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

摘 要

随着网络问政平台成为政府了解人民民意、人民发扬民主监督的重要渠道，各类留言文本量越来越庞大，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。

针对问题一（群众留言的分类）：这是一个文本分类的问题，首先进行数据预处理，去除重复值、干扰词，中文分词，过滤停用词，然后检查数据的分词效果，再对预处理的结果进行二次停用词去除、留言词库构建、词云构建，最后通过建立 LDA 主题模型，实现对留言数据的分类。

针对问题二（热点问题的文本挖掘）：在对数据进行预处理后，利用特征词权值计算方法 TF-IDF 计算各个单词权重，以构成一个向量空间模型用于 K-means 聚类。以类平均相似度为指标，聚类之后再分别定义，选出热点问题。

针对问题三（答复意见的评价）：针对相关部门对留言的答复意见，通过 Greedy matching（贪婪匹配法）、Embedding Average（向量均值法）、Vector Extrema（向量极值法）三个客观评价指标作为标准，建立对话系统的评价体系。利用从腾讯 AILab 下载的中文词向量包（Tencent_AILab_ChineseEmbedding）将答复文本字符串向量化，基于词向量的方法，通过这三个客观评价指标来计算单条文本之间的相似度。答复与留言语句之间相似度越高，可以判断答复的相关性、可解释性以及完整性越高。

关键词：自然语言处理　LDA 主题模型　TF-IDF　K-means　相似度　对话系统

目 录

“智慧政务”中的文本挖掘应用

一、问题重述与分析

1.1 背景重述

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。

本次挑战赛给出的数据收集于互联网公开的群众问政留言记录，包含相关部门对群众留言的答复意见，需要利用自然语言处理和文本挖掘方法处理下面给出的问题。

1.2 问题的提出与分析

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对留言进行分类，然后将分类后的留言交给不同部门进行处理。目前大部分政务系统仍然依靠人工，根据经验进行分类，存在工作量大、效率低，且差错率高等问题，于是问题一要求我们根据给出的数据，建立关于留言内容的一级标签分类模型，使用 F-Score 对分类方法进行评价。

这是一个文本分类的问题，属于自然语言的处理，根据题目所给数据的格式和数据量的大小，构建合适的分类模型即可，模型的选择较多，可以考虑传统机器学习方法或者深度学习方法。因为留言属于长文本，无意义表达、一词多义现象太多，本文使用 LDA 主题模型进行分类。LDA 是一种非监督机器学习技术，该方法假设每个词是从背后的一个潜在隐藏的主题中抽取出来，可以用来识别大规模文档集或语料库中的潜在隐藏的主题信息。

问题二：热点问题挖掘。及时发现某一时段内群众集中反映的问题，有助于相关部门进行针对性的处理，提升服务效率。问题给出包含特定地点或特定人群问题的留言，要求我们进行归类，依靠合理的热度评价指标给出评价结果，最后导出热点问题的表格。

根据题干要求，我们首先需要对问题进行识别，即从众多的留言中识别出相

似的留言。然后再将问题归类，把特定地点或人群的数据归并，即把相似的留言归为同一问题。最后进行热度评价，选取合适的指标，选出热度问题。这可以通过文本聚类完成。

问题三：答复意见的评价。相关部门对留言的答复意见各式各样，所以确定答复意见的质量评价标准是非常有必要的，答复意见的质量从相关性、完整性、可解释性方面进行考量。

我们主要从词向量的角度入手，引入相似度评价指标，通过不同的形式来计算提问和答复意见的相似度，进行全方面的答复质量评价。

二、问题一模型建立与求解

2.1 分析方法与过程

问题一的目的是根据给出的留言数据，建立留言内容的一级标签分类模型。

本次建模对群众留言的数据，首先进行数据预处理，包括去除重复值、去干扰词、中文分词、停用词过滤，再对预处理的结果进行二次停用词去除、构建留言词库等，然后检查数据的分词效果，对分词结果不断优化以达到预期设想目标，最后通过建立 LDA 主题模型，实现对留言数据的分类。LDA 模型即三层贝叶斯概率模型，包含文档（d）、主题（z）、词（w）三层结构，可以有效对文本进行建模。通过 LDA 主题模型，可以挖掘评论语料中的潜在主题，能有效解决一词多义和一义多词的问题，进而分析文本的集中关注点及其相关特征词。

模型建立步骤：

1） 数据预处理，进行文本去重、去除干扰词、中文分词、停用词过滤的操作；

2） 分词结果优化，主要是停用词二次过滤，构建留言词库；

3） 建立词云；

4） 根据词云、分词结果建立 LDA 模型；

5） 分析结果和总结。

数据库	数据抽取	数据探索与预处理	建模	分析结果
留言数据	按信息类别抽取留言详情和一级分类两列数据	去重去空 中文分词 停用词过滤 二次过滤、构建词云	构建 LDA 模型	结果评价

2.2 建模过程

2.2.1 数据预处理

题目所获的的群众留言数据，存在着许多重复的文本内容，且含有各种各样的符号、字母、数字等干扰数据，将这些干扰数据引进分词和词频统计或者模型当中，不仅消耗内存还会造成分析结果的偏差。所以在利用这些文本数据之前就必须对其进行预处理，初步将这些无价值的干扰数据去除。

文本数据的预处理主要包括六个部分：文本去重、中文分词、停用词过滤、二次过滤、构建词库。

1. 文本去重

在数据收集、获取、存储和提取的过程中，由于某些原因，造成了数据的重复等情况，所以要对原始数据进行去重的处理，如下表所示，存在两条记录完全重复的情况。针对留言数据只保存其中一条，其余删除。

留言编号	留言用户	留言主题	留言时间	留言详情	一级标签
303	U0007137	A1 区蔡锷南路 A2 区华庭楼顶水箱长年不洗	2019/12/6 14:40:14	A1 区 A2 区华庭小区高层为二次供水，楼顶水箱长年不洗，现在自来水龙头的水严重霉味，大家都知道，水是我们日常生活必不可少的用品，霉是一种强致癌物，我们住在这里连基本的健康保障都没有，请政府街道各领导重视起来，也请环保部门来检测，还我们一个健康安全的基本生活环境！	城乡建设
319	U0007137	A1 区蔡锷南路 A2 区华庭楼顶水箱长年不洗	2019/12/6 14:40:14	A1 区 A2 区华庭小区高层为二次供水，楼顶水箱长年不洗，现在自来水龙头的水严重霉味，大家都知道，水是我们日常生活必不可少的用品，霉是一种强致癌物，我们住在这里连基本的健康保障都没有，请政府街道各领导重视起来，也请环保部门来检测，还我们一个健康安全的基本生活环境！	城乡建设

2．中文分词

在中文中，一句话的含义往往通过一段连续的字符进行表达，字符串之间没有明显的标识将其断开。因此我们在对文本处理时，需要进行文本分词，并按照规则重新合成词序列。目前针对中文分词的方法有很多，如 KTDict Seg 分词、盘古分词器以及 jieba 分词等。本文采用 jieba 分词第三方库对产品评论进行处理。在 jieba 分词词库的原有基础之上可以引入其他字典，增加分词结果的可靠性，经过相关测试，jieba 分词精度可达到 90%以上。

以第 1000 条数据为例，查看分词效果。

原群众留言：

帝王广场栋开发商西地省佳煌房地产开发有限公司栋结构体系框剪结构本人在装修过程中发现上下层卧室同一位置出现度裂缝这种裂缝属于结构问题的裂缝还发现外墙承重抗震剪力墙靠内侧无受力钢筋网无梅花状拉勾施工单位没有按图施工偷要减料存在严重的安全隐患房屋隐患不排除房屋如出现脆性破坏一旦倒塌会造成群死群伤损失不可估量请市委市政府以安全第一预防第一核实真实情况谢谢

分词结果为：

['帝王', '广场', '栋', '开发商', '西地', '省佳煌', '房地产', '开发', '有限公司', '栋', '结构', '体系', '框剪', '结构', '本人', '在', '装修', '过程', '中', '发现', '上', '下层', '卧室', '同一', '位置', '出现', '度', '裂缝', '这种', '裂缝', '属于', '结构', '问题', '的', '裂缝', '还', '发现', '外墙', '承重', '抗震', '剪力墙', '靠', '内侧', '无', '受力钢筋', '网无', '梅花', '状', '拉勾', '施工单位', '没有', '按图', '施工', '偷要', '减料', '存在', '严重', '的', '安全隐患', '房屋', '隐患', '不', '排除', '房屋', '如', '出现', '脆性', '破坏', '一旦', '倒塌', '会', '造成', '群死群伤', '损失', '不可估量', '请', '市委', '市政府', '以', '安全', '第一', '预防', '第一', '核实', '真实情况', '谢谢']

3. 停用词过滤

文本经过分词后，就变为了一系列词语的集合，但其中一些词语包含的信息量较少，它们的存在不但不会提供关键信息，还会降低文本挖掘模型的准确率，所以去掉这些无用的词，是非常有必要的。通常一篇文本中的冠词、连词和介词等虚词以及在整个文本集中出现频率很高、但对区分类别作用不大的词，被称为停用词。英文的停止词有“the”、“and”、“of”等。中文停止词有“也”、“的”、“啊”等。去除停止词是文本预处理中不可缺少的步骤，所以我们可以引用停用词表，将以上所提到的词全部删除，还可根据数据所给特性，适当添加词语，去掉更多的无用词。处理成果为：

['帝王', '广场', '开发商', '西地', '省佳煌', '房地产', '开发', '有限公司', '结构', '体系', '框剪', '结构', '本人', '装修', '过程', '发现', '上', '下层', '卧室', '同一', '位置', '出现', '度', '裂缝', '这种', '裂缝', '属于', '结构', '问题', '裂缝', '发现', '外墙', '承重', '抗震', '剪力墙', '内侧', '无', '受力钢筋', '网无', '梅花', '状', '拉勾', '施工单位', '没有', '按图', '施工', '偷要', '减料', '存在', '严重', '安全隐患', '房屋', '隐患', '不', '排除', '房屋', '出现', '脆性', '破坏', '一旦', '倒塌', '造成', '群死群伤', '损失', '不可估量', '市委', '市政府', '安全', '第一', '预防', '第一', '核实', '真实情况', '谢谢']

4．二次清洗

经过分词和去停用词之后，数据会变得精简，但为了模型精度和构建词云，本文在进行了一次词频统计，在 python 中计算出现频率前 100 的词语，然后筛选，将其中一些常见的，把将对语意不产生影响的词语添加入停用词表，这样可以使文字变得更加精简。加强模型精度。

下表中统计初次分词结果频率前三十的词语：

词语	频数	词语	频数	词语	频数	词语	频数	词语	频数
领导	5448	业主	2634	居民	2134	教育	1935	办理	1666
公司	4055	学生	2634	教师	2107	职工	1860	管理	1623
学校	4013	医院	2616	企业	2051	尊敬	1860	退休	1613

政府	3518	单位	2389	生活	2049	时间	1815	请问	1597
国家	2713	政策	2295	孩子	2019	老百姓	1796	发展	1571
情况	2634	解决	2266	工资	1971	老师	1787	影响	1547

根据频数表，我们可以将类似于“尊敬”，“请问”这种词语加入停词表，提高分词的效果。以上部分为数据清洗的过程。

2.2.2 数据探索--绘制词云图

构建模型之前我们可以简单的分析一下数据，根据留言的数据构建词云，并提取关键词，文本数据经过了预处理就会产生大量的重复字词，某个字词在留言文本中出现的次数越多，意味着该词就越可能是关键词。下一步工作就是找到目前处理的关键词，看是否符合预期。

词云图可以直观的看到文本数据的高频词汇，过滤掉大部分的低频词汇，使得阅读这一眼就可以知道该类留言的主体信息。根据所给留言数据表 2 有七个类别，对每个类别分别绘制词云。如下表：

一级标签	词云图
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	

2.2.3 分类模型构建

根据留言数据构建的的词云图，针对文本的分类构建模型。

1. 文档主题生成模型（LDA）

LDA 主题模型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。所谓生成模型，就是说，我们认为一篇文章的每个词都是通过“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”这样的一个过程得到。文章到主题服从多项式分布，主题到词服从多项式分布。

LDA 是一种无监督机器学习技术，可以用来识别大规模文档集或语料库中潜藏的主题信息。采用了词袋的方法，这种方法将每一篇文档视为一个词频向量，从而将文本信息转化为易于建模的的数字星系。但是词袋方法没有考虑词与词之

间的顺序，简化了问题的复杂性，同时也为模型的改进提供了契机。每一篇留言代表了一些主题所构成的一个概率分布，而每一个主题有代表很多单词所构成的一个概率分布。

主题模型打破了传统空间向量文档—词的思维定向，将文档映射到主题空间上，表示为文档—主题—词。主题模型认为文档是由若干主题组合而成，而主题又是由单词组合而成。用主题描述文档，有效的降低了维度，即主题模型克服了空间向量模型的缺点。

LDA（Latent Dirichlet Allocation）模型是 Blei 等人在 2003 年提出的，他们在 pLSI 的基础上加入先验分布 Dirichlet 分布得到 LDA 模型。LDA 模型有 3 层生成式贝叶斯网络结构[33]，该模型的前提假设是：文档是由若干个隐含主题构成，而这些隐含主题[13](潘耘等,2009)是由文档中若干特定词汇构成，这些词汇在文档中出现的句法结构和先后顺序可以忽略。

2. 主题分析

对留言进行 LDA 主题分析，本文针对七个大类，对每个类别分别得到四个潜在的主题，每个主题下生成 8 个最有可能出现的词语及相应的概率，下表显示了关于“卫生计生”类别的潜在主题。

主题 1	主题 2	主题 3	主题 4
医院	医院	医院	医院
医生	医生	医生	医生
手术	国家	领导	政策
医疗	领导	小孩	生育
情况	办理	病人	领导
患者	政策	生育	子女
孩子	独生子女	医疗	国家
家属	生育	证明	家庭

根据“卫生计生”类别主题结果，可以看出“医院”、“医生”等词在四个主题中全都有，也可以通过此检验模型的准确性，按留言内容来说，出现“医院”、“医生”等词汇的留言确实属于“卫生计生”等分类中。其余六类则各选择一个主题作纪录，如下表说示：

城乡建设	环境保护	交通运输	教育文体	劳动和社保	商贸旅游
业主	污染	出租车	学校	公司	电梯
房屋	公司	快递	学生	工资	景区
物业	企业	交通	家长	领导	旅游
开发商	政府	的士	教育	政策	游客
社区	生活	公交	孩子	职工	收费

居民	环保	客运	教师	单位	电话
工程	噪音	车辆	孩子	退休	投诉
学校	村民	城市	成绩	缴纳	业主

从各类别的主题结果中可以看出不同的类别中有相同的关键词，但是此类关键词对于在模型主题中占有一定的地位，无法直接删去，这也是造成模型误差的原因之一。

2.3 模型评价

本文中的模型对各个类别的留言数据都会提取 3～4 个主题，对于留言数据的 9200 条数据，拆分为训练集和测试集，训练集占比为 75%，测试集占比为 25%，建立关于留言内容的一级标签分类模型。使用 F-Score 对分类方法进行评价。在 pycharm 中实现最终精度结果为：0.86626

通过 F1score 来计算模型得分，查准率和查全率如下图所示：

类别序号	查准率	查全率
1	0.859	0.90
2	0.863	0.88
3	0.787	0.86
4	0.807	0.80
5	0.816	0.78
6	0.824	0.75
7	0.829	0.70

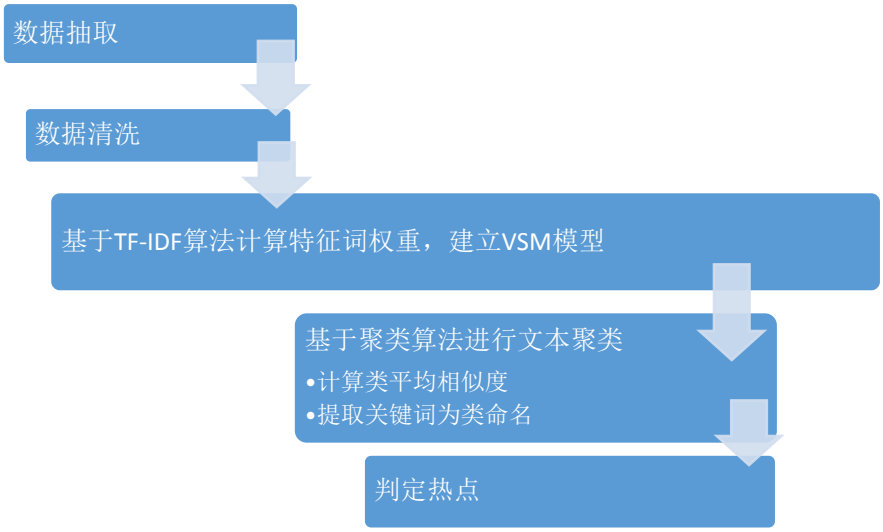
根据 python 中的 sklearn.Metrics 库中的 F1score 函数可以计算最终得分为：0.816

三、问题二模型建立与求解

3.1 问题分析与实验流程

热点问题的挖掘，其问题实质就是计算文本间的相似度，以相似度为指标，将所需数据聚类。聚类之后再定义热点问题。

实验流程如下：



3.2 理论方法与过程

1．中文分词及词性标注

中文分词就是将汉字序列切分成有意义的词，以字为单位，句和段则通过标点等分隔符来划界。目前主流的中文分词算法分为四类:基于字符串匹配的分词，基于理解的分词，基于统计和基于语义的分词。

词性标注是根据句子上下文环境给句中的每个词标记一个正确的词性，主要是机器针对多标记词 (即有多种词性的词)和未登录词(即在训练语料中未出现的词)标记词性。词性标注技术与分词技术一样，在自然语言处理、机器翻译、文本自动检索及分类、文字识别、语音识别等实际应用中占有重要地位。目前比较典型的标注算法归纳起来有:基于规则的方法，基于统计的方法，规则与统计相结合的 方法。本文选用的是规则与统计相结合的方法。

2．特征词权重提取

在一众提取特征词、计算权重的方法中，TF-IDF 综合考虑了特征字词项的在某一个文本中出现的频率以及该特征字词项在任意文本所具代表性的重要程度，

较全面的考虑了特征项与文本间的关系，且无需任何先验知识，因此本文采用TF-IDF 权重法来计算文本集中各个特征项的权重。

TF-IDF 模型的主要思想是：如果词 W 在一篇文本 D 中出现的频率高，并且在其他文档中很少出现，则认为词 W 具有很好的区分能力，适合用来把文本 D 和其他文章区分开来。

3．向量空间模型

向量空间模型 (Vector Space Model，VSM)是一个应用于信息过滤、信息撷取、索引评估相关性的代数模型 ，文本分析对象通常是以词为单位的 VSM 数据。运用这个模型把文本表示为向量，就可以将文本处理简化为向量空间中的向量运算。当文档转化为向量时，文档中每个词对应向量的每个特征项维度，所有文档中的词所对应的维度构成了整个空间，而特征权重则是每个词对应每一维的取值。于是，一个文档D_j转化为特征向量$\overrightarrow{D_j}$可表示为:

$$\overrightarrow{D_j} = \left(w_{i1}t_{i1}, w_{i2}t_{i2}, \dots, w_{ij}t_{ij}\right), 1 \le j \le M$$

其中t_{ij}是特征项，是特征权重，M 是文本t_{ij}中的特征项总数。另外，文本中作为特征项的词不能重复，即各特征项t_{ij}互异，且文本的内部结构不需要考虑，因此特征项t_{ij}无先后顺序。

(4)K-means 文本聚类

文本聚类的算法有许多，如 K-means、CLARANS、CURE 等，本文最终使用K-mean 算法进行聚类，它算法简洁且运行速度快。

K-means 算法欧式距离作为相似性的评价指标，即认为两个对象的距离越近，其相似度就越大， 得到紧凑且独立的簇是聚类的最终目标 。K -means 算法中距离的计算公式如下:

$$v = \sum_{i=1}^{k} \sum_{x_i \in S_I} (x_j - u_i)^2$$

K-means 算法流程

第一步，从数据对象中任意选择 K 个对象（K 值需要预先设定）作为初始聚类中心。第二步，计算剩下的对象与这些聚类中心的相似度（距离），并分别将它们分配给最相似的（聚类中心所代表的）类。第三步，重新计算每个新类的聚类中心（该聚类中所有对象的均值）。第四步，不断重复第二、三步，直到标准测度函数开始收敛为止，一般采用均方差作为标准测度函数。

该算法在处理大数据时是相对高效和可伸缩的，计算的复杂度为$O(N_{kt})$,其中 N 是数据对象的数目，t 是迭代的次数，一般$K \le N, t \le N$,同时算法对顺序不

太敏感，因此较适合对 VSM 表示的文本集进行聚类。本文聚类效果的验证采用类平均相似度，公式为：

$$AVG_T(SIM) = \frac{\sum C_{T_{i=1}}(avg(sim))}{C_T} \qquad (t \in T)$$

其中$AVG_T(SIM)$表示类 T 的平均相似度；C_T表示类 T 所包含的微博条数；$f_t(avg(sim))$表示类 T 中单条留言主题 t 的个体平均相似度。将类中所有主题的个体平均相似程度之和取平均值。将类中的所有主题的个体平均相似度之和取一次平均值，从而得到类的平均相似度。

3.3 建模过程

1.数据预处理

首先对留言标题进行文本预处理，与第一问流程类似，即进行去重、分词、无效信息过滤、降维等操作。实验中依然使用 python 中的 jieba 分词系统对留言标题进行分词处理，保留名词及名词性词语，再将所有的单字过滤，再去除所有的英文字符、数字和一系列数学符号等非中文词，只留下有意义的中文词语。以下所示为分词结果：

[西湖，建筑，集团，施工，安全隐患]
[在水一方，大厦，人为，烂尾，多年，安全隐患]
[投诉，市区，物业，违规，停车费]
[蔡锷，南路，区华庭，楼顶，水箱，长年，不洗]
[华庭，自来水，好大，一股，霉味]

2.话题识别

文本预处理后，针对每个留言标题，利用特征词权值计算方法 TF-IDF 计算各个单词权重，以构成一个向量空间模型用于聚类。实验中，K 值在最大值范围内通过多次实验结果验证来选取。经过多次实验，最终将留言的内容分为 10 类，并对各类进行关键词提取，结果如下表：

类别	留言条数	类平均相似度	关键词
1	145	0.41	学校，补课，收费，投诉
2	134	0.37	养老保险，中医院，百姓，健康，消费者
3	127	0.34	退休，工资，社会，请求，解决
4	124	0.28	涉嫌，非法，上户，缴费，职业
5	115	0.27	服务，乱收费，公平，污染，涨价
6	106	0.24	市中，报告，请求，机关，县乡镇
7	103	0.25	公司，拖欠，三无，疑问，建议
8	99	0.23	咨询，发放，农村，医疗补贴

9	96	0.20	小孩，孩子，教育，小学，老百姓
10	91	0.21	养老保险，企业，咨询，重视

最终根据类的平均相似度和留言的条数来选取热点问题。

3.排名热点问题

根据模型的最终结果，选取出排行第前五的热点问题，热度指标进行加权平均，结果如下表所示：

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	100	2017-2019	学校	学校补课乱收费
2	2	45	2018-2019	医院	医院建设
3	3	38	2015-2019	社区	社区垃圾处理
4	4	37	2012-2019	物业	道路检修小区建造
5	5	30	2016-2019	公司员工	工资拖欠

四、问题三模型建立与求解

4.1 问题分析与实验流程

附件四里相关部门的答复有非常完整的：

网友：
你好，收悉你的来信后，我镇党委政府高度重视，迅速成立了工作调查组对你所反映的问题进行了详细调查。现就有关情况答复如下：

2016 年 2 月 7 日（农历 12 月 29 日）晚上 7 点半左右，我镇共有 40 个左右的村相继停电。经查，停电的原因系春节来临城乡居民用电负荷陡增，供电设施配备不足，特别是变压器因严重超负运行烧坏所致。我镇主要负责人在得知部分村停电后，立即和国家电网桥头河供电所联系，要求其派出全部技术力量迅速修复，同时安排镇机关值班人员和相关村定补干部做好群众的情况说明、情绪疏导、矛盾化解等工作。 经过桥头河供电所 3 台次吊车、4 台次运输车和 30 余名技术人员的抢修，共更换 3 台变压器，到晚上 11 点，除桂花群工站 9 个村外全部恢复供电。截至 2016 年 2 月 8 日上午 10 时，除大水、野鸭塘两个村有半数用户外（因还需更换 2 台变压器），其余全部恢复供电。

对停电给广大群众生活带来的不便我们深表歉意，在以后的工作中将以此为鉴，不断总结汲取经验教训，不断完善各类应急预案，不断听取人民群众意见，确保人民群众过一个欢乐、祥和的春节！

2016 年 2 月 17 日

也有在“打太极”，不清晰的：

您好，您所反映的问题，已转交相关部门调查处置。

好的答复需要紧贴提问内容，提出相关的解决办法，所以定义答复意见质量标准如下：

用答复与提问之间的相似度来度量答复的相关性、可解释性和完整性。

通过对每一个词的分析来判断居民提问与相关部门答复的相关性，答复与提问语句之间相似度越高，可以判断答复的“相关性”、“可解释性”以及“完整性”越强。先通过对表格数据集中的词进行向量表示，再通过比较提问与回复二者表示的差距来达到相似度的比较。常用的有三种基于词向量的评价矩阵：Greedy Matching，Vector Extrema，Embedding Average。

这三种评价矩阵来源于深度学习对话系统，用于评价对话系统产生的自动回答效果的好坏，我们借此矩阵用来对相关部门的答复意见进行评价。

4.2 理论方法与过程

Greedy Matching 的计算方法如下 G（r，r^），对于提问的 r 和回复 r^ 中的

每个词都转化为词向量，分别计算两个句子的句向量， 然后计算两个词向量之间的余弦相似度， 将所有词的运算结果进行均值计算，再将两句话的顺序交换之后用同样方法计算，最后得到的 Greedy Matching（GM） 是两次结果的均值。

$$G\left(r,\hat{r}\right)=\frac{\sum_{w\in r;}\max_{\hat{w}\in\hat{r}}\cos\left(e_w,e_{\hat{w}}\right)}{|r|}$$

$$GM\left(r,\hat{r}\right)=\frac{G\left(r,\hat{r}\right)+G\left(\hat{r},r\right)}{2}$$

Embedding Average 是将一句话中每个词的词向量均值当作该句话的句子向量表示用于计算两句话的相似度，具体计算方法如下面公式所示，公式中的 e 表示句子 r 的句子向量。分别对 r 和 rˆ 计算句子向量 e 和 eˆ ，再通过 EA 公式计算标准答案与生成回复的相似度指标 Embedding Average（EA）。

$$\bar{e}_r=\frac{\sum_{w\in r}e_w}{|\sum_{w'\in r}e_{w'}|}$$

$$EA:=\cos\left(\bar{e}_r,\bar{e}_{\hat{r}}\right)$$

Vector Extrema 是通过向量极值来表示句子向量的方法，将词向量每一维中极值最大的一维抽取出来当作句子向量的表示

$$e_{rd}=\begin{cases}\max_{w\in r}e_{wd} & if\ e_{wd}>|\min_{w'\in r}e_{w'd}|\\ \min_{w\in r}e_{wd} & otherwise\end{cases}$$

公式中e_{wd}是词 w 对应词向量中的第 d 维，将提到的两个句子之间的词向量按照$VE:=\cos\left(\bar{e}_{\hat{r}},\bar{e}_r\right)$进行余弦相似度计算可以得到提问与答复之间的相似度指标 Vector Extrema（VE）。

余弦值越接近 1，就表明夹角越接近 0 度，也就是两个向量越相似，夹角等于 0，即两个向量相等。

4.3 评价方案的实现

1. 词向量

通过了解每一个词的意思来判断回复的相关性，词向量是实现这种评价方法的基础。下载腾讯 AILab 的中文词向量，然后给句子里每个词分配一个词向量，得到答复和提问语句的句向量。

2. 相似度指标

通过余弦距离得到两者相似度进行比较。

其中 Greedy Matching 指标，将得出的答复的词向量与提问语句的每个词向量计算余弦相似度，并取最大值。这个指标特点是仅在词向量基础上语句的相似性，很难捕捉长距离的语义。

Vector Extrema 指标，原理是在句向量上计算相似度的向量极值法。通过筛选词向量的每一维来选择整句话中极值最大的一维作为这个句子的向量表示，当然想要更准确的表达两个回复的相似度，仅计算向量极值是不够的，还需要计算回复之间的余弦距离才能更好的表示它们之间的相似程度。

Embedding Average 指标，利用向量均值法（通过句子中的词向量计算一个句子特征向量的方法），通过对句子中每一个词的向量求均值来计算句子的向量。这种方法在除对话系统之外的很多 NLP 领域内都应用过(例如计算文本相似度的任务)。

以上三个指标可以较好地对答复意见的质量作出评价，其中余弦值越接近 1，就表明夹角越接近 0 度，也就是两个向量越相似，夹角等于 0，即两个向量相等。

本文从附件 4 中分别选取两组数据作为实验检验，主要截取数据中的留言留言详情和答复意见为实验检验，数据如下:

留言详情	答复详情
梅溪湖至今没有一个图书馆，这与梅溪湖品位极不相称。建议在艺术中心先期借一个小馆开办读书馆。方便住在梅溪湖的市民借阅。	网友“UU008706”您好！您的留言已收悉。现将有关情况回复如下：梅溪湖一期引进 A 市图书馆分馆，位于梅溪湖创新中心，已开馆营业。梅溪湖二期金菊路与雪松路东南角规划有西地省图书馆新馆，目前正在进行前期筹备工作，具体开馆时间待定。感谢您对我们工作的支持、理解与监督!2019 年 1 月 9 日
我想咨询一下，我小孩是外地户口。我们在 L5 县经商四年了，房子租在县幼儿园附近.我的孩子可以在县幼儿园就读吗？怎样可以报名？需要准备什么资料呢？	网友：您好！首先感谢您对县幼儿园的关注，2019 年招生工作尚未启动，可参照 2018 年的招生简章的条件：1、招生对象：适龄幼儿（以当年招生公告为准），身体健康，可正常参加集体活动的幼儿。2、招生范围：县城卢峰镇中心片辖区各街道的居民幼儿，并具有有效户口本（幼儿和监护人所在户口）和监护人房产证（房产证和户口本必须在招生范围）。招生形式采用网上学位申请 + 现场原件审

	核（一律不接受现场排队报名）。如果招生条件有变化请以 2019 年招生公告为准。如还有疑问请到幼儿园现场咨询，也可以通过招生公告公示的咨询电话进行咨询！感谢你对我园工作的关心与支持！ 2019 年 7 月 18 日

结果如下：

留言	EA	VE	GM
留言 1	0.0354	0.0421	0.0375
留言 2	0.3521	0.4522	0.5631

根据上表中的结果，可以明显的看出留言 2 的回答从相关性、完整性和可解释性这三个方面来说要比留言 1 好很多。由此可见，可以用 EA、VE、GM 这三个指标来评价答复意见。

参考文献

[1] 童昱强. 基于数据挖掘的网络新闻热点发现系统设计与实现[D].北京邮电大学,2019..

[2] 徐文海，温有奎．一种基于 TFIDF 方法的中文关键词抽取算法[J]．情报理论与实践，2008，31(2)：298-302．

[3] 彭云,万常选,江腾蛟,刘德喜,刘喜平,廖国琼.基于语义约束 LDA 的商品特征和情感词提取[J].软件学报,2017,28(03):676-693.

[4] 李欣雨，袁方，刘宇，等. 面向中文新闻话题检测的多向量文本聚类方法[J]. 郑州大学学报（理学版），2016，48(2)：47-52．

[5] Rong X. word2vec Parameter Learning Explained[J]. Computer Science, 2014.

[6] 张杨子. 面向对话系统回复质量的自动评价研究[D].哈尔滨工业大学,2018.