城 长期 停水 得不到 解决
8	反映收取城垃圾处理费不平等的问题	反映 收取 城 垃圾处理 费 平等 问题
9	魏家坡小脏乱差	魏家坡 脏乱差
10	魏家坡小脏乱差	魏家坡 脏乱差
11	泰华一村小第四届非法业委会涉嫌侵	泰华 一村 第四届 非法 业委会 涉嫌 侵占 小业主 公共 资金
12	梅溪湖壹号御湾业主用水难	梅 溪湖 壹号 御湾 业主 用水 难
13	鸿涛翡翠湾强行对入住的业主关水限	鸿涛 翡翠 湾 强行 入住 业主 关水 限电
14	地铁号线施工导致锦楚国际星城小三	地铁 号线 施工 导致 锦楚 国际 星城 三期 一个月 停电 来次
15	润和紫郡用电的问题能不能解决	润 紫 郡 用电 问题 不能 解决
16	锦楚国际新城从月份开始停电好多次	锦楚 国际 新城 月份 停电 多次
17	给城南西片城铁站设立的建议	城南 西片 城铁 站 设立 建议
18	请政府加大对滨水新城的绿化建设	请 政府 加大 对滨水 新城 绿化 建设
19	楚府线几个小经常停电	楚府 线 几个 经常 停电
20	请调查西地省建望集团及西地省辉东	请 调查 西地省 建望 集团 及西地 省辉 东安 建 工程 有限公司 违法行为
21	山水嘉园栋三单元群租房扰民	山水 嘉园栋 三 单元 群 租房 扰民

```
Out[9]: array([[0., 0., 0., ..., 0., 0., 0.],
               [0., 0., 0., ..., 0., 0., 0.],
               [0., 0., 0., ..., 0., 0., 0.],
               ...,
               [0., 0., 0., ..., 0., 0., 0.],
               [0., 0., 0., ..., 0., 0., 0.],
               [0., 0., 0., ..., 0., 0., 0.]])
```

3. 使用 F-Score 对分类方法进行评价结果

```
model=GaussianNB ()         #模型
model.fit(X_tr,labels_tr )   #对模型进行训练
model.score(X_te,labels_te )   #对模型进行检测
```

```
Building prefix dict from the default dictionary ...
Loading model from cache C:\Users\64430\AppData\Local\Temp\jieba.cache
Loading model cost 0.702 seconds.
Prefix dict has been built successfully.
```

```
Out[3]: 0.7642857142857142
```

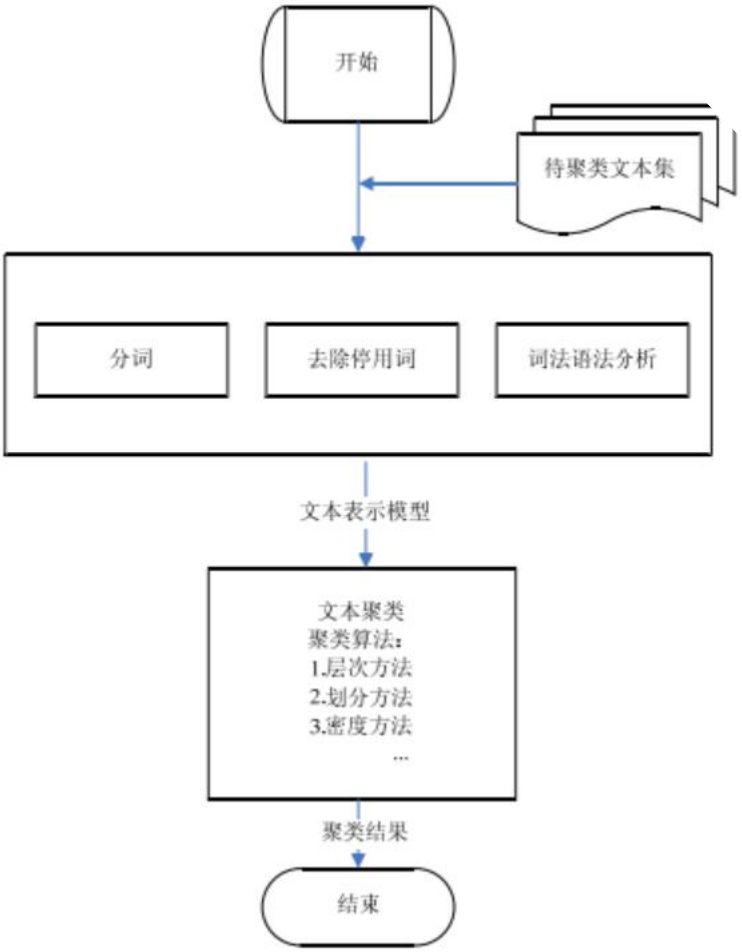
2.2.问题二的分析方法与过程

2.2.1. 问题的三要素提取

按照数据预处理模块中对群众留言信息进行的分词、关键词生成结果，提取问题的三要素，即群众留言内容中包含的特定时间、特定地点以及发生的问题。

2.2.2. 问题归并

通过基于文本相似度计算的文本聚类算法实现主题相似的留言问题的聚类和归并



2.2.3. 量化评价指标

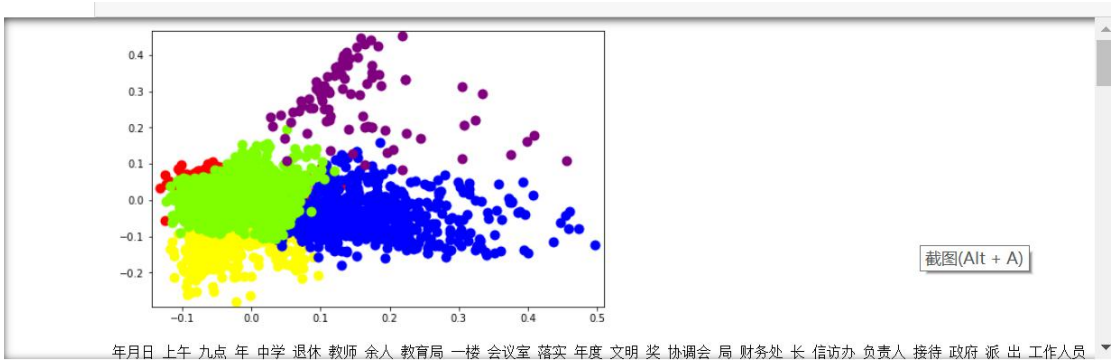
根据留言内容的文本特征，我们选取留言时间和关键词内容出现的频次作为热度的评价标准。

将每个文档的主题按照概率分布的形式给出每条留言详情中的词语按照一定的概率先选择一个主题，再通过这个主题按照一定的概率去选择词语。

```
shujv_model.print_topic(0)
 '0.018*"到户" + 0.015*"解决" + 0.010*"自来水" + 0.010*"城" + 0.010*"更换" + 0.010*"政府" + 0.010*"水泥路" + 0.008*"外排"
```

```
shujv_model.print_topic(0)

 '0.024*"车位" + 0.018*"捆绑" + 0.016*"认购" + 0.016*"景园" + 0.016*"滨河" + 0.016*"苑" + 0.013*"销售" + 0.013*"购房" + 0.013*"伊"
```



2.3.问题三的分析方法与过程

2.3.1. 答复的相关性

从答复的相关性来分析答复的质量，采用计算答复意见和留言详情的匹配程度，在对答复意见和留言详情进行完分词之后，开始进行词语的配对，计算词重复出现的次数以及两组信息中匹配的次数作为相关性的衡量标准。

例：我是不是你

　　我不是你

“是”在第一个句子中出现了两次，在第二个句子中出现了一次，那么“是”的相关度为 2，其余“我”“不”“是”的相关度都为 1，则该两个句子的相关度为 5，设 d=相关度，l=总数，则相关性 r=d/l,例子中 r=5/9,由此可以计算相关性。

对于答复意见的相关性，在对答复意见和留言详情进行完计算后，对于每条信息进行平均，由此给出评判标准。

r<0.05	0.05<r<0.1	0.1<r<0.15	0.15<r<0.2
相关性差	相关性一般	相关性良好	相关性好

2.3.2. 答复的完整性

基于词重叠率的评价来分析答复的质量。

首先对留言详情和答复意见分别进行数据的预处理。对所得的答复意见数据项进行 n-gram 的精度计算。公式如下：

$$P_n(r,\hat{r}) = \frac{\sum_k \min(h(k,r), h(k,r_i))}{\sum_k h(k,r_j)}$$

其中，Pn 代表数据 n-gram 的精度，r 代表留言详情，rj 是指答复意见，k 是每一个长度为 n 的序列 n-gram 值，h(k,r)表示在留言详情种 k 的个数。
最终所计算的 Pn 值代表的就是二者的重合程度，得到的数值越大，则表示答复意见中含括的留言详情内容更多，完整性也就越高。

2.3.3. 答复的可解释性

对答复意见的数据预处理后，主要统计某些专有性名词的词频数量和所占之比。比如‘法律’、‘规定’、‘会议’等词汇的出现，可以提高答复意见的可解释性。
我们需要提前将这些词汇提取，再分别对每一条答复意见进行词向量的构造，主要利用 one-hot 向量，统计所得的词向量非零个数。依靠非零个数的数量来评判可解释性的高低。

参考文献

[1] 吴柳，程恺，胡琪. 基于文本挖掘的论坛热点问题时变分析[J]. 软件，2017，38（4）
[2] 耿倩. 基于基于文本相似度计算的文本聚类算法研究与实现[D].哈尔滨：哈尔滨工业大学,2010
[3] 朱少杰. 基于深度学习的文本情感分类研究[D].哈尔滨：哈尔滨工业大学,2014
[4] 石晶，范猛，李万龙.基于 LDA 模型的主题分析[J].自动化学报，2009，35（12），1587-1592