智慧政务中的文本挖掘应用

摘要

本文旨在基于来自互联网收集公开的群众问政留言记录，以及相关部门对部分群众留言的答复意见；在大数据、云计算、人工智能等技术高速发展的背景下，建立智慧政务系统；从而提高政府的管理水平和施政效率。进而让政府更好的了解民意、汇聚民智、凝聚民气。

针对问题一，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，在实验的过程中不断提高数据的查准率，最后通过词云可视化，将留言内容中的一级标签成功分类。

针对问题二，对于群众集中反映的问题，基于 K-means 聚类对相似的留言数据进行归类，使用 Word2vec 模型和 skip-gram 模型进行热点问题的挖掘。并根据题目要求把排名前五的热点问题形成表格。

针对问题三，对于相关部门的答复意见，使用 python 的 Jieba 分词和缀词典实现高效的词图扫描为基础，结合 TF-IDF 算法和 CNN-DSSM 算法进行数据转化，从答复的相关性，完整性，可释性，不断地调用和训练对模型进行优化，最后得出预测结果。

关键词： K-means 聚类；Word2vec 模型；skip-gram 模型；TF-IDF 算法；CNN-DSSM 算法

Abstract

This paper aims to collect public records of people's political messages from the Internet, as well as the comments of relevant departments on some public messages. In the background of the rapid development of big data, cloud computing, artificial intelligence and other technologies, the establishment of intelligent government system; So as to improve the level of government management and efficiency. In this way, the government can better understand public opinion, pool people's wisdom and pool people's spirit.

In view of question 1, the first-level label classification model of message content was established, and f-score was used to evaluate the classification method. During the experiment, the accuracy rate of data was constantly improved. Finally, the first-level label of message content was successfully classified through word cloud visualization.

For the second problem, for the problems that the masses have concentrated on, similar message data are classified based on k-means clustering, and the Word2vec model and skip-gram model are used to dig hot issues. And according to the requirements of the topic to the top five hot issues form a table.

Based on the efficient word graph scanning based on Jieba word segmentation and dictionary of python and combined with the tf-idf algorithm and cnn-dssm algorithm, the model was optimized through continuous invocation and training based on the relevance, integrity and interpretability of the answers, and the predicted results were finally obtained.

Key words: k-means clustering; Word2vec model; Skip - gram model; TF - IDF algorithm; CNN - DSSM algorithm

目录

一、 引言

政府是一个城市的“大脑”，建设“智慧城市”的首要任务是建设“智慧政府”。“智慧政府”不仅强调新一代信息技术应用，也强调以用户创新、大众创新、开放创新、共同创新为特征的创新 2.0，将实现作为平台的政府架构，并以此为基础实现政府、市场、社会多方协同的公共价值塑造，实现从生产范式向服务范式的转变。“智慧政府”先行，可以带动经济、社会领域的智慧化建设，如智慧企业、智慧学校、智慧医院、智慧社区等。为此，一方面，各地在编制智慧城市规划时，要把“智慧政府”作为重要内容。另一方面，“智慧政府”代表着电子政务新的发展方向，各地在编制电子政务发展规划时，也要把“智慧政府”作为重要内容，强调社会各方的广泛参与。

在智能决策方面，采用数据仓库、数据挖掘、知识库系统等技术手段建立智能决策系统，该系统能够根据领导需要自动生成统计报表；开发用于辅助政府领导干部决策的“仪表盘”系统，把经济运行情况、社会管理情况等形象地呈现在政府领导干部面前，使他们可以像开汽车一样驾驭所赋予的本地区、本部门职责。

所以，本文的目的就是在智慧政务中的文本挖掘应用，做到更高效，更准确，更方便地将群众留言加以分类。从而让相关部门能“对症下药”，更好地解决民生问题，进而构建完善的“智慧政府”。

二、 问题分析

2.1 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。而各类网络问政平台也受到了群众的喜爱，留言数量和留言质量都不断提升，为政府解决民生问题提供了重要的参考作用和价值。

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。党的十九届四中全会上，“创新互联网时代群众工作机制”被写入全会《决定》，是对各级领导干部践行网上群众路线、提升社会治理能力的重要指示。倘若政府处理留言的效率不高，会严重影响到政府的社会治理能力。所以，群众留言分类效率显得格外重要。

2.2 问题解释

问题给出了四个附件数据，附件一总结了分析留言文本分类需要用到的三个三级分类的标签，一级分类为根标签包括了城乡建设、环境保护、交通运输、教育文体、劳动和社会保障、商贸旅游和和卫生计生为主要一级标签，根一级标签再进行划分为二级标签和三级标签，附件二给出了用户进行留言时候的数据，包括留言主题、留言详情、分类的一级标签、留言时间和留言编号，附件三与附件二类似，增加了各个留言相应的反对数和点赞数，附件四采集了针对留言问题给出了系统答复的数据，主要包括留言详情、答复意见、留言和答复时间。

任务一关于留言的一级标签分类，为了更好地将群众留言派分至相应的职能部门处理，就要解决人工分类中工作量大、效率低、差错率高的问题。根据所给数据，建立关于留言内容的一级标签分类模型，在实际分类中需要结合附件一的分类标签进行，需要进行数据的特征构建观察数据的特征，结合标签与特征进行理解。

问题所给的任务二要求在众多的留言问题中，提取出社会热点问题方便工作人员进行加急处理，在附件三中难免有许多相似的问题，这些问题因为留言次数多，来自不同用户的多次反馈，根据轻重缓急来划分，就成为了社会热点问题。为了政府更加快捷地解决热点问题，提高相关部门的工作效率。要对群众的留言进行热点问题的挖掘。

相关部门对于留言的答复意见，在一定程度上可以体现镇府的服务态度。留言的回复率可以衡量一个地方“网上问政”水平。于是需要从答复的相关性、完整性、可释性等角度对答复意见进行评价。

三、 数据准备

3.1 数据的清洗

通过观察样本和程序运行结果，数据中不存在空值的情况；文本中存在留言不符合规范的语句可以视作异常样本，适当进行剔除。问题一要求的是对留言的一级分类标签，异常样本过少，分析异常样本可以进行保留；问题二对留言的热度分析，对相似的留言进行归类，异常样本可以视作离群点剔除。问题一二主要用到的留言详情进行过滤，删除不必要的标点符号、空格、字母、数字，过滤出完整的数据。

3.2 删除与分析无关的指标

通过问题分析明确所需要的指标，不同问题需要的指标不同，处理任务一时可以删除留言编号、留言用户和留言时间，处理问题二的时候附件三中指标均有所用到不做处理。

3.3 构造分析需要的指标

基于文本分类分析需要对问题一二留言详情进行分词处理，程序中构造了分词列名为 cut_clean_data，运用网上常用的分词表往往把留言数据的某些特定词语进行了分词，所以分词前需要保留不被分词的词语；针对问题一一级分类标签还需要进行转换，程序中构造了 tag 列，把相应的一级标签转换为了数字一到七，方便后续对构造器的训练。

3.4 向量化处理

为了探索数据的特征和方便后续对构造器的训练，对分词后的数据进行向量化的处理。

四、 问题解决

4.1 问题一

4.1.1 实验处理

对数据进行探索，将分词数据进行可视化，构建出每个分类词频率最高的前 50 个词语

城乡建设

环境保护

交通运输

教育文体

劳动和社会保障

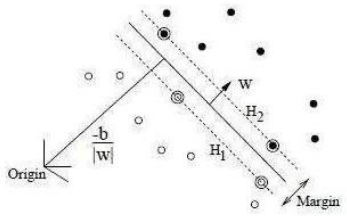
商贸旅游

卫生计生

附件一中的二三级分类可以视为一级分类的重点词频，进行筛选出类似的重点词频，提高后续模型的准确率。运用 LineSVC 分类器进行文本标签的一级分类，详细 LineSVC 原理如下：

线性支持向量机原理：

（1）线性支持向量机思想是对给定的训练样本，找到一个超平面去尽可能的分隔更多正反例。不同的是其选择最优的超平面是基于正反例离这个超平面尽可能远。



线性支持向量机模型

从上图可以发现，其实只要我们能保证距离超平面最近的那些点离超平面尽可能远，就能保证所有的正反例离这个超平面尽可能的远。因此，我们定义这些距离超平面最近的点为支持向量（如上图中虚线所穿过的点）。并且定义正负支持向量的距离为 Margin。

（2）函数间隔和几何间隔

对 SVM 思想有一定理解之后，设超平面为。我们讲解一下函数间隔和几何间隔的区别。给定一个样本，表示点 x 到超平面的距离。通过观察和是否同号，我们判断分类是否正确。所以函数间隔定义为：$\gamma' = y * (w^T x + b)$

而函数间隔不能正常反应点到超平面的距离，因为当我们等比例扩大和的时候，函数间隔也会扩大相应的倍数。因此，我们引入几何间隔。几何间隔就是在函数间隔的基础下，在分母上对加上约束（这个约束有点像归一化），定义为：

$$Y = \frac{y * (w^T x + b)}{\|w\|_2}$$

其实参考点到直线的距离，我们可以发现几何间隔就是高维空间中点到超平面的距离，才能真正反映点到超平面的距离。

将分词好的数据生成 TF-IDF 特征向量后进行训练 LinearSVC 分类器进行训练观察得出的结果。

4.1.2 实验总结

对问题的评价指标理解：

精确率和召回率是对于分类任务来说的

用 P 代表我们预测的正类，N 代表我们预测的负类，T 代表真正的正类，F 代表真正的负类

精确率(将正类样本预测成正类样本的个数对上全部预测为正类样本的比例)

精确率就是预测正确的结果中，有多少是真正正确的，也就是预测正确的样本中，有多少是正确的对上你认为正确的样本数 $TP/(TP+FP)$

召回率(将正类样本预测成正类样本的个数对上全部真正正确的样本的比例)召回率就是全部正确的样本中有多少被识别了出来$TP/(TP+FN)$

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

在实验的过程中，最大的难度就是提高模型的正确率。最后通过分词的方法，把数据的正确率提高到了 90%以上。

```
(7368, 72219)
F1值为 0.9078309845123312
正确率 0.9087947882736156
```

4.2 问题二

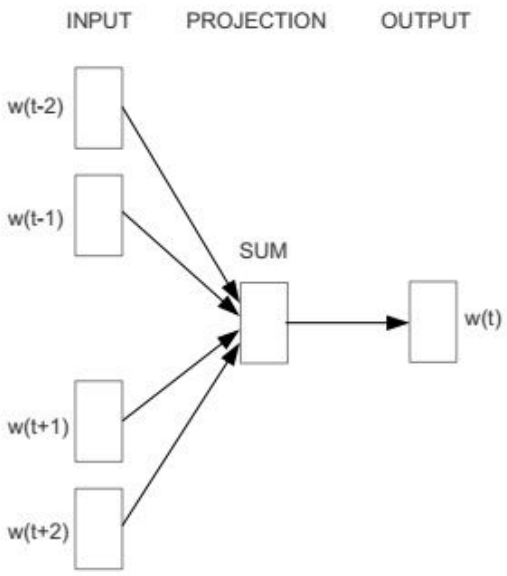
4.2.1 实验处理

基于 K-means 聚类对相似的留言数据进行归类

对数据进行探索时发现附件三留言文本存在大量相似数据，构建Word2vec进行文本的相似度探索。

Word2vec 模型具体原理如下：

（1）Continuous Bag-of-Words (CBOW)



CBOW 给定目标单词的上下文（前 c 个词以及后 c 个词）预测该目标单词是什么，可以用条件概率来建模这个问题，所以，我们的模型是求：

$$P(w_t|w_{t-c}:w_{t+c})$$

对于给定的一句话 w1,w2,w3...wT，该模型的目标函数就是最大化上式的对数似然函数：

$$L=\frac{1}{T}\sum_{t=1}^{T}logP(w_t|w_{t-c}:w_{t+c})$$

T: 句子长度

wt: 要预测的目标单词

c: 上下文大小

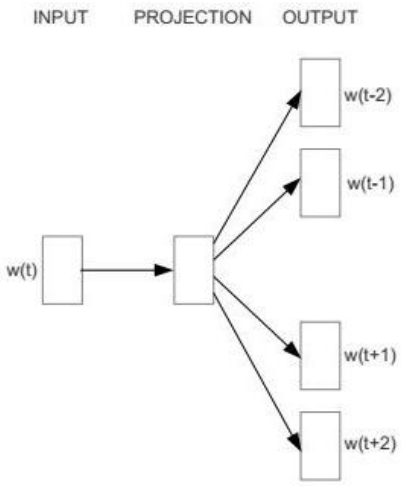
条件概率由 softmax 给出：

$$P(w_t|w_{t-c}:w_{t+c})=\frac{exp(\bar{v}^T v_{w_t})}{\sum_{n=1}^{N} exp(\bar{v}^T v_n)}$$

$$\bar{v}=\frac{1}{2c}\sum_{-c<=j<=c,j\neq 0} v_j$$

该模型的计算复杂度：n * P + P * V

（2）Skip-gram



skip-gram 模型与 CBOW 恰好反过来，它是在给定一个单词的条件下，预测其上下文单词最有可能是哪些。

给定一句话 w1,w2...wT 该模型的目标函数：

$$L=\frac{1}{T}\sum_{t=1}^{T}\sum_{-c<=j<=c,j\neq 0} logP(w_{t+j}|w_t)$$

$$P(w_{t+j}|w_t)=\frac{exp(w_{t+j}^T w_t)}{\sum_{n=1}^{N} exp(w_{t+j}^T w_n)}$$

该模型的计算复杂度：

C * (P + P * V)

在用梯度下降对 L 求导时，计算代价正比于整个单词表大小 N 。对此，常见的优化手段包括：

Hierarchical softmax: 不是直接计算 softmax，而是通过构建一颗二叉树，把问题转化为 Log N 次二分类问题。spark 上 word2vec 实现采用此方法。

Negative sampling： 负采样，完全抛弃了 softmax，从词汇表随机采样 neg 个(context(w),wi)构成负样本，训练里存在的(context(w)， w)构成正样本，将多分类问题转化为 neg 个二分类问题。tensorflow 上的 word2vec 是负采样实现的。

将分词后的数据用 Word2Vec 进行训练，设置参数 size = 400, min_count = 15，表示向量维度为 400，词汇至少出现 15 次，构建出词表如下图，训练完后数据的维度为(4010210, 5328600)

dict_keys(['A市', 'A3', '春天', '栋', '一家', '据说', '一个', '居民楼', '内部', '时间', '请', '税务局', '工商局', '查', '一下', '看看', '有没有', '正常', '没有', '应该', '会', '操作', 'A6', '区', '道路', '规划', '已经', '初步', '公示', '文件', '成为', '正式', '希望', '加快', '完成', '规范', '安装', '变更', '路', '及时', '更换', '农村', '10', '年', '未曾', '统一', '现在', '找', '地方', '只能', '说', '路口', '作用', '调整', '完毕', '更新', '同步', '开展', '春华', '组', '村民', '不知', '是否', '相关', '水泥路', '到户', '政策', '自来水', '政府', '主导', '投资', '部分', '集资', '提出', '个人', '意见', '建立', '解决', '民生问题', '之上', '大部分', '没', '很多', '通', '路灯', '天', '黑', '晚上', '有人', '出行', '地区', '农田', '整改', '二', '家庭', '用水', '相当', '紧张', '真正', '需要', '不能', '依靠', '资金', '用于', '觉得', '农民', '干', '实事', '步行街', '城', '南路', '街道', '一步', '两', '小区', '停车场', '东面', '围墙', '外', '第一', '单元', '住户', '卫生间', '直接', '人行', '马路', '一栋', '楼', '户', '居民', '长期', '臭气熏天', '非常', '恶心', '进行', '投诉', '处理', '一次', '仅仅', '表面', '根本', '问题', '主要', '原因', '约', '长', '外面', '管道', '堵塞', '开挖', '路边', '重新', '完全', '未能', '再次', '强烈呼吁', '部门', '尽快', '此事', '百姓', '生存环境', '城市', '特别', '创建', '全国', '文明城市', '更', '细节', '共同', '建设', '美好', '家园', '中海', '国际', '社区', '三期', '四期', '中间', '蓝天', '洲', '幼儿园', '旁边', '块', '空地', '一直', '处于', '状态', '物业', '城管', '市政', '去年', '周围', '建筑工地', '这块', '土', '过来', '挖', '目前', '每天晚上', '每天', '十点', '工作', '凌晨', '五点', '噪音', '高达', '70', '分贝', '多位', '自称', '干部', '工作人员', '打电话', '业主', '电话', '人员', '接到', '现场', '亲自', '了解', '情况', '恶劣', '充满', '无奈', '表明', '太', '专业', '夜间', '施工',

训练完后的模型实质，观察词向量的相似度，输入学生观察，与在校，补课，上课等词语相关，依据现实分析是比较合理的

[('在校', 0.8708213567733765), ('接送', 0.8526713848114014), ('补课', 0.8473405838012695), ('上课', 0.8397630453109741), ('放学', 0.8352636694908142), ('家长', 0.8225957155227661), ('周末', 0.8156570792198181), ('小朋友', 0.8155423402786255), ('幼儿', 0.8126198053359985), ('训练', 0.8039258122444153)]

5.2.2 K-Means 算法原理

K-Means 算法是一种无监督分类算法，假设有无标签数据集：

$$X = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(m)} \end{bmatrix}$$

该算法的任务是将数据集聚类成个簇，最小化损失函数为：

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2$$

其中为簇的中心点：

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in Ci} x$$

要找到以上问题的最优解需要遍历所有可能的簇划分，K-Mmeans 算法使用贪心策略求得一个近似解，具体步骤如下：

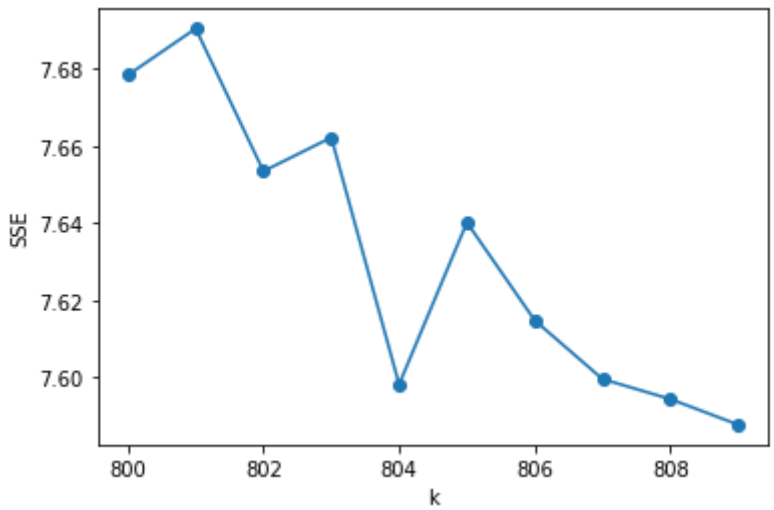
（1）在样本中随机选取个样本点充当各个簇的中心点。

（2）计算所有样本点与各个簇中心之间的距离，然后把样本点划入最近的簇中。

（3）根据簇中已有的样本点，重新计算簇中心。

重复 2、3 步骤，直到质心的位置不再发生变化或者达到设定的迭代次数。

通过分析文本的相似度对留言文本进行聚类，聚类过程中发现结果受噪声影响过大，通过增大 K 值的分类数进行减缓噪声，绘制 K 值与评价参数 inertia_值得关系可以进行 K 值得选取，其中 inertia_评价参数表示的是簇中某一点到簇中距离的和，通过绘制图形选取了 K 值为 810 时，评价参数 inertia_值较小，说明聚类后的簇结果更加的精细。



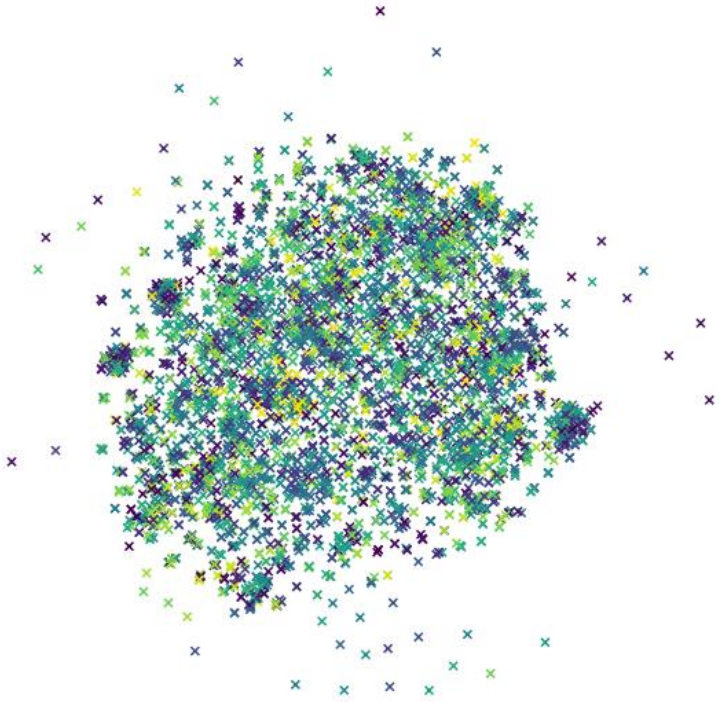
根据数据构建出 600 个分类，选取了聚类中心 600 个，对应数据的标签有 4326

```
聚类中心：[[-2.05019932e-02 -1.23732150e-02 -2.10720068e-02 ... -2.57242184e-02
  -1.10671409e-02  1.38169240e-06]
 [-7.36230041e-02 -1.89769030e-01  3.66857655e-01 ... -6.55990613e-03
  -3.40614300e-02 -1.20670669e-02]
 [ 4.67549943e-02 -3.23100579e-02 -2.50850055e-02 ... -6.17742823e-03
  -1.54825890e-02  5.58596045e-02]
 ...
 [ 7.31928637e-02 -1.06181394e-01 -6.59774916e-02 ... -6.59027615e-02
  -4.34407538e-02  3.61769927e-02]
 [-1.63243854e-02  9.84019309e-02  3.35529581e-02 ... -1.09215844e-01
   3.76697922e-03 -2.40387672e-03]
 [-9.94056468e-02 -5.12285216e-02  4.38534444e-02 ...  3.98844070e-02
  -5.97275219e-02  6.02529979e-02]]
```

各个类别的示例个数统计：

```
各个类别的数目 26      59
369     34
494     33
516     33
16      33
        ..
3        1
341      1
328      1
79       1
585      1
```

聚类结果分成了 600 个类可视化聚类结果

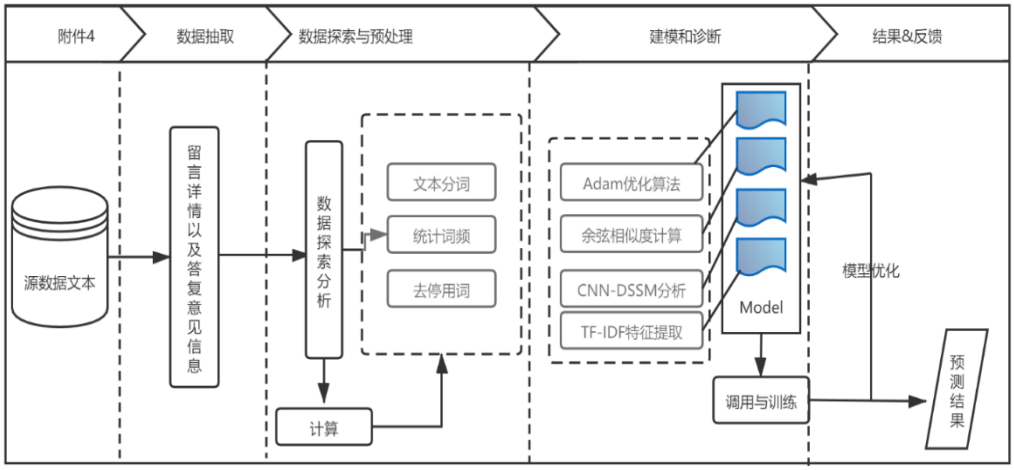


对 k-means 算法进行评价，通过 CH 指标计算类中各点与类中心的距离平方和来度量类内的紧密度，通过计算各类中心点与数据集中心点距离平方和来度量数据集的分离度，CH 指标由分离度与紧密度的比值得到。得到的 CH 值为 88，数据集中心点与数据集的分离度紧密程度合理，将得到的标签整理到数据，命名列名为问题 ID，定义热点问题用所分类和所有数据进行除权，可以大概评估热点数据需要 13 条，在相似数据中过滤出不满足 13 条数据的标签定义为离群点数据，整理出热点问题留言明细表。根据社会舆论热度评价指标定义相似数据的条数*0.9+点赞数和反对数的差*0.1 来定义每条类的热度保存为热度表，观察热度表和热点问题命题留言明细表筛选出排名前 5 的热点问题，保存为热点问题表.xls。

4.3 问题三

4.3.1 分析流程与方法

（1）流程分析图



（2）数据预处理

对附件 4 进行内容筛选

通过数据信息给出的问题，这里我们为考虑到后续实验的进行以及容错率的提升，这里我们采用了人工处理手段，直接将留言详情和答复意见保留，其他信息进行删除；同时为了便于统计该数据表的问题数，添加了序号列。删改后的文件数据保存为 data_3.xlsx 中。

（3）对附件 4 进行中文分词

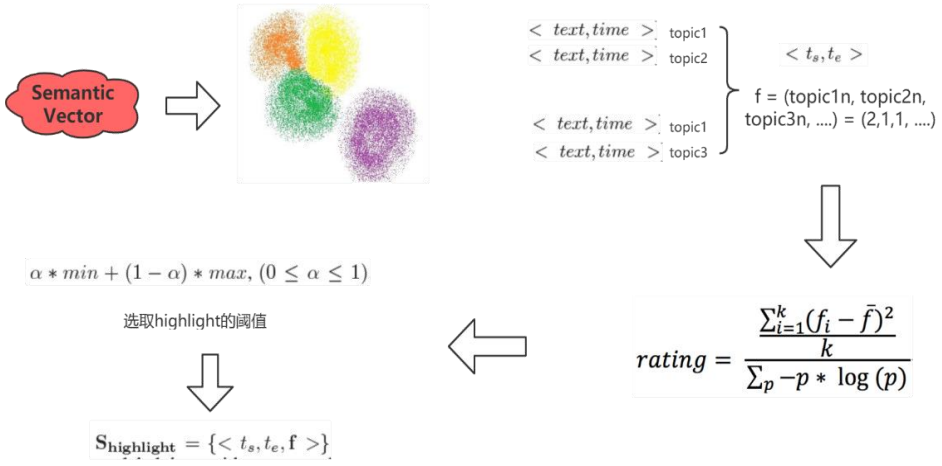
在进行针对附件 4 的数据分析和内容筛选之前，我们先要统计出该文本数据的主要关键词和内容，以便后续操作的运行和实现。在附件 4 表中，主要以中文文本的形式给出了具体的数据内容，而我们所需要的是留言详情和答复意见的内容，所以我们主要对这两块信息进行中文分词和和词频统计。其分词工具主要由 python 给出的 jieba 分词来进行操作。Jieba 分词以前缀词典实现高效的词图扫描为基础，与此同时在生成句子中汉字所有可能成词情况所构成的有向无环图 (DAG)采用了动态规划查找最大概率路径， 随后找出以词频的最大切分组合为基础；同时对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法来进行操作，这让中文分词的效果更加明显。分词后的文件分别保存为答复意见词频.csv 和留言详情词频.csv 文件中，主要用到停用词表为 stword.txt 文件。

（4）对 data_3 表进行数据转化

在对数据分析之前，考虑到如何让数据更加高效的被数据算法所调用处理，也在某种程度上是为了提高数据信息的关联性。首先，这里主要采用的是将每一行数据形成一个列表，然后将列表形成字典集。通过转化，得到的主要数据元素包括：留言集，答复集以及字符集[index]。然后，得到的数据将有效的转化为 QA 对形式，这样使得数据更加有效的被利用和分析。最后，得到的数据文件为 practice.json 数据文件。

4.3.2 CNN-DSSM 算法

CNN-DSSM 神经网络图：



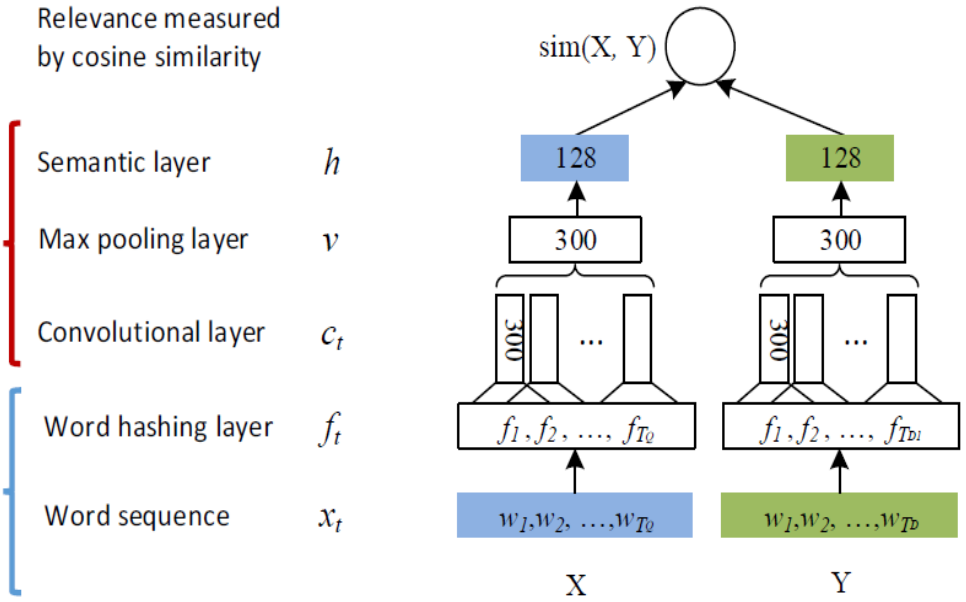
CNN-DSSM，通过搜索引擎里 Query 和 Doc 的海量的点击曝光日志，用 DNN 把 Query 和 Doc 转化为低纬语义向量，并通过 cosine 距离来计算两个语义向量的距离，并且根据用户的点击选择 Doc 当做标签值进行有监督学习，最终训练出语义相似度模型。该模型既可以用来预测两个句子的语义相似度，又可以获得某句子的低纬语义向量表。

CNN-DSSM 从下往上可以分为三层结构：输入层、表示层、匹配层

（1）输入层：为保证可控性，这里的输入层主要采用的是 DSSM 算法的输入模式

输入层做的事情是把句子映射到一个向量空间里并输入到 DNN 中,本算法主要运用于中文文本。由于中文分词的不可控性，所以这里出于向量空间的考虑，采用字向量（one-hot）作为输入。

（2）表示层： CNN-DSSM 的表示层由一个卷积神经网络组成，如下图所示：



卷积层——Convolutional layer

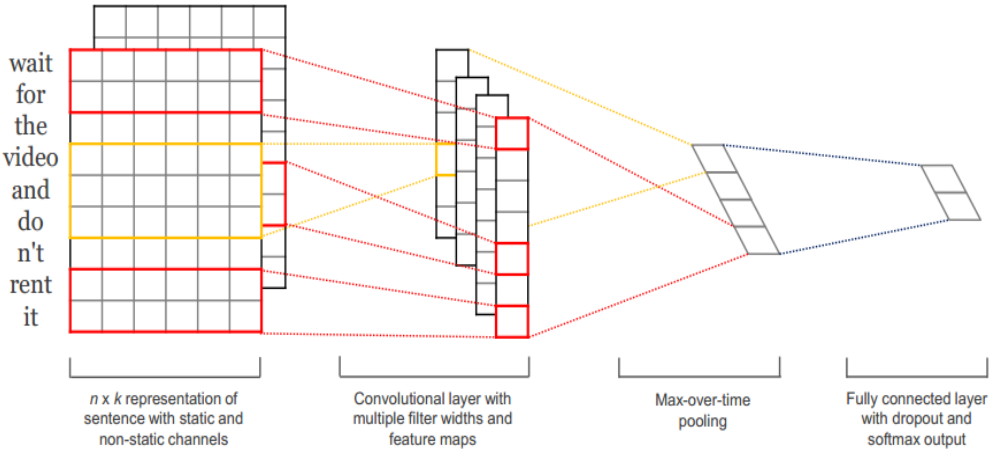


Figure 1: Model architecture with two channels for an example sentence.

池化层——Max pooling layer

全连接层——Semantic layer

最后通过全连接层把一个 300 维的向量转化为一个 128 维的低维语义向量。全连接层采用 tanh 函数：

$$tanh(x) = \frac{1 - e^{-2x}}{1 + e^{-2x}}$$

（3）匹配层

Query 和 Doc 的语义相似性可以用这两个语义向量(128 维) 的 cosine 距离来表示：

$$R(Q,D) = \text{cosine}(y_Q, y_D) = \frac{y_Q{}^T y_D}{\|y_Q\|\|y_D\|}$$

通过 softmax 函数可以把 Query 与正样本 Doc 的语义相似性转化为一个后验概率：

$$P(D^+|Q) = \frac{exp(\gamma R(Q,D^+))}{\sum_{D' \in D} exp(\gamma R(Q,D'))}$$

其中 r 为 softmax 的平滑因子，D 为 Query 下的正样本，D-为 Query 下的负样本（采取随机负采样），D 为 Query 下的整个样本空间。

在训练阶段，通过极大似然估计，我们最小化损失函数：

$$L(W,b) = -log \prod_{(UV,IV^+)} P(IV_i^+|UV)$$

残差会在表示层的 DNN 中反向传播，最终通过随机梯度下降（SGD）使模型收敛，得到各网络层的参数$\{W_i, b_i\}$。

4.3.3 TF-IDF 算法

为有效降低 CNN-DSSM 匹配层维度问题，过 TF-IDF 过滤掉最常用的特征。

原理：TF 是指归一化后的词频，IDF 是指逆文档频率。给定一个文档集合 D，有 d1,d2,d3,.....,dn∈D。文档集合总共包含 m 个词 w1,w2,w3,.....,wm∈W。我们现在以计算词 wi 在文档 dj 中的 TF-IDF 指为例。TF 的计算公式为：

TF=freq(i,j)maxlen(j)

在这里 freq(i,j) 为 wi 在 dj 中出现的频率，maxlen(j)为 dj 长度。

TF 只能时描述词在文档中的频率，但假设现在有个词为”我们“，这个词可能在文档集 D 中每篇文档中都会出现，并且有较高的频率。那么这一类词就不具有很好的区分文档的能力，为了降低这种通用词的作用，引入了 IDF。

IDF 的表达式如下：

IDF=log(len(D)n(i))

在这里 len(D)表示文档集合 D 中文档的总数，n(i)表示含有 wi 这个词的文档的数量。

得到 TF 和 IDF 之后，我们将这两个值相乘得到 TF-IDF 的值：

TF-IDF=TF*IDF

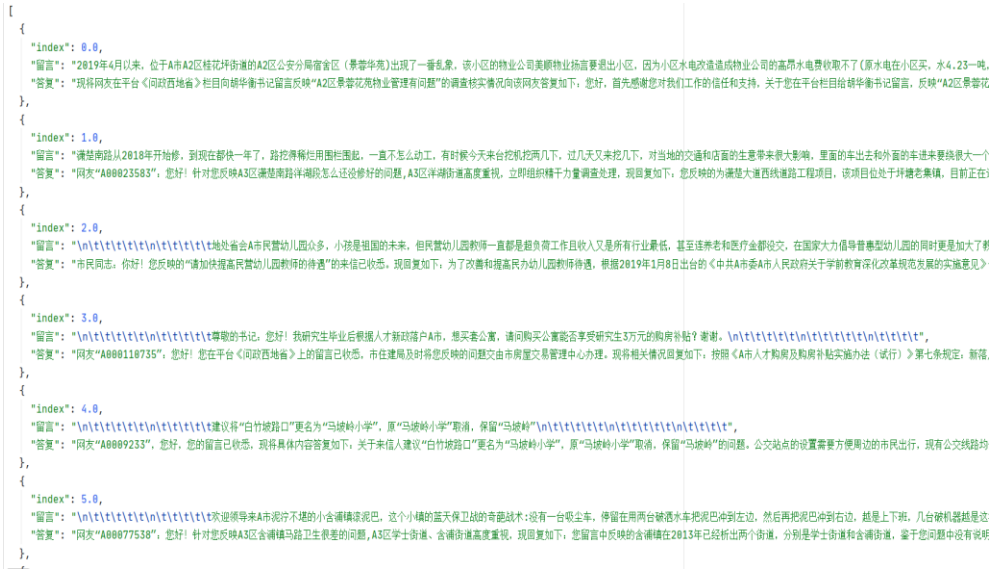
TF 可以计算在一篇文档中词出现的频率，而 IDF 可以降低一些通用词的作用。因此对于一篇文档我们可以用文档中每个词的 TF-IDF 组成的向量来表示该文档，再根据余弦相似度这类的方法来计算文档之间的相关性。

4.3.4 结果分析

（1）数据预处理结果

首先需要解决的问题就是如何对一个句子或一篇文章进行表示，将其转换为对应的特征向量，常见的表示方法可通过 Bi-RNN，Attention 方式实现。DSSM 与 Q-Q match 的主要区别就在于，DSSM 得到文章和问题的表示以后，通过计算两者之间的余弦相似度来计算相关性，而 Q-Q match 则是将两个表示拼接起来，带入到一个 MLP 中，进行分类，1 则为相关，0 则为不相关，或者可以进行更细粒度的分类。

而基于关键词的信息检索，即通过深度学习的方法，以一种端到端的方式，来计算问题和答案、问题和文章、甚至是词语与词语的相关联系，一个完整的问答系统处理流程主要分为三个部分：1.解析；2.匹配；3.生成。所以我们为了能够达到这样的机制，使得信息检索的过程更将便利。所以，通过对文本进行删改处理以及分词转化操作最后得到的结果如下图所示。



```
[
  {
    "index": 0.0,
    "留言": "2019年4月以来，位于A市A2区桂花坪街道的A2区公安分局宿舍区（景蓉华苑）出现了一番乱象，该小区的物业公司美顺物业扬言要退出小区，因为小区水电改造造成物业公司的高昂水电费收取不了(原水电在小区买，水4.23一吨，
    "答复": "现将网友在平台《问政西地省》栏目向胡华衡书记留言反映“A2区景蓉花苑物业管理有问题”的调查核实情况向该网友答复如下：您好，首先感谢您对我们工作的信任和支持，关于您在平台栏目给胡华衡书记留言，反映“A2区景蓉花
  },
  {
    "index": 1.0,
    "留言": "满楚南路从2018年开始修，到现在都快一年了，路挖得稀烂用围栏围起，一直不怎么动工，有时候今天来台挖机挖两几下，过几天又来挖几下，对当地的交通和店面的生意带来很大影响，里面的车出去和外面的车进来要绕很大一个
    "答复": "网友“A0023583”，您好！针对您反映A3区满楚南路洋湖段怎么还迟修好的问题,A3区洋湖街道高度重视，立即组织精干力量调查处理，现回复如下：您反映的为满楚大道西线道路工程项目，该项目位处于坪塘老集镇，目前正在
  },
  {
    "index": 2.0,
    "留言": "\n\t\t\t\t\t\t\n\t\t\t\t\t\t地处省会A市民营幼儿园众多，小孩是祖国的未来，但民营幼儿园教师一直都是超负荷工作且收入又是所有行业最低，甚至连养老和医疗金都没交，在国家大力倡导普惠型幼儿园的同时更是加大了教
    "答复": "市民同志：你好！您反映的“请加快提高民营幼儿园教师的待遇”的来信已收悉。现回复如下：为了改善和提高民办幼儿园教师待遇，根据2019年1月8日出台的《中共A市委A市人民政府关于学前教育深化改革规范发展的实施意见》
  },
  {
    "index": 3.0,
    "留言": "\n\t\t\t\t\t\n\t\t\t\t\t\t尊敬的书记，您好！我研究生毕业后根据人才新政落户A市，想买套公寓，请问购买公寓能否享受研究生3万元的购房补贴？谢谢。\n\t\t\t\t\t\n\t\t\t\t\t\t\n\t\t\t\t\t\t",
    "答复": "网友“A0001110735”，您好！您在平台《问政西地省》上的留言已收悉，市住建局及时将您反映的问题交由市房屋交易管理中心办理。现将相关情况回复如下：按照《A市人才购房及购房补贴实施办法（试行）》第七条规定，新落
  },
  {
    "index": 4.0,
    "留言": "\n\t\t\t\t\t\n\t\t\t\t\t\t建议将“白竹坡路口”更名为“马坡岭小学”，原“马坡岭小学”取消，保留“马坡岭”\n\t\t\t\t\t\n\t\t\t\t\t\n\t\t\t\t\t\t",
    "答复": "网友“A0009233”，您好，您的留言已收悉，现将具体内容答复如下：关于来信人建议“白竹坡路口”更名为“马坡岭小学”，原“马坡岭小学”取消，保留“马坡岭”的问题。公交站点的设置需要方便周边的市民出行，现有公交线路均
  },
  {
    "index": 5.0,
    "留言": "\n\t\t\t\t\t\n\t\t\t\t\t\t欢迎领导来A市泥泞不堪的小含浦镇视泥巴，这个小镇的蓝天保卫战的奇葩战术:没有一台吸尘车，停留在用两台破洒水车把泥巴冲到左边，然后再把泥巴冲到右边，越是上下班，几台破机器越是这
    "答复": "网友“A0007538”，您好！针对您反映A3区含浦镇马路卫生很差的问题,A3区学士街道、含浦街道高度重视，现回复如下：您留言中反映的含浦镇在2013年已经析出两个街道，分别是学士街道和含浦街道，鉴于您问题中没有说明
  },
```


（2）模型分析

将输入的留言集(q)和答复集(t)转转化为每个字符对应的字向量，形成一个三维矩阵 q-t,从而获取留言集和答复集的实际长度及最大长度。

```
Initializing------
留言集的最大长度为 3084
答复集的最大长度为 7883
```

在本模型中，输入数据通过引入一个双向 GRU 来对答复集每句话的上下文信息来进行实施编码操作，在原先存在的语义向量基础上规则性的加上每个词语在句子的位置编码向量。

数据编码操作后，进行 transformer 操作，通过实现类的计算，最后得到的结果进行封装。

通过负采样以及余弦近似度计算，最终的输出是一个二维矩阵数列，矩阵中的每一行代表一个留言与其所对应意见。

（3）对留言以及答复意见，提出建议

从数据分析中不难看出主要的民生问题，第一个主要在于住所、社会保障问题和劳动者工作环境的问题；第二个主要在于就业指标、收入分配以及教育问题；第三个则主要在于能源、资源问题。

首先，针对住所等问题。要增加投入，转换机制，及时、有效地做好前期介入工作。在面对与业主的问题时，要努力为业主入伙提供一条龙服务，尤其是要求开发商必须在入伙现场设立答疑组和维修小组，随时为业主答疑，并对联合验房时发现的问题及时进行修复或对业主做出明确的维修承诺等；工作人员应经常性开展业主意见调查，主动及时了解业主的需求变化，找出近期服务的弱项或盲点，制定整改措施并迅速纠正，同时向业主公示，请求业主监督实施。主动发现和解决问题与被业主投诉后才去处理相比，那怕后者处理得再好，两者的服务效果是大不相同的，作为物业管理的项目经理应时常提醒自己这一点。

其次，在针对就业等社会现象时，国家应做到加强宏观调控，扩大就业门路，发展第三产业，完善社会就业服务体系。企业应做到调整国有企业布局，发展劳动密集型企业。实时关注学生教育以及就业现象问题，抓紧中小幼的教育问题，让学生成长，让家长放心；加大对大学生就业市场的调控，使大学生就业问题的压力不在过于沉重。

最后，就目前我们国家而言能源资源存在的问题是使用效率低，节约意识容不强。从这几年的数据分析中我们可以看出，我们国家的资源能源的发展出现了停滞不前的现实问题，所以我们要从推动我国经济社会持续发展和人民生活水平不断提高的全局出发,全面分析能源资源形势,深入研究能源资源问题,全面做好能源资源工作。

当然，除了以上的民生问题，我国还存在着各式各样的社会问题，作为政府应该时时刻刻关注社会问题，政府要措施是增加投入，政府还要积极转变政府职能，使之尽快履行公共服务者的职能。民生问题是政府公共职能的重要体现，应该积极建立政府的公共服务型政府职能模式，政府不仅要加大对公共领域的财政投入，还要从政策、制度、公务员队伍建设等方面人手。此外，还应严格规范政府行为，防止政府一些部门与民争利，防止利于民生的政策走样变形。

政府要明白解决民生问题，法治是必经之路。总而言之，政府要努力做好人民的好公仆。

五、参考文献

[1]. Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, Larry Heck Learning deep structured semantic models for web search using clickthrough data.2013

[2].施聪莺，徐朝军，杨晓江 TFIDF 算法研究综述 2009

[3]. Shen, Yelong, et al. "A latent semantic model with convolutional-pooling structure for information retrieval." Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. ACM, 2014.

[4]. Palangi, Hamid, et al. "Semantic modelling with long-short-term memory for information retrieval." arXiv preprint arXiv:1412.6629 2014.

[5]. Gers, Felix A., Schraudolph, Nicol N., and Schmidhuber, J¨urgen. Learning precise timing with lstm recurrent networks. J. Mach. Learn. Res., 3:115–143, March 2003.

[6]. 程军年 五位一体助推民生国家体系建设 2020

[7] https://github.com/makeplanetoheaven

[8] https://scikit-learn.org/stable/

[9] https://github.com/fxsjy/jieba