“智慧政务”中的文本挖掘应用

摘要

在如今这个大数据时代，互联网的高速发展下的信息交互越来越频繁，网上平台已成为政府了解民意的重要渠道。公共便民业务持续完善，网络民生交互渠道的多样化，反映民情文本数据量剧增，使得如何处理民众网上的留言已成为难题，“智慧政务”随之运营而生，对民众留言的高效处理有着重要的意义。

对于问题一，在数据预处理阶段通过 jieba 分词工具对附件二的留言详情进行中文分词，再进行停用词过滤、删除文本中的标点符号和特殊符号的预处理。然后使用 Gensim 的 Doc2Vec 算法对数据进行量化处理，配置分布式词袋(DBOW)模型，接着创建文本的特征向量，再用 sklearn 的逻辑回归进行预测，最后使用 F-Score 对建立的模型进行评估。

对于问题二，我们先对附件三的留言主题进行中文分词，过滤停用词，接着在热点话题的挖掘过程中，将分词后的文本的词语用 CountVectorizer 函数转化为词频矩阵，统计词频，用 TfidfTransformer 函数统计词频矩阵中每个词语出现的 TF-IDF 值。然后，采用 K-Means 方法对 TF-IDF 权重向量进行聚类，通过观察聚类后的结果对停用词进行修改从而改进结果，反复迭代后选取出最优的分类结果并作统计。

对于问题三，我们对附件 4 的留言详情和留言答复进行分词，用 dictionary 方法获取分词后的留言详情中的词语放进词袋并将词袋中的所有词进行编号，然后使用 doc2bow 制作语料库把文本转换为二元组（编号、频次数）的向量。使用 TF-IDF 模型对语料库进行建模，对每一条答复分析整体答复意见的相关性。

关键词：中文分词；Doc2Vec；分类模型；K-Means 聚类；TF-IDF 算法

Abstract

In today's era of big data, with the rapid development of the Internet, information interaction becomes more and more frequent, and online platform has become an important channel for the government to understand public opinion. With the continuous improvement of the public convenience business, the diversification of the interactive channels for people's livelihood on the network, and the sharp increase of the amount of text data reflecting the people's situation, how to deal with the message on the public network has become a difficult problem, "smart government" comes into being with the operation, which is of great significance for the efficient processing of the message.

For the first problem, in the data preprocessing stage, we use the Jieba word segmentation tool to segment the message details of Annex II in Chinese, and then filter the stop words, delete punctuation and special symbols in the text. Then we use the doc2vec technology of gensim to quantify the data, configure the distributed word bag (dbow) model, then create the feature vector of the text, then use the logic regression of sklearn to predict, and finally use F-score to evaluate the established model.

For question 2, we first segment the message subject in Annex 3 in Chinese and filter the stop words. Then, in the process of mining hot topics, we transform the words of the segmented text into word frequency matrix with countvectorizer function, count word frequency, and count the TF-IDF value of each word in the word frequency matrix with tfiftransformer function. Then, the weight vector of TF-IDF is clustered by K-Means method, and the stop words are modified by observing the clustering results to improve the results. After repeated iterations, the optimal classification results are selected and counted.

For question 3, we segment the message details and response in Annex 4. We use the dictionary method to get the words in the message details after segmentation and put them into the word bag and number all the words in the word bag. Then we use doc2bow to make the corpus to convert the text into a vector of two tuples (number, frequency). The TF-IDF model is used to model the corpus, and the relevance of the overall response is analyzed for each response.

Key words:Chinese participle,Doc2Vec,Classification model,K-Means clustering, TF-IDF algorithm

目录

一、问题重述

1. 根据附件 2 中的群众留言的数据，建立关于留言内容的一级标签分类模型，对分类方法进行评价。

2. 根据附件 3 的内容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，定义合理的热度评价指标，并给出评价结果。在一个“热点问题表.xls”文件中，按照相应的格式给出排名前 5 的热点问题。在一个“热点问题表留言明细表.xls”文件中，按照相应的格式给出相应热点问题对应的留言信息。

3. 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。

二、问题一分析方法与过程

1. 流程图

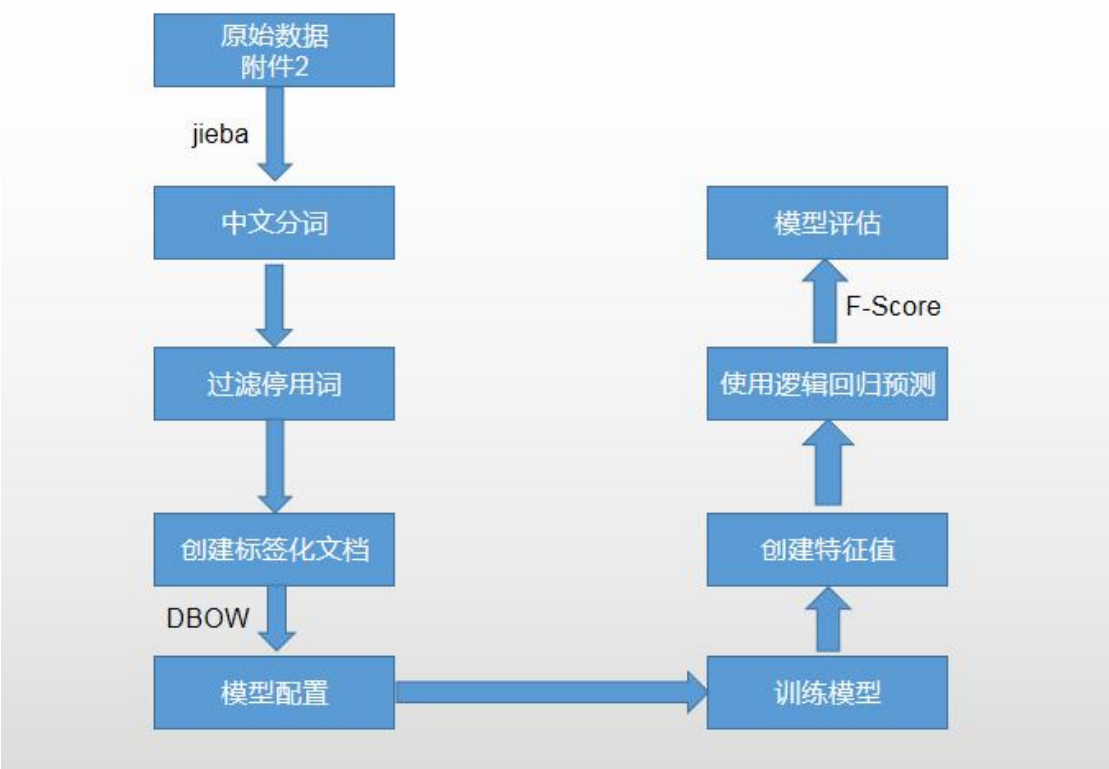

图 1 问题 1 流程图

2. 数据预处理

(1) jieba 分词

jieba 分词是基于 Trie 树结构实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图，采用了动态规划查找最大概率路径，找出基于词频的最大切分组合，使得分词有很好的效果。

(2) 对留言内容进行中文分词

在对附件 2 进行建立关于留言内容的一级标签的分类模型之前，先要把非结构化的文本信息转换为计算机能够识别的结构化信息。在附件 2 的留言详情中，以中文文本的方式呈现。为了便于转换，先对这些信息进行中文分词。这里采用 Python 中 jieba 包进行分词。

(3) 过滤停用词

文本分词之后还会存在很多无关的词语，需要将这些词语进行过滤处理。对定义过滤停用词函数，在分词得到结果后，通过调用过滤停用词函数进行停用词处理。这里的停用词表是收集了常用的 1893 个词，包括一些特殊符号、单个数字、方位词、语气词等等。

(4) 创建训练集合测试集

预处理后的数据，按照比例 7：3 来拆分数据，创建训练集和测试集的标签化文档。

3. 模型配置与训练

(1) DBOW 模型

Doc2Vec 是一种非监督式的算法，它可获取句子/段落/文档的向量表达，学习出来的向量可以通过计算距离来找句子/段落/文档之间的相似性。对有标签的数据，可以用监督学习的方法进行文本分类。

分布式词袋（DBOW）是 Doc2Vec 其中的一种训练方式，它通过训练神经网络来获得文档向量，该神经网络用于预测段落中的单词的概率分布，从而得到段落的随机采样的单词。

(2) 模型训练

DBOW 可以指定工作线程的数量来对模型进行训练，在这里用 multiprocessing 包来获取 CPU 的数量，数量为 4 个线程。然后对模型训练 10 个周期。

(3) 预测和模型评估

模型训练完毕后，需要创建生成文本的特征向量，然后进行预测。这里使用 sklearn 中的逻辑回归来进行预测。在预测完成后使用 sklearn 中的 f1_score 对预测结果进行评估。

三、问题二分析方法与过程

1. 流程图

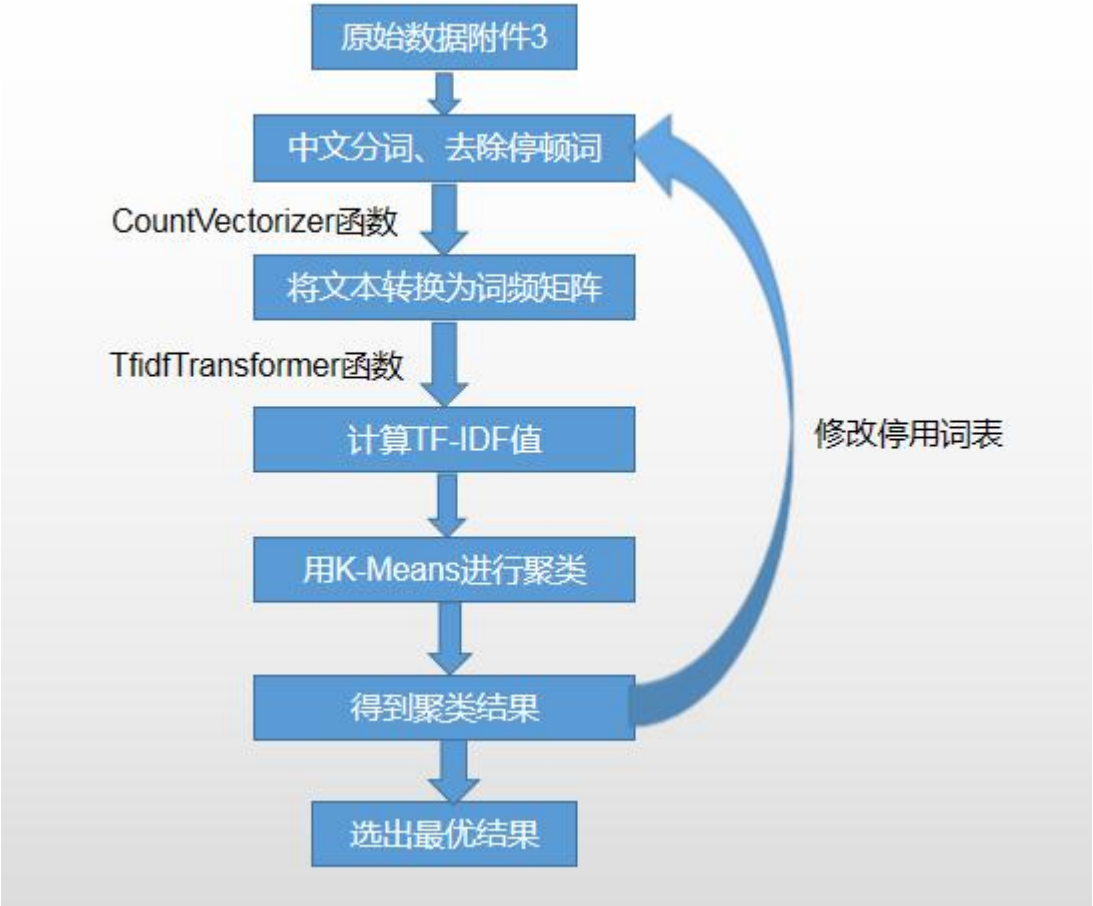

图 2 问题 2 流程图

2. 数据预处理

这里对附件 3 的留言主题进行中文分词，过滤停用词并统计词频，根据词频进行倒排排序，这里也生成词云图更直观地看出词语在文档中占有的比例高，对后续进行聚类调优作基础。

3. TF-IDF 算法原理

计算词频，即 TF 权重（Term Frequency）。

$$词频（TF）= 某个词在文本中出现的次数$$

考虑到文章有长短之分，为了便于不同文章的比较，进行“词频”标准化。

$$词频(TF)=\frac{某个词在文本中出现的次数}{文本的总词数}$$

或者

$$词频(TF)=\frac{某个词在文本中出现的次数}{该文本中出现次数最多的词的出现次数}$$

计算 IDF 逆文档频率（Inverse Document Frequency），需要建立一个语料库（corpus），用来模拟语言的使用环境。

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

计算 TF-IDF 值（Term Frequency Document Frequency）。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。

$$TF-IDF=词频(TF)\times 逆文档频率(IDF)$$

4. K-Means 聚类算法

算法有两个输入信息，一是 k,表示选取的聚类个数，另一个是训练数据集 $x^{(1)},x^{(2)},\cdots,x^{(m)}$。

① 随机选择 k 个聚类中心 $u_1,u_2,\cdots,u_k$

② 从 1 到 m 中遍历所有的数据集，计算 $x^{(i)}$ 分别到 $u_1,u_2,\cdots,u_k$ 的距离，记录距离最短的聚类中心点 $u_j(1\le j\le k)$，然后把 $x^{(i)}$ 这个点分配给这个聚类，即令 $c^{(i)}=j$。计算距离时，一般使用 $\|x^{(i)}-u_j\|$ 来计算。

③ 从 1 到 k 中遍历所有的聚类中心，移动聚类中心的新位置到这个聚类的均值处。即 $u_j=\frac{1}{c}\left(\sum_{d=1}^{c}x^{(d)}\right)$，其中 c 表示分配给这个聚类的训练样例点的个数，$x^{(d)}$ 表示属于 u_j 这个类别的点。

④ 重复步骤（2），直到聚类中心不再移动为止。

5. 生成 TF-IDF 向量

中文分词后，需要计算出每个词语出现的 TF-IDF 值，先用 CountVectorizer

函数将文本转化为词频矩阵，再用 TfidfTransformer 函数统计词频矩阵中每个词语出现的 TF-IDF 值。

6. 聚类过程

生成 TF-TDF 向量之后，用 K-Means 聚类方法将 TF-IDF 值进行聚类，这里设置聚类的类数为 6，将结果聚成 6 个类，并通过循环打印出文本结果。观察聚类的文本结果，如下图所示：

```
请问 市暮坪 大桥 开工 建设
请 市 加快 申报 国家 中心 城市 建设
请问 A7 县 深业 睿 城 f 区 地下 车位 建设 成本 承担
请 市 加快 流动 摊位 试点
请 加快 市 一圈 场 三道 步伐 力度
A4 区 城管 纵容 圣爵 菲斯 别墅 业主 违章 建设
市 城市 建设 开发 公司 中层干部
请 加快 市 国际 消费 中心 城市 创建
请 加快 市园 博园 建设 力度
希望 提高 市 规划 建设 水平
市 银星 路 银杉 路 交叉口 常年 扬尘 扰民 请 加快 雾 霾 整治 力度
请 加快 市 创建 国家 中心 城市 建设 步伐
A4 区 洪山 公园 开工 建设
请 加快 市 国家 中心 城市 建设 刻不容缓
市 提升 规划 建设 水平 带动 经济 发展 些许 建议
市 加快 招商引资 步伐 有何
启动 市梅 溪湖 路 路过 江 通道 建设
```

图 3 聚类结果 1

通过图 3 的聚类结果显示，有些词语在不同留言主题中都存在，需要对聚类进一步的优化，通过词云图可以了解词语的频率，如图 4 所示：

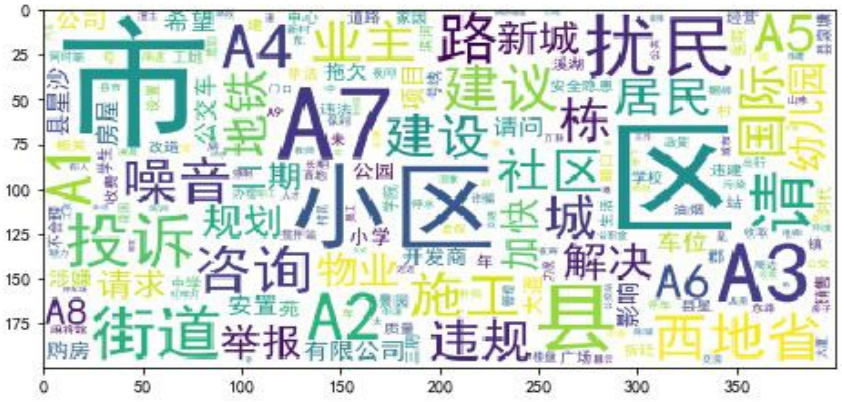

图 4 频繁词词云图

通过词云图，可以判断例如“市”、“区”等等词为很多不同话题之中共有的词语，对这些词，添加到停用词表中，将其过滤，再进行分类。当分类结果为类与类之间的信息联系不大，而类内之间的联系信息大，就可以判断为筛选出的热点问题，直到出现 5 个类都是热点问题，另外一个类不是热点问题，就进行对

5 个热点问题进行统计。

7. 热点问题评价

在留言问题中，一段时间段内反映的次数越多，问题的热度越大。根据附件 3 中所收集的问题的特点，留言问题的时间基本都是在 2019 年到 2020 年一年内。可以通过对相应的热点问题反映的数量作为热点指数。即如果问题 A 出现的次数为 1，那么热点指数为 1。

四、问题三分析方法与过程

1. 问题分析

(1) 相关性：答复与问题是否相关，如果回答内容与问题越相符合，那么答复与问题就越相关。

(2) 完整性：答复是否完整地回答了问题，遗漏要回答的问题越多就越不完整。可通过标记各个答复意见和对应问题之间的完整度，再用统计方法描述来反映整体完整度。

(3) 可解释性：答复是否为提问者难以理解，例如附件 4 的一些答复中只有一个日期，如图 5 所示，这种情况下是比较难理解一个日期的代表含义。通过筛选出类似图 5 的情况的数量（设为 a），再求出这个数（a）与所有答复的数量（设为 b）的比值，通过比例（a/b）反映出答复的可解释性程度，比例越高，整体答复的可解释性越差，比例越低，整体答复的可解释性越好。

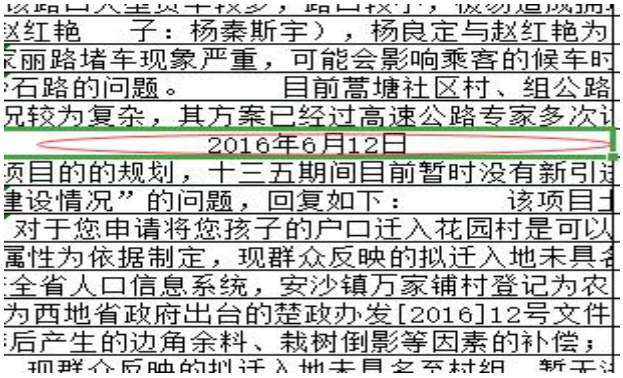

图 5

2. 答复质量评价方案的实现

(1) 流程图

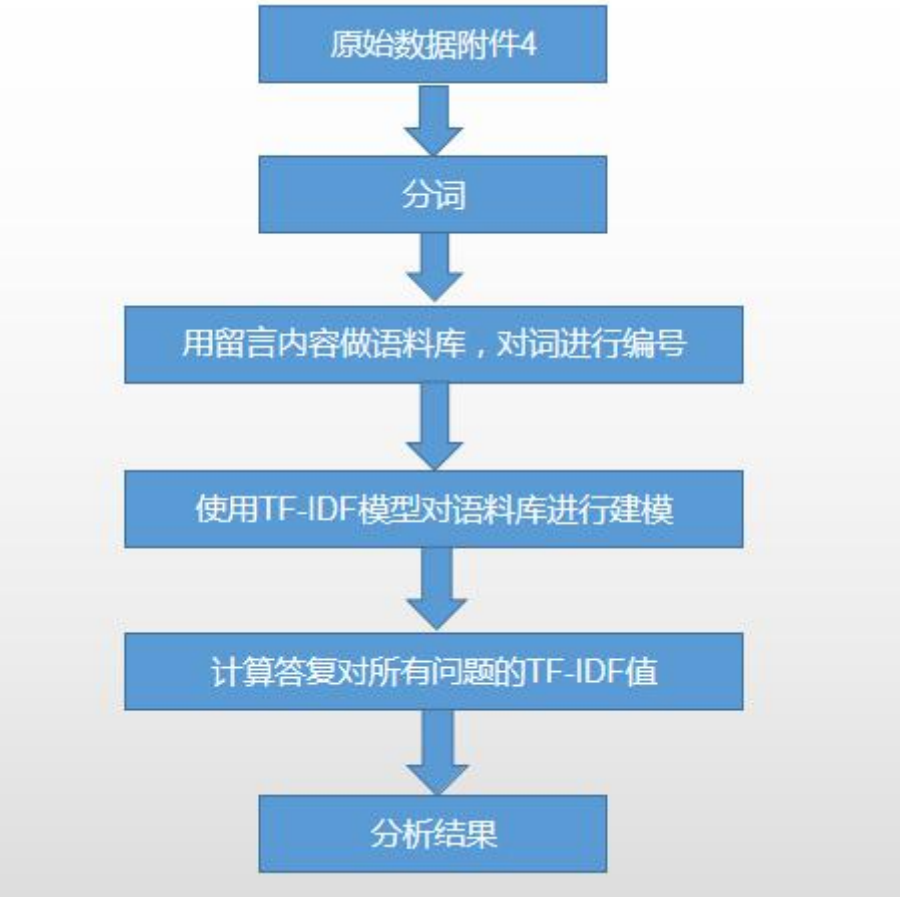

图 6 答复的相关性评价实现流程图

(2) 分析结果过程

对计算所得的 TF-IDF 值收集所对应的留言的序号，例如序号为 1 的答复所对应的留言问题的 TF-IDF 最大值的序号也为 1，就标记为 1，如果序号为 1 的答复所对应的留言问题的 TF-IDF 在次最大值的序号为 1，那么就标记为 2。依次类推，标记到 3，其余标记为 0。得到的标记结果如图 7 所示。

Index	Type	Size	Value
0	int	1	2
1	int	1	0
2	int	1	1
3	int	1	2
4	int	1	1
5	int	1	0
6	int	1	3
7	int	1	0
8	int	1	0
9	int	1	1
10	int	1	0

图 7

五、问题 1 结果分析

通过 F-score 对模型进行评价,得到的结果为：

```
FutureWarning: Default multi_class will be
changed to 'auto' in 0.22. Specify the
multi_class option to silence this warning.
  "this warning.", FutureWarning)
Testing F1 score: 0.8457020376223624
```

图 8 模型评价结果

如图 8 所示：F-score 的值约为 0.8457，即所建立的模型的准确率在 84.57%左右。

六、问题 2 结果分析

通过对附件 3 中的留言内容聚类得到下类似的结果如图 9 所示：

热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
1	1	40	2019/7/7至2019/9/1	A市广铁集团职工	A市伊景园滨河苑捆绑销售车位
2	2	38	2019/11/13至2020/1/25	A2区丽发新城小区居民	A2区丽发新城小区附近搅拌站污染和噪音扰民
3	3	22	2018/11/15至2019/12/2	A市新政面向人群	A市人才新政住房补贴问题
4	4	10	2019/2/14至2019/12/2	A7县凉塘路附近群众	A7县星沙塘路旧城改造拖延
5	5	9	2019/1/9至2019/9/12	A3区茶场村附近居民	A3区茶场村拆迁拖延

图 9 热点问题表截图

通过热点问题结果，可以知道 A 市的伊景园滨河苑的销售捆绑销售车位情况遭到大量的投诉，这种不合理的做法对广铁集团的职工是不公平的对待，两个月不到的时间内就有 40 起这类问题的反映，可见伊景园滨河苑的销售行为引起了群众的不满，合理的销售才可以使人称心满意，政府有关部门可以针对这类现象进行公告和警示。对于 A2 区丽发新城小区附近的搅拌站污染环境和噪音扰民也是引起广大群众的反映，居民生活受到影响，生活质量下降，政府有关部门应该及时制止并建议对搅拌站进行迁移。对于人才新政下的住房补贴问题也是得到群众的广大热议，住房补贴可以改善群众的生活，政府有关部门可在这方面增设人员以帮助有关群众了解政策。第 4、5 个热点问题都以城建拖延有关，有关部分应该规划好拆迁或改造方案，按时完成工程。

七、问题 3 结果分析

通过对上述图 7 中的结果进行可视化，如图 10 所示。其中，第二类是答复所对应的 TF-IDF 最大值为相应的留言问题。第三类为答复所对应的 TF-IDF 第二

大值为相应的留言问题，第四类为答复所对应的 TF-IDF 第三大值为相应的留言问题。第一类为上述情况的补集。

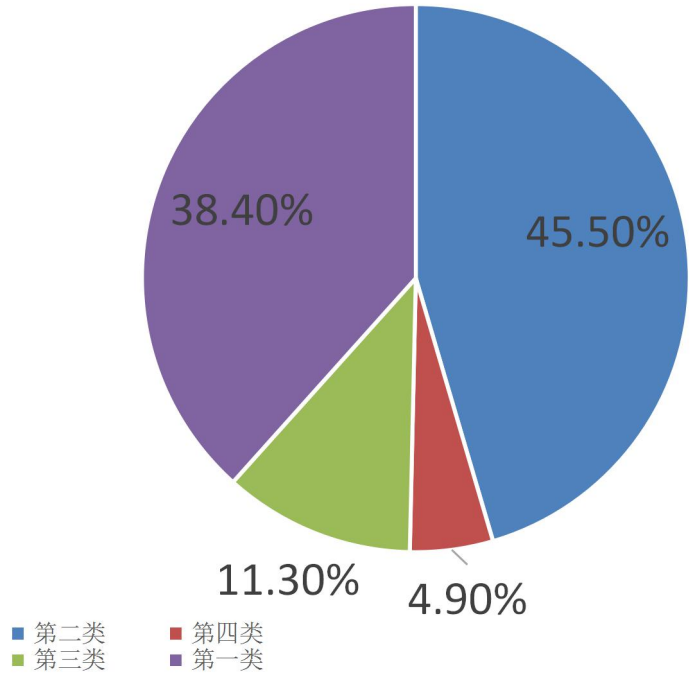

图 10

通过对图 10 的分析，答复所对应的 TF-IDF 最大值为相应的留言问题的比例为 45.5%，答复所对应的 TF-IDF 第二大值为相应的留言问题为 11.3%，第四类为答复所对应的 TF-IDF 第三大值为相应的留言问题为 4.9%。整体的答复意见是对于相应问题超过 60%。第三、第四类出现占比达到 15%以上的原因是存在相似的问题，一个答复意见可以对应多个问题。

八、参考文献

[1] 黄永昌，《scikit-learn 机器学习：常用算法原理及编程实战》，北京：机械工业出版社，2018
[2] -派神-，《文本多分类之 Doc2Vec 实战篇》，https://blog.csdn.net/weixin_42608414/article/details/88391760,访问时间（2020 年 4 月 24 日）
[3] 爱吃橙子的人吖，《Python 使用 k-means 方法将列表中相似的句子归类》，https://m.jb51.net/article/167248.htm，访问时间（2020 年 4 月 27 日）
[4] 番薯要吃肉，《用 Python 进行简单的文本相似度分析》，https://blog.csdn.net/xiexf189/article/details/79092629,访问时间（2020 年 5 月 4 日）