所选题目： C：”智慧政务”中的文本挖掘应用

综合评定成绩：

评委评语：

评委签名：

“智慧政务”中的文本挖掘应用

摘要

近年来，随着物联网、云计算、大数据分析等互联网信息技术的发展，政府以提高办公、监管、决策的智能水平，形成高效敏捷公开便民的新型政府为目标提出了智慧政务的概念，文本分析和数据挖掘技术则对推进“智慧政务”工程起着重要作用。

对于问题一，通过pandas对群众留言进行简单的去重。然后从清洗的文档内容进行提取构建新的文档。再按照训练集：验证集：测试集为8:1:1的比例。构建词汇表并将其转换成可训练数据集。构建字符级CNN卷积神经网络模型，建立卷积层和最大池化层，然后进行模型训练，进行多轮迭代每轮迭代利用验证集测试该次的准确率。若准确率在多轮后没有继续增加，保存在验证集上表现最好的模型，再利用该模型对测试集进行测试，输出每个类别和总体的准确率：F1-score。

对于问题二，首先利用进行分词处理，再通过TF-IDF对文本进行向量化，得到文本的TF-IDF权重，通过jieba中的词性库对前面分词的结果的词性进行记录，针对不同的词性给予其TF-IDF权重以不同的乘数以突出某些类型词汇的重要性。将新的权重矩阵放入DBSCAN模型中进行投喂，训练。根据聚类后的结果，基于各类别的留言条数，留言用户数，点赞数，反对数通过因子分析法给定不同的权值进行热度评价的构建。

对于问题三，相似度方面：先进行 jieba 分词，载入词典和停词表提高准确率，然后通过 Word2Vec、Sent2Vec 等方法将句子转换成向量表示，利用余弦相似度相加并加权即可得出句子的相似性。完整性方面：同样先分词，在构建词向量。再建立 LSTM 神经网络模型进行训练和预测，选取概率最大的标签作为预测结果。可解释性方面：首先建立建立隐含狄利克雷 LDA 模型，结合矩阵分解模型对回复进行解释，从主题层次进行分析答复的可解释性。

关键词：jieba 分词 cnn 卷积神经网络 TF-IDF 算法 Dbscan 聚类算法 Word2Vec 模型 perplexity 指标 隐含狄利克雷模型

The thesis title

Abstract

In recent years, with the Internet of things, cloud computing, big data analysis of the development of the Internet information technology, such as the government to improve the intelligent level of office, supervision and decision, to form effective agile open for the convenience of the new government to target puts forward the concept of political wisdom, text analysis and data mining technology is to promote the "e-government" wisdom engineering plays an important role.

For question one, simply de-weigh the message of the masses through Position.The new document is then built from the cleaned document content.Then according to the training set: verification set: test set for the ratio of 8:1:1.Build the vocabulary and transform it into a trainable data set.A character-level CNN convolutional neural network model was constructed, the convolutional layer and the maximum pooling layer were established, and then model training was carried out, and the accuracy of each iteration was tested by verification set.If the accuracy does not continue to increase after several rounds, the model with the best performance on the verification set is saved, and then the model is used to test the test set, and the accuracy of each category and the whole is output: f1-score.

For the second problem, first use the word segmentation processing, then use TF-IDF to vectorize the text, get the TF-IDF weight of the text, record the part of speech of the results of the previous word segmentation through the part of speech database in Jieba, and give the TF-IDF weight to different parts of speech with different multipliers to highlight the importance of some types of words. The new weight matrix is put into DBSCAN model for feeding and training. According to the result of clustering, based on the number of messages, the number of users, the number of likes and dislikes of each category, different weights are given by factor analysis to construct the heat evaluation.

For the third question, similarity: first, jieba word segmentation, loading dictionary and stop word list to improve the accuracy, and then Word2Vec, Sent2Vec and other methods to convert the sentence into vector representation, using cosine similarity sum and weight to get the similarity of the sentence.Integrality: again, word segmentation, in the construction of word vectors.Then the LSTM neural network model was established for training and prediction, and the label with the highest probability was selected as the prediction result.Interpretability: firstly, the implicit dirichlet LDA model is established, then the response is interpreted by combining with the matrix decomposition model, and the interpretability of the response is analyzed from the subject level.

Key words: Jieba participle CNN convolutional neural network TF - IDF algorithm Dbscan clustering algorithm Word2Vec model Perplexity indicators Latent Dirichlet allocation model

目 录

1.挖掘目标

本次建模目标是利用互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见数据，利用 TensorFlow 中实现 CNN 进行文本分类、Dbscan 算法和答复评判模型，达到以下几个目标：

1） 利用该分类算法，对群众留言进行划分，提高划分准确率、政府管理水平和施政效率，以便后续将群众留言分派至相应的职能部门处理。

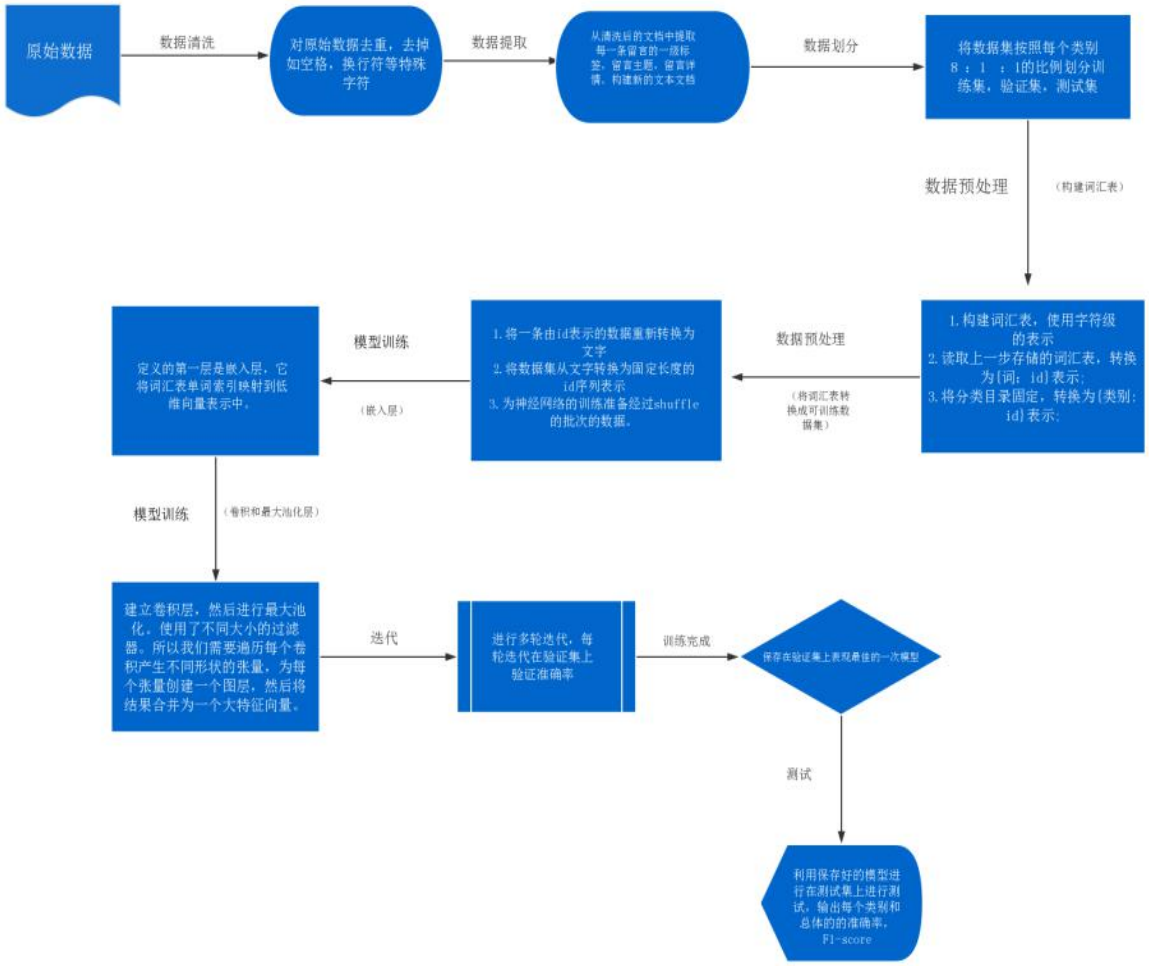
2） 构建热度评估模型分析群众留言不同类的热度。利用该模型使政府部门进行针对性处理提高服务效率。

3） 根据给定的群众问题——政府答复数据集，根据答复相似性完整性和可解释性几个方面建立答复评估模型，利用该模型对政府答复进行分析，有利于政府部门完善工作，提高群众的满意度，构建更加和谐的社会。

2.分析方法与过程

2.1 问题 1 分析方法与过程

2.1.1 流程图



2.1.2 数据预处理

2.1.2.1 数据清洗

2.1.2.2 构建词汇表

1、给定每个句子的最大长度，给定句子相同的长度是十分必要的，因为它有利于我们后续对数据进行批处理，因为批处理中的每个示例必须具有相同的长度。

2、构建词汇表，使用字符级的表示，将每个单词映射成整数，从而使每个句子都成为整数的向量.

3. 读取上一步存储的词汇表，转换为`{词：id}`表示，并且将分类目录固定，转换为`{类别: id}`表示；然后将一条由 id 表示的数据重新转换为文字;将数据集从文字转换为固定长度的 id 序列表示。

4.处理完后的数据格式如下所示：

Data	Shape	Data	Shape
x_train	[7360, 600]	y_train	7360, 10]
x_val	[920, 600]	y_val	[920, 10]
x_test	[920, 600]	y_test	[920, 10]

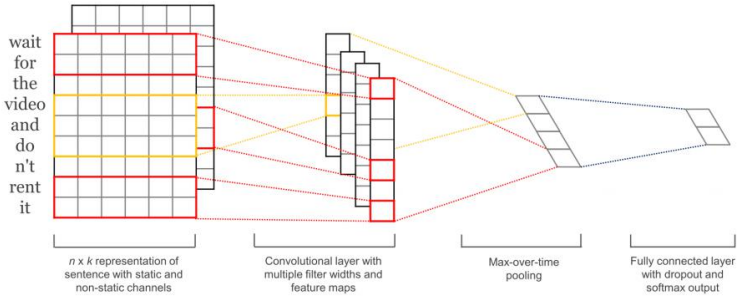
注：相关处理文件在相关程序的 data-process/loader.py 中

2.1.3 卷积神经网络

2.1.3.1 卷积神经网络模型构建[1]

该模型

我们将构建的网络大致如下：第一层将单词嵌入到低维向量中。下一层使用多个滤波器大小对嵌入的单词向量执行卷积。接下来，我们将卷积层的结果最大池化为一个长特征向量，添加掉落正则化，然后使用 softmax 层对结果进行分类。



输入占位符

我们首先定义传递到网络的输入数据：tf.placeholder 创建一个占位符变量，当我们在训练或测试时执行它时，我们将它提供给网络。第二个参数是输入张量的形状。None 意味着这个维度的长度可以是任意的。在我们的例子中，第一个维度是批处理大小，并使用 None 允许网络处理任意大小的批处理。将神经元保存在退出层的概率也是网络的输入，因为我们只允许在训练期间退出。我们在评估模型时禁用它

嵌入层

我们定义的第一层是嵌入层，它将词汇索引映射为低维向量表示。它本质上是我们从数据中学习到的查找表。

卷积和池化层

准备好构建我们的卷积层，然后是池化层。使用不同大小的过滤器。因为每个卷积产生不同形状的张量，需要迭代它们，为它们创建一个层，然后将结果合并成一个大的特征向量。

脱落层

脱落也许是最流行的方法来正则化卷积神经网络。一个脱落层随机地“破坏”了它的一小部分神经元。这阻止神经元共同适应，迫使他们学习个别有用的特征。

2.1.3.2 卷积神经网络模型可视化

检查点

我们使用的另一个 TensorFlow 功能检查点-保存模型的参数，以便稍后恢复它们。检查点可以用来在以后继续训练，或者使用早期停止来选择最佳的参数设置。以便于迭代过程中保存最佳状态。

初始化

cnn 模型参数的配置：

```
"""CNN配置参数"""

embedding_dim = 100  # 词向量维度
seq_length = 600  # 序列长度
num_classes = 7  # 类别数
num_filters = 256  # 卷积核数目
kernel_size = 5  # 卷积核尺寸
vocab_size = 8000  # 词汇表小

hidden_dim = 128  # 全连接层神经元

dropout_keep_prob = 0.5  # dropout保留比例
learning_rate = 1e-3  # 学习率
# 64
batch_size = 64  # 每批训练大小
num_epochs = 20  # 总迭代轮次

print_per_batch = 100  # 每多少轮输出一次结果
save_per_batch = 10  # 每多少轮存入tensorboard
```

训练回路

最后，编写我们的训练循环。我们多次迭代数据，调用 train_step 函数，并评估和检查我们的模型

在 Tensorboard 中可视化结果

我们的培训脚本将摘要写入输出目录，并通过指向 Tensorboard 在这个目录中，我们可以可视化我们创建的图表和摘要。

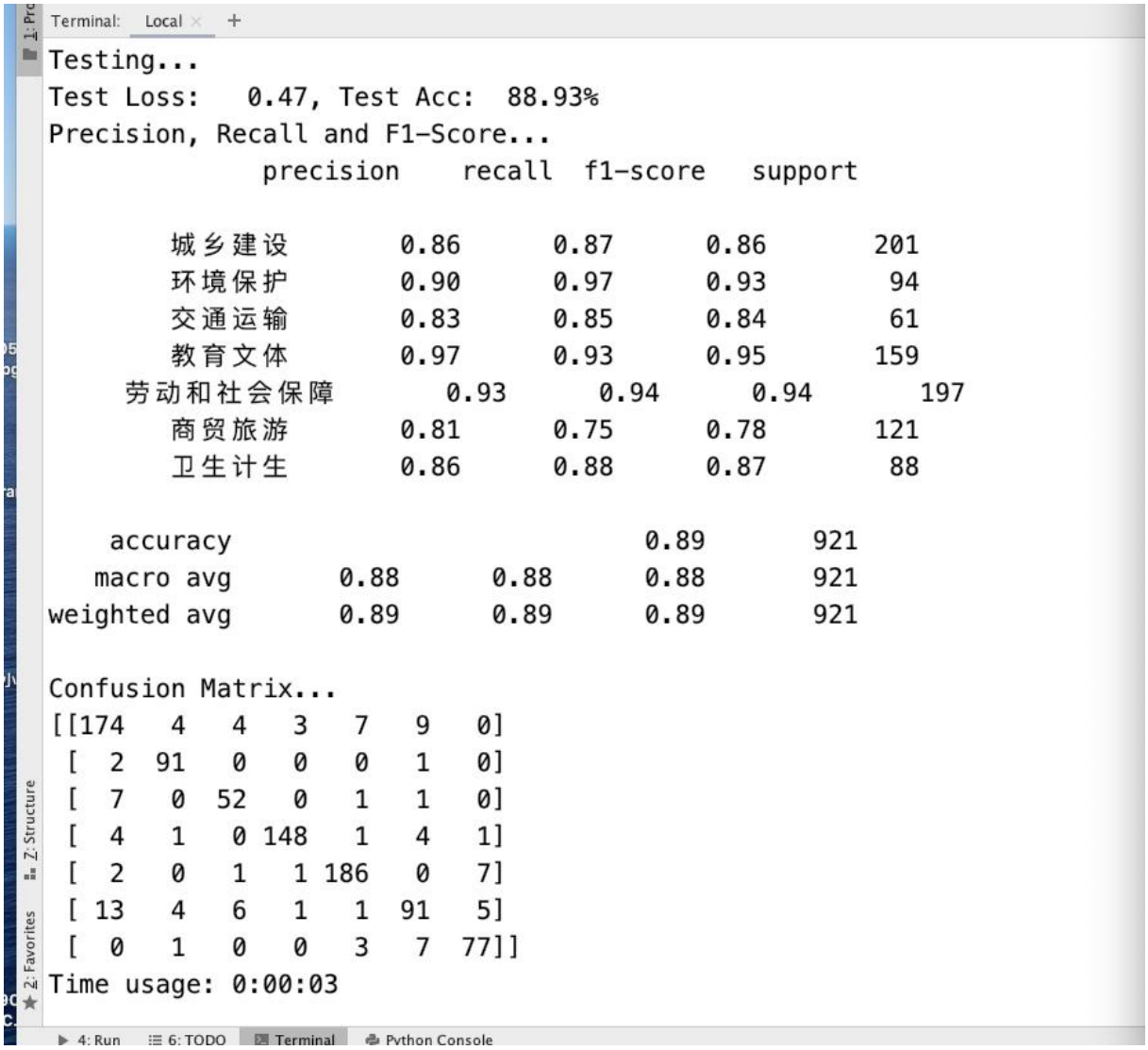
2.1.4 结果分析:

如图所示，经过 17 轮训练迭代后，模型在验证集上的准确率不再增加。在验证集上的最高准确率达到 88.08%。保存改表现最佳的模型。

```
Iter:   1500, Train Loss: 0.00082, Train Acc: 100.00%, Val Loss:    0.5, Val Acc:  88.08%, Time: 0:04:57
Epoch: 14
Iter:   1600, Train Loss: 0.00092, Train Acc: 100.00%, Val Loss:   0.53, Val Acc:  87.54%, Time: 0:05:17
Epoch: 15
Iter:   1700, Train Loss: 0.0019, Train Acc: 100.00%, Val Loss:   0.59, Val Acc:  87.54%, Time: 0:05:37
Epoch: 16
Iter:   1800, Train Loss: 0.0018, Train Acc: 100.00%, Val Loss:    0.6, Val Acc:  87.54%, Time: 0:05:57
Epoch: 17
Iter:   1900, Train Loss: 0.0034, Train Acc: 100.00%, Val Loss:   0.62, Val Acc:  85.81%, Time: 0:06:19
No optimization for a long time, auto-stopping...
```

在测试集上进行测试，如图所示。在测试集上的表现效果 F1-score 达 89%

混淆矩阵和各类别的准确率如图所示



2.2 问题 2 分析方法与过程

2.2.1 流程图

2.2.2 数据预处理

对用户留言主题和留言详细内容利用 jieba 进行中文分词。并记录每个词的词性。便于后面的词向量训练和词性分析

2.2.3 TF-IDF

TF-IDF（词频-逆文档频率）算法是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。该算法在数据挖掘、文本处理和信息检索等领域得到了广泛的应用，如从一篇文章中找到它的关键词。

TFIDF 的主要思想是：如果某个词或短语在一篇文章中出现的频率 TF 高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。TF-IDF 实际上就是 TF*IDF，其中 TF（Term Frequency），表示词条在文章 Document 中出现的频率；IDF（Inverse Document Frequency），其主要思想就是，如果包含某个词 Word 的文档越少，则这个词的区分度就越大，也就是 IDF 越大。对于如何获取一篇文章的关键词，我们可以计算这边文章出现的所有名词的 TF-IDF，TF-IDF 越大，则说明这个名词对这篇文章的区分度就越高，取 TF-IDF 值较大的几个词，就可以当做这篇文章的关键词。

TF-IDF 实际上是 TF 和 IDF 的组合。TF 即词频（Term Frequency），IDF 即逆向文档频率（Inverse Document Frequency）。

2.2.4 生成 TF-IDF 词向量权重矩阵

1.使用 TF-IDF 对文本进行向量化，得到文本的 TF-IDF 权重。
2.统计每个词语的 tf-idf 权值，将文本转为词频矩阵。
3.将 tf-idf 矩阵抽取出来，元素 w[i][j]表示 j 词在 i 类文本中的 tf-idf 权重

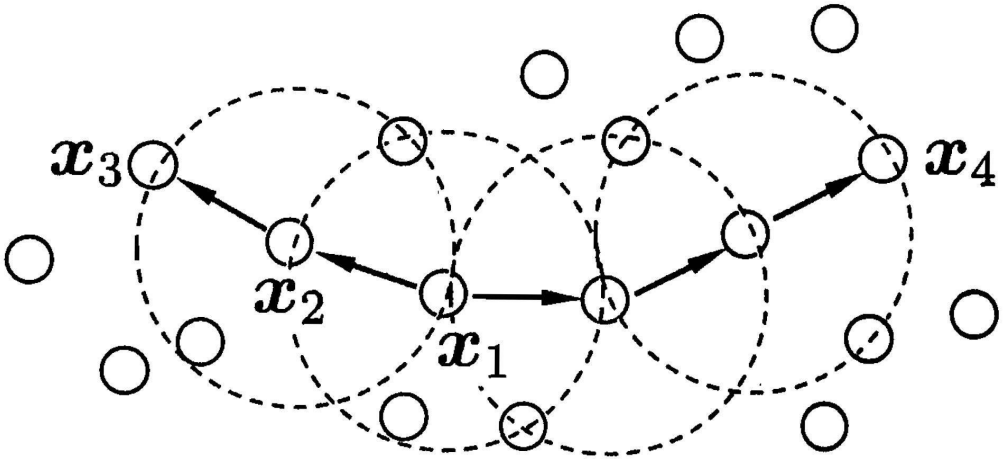
2.2.5 构建基于词性的新权重

1.因为在一条留言中不是所有的词语能特征表现这句留言的特点。所以基于不同的词性给定不同的权值会给聚类带来比较好的效果。
2.部分地名提取，对于留言中的区域地名如 A 市，A1 区，A3 县这样的地名在结巴词库中是没有的。所以，我们通过正则表达式提取出相关的地名共 94 条保存在 newns.txt 中并且设定词性为 ns（地名）。通过 jieba 的加载自定义词典将这些新地名加载进去。
3.前面得到了分词的结果，并对词性进行了记录，接下来可以针对不同词汇的词性吗，给与其 TF-IDF 权重以不同的乘数，这样可以突出某些类型的词汇的重要性。
4.在本题中我们给定的乘数是：名词（n）*1.2　地名（ns）*1.2 动名词（vn）*1.1

2.2.6 DBSCAN 模型

2.2.6.1 DBSCAN 原理

DBSCAN(Density-Based Spatial Clustering of Applications with Noise)是一个比较有代表性的基于密度的聚类算法。与划分和层次聚类方法不同，它将簇定义为密度相连的点的最大集合，能够把具有足够高密度的区域划分为簇，并可在噪声的空间数据库中发现任意形状的聚类。



DBSCAN 中的几个定义：

1.E邻域：给定对象半径为E内的区域称为该对象的E邻域；

2.核心对象：如果给定对象E邻域内的样本点数大于等于 MinPts，则称该对象为核心对象；

3.直接密度可达：对于样本集合 D，如果样本点 q 在 p 的E邻域内，并且 p 为核心对象，那么对象 q 从对象 p 直接密度可达。

4.密度可达：对于样本集合 D，给定一串样本点 p1,p2….pn，p= p1,q= pn,假如对象 pi 从 pi-1 直接密度可达，那么对象 q 从对象 p 密度可达。

5.密度相连：存在样本集合 D 中的一点 o，如果对象 o 到对象 p 和对象 q 都是密度可达的，那么 p 和 q 密度相联。

步骤：

1.DBScan 需要二个参数： 扫描半径 (eps)和最小包含点数(minPts)。 任选一个未被访问(unvisited)的点开始，找出与其距离在 eps 之内(包括 eps)的所有附近点。

2.如果附近点的数量 ≥ minPts，则当前点与其附近点形成一个簇，并且出发点被标记为已访问(visited)。然后递归，以相同的方法处理该簇内所有未被标记为已访问(visited)的点，从而对簇进行扩展。

3.如果 附近点的数量 < minPts，则该点暂时被标记作为噪声点。

4.如果簇充分地被扩展，即簇内的所有点被标记为已访问，然后用同样的算法去处理未被访问的点。

2.2.6.2 DBSCAN 实现。

1.DBSCAN 需要两个参数 eps，和 min-samples.其中 eps 为聚类半径。尝试通过对 eps 和 min_samples 的各自组合进行拟合计算。得出 eps 的最佳值在 1-1.08 之间。经过多次试验。最终选定 1.03 为 eps；min_samples 我们将它设置为 2，以发现更多的类。

2.将前面训练好的新的权重矩阵放入 dbscan 进行训练。

3.训练成功后剔除标签为-1 的点及离群点，共 2989 条。以便后面简化操作。

4.聚类完成

2.2.6.3 DBSCAN 聚类评价

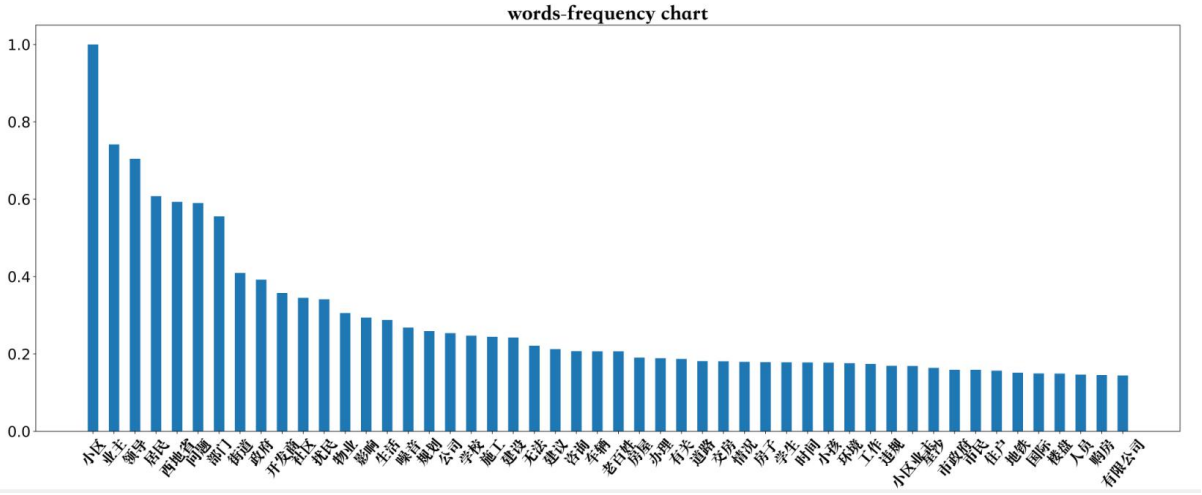
DBSCAN 的主要优点有：

1.可以对任意形状的稠密数据集进行聚类，相对的，K-Means 之类的聚类算法一般只适用于凸数据集。

2.可以在聚类的同时发现异常点，对数据集中的异常点不敏感。

3.聚类结果没有偏倚，相对的，K-Means 之类的聚类算法初始值对聚类结果有很大影响

对于此数据集来说数据量大，离群点多且类别多，dbscan 是比 k-means 更好的选择。



2.2.7 热度分析:

2.2.7.1 数据提取

1.首先提取出各类别的留言数目，记为 counts 以字典形式储存。

2.其次提取出各类别的点赞数之和-反对数之和，记为 X，同样以字典形式储存。

3.并且提取出各类别的留言用户数，记为 users_counts，以字典形式储存。

4.最后提取各类别中的时间维度即最早时间和最晚时间之差，记为 times-span

2.3.7.2 热度算法构建

1.指标

F1	F2	F3
用户数	留言数量	点赞数
留言数量	时间跨度	反对数

2.权重分配

1.利用每个类别中留言数量和用户数量之差 判断该类留言是否是一个用户多次留言。以此来降低此类问题的热度。

2.通过留言数量和时间跨度（天数）之比来判定是否是一段时间内的多数问题。

3.通过点赞数和反对数之差来判断该类留言是否为大众支持的言论。

4 构建数学模型:

根据统计 counts 的最大值为 40，user_counts 的最大值为 9，X 的最大值为 609 且与平均数相差较大。

基于该文档中的留言数量。以该分类中的留言数目为主要评判标准。利用对数函数将 X 进行平滑。最终的线性模型如下:

F1=0.7*counts-0.3*(counts-user_counts)

F2=counts/times-span

F3=ln（1+X)

Hot_vaule=0.6*F1+0.2*F2+0.2*F3

结果分析：

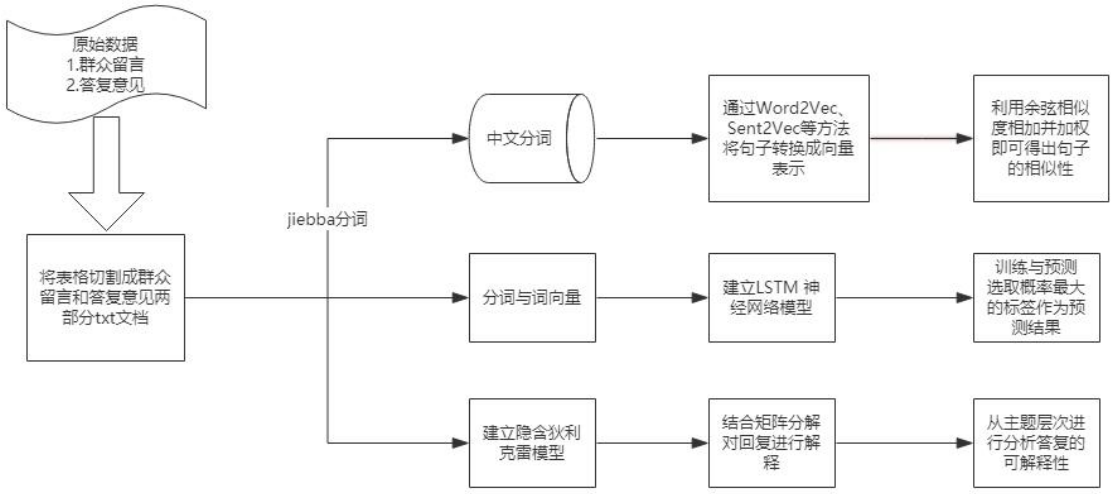
利用 dbscan 算法聚类的效果相较于 k-means 较好。事先设定较小的 eps，来排除较多的离群点。和对不同的词性给定不同的权值来加强文本特征有利于对热点问题的挖掘。部分聚类结果展示：

A市伊景园滨河苑强行捆绑车位销售，诱骗购房者交付车位定金，还不写入协议，不给签合同。请有关领导帮忙查处。	0	0	17
是与中央文件及法律法规不符，但辛辛苦苦一辈子的老百姓房都买不起，要贷款，何来钱买车位？并且买这套房还要将原福利房腾退，这些霸王条款是谁给	0	0	17
受害了购房职工的合法权益，在不签购房合同的情况下，以取消购房资格相要挟，逼迫职工交钱，违法捆绑销售车位。其中的猫腻不得而知！民愤难平！请	0	0	17
而且很大部分车都没有，买了车位也只是空在那里，希望能取消车位捆绑。2.希望这个房子对职工来说能够真正的得到实惠，因为现在要我们统一交18.5万	0	0	17
苦攒钱买个房这么多事！先是强行收我首期认购金，还不给我合同，说什么预购不用！后面就没有消息了，还停工？现在又想强迫我购买车位，有病吗？	0	0	17
什么，但是对于一名普通职工来说，要多拿出12万的车位费用是伤筋动骨的，购房者中有一部分人是退休职工，一个月拿的退休金还不一定够还房贷，而且	0	0	17
文件内容为广铁集团与A市政府及A市政工程有限公司协商要求职工购买房子的同时一对一购买所谓按成本价销售的12万一个的车位。请问A市政府为何会允	0	0	17
定的房子同时一对一购买所谓按成本价销售的12万一个的车位，这个还是和谐社会么？政府的正经职能是协助房地产公司卖房么？和谐社会，不是依法治国	0	0	17
无辜职工获取利益，已经收了18.5万的认购款却没网签、没合同，现在还要求一户一车位，捆绑销售车位，12万的车位如果不买就取消购房资格，这是欺骗	0	1	17
工程停工快一年了，没有任何解释和交代，并强行要求捆绑车位，请有关政府部门为民做主！2803名购房职工的合法权益受到侵害，捆绑销售车位合法吗？	0	0	17
共有，但属于城镇公共道路的除外。建筑区划分内的绿地属于业主共有，但属于城镇公共绿地和明示属于个人的除外。建筑区划内的其他公共场所，公用设	0	0	17
广铁集团伊景园滨河苑商品房本来是政府为铁路职工购房提供福利的项目，但现在却要求购房者在买房时还要购买高额的车位，开发商是这样配合政府政策	0	2	17
目原本是为了解决广铁集团基层职工住房问题的，但广铁集团在项目中非法绑定高价车位出售，单个车位高达12万占总价的40%。如此高价是非常不合理的	0	0	17
逼迫职工缴纳18.5万购房款且不签正式购房合同，工地停工1年多，现广铁集团又强烈要求职工捆绑购买12万车位一个，不买就取消购房资格，房子都要贷	0	0	17
都抢12万,在未取得预售资格强行逼迫职工缴纳18.5万购房款且不签正式购房合同，工地停工1年多，现广铁集团又强烈要求职工捆绑购买12万车位一个，不	0	0	17
伊景园滨河苑收取购房者18万5千元的认购金后不与购房者签订合同，恶意诈骗购房者钱财。	0	0	17
堵大的广铁集团，竟不顾职工死活，打着为职工谋福利的牌子，违法捆绑车位销售，难道就没地方讨回公道了？请别让老百姓对这个社会失望，请求相关工	0	1	17
与A市政府及A市政工程有限公司要求职工购买房子的同时一对一购买所谓按成本价销售的12万一个的车位。虽然伊景园滨河苑项目是定向福利房。明文要求	0	0	17
新城片区的伊景园滨河苑是我们的定向限价商品房,但现在楼盘还未建成,广铁集团却要求捆绑购买12万的车位，否则取消购房资格，我们不需要车位.请领导	0	0	17
A市伊景园 · 滨河苑强制要求购房者捆绑购买其根本不需要的车位，不买就取消购房资格，2千多购房者饱受欺压，无处伸冤。	0	0	17
签购房合同，取消购房资格，2018年6月29日，国家七部委发文整治房地产市场乱象，A市七部委联合行动打击违法行为，怎么开发商置国家文件形同虚设，	0	0	17
发文件内容为广铁集团与A市政府及A市政工程有限公司协商要求职工购买房子的同时一对一购买所谓按成本价销售的12万一个的车位，是是什么情况。强	0	0	17
A市伊景园 · 滨河苑违规涨价18.5万元，捆绑价值12万的车位，恶意坑害广大购房者权益。	0	0	17
职工谋福利呢！没想到竟然强行要求我们职工缴纳数十万的预售购房款，还不带合同！原来以为集团不坑我们的，没当回事，后来要求我们同时要购买车	0	0	17
本人购买伊景园滨河苑楼盘房产，开发商联合广铁集团强制要求铁路职工买房必须购买车位要不然不签合同。根据A市政策这是属于违法的。请政府主持	0	0	17
捆绑房子和12万的车位一对一销售，否则取消职工购房资格。我们查阅了《物权法》的相关规定，发现小区内的车位应该属于业主共有，所以这样捆绑销售	0	2	17
部分人已经没有能力再拿出12万购买车位了，并且很多都是没有车的并没有购买车位的需求。网上查了《反不正当竞争法》第十二条中规定：经营者销售商	0	0	17
产权车位，不交满12万的车位款不给业主网签购房合同，请国家政府部门立即制止开发商这种违法捆绑车位销售行为，维护2000多名业主的合法权益！证据	0	0	17

热点问题挖掘结果在附件热点问题表与热点问题明细表中

2.3 问题 3 分析方法与过程

2.3.1 流程图



2.3.2 数据预处理

2.3.2.1 对政府答复进行中文分词

在对政府答复进行相似度分析之前，需要将非结构化的文本信息转换成计算机能够偶识别的结构化信息。在附件中，以中文文本方式给定了数据，为了方便转换，首先对答复信息进行中文分词。例如“联想移动通信科技有限公司”，我们希望将其切分为“联想 移动 通信 科技 有限 公司”。python 提供专门的中文切词工具“jieba”，它可以将中文长文本划分为若干个单词。我们利用 jieba.load_userdict(file_name)函数导入自定义词典，词典中有大量和政务办公相关的词组，增加分词的准确率，为了进一步提高分词的准确率，我们还需要考虑干扰因素：一是英文字母大小写的影响，为此我们将英文字母统一转换为大写；二是例如 “有限”、“责任”、“股份”、“公司”等通用的词汇，我们将这样的词汇连同“（）”、“-”、“/”、“&”等符号作为停用词，将其从分词结果中去除掉，最后得出有效的词汇组合。

2.3.2.2 Word2Vec 模型

简介：Word2vec，是一群用来产生词向量的相关模型。这些模型为浅而双层的神经网络，用来训练以重新建构语言学之词文本。网络以词表现，并且需猜测相邻位置的输入词，在 word2vec 中词袋模型假设下，词的顺序是不重要的。训练完成之后，word2vec 模型可用来映射每个词到一个向量，可用来表示词对词之间的关系，该向量为神经网络之隐藏层。

依赖：词袋模型

词袋模型（Bag-of-words model）是个在自然语言处理和信息检索(IR)下被简化的表达模型。此模型下，像是句子或是文件这样的文字可以用一个袋子装着这些词的方式表现，这种表现方式不考虑文法以及词的顺序。最近词袋模型也被应用在计算机视觉领域。词袋模型被广泛应用在文件分类，词出现的频率可以用来当作训练分类器的特征。关于"词袋"这个用字的由来可追溯到泽里格·哈里斯于 1954 年在 Distributional Structure 的文章。

有关术语：词向量

词向量具有良好的语义特性，是表示词语特征的常用方式。词向量每一维的值代表一个具有一定的语义和语法上解释的特征。所以，可以将词向量的每一维称为一个词语特征。词向量具有多种形式，distributed representation 是其中一种。一个 distributed representation 是一个稠密、低维的实值向量。distributed representation 的每一维表示词语的一个潜在特征，该特 征捕获了有用的句法和语义特性。可见 ，distributed representation 中的 distributed 一词体现了词向量这样一个特点：将词语的不同句法和语义特征分布到它的每一个维度去表示。

2.3.2.2 生成 Word2Vec 词向量

第一步：分词和统计词频，统计词频时除了去除停用词之外，还需要去掉非常高频的词和非常低频的词。去掉高频词是因为没有特殊性，去掉低频词是因为没有普适性。

第二步：构建 huffman 树。所有的非叶节点存储有一个参数向量，所有的叶节点分别代表了词典中的一个词。参数向量初始值为 0。构建完 huffman 树之后，将对应的 huffman 码分配给每个单词。此外，还需要随机初始化每个单词的词向量。

第三步：训练模型，采用 CBOW 模型，CBOW 中训练集的输入是周围几个单词的词向量之和，输出是中间那个单词。从根节点开始，沿着 huffman 树不停地进行 logistic 分类，每进行一次分类就沿着 Huffman 树往下一层并更正词向量，直到最后达到叶节点。

2.3.3 相似性分析

2.3.3.1 利用余弦相似度计算文本相似度

余弦相似度，又称为余弦相似性，是通过计算两个向量的夹角余弦值来评估他们的相似度。余弦相似度将向量根据坐标值，绘制到向量空间中，如最常见的二维空间。相比欧氏距离度量，余弦相似度更加注重两个向量在方向上的差异，而非距离或长度上的差异。余弦值的计算公式如下：

$$\cos\theta = \frac{\sum_1^n (A_i * B_i)}{\sqrt{\sum_1^n A_i^2} * \sqrt{\sum_1^n B_i^2}} \quad (1)$$

相对于欧氏距离，余弦相似度更适合计算文本的相似度。首先将文本转换为权值向量，通过计算两个向量的夹角余弦值，就可以评估他们的相似度。余弦值的范围在[-1,1]之间，值越趋近于 1，代表两个向量方向越接近；越趋近于-1，代表他们的方向越相反。为了方便聚类分析，我们将余弦值做归一化处理，将其转换到[0,1]之间，并输出，值越小表示距离越接近，相似度越高。

我们从附件四中任意选择五条留言详情和对应的答复意见，对其进行余弦相似度计算并输出结果。

相似度分析结果：

----------The result of bm25 method----------
The computing cost 4.012 seconds
{
"A2区景蓉花苑物业管理有问题：小区停车收费问题，物业公司去留问题": [
"A2区景蓉花苑物业管理有问题",
19.468229495648778
],
"幼儿园教师待遇低，没有养老保险": [
"保障民办幼儿园教师待遇",
7.108963087847906
],
"研究生购房能否享受3万元补贴": [
"榔梨街道购房补贴由10万元",
6.54910221462265
],
"大唐印象工程施工的噪音影响人们生活": [
"该工地是大唐印象三期工程",
11.728214818455868
],
"的士司机不打表计费，存在宰客行为": [
"确实存在不打表计费的情况",
12.968688532468844
]
}

----------The result of jaccard method----------

2.3.4 完整性分析

2.3.4.1 基于双层的 Bi-LSTM 循环神经网络

本文提出的模型采用基于双层的 Bi-LSTM 循环神经网络[4]，结构如图所示。首先，模型的输入为原始文本经过预处理后的词序列，将其映射为相应的词向量并标注,经过循环滑动窗口和欠采样处理后作为 BiLSTM 的输入。然后通过双层 Bi-LSTM 更加准确地学习特征,最终通过分类器输出相应标签概率。

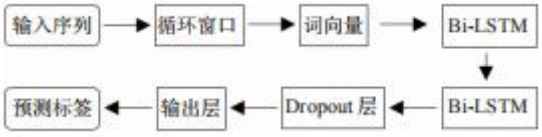


图 1　基于循环神经网络的语义完整性分析方法架构

$$\text{macro}-P=\frac{1}{n}\sum_{i=1}^{n}P_i \tag{2}$$

$$\text{macro}-R=\frac{1}{\text{n}}\sum_{\text{i}=1}^{n}\text{R}_{\text{i}} \tag{3}$$

$$\text{macro}-F1=\frac{2*macro-P*macro-R}{macro-p+macro-R} \tag{4}$$

本文属于多分类问题，我们采用准确率（*A*）、宏查准率（*macro-P*）、宏查全率（*macro-R*）以及宏*F*1（*macro- F*1）作为评价模型效果的指标. *A* 为模型整体的准确率。其他指标计算方式如下，其中*n* 为类别数，*Pi*、*Ri* 分别表示第 *i* 个类别的 *P* 值和 *R* 值.

2.3.5 可解释性分析

2.3.5.1 基于主题的评判方法

建立隐含狄利克雷模型，隐含狄利克雷模型 LDA（Latent Dirichlet allocation）[3]是可解释推荐研究中提取文本评论主题信息时常用的主题模型，它是一种基于贝叶斯概率的文档主题生成模型。它将一篇文档看成是词的集合，词和词之间没有先后顺序，一篇文档中包含多个主题，文档中的每一个词都根据其中的一个主题生成，它可以将文档集中每篇文档的主题按照概率分布的形式给出。该模型包含三层结构：文档、主题和词，文档和主题之间、主题和词之间服从多项式分布。由于它是一种无监督的学习算法，在训练的时候不需要手工标注训练集，只需给定文档集和主题的数量就可以实现模型的训练。其优点是对于文档中的每一个主题均可找出描述它的相关词语。由于主题模型是根据文档、主题、词三者的概率分布计算词和词的共现规律，当文档很短的时候就不利于统计这种规律。

通过矩阵分解模型和隐含狄利克雷模型相结合，通过主题对回复进行解释，阐述了答复的可解释性。

2.4 问题三结论

答复质量评判在众多生成模型出现的现状下显得格外重要，本论文对答复质量评判的研究着眼于文本的相似性、完整性和可解释性，设计了多个非常有针对性的模型，对目前已有对话系统回复质量评价的方法进行了总结和分析，也大胆的对回复质量评价模型的发展进行了探索和尝试。本文通过对开放型对话系统的特点分析，构造了合理的指标和实验过程，并通过不同的特征抽取证明了方法的可行性，在实验中还注意保证了数据来源的多样性以此来保证得到的实验结果具有一定的普适性。

3. 参考文献

[1]Implementing a CNN for Text Classification in TensorFlow – WildML
[2]《基于文本评论的可解释推荐研究》赵丽娅 四川大学计算机学院

[3]Yang，L.，Qiu，M.，Gottipati，S.，Zhu，F.，Jiang，J.，Sun，H.，Chen，Z.:Cqarank:Jointly Model Topics and Expertise in CommunityQuestion Answering.In: Proceedings of the22nd ACM International Conference on Information & Knowledge Management，ACM，2013:99-108

[4]《基于循环神经网络的语义完整性分析》刘京麦野， 刘 新， 郭炳元， 孙道秋
[5]L.Bottou,F.FogelmanSoulie,P.Blanchet,andJ.Lienard.Experiments with time delay networks and dynamic time warping for speaker independent isolated digit recognition.In Proceedings of EuroSpeech 89,volume2,pages 537-540,Paris,France,1989.
[6]Y.-L.Boureau,F.Bach,Y.LeCun,andJ.Ponce.Learning mid-level features for recognition.In Computer VisionandPattern Recognition (CVPR),2010 IEEE Conference on,pages 2559-2566.IEEE,2010.
[7]Y.-L.Boureau,J.Ponce,andY.LeCun.A theoretical analysis of feature pooling in visual recognition.In Proceedings of the 27th International ConferenceonMachine Learning (ICML-10), pages 111-118,2010.
[8]基于新浪热门平台的微博热度评价指标体系实证研究 梁昌明 1 李冬强 2
[9].山东师范大学历史与社会发展学院，济南 250014；2. jt 京科技大学图书馆，北京 100083)