“智慧政务”中的文本挖掘应用

摘要

近年来，随着互联网和移动互联网的飞速发展，大数据时代的标签早已悄无声息地融入到人类的生活中，人们通过互联网产生的数据急剧性增长，数据开始成为人们交流沟通的重要载体，互联网平台也成为了人们传输信息的重要平台。人们在互联网上表达自己意见的倾向日渐增长，网络问政平台上的各类民意信息成为了政府获取数据的重要来源，对政府提升监管、服务和决策的智能水平，形成高效、公开、便民的“智慧型政府”具有重要意义。

对于问题 1：通过 CNN 算法，建立神经网络，实现自动对数据进行整理和预测。首先将留言主题和对应的一级标题提取成结构化数据。利用 jieba 中文分词工具对信息进行分词，并进行停用词清洗，对文本进行去重。随后进行 word2vec 模型训练，得出每个词的词向量，再将每个词向量作为输入数据打进卷积神经网络进行训练测试，得出预测的以及标题。

对于问题 2：同样先将留言主题提取成结构化数据，再用 jieba 中文分词工具对信息进行分词，并进行数据清洗去重。随后进行 word2vec 模型训练，得出每个词的词向量。再对候选关键词进行 K-Means 聚类，得到三十个聚类中心后，计算各类别下，组内词语与聚类中心的距离（欧几里得距离），按聚类大小进行升序排序。最后对候选关键词计算结果得到排名前 Top5 个词汇作为文本关键词。

对于问题 3：通过对留言详情和答复意见的分词处理，进行关键词匹配度的观察和答复涉及留言关键词百分比的计算，实现文本相关性和完整性的预测。再结合观测答复时间差，制定答复评价模型。

关键词：CNN 算法（卷积神经网络）、word2vec、中文分词、K-means 聚类分析

Text mining application in "smart government"

abstract

In recent years, with the rapid development of the Internet and mobile Internet, the tags in the era of big data have been quietly integrated into human life. The rapid growth of data generated by people through the Internet has made data become an important carrier for people to communicate, and the Internet platform has become an important platform for people to transmit information. All kinds of public opinion information on the Internet platform has become an important source for the government to obtain data. It is of great significance for the government to improve the intelligent level of supervision, service and decision-making, and to form an efficient, open and convenient "intelligent government".

For problem 1: through CNN algorithm, establish neural network to realize automatic data sorting and prediction. Firstly, the subject of the message and the corresponding first level title are extracted into structured data. Using the Chinese word segmentation tool of Jieba to segment the information, clean the stop words, and de duplicate the text. Then we train the word 2 VEC model and get the word vector of each word. Then we use each word vector as the input data into the convolution neural network for training and testing, and get the predicted and the title.

For question 2: we also extract the subject of the message into structured data, and then use the Chinese word segmentation tool of Jieba to segment the information, and clean the data. Then we train the word 2 VEC model to get the word vector of each word. Then, K-means clustering is applied to the candidate keywords, and clustering centers are obtained. Then, the distance between words and clustering centers (Euclidean distance) in each category is calculated, and the clustering is sorted in ascending order according to the clustering size. Finally, the top 5 words are selected as text keywords.

For question 3: Through the word segmentation processing of message details and reply comments, the observation of keyword matching degree and the calculation of the percentage of key words involved in the reply are carried out to achieve the prediction of text relevance and integrity. Combined with the difference of observation response time, the response evaluation model is established.

Keywords: CNN algorithm (convolutional neural network), word2vec, Chinese word segmentation, K–means clustering analysis

目　录

1.挖掘目标

近年来，随着互联网和移动互联网的飞速发展，大数据时代的标签早已悄无声息地融入到人类的生活中，人们通过互联网产生的数据急剧性增长，数据开始成为人们交流沟通的重要载体，互联网平台也成为了人们传输信息的重要平台。在网络问政平台上的各类民意信息成为了政府获取数据的重要来源，对政府提升监管、服务和决策的智能水平，形成高效、公开、便民的“智慧型政府”具有重要意义。

本次数据挖掘的建模目标是利用互联网公开来源的群众问政留言记录以及相关部门对群众留言的答复意见的信息数据，运用中文分词工具、CNN 神经网络技术、K-Means 聚类分析，达到以下目标：

1) 利用中文分词工具对群众留言信息数据进行文本挖掘，提取留言信息中的结构化数据，构建神经网络模型，结合一级标题中的分类标准对留言信息进行分类，将人工处理智能化，提高政府办公效率，降低回复差错率。
2) 利用中文分词工具和文本聚类分析的方法对群众反映的问题信息数据进行文本挖掘，提取结构化数据，结合留言区间和点赞数量构建热度评价指标，进行热点问题挖掘，形成热点问题表。
3) 根据相关部门对群众留言信息的答复信息数据，结合答复时间、答复相关程度、答复完整性的指标构建合理的答复评价模型，对工作人员答复进行分析做出评价，了解部门答复意见的质量，明确后续政务工作改进方向。

2.总体流程图

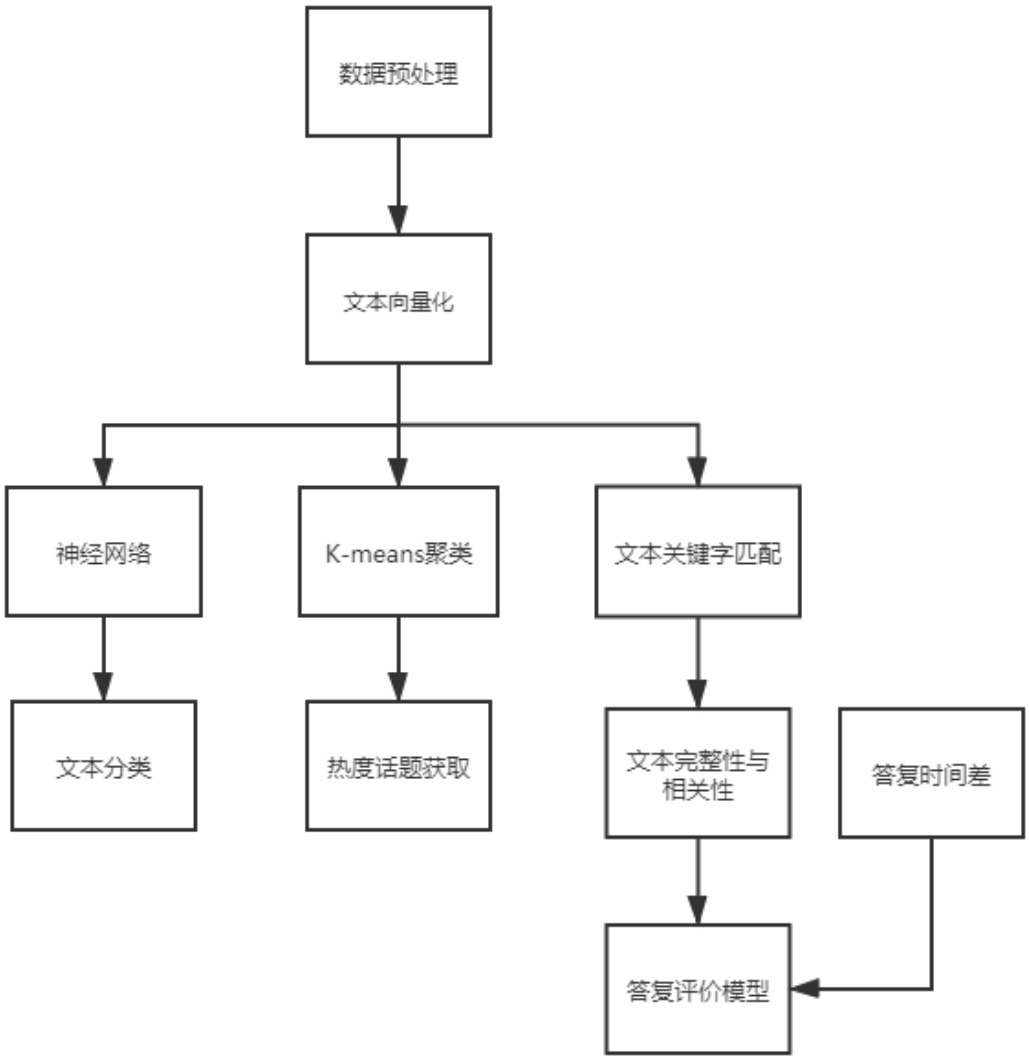

图 1

3.数据预处理

3.1 文本预处理

3.1.1 结构化文本数据

由于已知数据集中都是非结构化的文本，以及对应的相关编号，标签和时间，

为方便接下来的文本分词，词向量构建的步骤，这里提取“附件一”，“附件二”，“附件三”的部分数据构成简单的结构化数据以便后续使用 python 进行进一步分析。例如，将“附件二”中的编号一列及对应的留言主题保存“discussdata.txt”中，如图 2。

188006 A3区一米阳光婚纱艺术摄影是否合法纳税了？
188007 咨询A6区道路命名规划初步成果公示和城乡门牌问题
188031 反映A7县春华镇金鼎村水泥路、自来水到户的问题
188039 A2区黄兴路步行街大古道巷住户卫生间粪便外排
188059 A市A3区中海国际社区三期与四期中间空地夜间施工噪音扰民
188073 A3区麓泉社区单方面改变麓谷明珠小区6栋架空层使用性质
188074 A2区富绿新村房产的性质是什么？
188119 对A市地铁违规用工问题的质疑
188170 A市6路公交车随意变道通行
188249 A3区保利麓谷林语桐梓坡路与麓松路交汇处地铁凌晨2点施工扰民
188251 A7县特立路与东四路口晚高峰太堵，建议调整信号灯配时
188260 A3区青青家园小区乐果果零食炒货公共通道摆放空调扰民
188396 关于拆除聚美龙楚在西地省商学院宿舍旁安装变压器的请求

图 2

3.1.2 基于 jieba 的文本分词

中文分词是中文文本处理的一个基础步骤，也是中文人机自然语言交互的基础模块。本文使用现在非常流行的且开源的分词器 jieba 分词器。jieba 分词算法使用了基于前缀词典实现高效的词图扫描，生成句子中汉字所有可能生成词情况所构成的有向无环图(DAG)，再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，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。结果如图 3

```
地铁5号线施工导致A市锦楚国际星城小区三期一个月停电10来次
Loading model cost 0.638 seconds.
Prefix dict has been built succesfully.
['地铁', '5', '号线', '施工', '导致', 'A', '市锦楚', '国际', '星城', '小区', '三期', '一个月', '停电', '10', '来次']
```

图 3

3.1.3 词语筛选

将文本分词之后，会出现许多停用词和符号的情况，如“的”，“之”，“尽管”等。将它们一并输入到后续的神经网络或者其他算法模型时会影响最后输出结果

的效果。本文使用停用词表的方法去除文本分词后的无意义词语，在 github 上下载通用停用词表 stopword.txt 并保存在 data 文件夹中。停用词筛选的效果如图 4 所示。

```
A市何时能实现冬季集中供暖？
未筛选前： ['A', '市', '何时能', '实现', '冬季', '集中', '供暖', '？']
筛选后： ['何时能', '实现', '冬季', '集中', '供暖']
```

图 4

3.2 文本向量化

3.2.1 word2vec 模型

Harris 在 1954 年提出的分布假说（distributional hypothesis）为这一设想提供了理论基础：上下文相似的词，其语义也相似。而基于分布假说的词表示方法，根据建模的不同，主要可以分为三类：基于矩阵的分布表示、基于聚类的分布表示和基于神经网络的分布表示。而 word embedding 一般来说就是一种基于神经网络的分布表示。

word2vec 是只有一个隐层的全连接神经网络，用来预测给定单词的关联度大的单词。或者说是一个语言模型。模型如图 5 所示

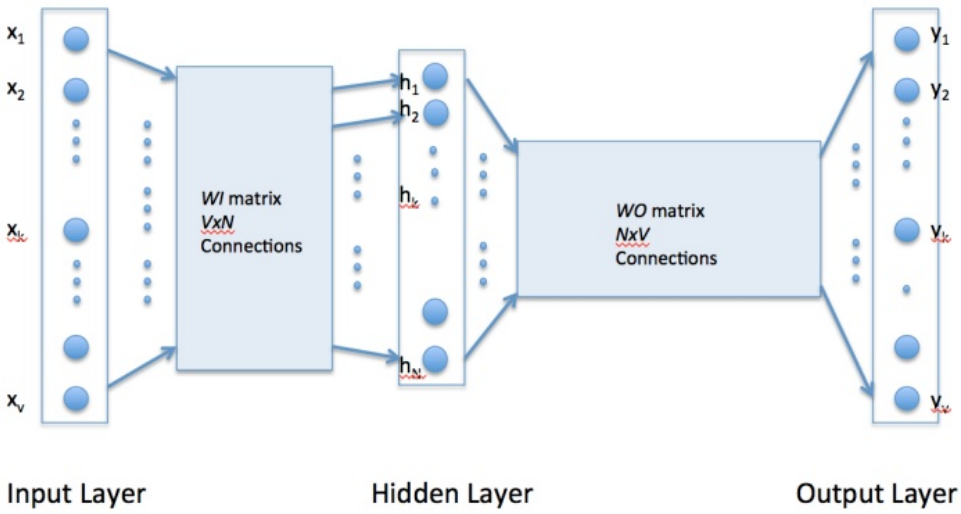

图 5

- 在输入层，一个词被转化为 One-Hot 向量。
- 然后在第一个隐层，输入的是一个（就是输入的词向量，和是参数），做一个线性模型。注意这里只是简单的映射,并没有非线性激活函数,当然一个神经元可以是线性的，这时就相当于一个线性回归函数。
- 第三层可以简单看成一个分类器，用 Softmax 回归，最后输出每个词对应的概率

3.2.2 word2vec 模型原理

Word2Vec 有两种训练方法，一种叫 CBOW，核心思想是从一个句子里将一个词抠掉，用该词的上文和下文去预测被抠掉的词；第二种叫做 Skip-gram，和 CBOW 正好相反，输入某个单词，要求网络预测它的上下文单词。

1) CBOW 模型

CBOW 模型的训练输入是某一个特征词的上下文相关的词对应的词向量，而输出就是这特定的一个词的词向量。模型如图 6 所示。

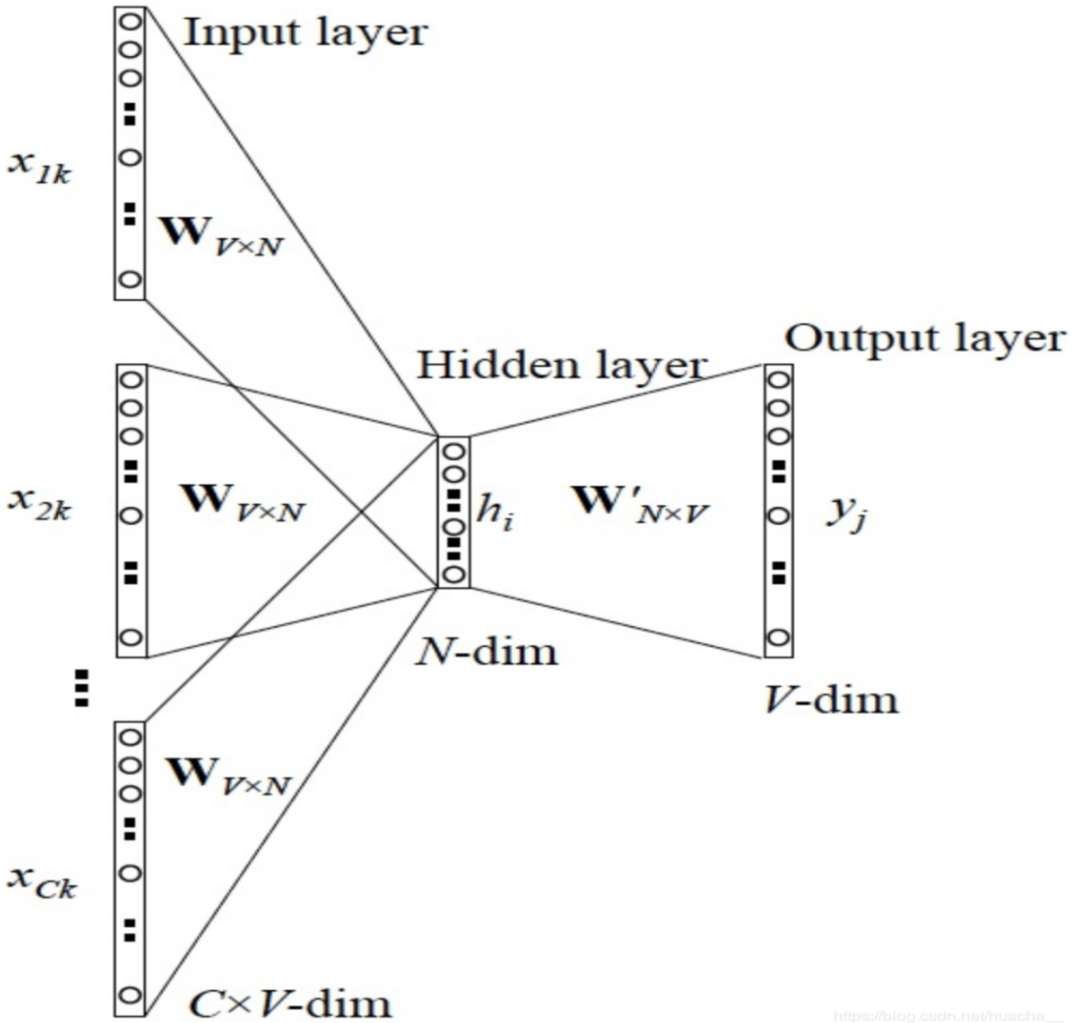

图 6

2) Skip–gram 模型

Skip–Gram 模型和 CBOW 的思路是相反的，即输入是特定的一个词的词向量，而输出是特定词对应的上下文词向量。模型如图 7 所示。

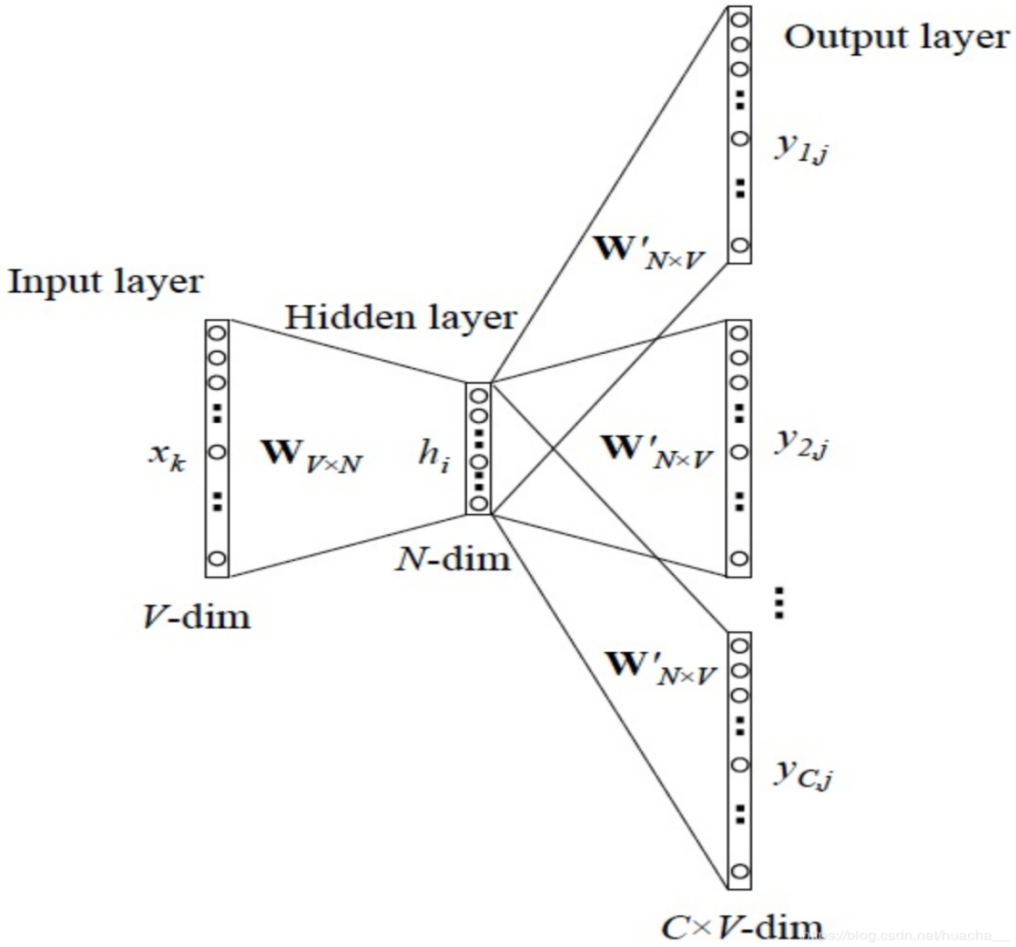

图 7

3.2.3 基于 gensim 的 word2vec 应用

在 python 的第三方库 gensim 中有自带的 Word2Vec 函数来训练自己语料库的词向量,本文的语料库数据是基于每个题目相关的文本数据集中，每行 excel 中的一条文本数据，是经过分词和去停用词之后的数据，本文默认采 skip-gram 来训练模型。最后输出结果是每行为“词语 + 100 维的词向量”的 txt 文本,部分结

果如图 8 所示。

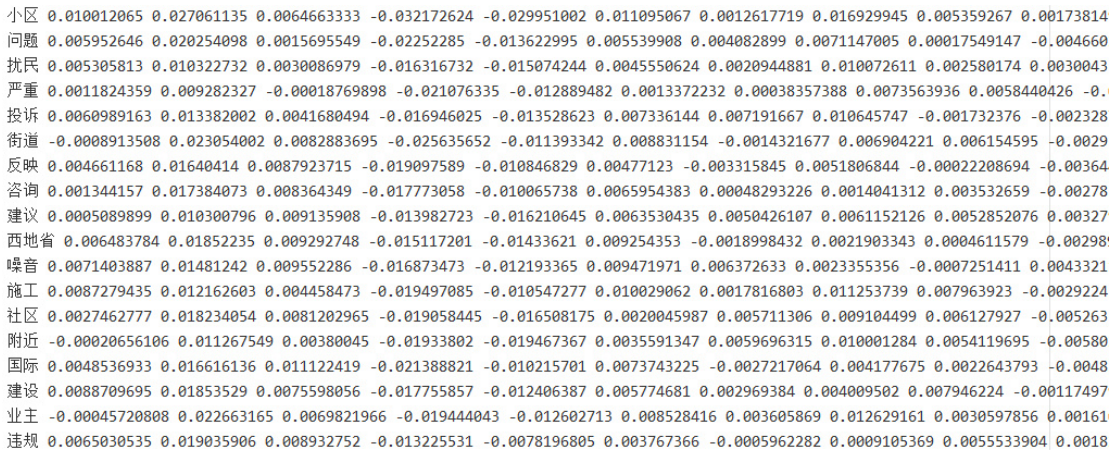
小区 0.010012065 0.027061135 0.0064663333 -0.032172624 -0.029951002 0.011095067 0.0012617719 0.016929945 0.005359267 0.00173814
问题 0.005952646 0.020254098 0.0015695549 -0.02252285 -0.013622995 0.005539908 0.004082899 0.0071147005 0.00017549147 -0.004660
扰民 0.005305813 0.010322732 0.0030086979 -0.016316732 -0.015074244 0.0045550624 0.0020944881 0.010072611 0.002580174 0.0030043
严重 0.0011824359 0.009282327 -0.00018769898 -0.021076335 -0.012889482 0.0013372232 0.00038357388 0.0073563936 0.0058440426 -0.
投诉 0.0060989163 0.013382002 0.0041680494 -0.016946025 -0.013528623 0.007336144 0.007191667 0.010645747 -0.001732376 -0.002328
街道 -0.0008913508 0.023054002 0.0082883695 -0.025635652 -0.011393342 0.008831154 -0.0014321677 0.006904221 0.006154595 -0.0029
反映 0.004661168 0.01640414 0.0087923715 -0.019097589 -0.010846829 0.00477123 -0.003315845 0.0051806844 -0.00022208694 -0.00364
咨询 0.001344157 0.017384073 0.008364349 -0.017773058 -0.010065738 0.0065954383 0.00048293226 0.0014041312 0.003532659 -0.00278
建议 0.0005089899 0.010300796 0.009135908 -0.013982723 -0.016210645 0.0063530435 0.0050426107 0.0061152126 0.0052852076 0.00327
西地省 0.006483784 0.01852235 0.009292748 -0.015117201 -0.01433621 0.009254353 -0.0018998432 0.0021903343 0.0004611579 -0.00298
噪音 0.0071403887 0.01481242 0.009552286 -0.016873473 -0.012193365 0.009471971 0.006372633 0.0023355356 -0.0007251411 0.0043321
施工 0.0087279435 0.012162603 0.004458473 -0.019497085 -0.010547277 0.010029062 0.0017816803 0.011253739 0.007963923 -0.0029224
社区 0.0027462777 0.018234054 0.0081202965 -0.019058445 -0.016508175 0.0020045987 0.005711306 0.009104499 0.006127927 -0.005263
附近 -0.00020656106 0.011267549 0.00380045 -0.01933802 -0.019467367 0.0035591347 0.0059696315 0.010001284 0.0054119695 -0.00580
国际 0.0048536933 0.016616136 0.011122419 -0.021388821 -0.010215701 0.0073743225 -0.0027217064 0.004177675 0.0022643793 -0.0048
建设 0.0088709695 0.01853529 0.0075598056 -0.017755857 -0.012406387 0.005774681 0.002969384 0.004009502 0.007946224 -0.00117497
业主 -0.00045720808 0.022663165 0.0069821966 -0.019444043 -0.012602713 0.008528416 0.003605869 0.012629161 0.0030597856 0.00161
违规 0.0065030535 0.019035906 0.008932752 -0.013225531 -0.0078196805 0.003767366 -0.0005962282 0.0009105369 0.0055533904 0.0018

图 8

4.具体分析方法与过程

4.1 问题 1 分析方法与过程

4.1.1 流程图

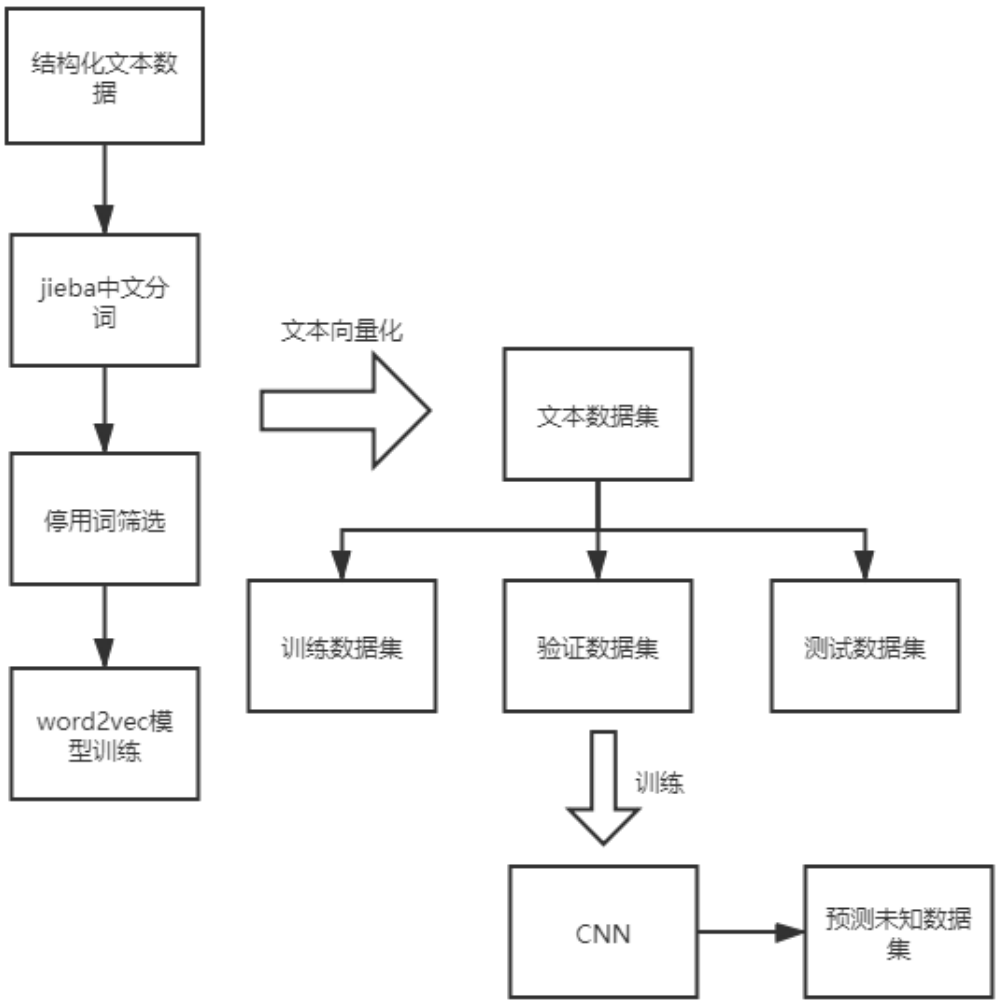

图 9

4.1.2 数据集处理

提取附件 2.xlsx 中的一级标签列与留言详情列，将留言详情列进行中文分词，停用词筛选与 word2vec 向量化，具体实现步骤见数据准备。然后将其首先

进行 8.5:1.5 比例的随机分配，将数据集分别保存为 train_data.txt 与 temp.txt。再将 temp.txt 以 1:1 比例随机分配为 val_data.txt 与 test_data.txt 数据集，作为神经网络的训练集，验证集与测试集，都保存在 data 文件夹下。

4.1.3 卷积神经网络

1) 假定输入 CNN 的数据是二维的，其中每一行表示一个样本（即一个字词），如图中“I”、“like”等。每一个样本（字词）有 d 个维度，可以看成是词向量长度，即每个字词的维度，程序中用 embedding_dim 表示。
2) 使用 CNN 的卷积对这个二维数据进行卷积：在图像的 CNN 卷积中，卷积核的大小一般是 3*3，5*5 等，但在 NLP 中不能这样进行运算，因为这里的输入数据每行是一个样本。假设卷积核的大小[filter_height,filter_width],那么卷积核的高度 filter_height 可以为 1，2，3 等任意值，而宽度 filter_width 只能是 embedding_dim 的大小，这样完整的样本框才能输入进去。模型如图 10 所示。

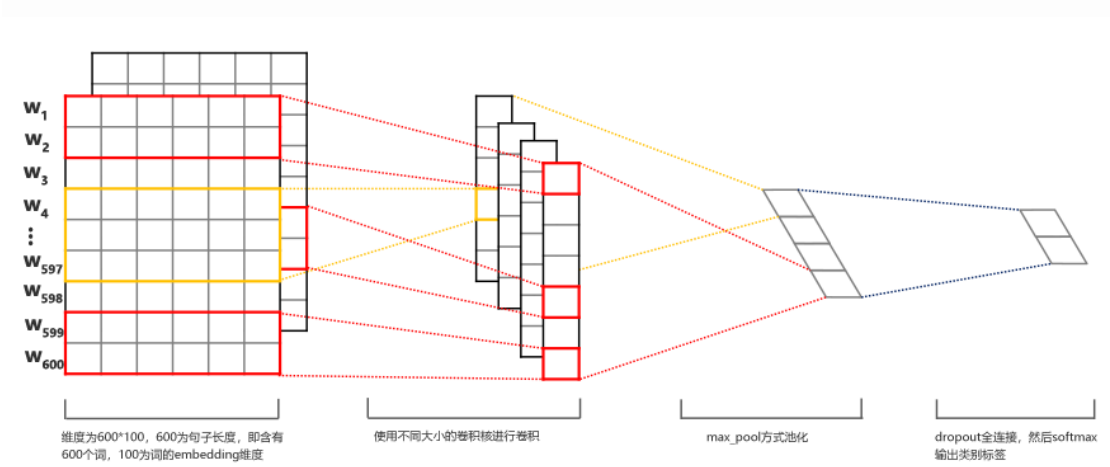

图 10

模型的 label 在本文中就是指一级标签的种类，输入数据就是文本分词后每个单词的词向量。

4.2 问题 2 分析方法与过程

4.2.1 流程图

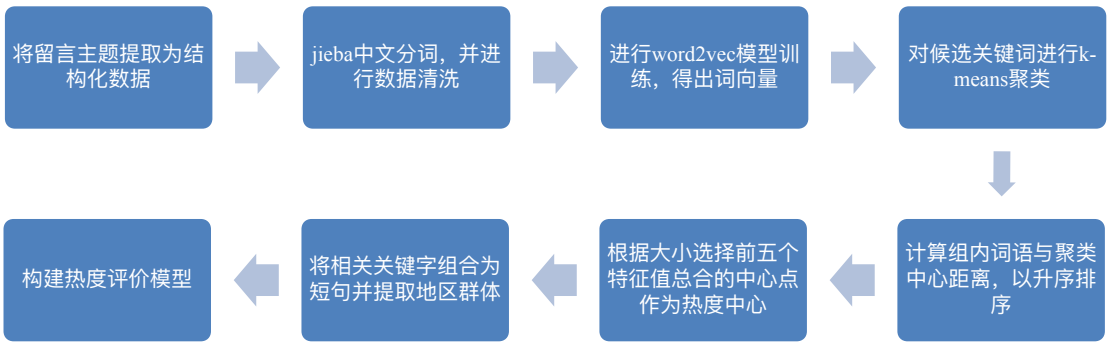

图 11

4.2.2k-means 算法

在生成留言信息的数据向量之后，根据向量权重，对留言进行分类。这里采用 k-means 聚类分析算法将留言信息分类。

4.2.2.1k–means 聚类分析的原理

k-means 算法的输入为一个样本集（或称点集），通过该算法将样本进行聚类，具有相似特征的样本聚为一类。针对每个点，计算这个点距离所有中心点最近的那个中心点，然后将这个点归为这个中心点代表的簇。一次迭代结束之后，针对每个簇类，重新计算中心点，然后针对每个点，重新寻找距离自己最近的中心点。如此循环，直到前后两次迭代的簇类没有变化。

每个点到中心点的距离公式 ： （欧式距离）

4.2.2.2k-means 聚类分析的流程图

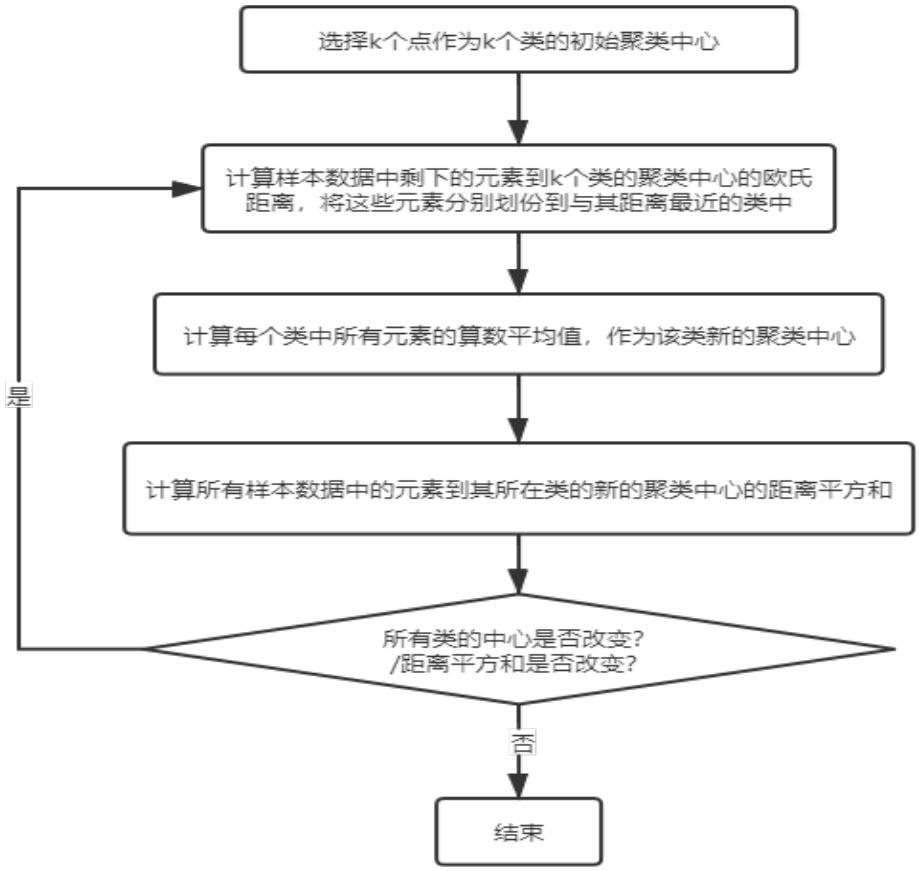

图 12

4.2.2.3k-means 聚类分析的算法步骤

1) 首先根据对数据的先验经验或交叉验证选择一个合适的 k 值。
2) 随机选择 k 个点作为 k 个类的初始聚类中心。
3) 分别计算样本数据中剩下的元素到 k 个聚类中心的欧氏距离，将这些元素分别划分到与其距离最近的类当中。
4) 根据聚类结果，计算各个类中所有元素的算术平均值，作为 k 个类各自新的聚类中心点。
5) 将 k 个类的新聚类中心点距离与原有中心点距离进行比较，如果中心点不再变化/聚类结果不再变化/迭代计算轮次达到最大值，则输出聚类结果。否则回到第 3 步继续进行迭代计算。

4.2.3 基于 sklearn 的 K-means 实现及热度关键词提取

本文使用 python 第三方库 sklearn 构建 K-means 模型。首先使用手肘法遍历 1 到 30 的聚类中心寻找最优 K 值。手肘法的核心思想是：随着聚类数 k 的增大，样本划分会更加精细，每个簇的聚合程度会逐渐提高，那么误差平方和 SSE 自然会逐渐变小。并且，当 k 小于真实聚类数时，由于 k 的增大会大幅增加每个簇的聚合程度，故 SSE 的下降幅度会很大，而当 k 到达真实聚类数时，再增加 k 所得到的聚合程度回报会迅速变小，所以 SSE 的下降幅度会骤减，然后随着 k 值的继续增大而趋于平缓，也就是说 SSE 和 k 的关系图是一个手肘的形状，而这个肘部对应的 k 值就是数据的真实聚类数。结果如图 13 所示。

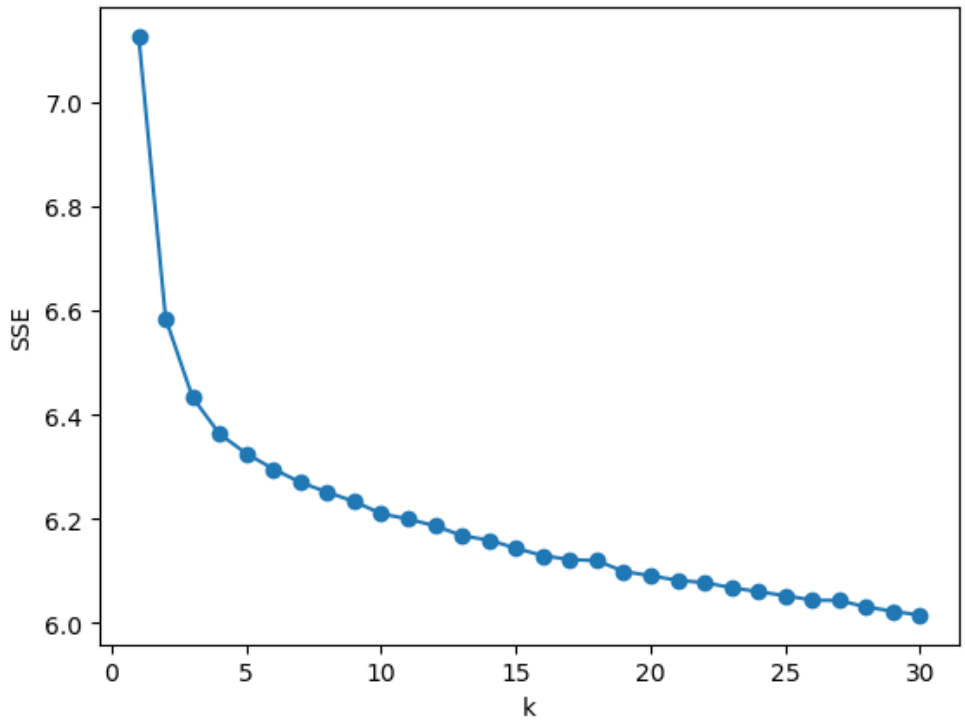

图 13

不难发现，当 K 值为三时，对于这个数据集的聚类而言，最佳聚类数应该选三。但题目中我们需要找出前五个热门话题，三个中心远远不够。这里选用 30 个聚类中心，我们需要从中筛选出文本特征值最高的前五个聚类中心。这里将所有文本特征值可视化，结果如图 14 所示。

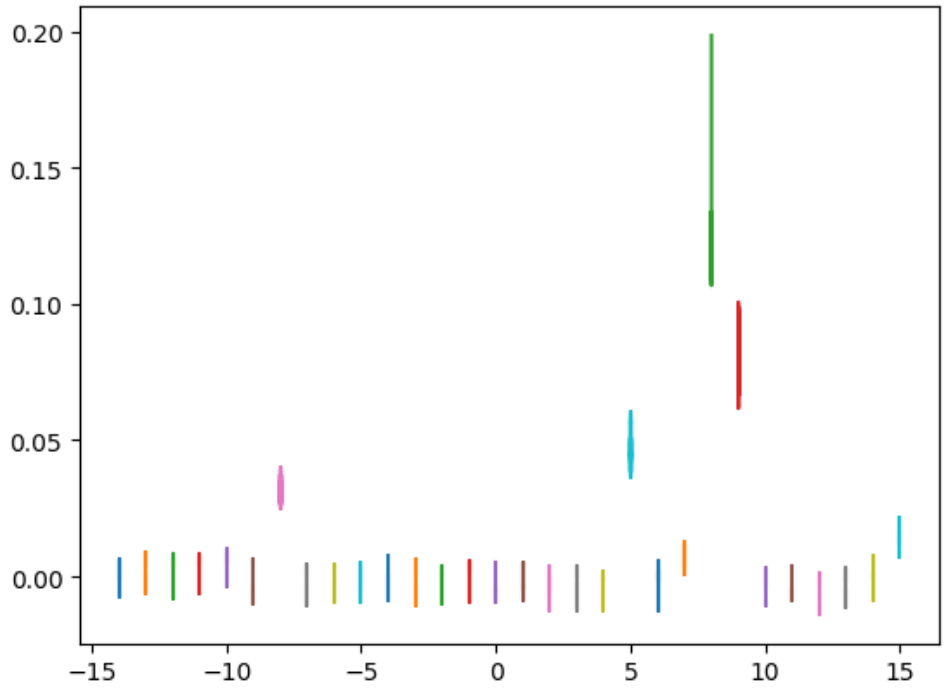

图 14

4.2.4 构建热度评价模型

为及时发现群众集中反映的热点问题，从而帮助政务部门更有效率地进行针对性处理，需要对留言信息进行热度分析和评价。在这里，将结合留言信息关键词出现频率、留言问题点赞和反对数量、留言问题时间区间这三个考察维度来构建群众留言中热点问题的热度评价模型。

4.2.4.1 留言信息关键词

首先将留言信息中的分词进行关键词提取，将其与聚类分析生成的热度关键词相比较，由于一个热度关键词可以对应多个主题，所以留言信息关键词出现频率作为热度评价模型中的一个指标。热度关键字如图 15 所示。

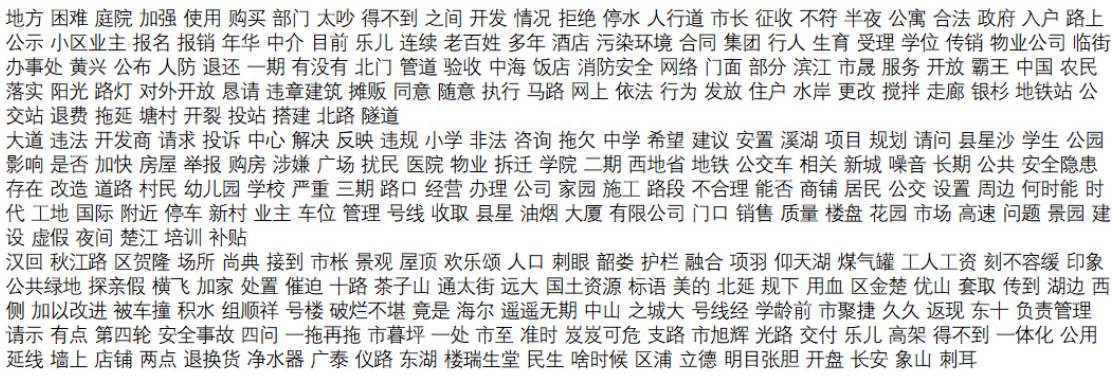

图 15

4.2.4.2 留言问题点赞和反对数量

留言问题的点赞和反对数量在一定程度上反映了其他群众对相关问题的关注程度，是热度分析的重要指标，因此将其纳入热度评价模型。这里将同类主题的点赞和反对数相加，其和作为一个评价指标。

4.2.4.3 留言问题时间区间

由于热点问题需要及时、集中地进行体现，因此留言时间区间也可作为热度评价模型的相关评价指标，部分留言时间过久的问题反馈可以分离纳入政务工作的后期单独处理，而留言时间较近的问题可以通过政务平台进行热度问题的集中答复和解决。

4.2.4.4 热度主题提取

步骤：

1) 将附件 3 中的反对数与点赞数相加组合成新的变量，根据新的变量进行初步评论热度排序，排序后的数据结果以 sort_data 文件保存。
2) 提取前五个评论热度最高的留言主题，结合 K-means 得出的热度关键词寻找与之类似的留言信息关键词，根据关键词相同的多少重新排序。
3) 提取同一类留言主题的最早发布时间与最晚发布时间。一般情况下，只有在第一步和第二步中有两个不同主题排序相同时，才会考虑发布时间的早晚。
4) 计算热度指数，这里就是，最后使用 python 的 pandas 第三方库根据热度指

数排序将数据保存在 excel 中。

4.2.4.5 地区人物归纳

由于地区人物本身具有不规则性，若要直接对其进行提取会十分困难。本文将问题简化，鉴于附件 3 中留言主题中的地区人物大多处于开头，这里使用 jieba 分词第三方库对每一行先分词再进行词性标注，然后选择前两个名词作为地区人物的概括。实现结果如图 16 所示。

```
标注文本： A7/eng,县/n,特立/b,路/n,与/p,东四/t,路口/s,晚/tg,高峰/nr,太/d,堵/v,，/x,建议/n,调整/vn,信号灯/n,配/v,时/ng
处理后：A7县特立路与东四路

Process finished with exit code 0
```

图 16

4.3 问题 3 分析方法与过程

4.3.1 流程图

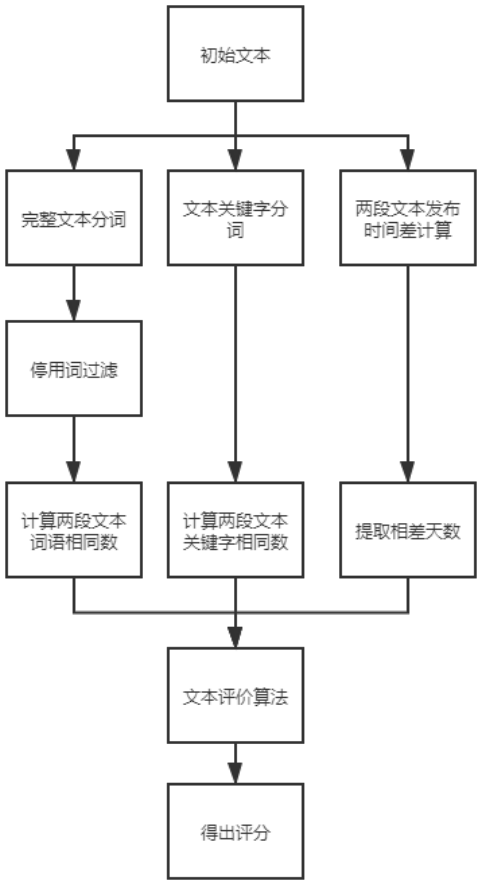

图 17

4.3.2 文本处理

对初始文本进行完整文本分词、关键字提取、同时计算文本发布时间差，通过计算两段文本的词语相同数，关键字相同数，建立文本评价算法。

基于 jieba 文本分词。jieba 分词算法使用了基于前缀词典实现高效的词图扫描，生成句子中汉字所有可能生成词情况所构成的有向无环图(DAG)，再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，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。将初始文本进行分词。

将文本分词之后，会出现许多停用词和符号的情况，如“的”，“之”，“尽管”等。使用停用词表的方法去除文本分词后的无意义词语。

4.3.3 相关性和完整性计算

运用之前的文本分词和关键字提取，通过计算两段文本的词语相同得到相关性结果，计算关键词相同数得到完整性结果

4.3.4 构建文本评价算法

构建评价得分模型

评价得分公式：

完整分词后词语相同数 * （1+ 0.1*关键词提取后词语相同数）/ 相隔天数

当相隔天数为 0 则乘以 1.1

结果如图 18 所示：

	留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	score
0	2549	A00045581	A2区景馨华	2019/4/25 9	2019年4月1	现将网友在	2019/5/10 1	2.333333333
1	2554	A00023583	A3区潇楚苑	2019/4/24 1	潇楚南路从	网友“A000	2019/5/9 9:4	0.285714286
2	2555	A00031618	请加快提高	2019/4/24 1	地处省会A	市民同志：	2019/5/9 9:4	0.771428571
3	2557	A00011073	在A市买公	2019/4/24 1	尊敬的书记	网友“A000	2019/5/9 9:4	0.928571429
4	2574	A0009233	关于A市公	2019/4/23 1	建议将“自	网友“A000	2019/5/9 9:5	0.78
5	2759	A00077538	A3区含浦钊	#########	欢迎领导来	网友“A000	2019/5/9 10	0.077419355
6	2849	A00010080	A3区教师村	2019/3/29 1	尊敬的胡书	网友“A000	2019/5/9 10	0.42
7	3681	UU00812	反映A5区为	2018/12/31	我做为一东	网友“UU00	2019/1/29 1	0.471428571
8	3683	UU008792	反映A市美	2018/12/31	我是美麓阳	网友“UU00	2019/1/16 1	1.0625
9	3684	UU008687	反映A市洋	2018/12/31	胡书记好！	网友“UU00	2019/1/16 1	1.21875
10	3685	UU0082204	反映A2区力	2018/12/30	我家住在A	网友“UU00	2019/3/11 1	0.2
11	3692	UU008829	A5区鄱阳村	2018/12/29	胡书记：您	网友“UU00	2019/1/29 1	0.433333333
12	3700	UU00877	A4区万国城	2018/12/29	尊敬的书记	网友“UU00	2019/1/14 1	0.625
13	3704	UU0081480	举报A市芒	2018/12/28	尊敬的领导	网友“UU00	2019/1/3 14	1
14	3713	UU0081227	建议增开A	2018/12/28	建议增开A	网友“UU00	2019/1/14 1	0.323529412
15	3720	UU008444	关于A市新	2018/12/27	2016年下半	网友“UU00	2019/3/6 10	0.882352941
16	3727	UU0081194	投诉A3区村	2018/12/27	12月16日上	网友“UU00	2019/1/3 14	0.471428571

图 18

5.运行结果分析

5.1 问题 1 结果分析

通过数据结构化后对数据进行中文分词、数据筛选处理后，将数据导入训练

集与验证集在 CNN 模型中训练，使 CNN 模型具备了对给定文本数据的自我学习判断其标签的能力，部分训练过程如图 19 所示。

```
Training and evaluating...
Epoch: 1
step: 100,train loss: 1.834, train accuracy: 0.312, val loss: 1.413, val accuracy: 0.459,training speed: 0.080sec/batch *

step: 200,train loss: 1.301, train accuracy: 0.344, val loss: 1.241, val accuracy: 0.540,training speed: 0.034sec/batch *

Epoch: 2
step: 300,train loss: 1.059, train accuracy: 0.719, val loss: 1.205, val accuracy: 0.562,training speed: 0.018sec/batch *

step: 400,train loss: 1.274, train accuracy: 0.406, val loss: 1.175, val accuracy: 0.564,training speed: 0.032sec/batch *

Epoch: 3
step: 500,train loss: 1.060, train accuracy: 0.562, val loss: 1.100, val accuracy: 0.612,training speed: 0.003sec/batch *

step: 600,train loss: 0.886, train accuracy: 0.688, val loss: 1.064, val accuracy: 0.606,training speed: 0.033sec/batch

step: 700,train loss: 0.821, train accuracy: 0.750, val loss: 1.001, val accuracy: 0.640,training speed: 0.033sec/batch *
```

图 19

最后，通过预测测试集中的文本数据计算 F-Score，得到本次数据一级标签分类的正确率，如图 20 所示。从图中可以看出该模型对所有留言信息数据按照一级标签的分类综合正确率为“85%”，其中对一级标签中“卫生计生”与“劳动和社会保障”两类预测效果最佳，其正确率达到 90%，其次是“交通运输”和“教育文体”两类标签，其正确率达到了“86%”。

```
Testing...
Test Loss:    0.5, Test Acc:  85.09%
Precision, Recall and F1-Score...
              precision    recall  f1-score   support

        城乡建设       0.82      0.81      0.82       150
        环境保护       0.77      0.88      0.82        57
        交通运输       0.86      0.75      0.80        57
        商贸旅游       0.83      0.70      0.76        81
        卫生计生       0.90      0.87      0.88        61
        教育文体       0.86      0.97      0.91       127
    劳动和社会保障       0.90      0.89      0.89       158

    accuracy                           0.85       691
   macro avg       0.85      0.84      0.84       691
weighted avg       0.85      0.85      0.85       691
```

图 20

5.2 问题 2 结果分析

通过数据结构化、中文分词后，将群众留言信息数据导入 word2vec 模型进行训练得到词向量，由 K-means 聚类分析算法将词向量划分到各个类并确定各自聚类中心，即各类关键词，将各类元素中与各自聚类中心的距离升序排列，根据“距离近则优”的原则选择前五个元素作为热度问题 Top5，结果如图 21 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	2643	2019.01.11–2019.07.08	A 市	A 市 58 车诈骗
2	2	2306.7	2019.01.23–2019.12.05	金路小区	金路小区有问题
3	3	1099.5	2019.01.02–2019.12.08	A4 区绿地海外滩小区	高铁问题
4	4	978.3	2019.01.15–2019.06.19	A 市富绿物业	富绿质量问题
5	5	867.7	2019.01.10–2019.07.15	西地省	地铁问题

图 21

从热点问题表（图 21）中可以看出排名前五的热度问题分别为：“A 市 58 车诈骗问题”、“金路小区有问题”、“高铁问题”、“富绿质量问题”和“地铁问题”。

5.3 问题 3 结果分析

分别对群众留言信息和政府答复意见信息进行数据结构化和中文分词后，对其进行 K-means 聚类分析得到每条信息数据的关键词，通过以下三个指标构建答复意见评价模型：

- 通过检测答复意见信息的关键词是否存在在对应群众留言信息的关键词中来考察答复意见的相关性。
- 通过检测群众留言信息的关键词是否都存在在对应答复意见信息的关键词中来考察答复意见的完整性。
- 最后在评价模型中加入答复时间和留言时间的时间差作为辅助指标来衡量答复的及时性。

最终得到的评分结果如图 22 所示。

留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	score
4233	UU0081297	积金新政的	11/2 18:	日将所有租	若您在11	11/20 9:	1.10
4236	UU0081736	区开发商/	11/2 15:	吗，是谁约	已张贴在	11/14 9:	2.95
4237	UU008323	餐饮的有/	11/2 12:	被告知000	的商业楼月	11/9 15:	0.63
4261	UU008885	水龙庭售机	10/30 17:	是人为邹丹	年5月7日有	12/7 14:	1.44
4267	UU0081605	户及购房事	10/30 13:	只要在A	核为准，	11/15 16:	1.72
4277	UU0081915	租赁有限	10/29 16:	我租赁使月	合同相对/	11/14 9:	1.90
4280	UU0082196	广场公寓核	10/29 11:	高出将近一	局了解，	11/14 9:	0.84
4287	UU0082299	对农村房月	10/28 20:	有权分吗？	法律法规，	11/9 15:	0.27
4295	UU0081422	在梅溪湖	10/27 15:	度过愉快的	通线网规划	11/14 9:	0.18
4325	UU008867	程花园酒	10/24 12:	过法律途径	及大多数	11/9 15:	0.41
4331	UU0082338	价格以及	10/24 10:	的计算表	位工程费月	11/6 10:	14.08

图 22

6 结论

网络问政的兴起给国家治理、社会发展带来的新的机遇，同时也伴随着新的挑战。通过网络问政平台全方位、全区域地通过群众留言了解群众对政府政策、社会热点、服务反馈等信息的坚实依托就是互联网背后复杂而强大的大数据处理分析，如何利用庞大的信息数据使政府层级管理者真正地了解群众感受、切实地解决群众反映地问题、及时地给予群众留言反馈也成为了当下“智慧型政府”转变的重要一环。

本文采用中文分词、word2vec 模型、CNN 神经网络构建、k-means 聚类分析算法，对网络问政平台平台上的相关群众留言和政府答复意见的信息数据进行深度挖掘，以更高效的方式将群众留言分类划归各部门处理，统计群众留言信息中最集中的热点问题，并定义系列评价指标对政府答复意见进行自我检查与分析，以助于后期网络问政服务的完善与进步。

7.参考文献

[1]张启宇,朱玲,张雅萍.中文分词算法研究综述[J].情报探索,2008(11):53-56.

[2]梁喜涛,顾磊.中文分词与词性标注研究[J].计算机技术与发展,2015,25(02):175-180.

[3]曾小芹.基于 Python 的中文结巴分词技术实现[J].信息与电脑(理论版),2019,31(18):38-39+42.

[4]毛郁欣,邱智学.基于 Word2Vec 模型和 K-Means 算法的信息技术文档聚类研究[J].中国信息技术教育,2020(08):99-101.

[5]杨锐,陈伟,何涛,张敏,李蕊伶,岳芳.融合主题信息的卷积神经网络文本分类方法研究[J].现代情报,2020,40(04):42-49.

[6]裴志利,阿茹娜,姜明洋,卢奕南.基于卷积神经网络的文本分类研究综述[J].内蒙古民族大学学报(自然科学版),2019,34(03):206-210

[7]何养明. 基于卷积神经网络结合词向量的中文短文本分类研究[D].重庆理工大学,2019.

[8]雷小锋,谢昆青,林帆,夏征义.一种基于 K-Means 局部最优性的高效聚类算法[J].软件学报,2008(07):1683-1692.

[9]王和勇,吴晓桦.聚类分析在网络客户评论研究中的应用研究[J].物流工程与管理,2014,36(04):86-89+115.

[10]Amit Kumar Sharma,Sandeep Chaurasia,Devesh Kumar Srivastava. Sentimental Short Sentences Classification by Using CNN Deep Learning Model with Fine Tuned Word2Vec[J]. Procedia Computer Science,2020,167.