“智慧政务”中的文本挖掘应用

摘要：近年来，随着微信、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，各类社情民意相关的文本数据量不断攀升，给以往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带来了极大挑战。同时，随着大数据、 云计算、人工智能等技术的发展，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。 本文将基于自然语言处理和数据挖掘技术对多条网络留言进行分析，提取信息。

针对问题一： 本文首先对附件 2 中的数据进行去空、中文分词、根据词性以及停用词表过滤等数据预处理，然后针对所有分词建立字典。接着划分训练集和测试集，把分类标签进行处理，对应成整数。本文使用 SVM（支持向量机）算法训练模型。

针对问题二：本文首先使用正则表达式对附件 3 数据进行初步处理，经过简单的信息提取，本文得到了留言中的地点信息。经过抽样观察和对比，本文发现，针对本问题，地点信息对于热点问题的提取无作用。本文对留言主题和留言详情基于 TF-IDF 算法提取关键词。进而使用 K-means 算法训练模型，数据聚类。接着根据得到的类别的频数、反对数和支持数，排序所有留言信息。最后对热度最前的五类留言进行总结，得出表一。

针对问题三：本文首先对附件 4 的数据进行分词并过滤处理，包括去停用词、去空等，然后建立字典。然后使用已经建立的字典把语料转化为 bow 形式的数据，训练 tf-idf 模型。根据训练得到的 tf-idf

模型，把留言信息和答复转换为 tf-idf 形式数据。本文采用余弦相似性来决定两者的相似性，并以此作为答复的评分标准。

目录

1 问题背景

近年来，随着微信、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，各类社情意相关的文本数据量不断攀升，给以往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带来了极大挑战。同时，随着大数据、云计算、人工智能等技术的发展，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。

2 问题分析

问题一：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（参考附件 1 提供的内容分类三级标签体系）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。通常使用 F-Score 对分类方法进行评价：

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

，

其中为第 i 类的查准率，为第 i 类的查全率。

针对本问题，附件 2 已提供的数据包括：留言编号、留言用户、留言主题、留言时间、留言详情和一级标签。本文根据对该部分数据分析，采用有监督式学习来训练模型。监督学习的常用学习方法有朴素贝叶斯法、决策树和支持向量机等。本文选取支持向量机作为本题的学习方法。本文需要首先对文本进行一系列预处理，获得向量化的数据，再训练模型。

问题二：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。

针对本问题，附件 3 中没有提供分类信息，适合无监督学习的训练模型。本文分别采用了三个方案来聚类热点问题，方案一：提取地点信息，相同地点内进行聚类；方案二：利用问题一训练得出的分类模型，对相同标签的信息在进行聚类；方案三：直接对所有信息聚类。对结果进行观察，方案三聚类效果最好，本文最终选用方案三的聚类标签。由于问题没有指定热度指数计算方法，所以文本自定义了热度指数评价方案：问题出现频率占据最大权重，点赞数增加热度指数，反对数降低热度指数但权重只有点赞数的一半，并假设时间跨度跟热度无关。

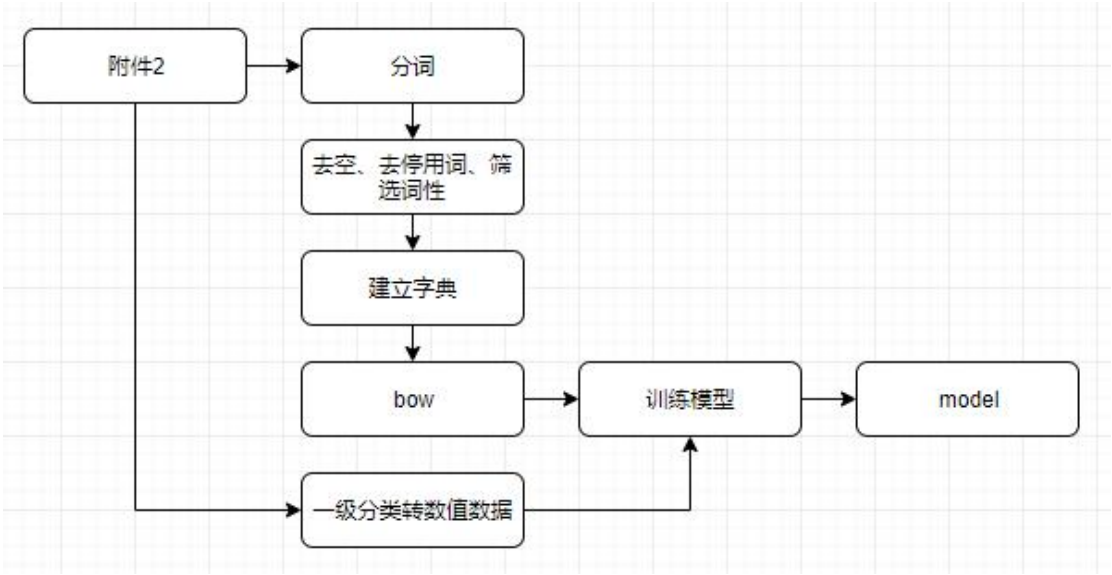
问题三：答复意见的评价

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。

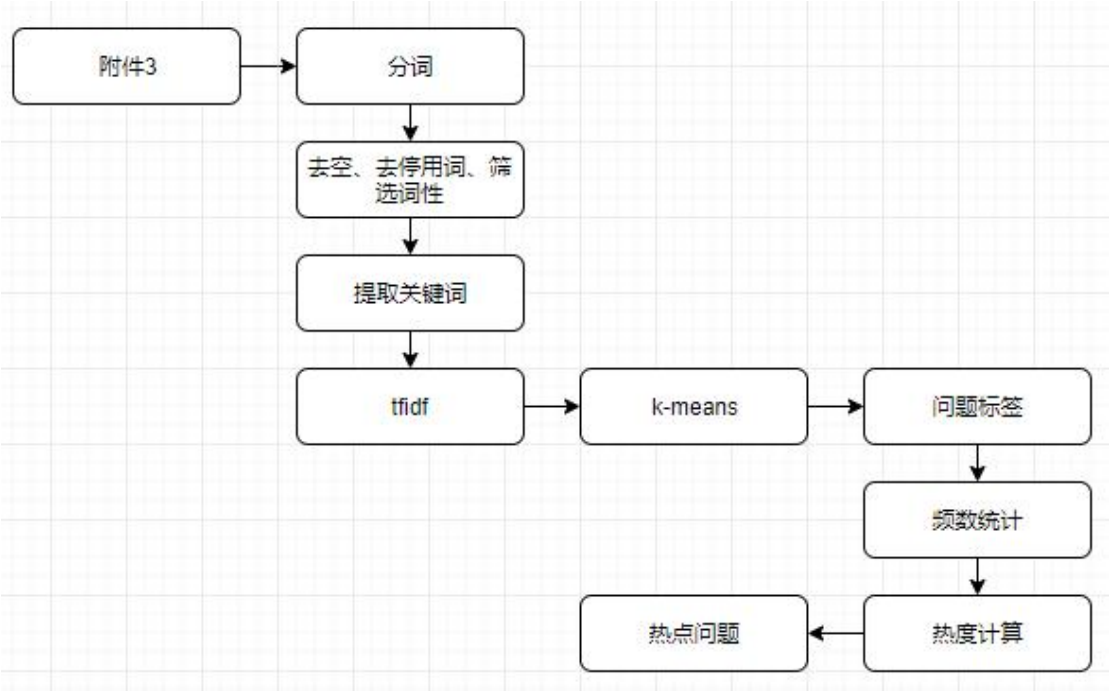
本文认为文本相似性能较为直观地反映答复意见的相关性和完整性，所以选择了余弦相似性来评价答复意见。

3 主要流程

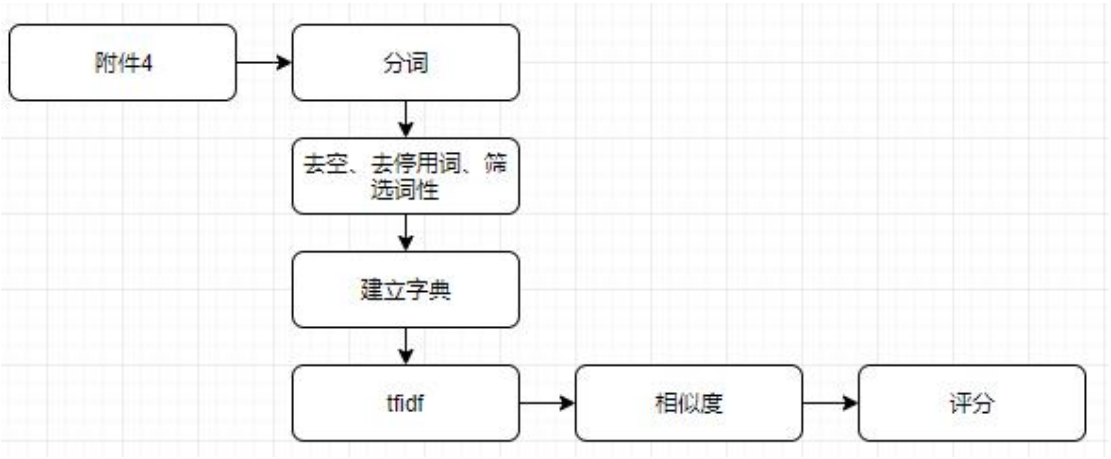
问题一：


附件2
分词
去空、去停用词、筛选词性
建立字典
bow
训练模型
model
一级分类转数值数据

问题二：


附件3
分词
去空、去停用词、筛选词性
提取关键词
tfidf
k-means
问题标签
频数统计
热度计算
热点问题

问题三：



4 相关模型

4.1 支持向量机

支持向量机（support vector machines，SVM）是一种二分类模型。它的基本模型是定义在特征空间上的间隔最大的线性分类器，间隔最大使它有别于感知机；支持向量机还包括核技巧，这使它成为实质上的非线性分类器。支持向量机的学习策略就是间隔最大化，可形式化为一个求解凸二次规划的问题，也就等价于正则化的和合页损失函数的最小化问题。支持向量机的学习算法是求解凸二次规划的最优化算法。[1]

4.2 k 均值聚类

k 均值聚类是基于样本集合划分的聚类算法。k 均值聚类将样本集合划分为 k 个子集，构成 k 个类，将 n 个样本分到 k 个类中，每个样本到其所属类的中心的距离最小。每个样本只能属于一个类，所以 k 均值聚类是硬聚类。[2]

4.3 TF/IDF

TF/IDF 算法（Term Frequency-Inverse Document Frequency，词频-逆文档频次算法）是一种基于传统的计算方法，常用于评估在一个文档集中一个词对某份文档的重要程度。[3]

5 相关工具

5.1 分词工具

本文分词使用Github社区的开源项目Jieba进行中文分词。Jieba提供三种分词模式：精确模式、全模式、搜索引擎摸索。

5.2 词性标注

本文利用 Jieba 自带字典标注词性，本文没有额外标注词性。

5.3 读取和修改表格数据

本文使用库 Pandas 读取附件的数据，并使用 Pandas 提供的方法修改和保存运行过程中产生的部分数据。

5.4 模型训练

本文使用库 gensim 和库 sklearn 提供的类和方法来训练模型，其中包括 SVC（）、Tfidfmodel（）、SparseMatrixSimilarity（）和 Kmeans（）。

6. 相关数据

6.1 模型 1

（1）测试集和训练集：

本文选取 0.85 的数据作为训练集，0.15 的数据作为测试集。

（2）一级分类对应的标签：

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

（3）测试结果报告：

	precision	recall	f1-score	support
0	0.98	0.75	0.85	133
1	0.82	0.93	0.87	303
2	0.87	0.72	0.79	180
3	0.92	0.68	0.78	122
4	0.95	0.88	0.91	235
5	0.92	0.52	0.67	109
6	0.66	0.92	0.77	300
accuracy			0.82	1382
macro avg	0.87	0.77	0.81	1382
weighted avg	0.85	0.82	0.82	1382

（4）模型参数：

```
SVC(C=1.0, break_ties=False, cache_size=200, class_weight=None, coef0=0.0,
    decision_function_shape='ovr', degree=3, gamma='scale', kernel='rbf',
    max_iter=-1, probability=False, random_state=None, shrinking=True,
    tol=0.001, verbose=False)
```

6.2 模型 2

（1）模型参数：

```
KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
       n_clusters=1000, n_init=10, n_jobs=2, precompute_distances='auto',
       random_state=None, tol=0.0001, verbose=0)
```

（2）分类标签：

见附件。

6.3 模型 3

（1）模型参数

```
TfidfModel(num_docs=5632, num_nnz=316032)
```

（2）相似度评分

见附件。

7 总结

本文主要运用自然语言处理的相关技术，对三个问题进行分析并尝试解决。本文实现了训练分类模型、热点问题提取、答复意见评分。

8 参考文献

[1] 李航．统计学习方法(第 2 版)．北京：清华大学出版社

[2] 李航．统计学习方法(第 2 版)．北京：清华大学出版社

[3] 涂铭，刘祥，刘春树.Python 自然语言处理实战：核心技术于算法．机械工业出版社.