
基于机器学习和 word2vec 模型的“智慧政务”留言分析处理与评价模型

摘要：

近年来，随着大数据、人工智能等技术的发展，以自然语言处理技术（NLP）为核心的智慧政务系统已经是社会治理创新发展的新趋势，利用计算机处理大量繁杂的网络上的群众留言信息拥有无可比拟的优势，也有着很大的发展空间。对建设服务型政府具有极大的推动作用。

本文旨在完整充分地使用多种机器学习领域的方法和知识，对不同算法的结果进行对比分析。

通过机器学习从特征提取，训练模型到预测的完整流程解决群众留言分类和聚类的操作。其中，在特征提取的过程中，以 TF-IDF 模型为桥梁实现文本数据到向量的转化。TF-IDF 作为传统机器学习领域进行特征提取的“终极产品”，虽然它在本质上提取的是一种优化后的词频特征，但在大样本学习和训练后，体现出了性能稳定可靠且对关键词敏感的强大优势。对于 TF-IDF 的缺陷，即大样本带来的词向量维数骤增的问题，我们使用 PCA 模型进行降维，并考虑到了降维对于精度带来的影响。

在对地点或人群进行提取时，我们以哈工大中文语言处理工具包为主要工具，通过词性筛选、句法分析等手段，灵活准确地挖掘出人群或地点。

在对答复意见评价中，以 word2vec 为主要工具，以求解词语与词语之间的相似度为基本手段，求解留言主题、留言详情以及答复意见之间的相似度，最终获得答复意见的相关性；我们通过留言中的子问题数与答复中的响应数之间的比例来衡量答复意见的完整性质量；最后，通过建立包含具有优良的可解释性的词汇的词表来评价可解释性。

关键词：TF-IDF、高斯聚类、word2vec 模型、自然语言处理

Summary of message analysis, processing and evaluation

model of "intelligent government" based on machine learning and word2vec

Abstract:

In recent years, with the development of big data, artificial intelligence and other technologies, intelligent government system with natural language processing technology (NLP) as the core has become a new trend of social governance innovation, and it has incomparable advantages to use computers to deal with a large number of messages left by the masses on the network. There is also a lot of room for development. It plays a great role in promoting the construction of service-oriented government.

The purpose of this paper is to make full use of a variety of machine learning methods and knowledge to compare and analyze the results of different algorithms.

The operation of mass message classification and clustering is solved through the complete process of feature extraction, training model and prediction through machine learning. Among them, in the process of feature extraction, TF-IDF model is used as a bridge to realize the transformation of text data to vector. As the "ultimate product" of feature extraction in the field of traditional machine learning, TF-IDF is essentially an optimized word frequency feature, but it shows a strong advantage of stable and reliable performance and sensitive to keywords after large sample learning and training. For the defect of TF-IDF, that is, the sudden increase of word vector dimension caused by large samples, we use PCA model to reduce the dimension, and take into account the impact of dimensionality reduction on the accuracy.

When extracting the location or population, we use the Chinese language processing toolkit of Harbin University of Technology as the main tool to dig out the crowd or place flexibly and accurately by means of part-of-speech screening, syntactic analysis and so on.

In the evaluation of response comments, we take word2vec as the main tool, solve the similarity between words and words as the basic means, solve the similarity between message subject, message details and response comments, and finally get the correlation of response opinions, and we measure the integrity and quality of responses by the ratio between the number of sub-questions in the message and the number of responses in the reply. Finally, the interpretability is evaluated by establishing a vocabulary containing excellent interpretable words.

Keywords: TF-IDF, Gaussian clustering, word2vec model, natural language processing

目录

引言.....	5
论文的结构安排	5
第一章 群众留言分类.....	5
1.1 问题描述与分析	5
1.2 数据预处理.....	6
1.2.1 预处理.....	6
1.2.2 数据效果增强	6
1.3 问题建模	7
1.3.1 特征提取.....	7
1.3.2 分类.....	7
1.4 结果分析	9
1.5 优化处理分析.....	11
1.6 优化处理	12
1.6.1 定义“高性能词”与“模糊词”	12
1.6.2 定义模糊词评判指标.....	14
1.6.3 使用稀疏度-峰均比与命中率优化结果对比.....	16
1.6.4 使用命中率指标去除停用词及效果展示.....	17
第二章 热点问题挖掘.....	19
2.1 问题描述与分析	19
2.2 数据预处理.....	19
2.2.1 pyltp 介绍.....	19
2.2.2 预处理.....	20
2.3 问题建模	20
2.3.1 特征提取.....	20
2.3.2 留言聚类:	20
2.3.3 热度计算.....	22
2.3.4 挖掘地点/人群, 并形成问题描述	25
2.4 聚类结果展示及分析	25
2.5 优化处理	28
2.5.1 使用峰均比优化相似度比较模型	28
2.5.2 使用 SVD 奇异值分解获得聚类数目优化 GMM.....	30

第三章 留言回复评价模型	32
3.1 word2vec-CBOW 模型简介	32
3.2 数据预处理.....	33
3.2.1 数据分析	33
3.2.2 数据预处理.....	34
3.3 问题建模	36
3.3.1 特征提取	36
3.3.2 去除冗余信息，获取子问题	36
3.3.3 相关性评估	38
3.3.4 完整性评估	39
3.3.5 可解释性评估	40
3.3.6 时效性评估	41
3.4 形成综合评价方案.....	41
第四章 总结与展望.....	42
4.1 总结	42
4.2 展望	42
参考文献.....	43
附录.....	43
(一) 实验环境:	43
(二) 答案数据:	44

引言

目前，基于 Python 的自然语言处理技术主要用于字符串处理、文本分类、文本摘要、文本聚类等操作，且绝大多数是面向英文处理的。将基于 Python 的自然语言处理技术运用到“智慧政务”中的文本挖掘领域主要存在以下困难：

1. 缺乏面向中文的功能全面的开源自然语言处理工具包。目前，诞生于宾夕法尼亚大学，以研究和教学为目的而生的 Python 自然语言处理工具 NLTK，全称“Natural Language Toolkit”，最为知名。但是它对中文并不友好，很多重要功能面向中文受限。

2. 一方面网络问政平台上的群众留言情况复杂，良莠不齐，另一方面汉语的语言习惯使得难以从文本中准确提取有效信息，识别命名实体，划分单独完整的表意单元。传统的用于主题建模或文档摘要的奇异值分解（SVD）、隐含语义分析（LSA）、迪利克雷分布（LDA）、TextRank 等算法会显得过于机械。

3. 我们放弃采用深度学习平台训练神经网络模型的思路。如很多学者使用 Tensorflow 建立循环神经网络在自然语言处理方面取得了重大突破。我们尝试以 word2vec 为有力工具，灵活便捷地使用，让其发挥到最佳效果。

本文对涉及到的数学模型作以简要介绍，重点着眼于数学算法与实际问题和编程实现的对接，以及选择该模型的原因和应用效果。

论文的结构安排

为紧扣问题展开论述，将本文共分为五大部分，各章节内容安排如下：

第一、二、三章分别对应解决“群众留言分类”、“热点问题挖掘”、“答复意见评价”。前三章中对于每一问大体按照问题描述与分析、数据预处理、问题建模、结果分析、优化处理的流程论述。

第四章为总结及展望。最后为参考文献、附录。

第一章 群众留言分类

1.1 问题描述与分析

查阅附件二，我们可以发现，理论上通过对留言的文本使用传统的有监督的机器学习分类算法，即可完成分类。

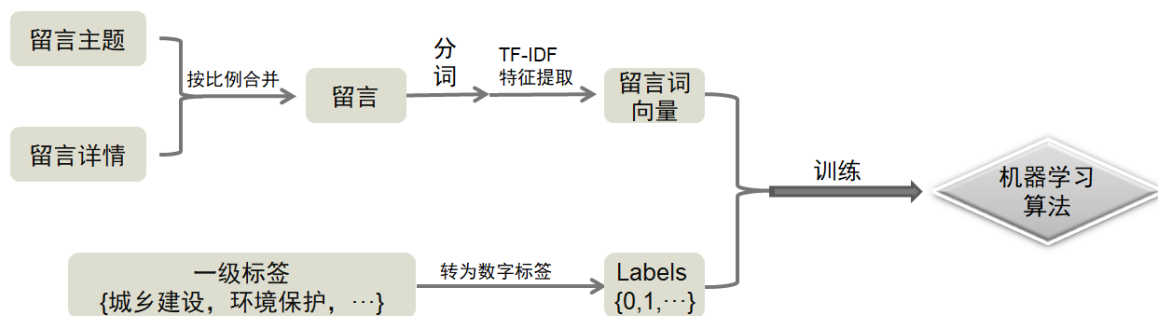


图 1-1 群众留言分类模型框架

本题中需要灵活处理的地方在于如何协调地同时使用留言主题与留言详情。因此，在数据预处理中进行了特殊处理。

1.2 数据预处理

1.2.1 预处理

第一步，考虑到留言主题对于分类的重要性程度更高，为保证留言主题相对于留言详情始终保持一个固定的较高的权重，对留言主题和留言详情进行固定文本长度比的合并。具体实现方法为：通过计算主题与详情的长度比，让主题重复出现多次，保证主题与详情的文本长度之比固定为 1：6。

第二步，针对合并后的留言主题和详情中的词语建立了停用词列“stopwords.txt”，以供操作。

第三步，利用 jieba 中文分词库模型对留言进行分词操作。

另外，根据第一次运行的实验结果的分析，将会在第二次运行前添加新的停用词至“stopwords.txt”，详见优化处理。

1.2.2 数据效果增强

为了在下一步提取中能够提取更多有用的特征信息，我们同时考虑使用了“附件一”中的二级分类、三级分类作为留言详情参与分词处理，以提高后期训练的效果。具体做法是利用“附件一”将二级分类和三级分类的标签加入到对应的一级分类中的“留言详情”中。

但在使用附件二完整数据训练后发现，分类效果变差。即虽然传统的机器学习算法对样本高度依赖，但在大样本训练后，会更趋于稳定可靠。

实现代码示例如下：

```

for i in range(0, len(dataload)):
    data['留言详情'][i] = round(len(dataload['留言详情'][i]) / len(dataload['留言主题'][i]) / 6) * dataload['留言主题'][i] + \
        dataload['留言详情'][i]
  
```

图 1-2 留言数据合并代码示例

1.3 问题建模

1.3.1 特征提取

使用模型：词频-逆向文件频率模型（TF-IDF）

TF：一个词语在一篇文章中出现次数越多，同时所有文档中出现次数越少，越能够代表该文章。

$$TF(x) = \frac{\text{该文章中出现该词的次数}}{\text{文章总词数}}$$

IDF：IDF 反应了一个词在所有文本中出现的频率，如果一个词在很多的文本中出现，那么它的 IDF 值应该低，比如中文中的“的”。

$$IDF(x) = \log \frac{N}{N(x)}$$

（N 代表语料库中文本的总数，而 N(x) 代表语料库中包含词 xx 的文本总数）

TF-IDF：某一特定文件内的高词语频率，以及该词语在整个文件集中的低文件频率，可以产生出高权重的 TF-IDF。因此，TF-IDF 倾向于过滤掉常见的词语，保留重要的词语。

$$TF-IDF(x) = TF(x) * IDF(x)$$

通过 TF-IDF 模型，将有利于分类的具有本类型特征的高权重词语保留而又过滤掉了在多类文本中共同出现的低权重词语。

1.3.2 分类

同时并列使用朴素贝叶斯、多项式朴素贝叶斯、随机梯度下降三种模型

1) 朴素贝叶斯 ComplementNB

首先介绍贝叶斯定理：即已知 $P(A|B)$ ，则 $P(B|A) = \frac{P(A|B) * P(B)}{P(A)}$ ，假设 A，B 独立，则 $P(AB) = P(A) * P(B)$ 。

朴素贝叶斯分类模型就是对于给定的待分类项，求解在此基础上每个类别出现的概率，最大的即作为最终分类的类别。其步骤为：

- 1、假设 $x = (a_1, a_2, \dots, a_n)$ 为待分类项， a_i 为每个待分类属性。
- 2、类别集合为 $\{y_1, y_2, \dots, y_n\}$ 。
- 3、计算概率 $p(y_1|x), p(y_2|x), \dots, p(y_n|x)$ 。
- 4、取第 3 步骤中概率值最大的为待分类项的类别。

具体操作是把上述第 3 步骤转化成求每个特征的条件概率。

$$p(y_i|x) = \frac{p(x|y_i) * p(y_i)}{x}$$

因为最终考虑的是在每个类别中概率值的大小，所以去掉除数 x ，不影响最终分类结果。其代码实现过程类似，即：

- 1、首先实现计算每个类别的概率。
- 2、计算每个特征在不同类下的条件概率（每个特征出现的次数/类别下特征的总个数）。
- 3、计算 $p(y_i) \prod p(a_j|y_i)$ 。
4. 判断大小进而判断类别。

2) 多项式朴素贝叶斯 MultinomialNB

朴素贝叶斯算法中一种常用的假设是多项式朴素贝叶斯（multinomial naive Bayes），它假设特征是由一个简单多项式分布生成的。多项分布可以描述各种类型样本出现次数的概率，因此多项式朴素贝叶斯非常适合用于描述出现次数或者出现次数比例的特征。

这个理念和高斯贝叶斯分布的一样，只不过模型数据的分布不再是高斯分布，而是用多项式分布代替而已。多项式朴素贝叶斯的特征矩阵经常是稀疏矩阵（不一定总是稀疏矩阵），并且它经常被用于文本分类。我们可以使用 TF-IDF 向量技术，也可以使用常见并且简单的单词计数向量手段与贝叶斯配合使用。这两种手段都属于常见的文本特征提取的方法，可以很简单地通过 sklearn 来实现。

3) 随机梯度下降 SGDClassifier

梯度下降算法包含多种不同的算法，有批量梯度算法，随机梯度算法，折中梯度算法等等。随机梯度下降（SGD）是一种简单但非常有效的方法，多用于支持向量机、逻辑回归等凸损失函数下的线性分类器的学习。并且 SGD 已成功应用于文本分类和自然语言处理中经常遇到的大规模和稀疏机器学习问题。

对于随机梯度下降算法而言，它通过不停的判断和选择当前目标下最优的路径，从而能够在最短路径下达到最优的结果。我们可以在一个人下山坡为例，想要更快的到达山底，最简单的办法就是在当前位置沿着最陡峭的方向下山，到另一个位置后接着上面的方式依旧寻找最陡峭的方向走，这样每走一步就停下来观察最下路线的方法就是随机梯度下降算法的本质。其算法理论基础如下：

在线性回归中，我们给出回归方程，如下所示：

$$h(\theta) = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n = \sum \theta_i x_i$$

我们知道，对于最小二乘法要想求得最优变量就要使得计算值与实际值的偏差的平方最小。而随机梯度下降算法对于系数需要通过不断的求偏导求解出当前位置下最优化的数据，那么梯度方向公式推导如下公式，公式中的 θ 会向着梯度下降最快的方向减少，从而推断出 θ 的最优解。

$$\begin{aligned}
\frac{\partial}{\partial \theta_j} J(\theta) &= \frac{\partial}{\partial \theta_j} \frac{1}{2} (h_{\theta}(x) - y)^2 \\
&= 2 \cdot \frac{1}{2} (h_{\theta}(x) - y) \cdot \frac{\partial}{\partial \theta_j} (h_{\theta}(x) - y) \\
&= (h_{\theta}(x) - y) \cdot \frac{\partial}{\partial \theta_j} \left(\sum_{i=0}^n \theta_i x_i - y \right) \\
&= (h_{\theta}(x) - y) x_j
\end{aligned}$$

https://blog.csdn.net/qq_37142346

因此随机梯度下降法的公式归结为通过迭代计算特征值从而求出最合适的值。 θ 的求解公式如下：

$$\theta = \theta - \alpha \cdot \frac{\partial J(\theta)}{\partial \theta}$$

<https://blog.csdn.net/42346>

其中 α 是下降系数，即步长，学习率，通俗的说就是计算每次下降的幅度的大小，系数越大每次计算的差值越大，系数越小则差值越小，但是迭代计算的时间也会相对延长。 θ 的初值可以随机赋值。

1.4 结果分析

由于当前无测试数据，故只给出评价分析：

分类结果评价模型：F1-score

在对分类模型的实现效果进行评价时，通常采用 F-Score 模型，下面将对这种模型做出介绍。

首先基于二分类(只有正负样本)做出介绍，下面给出几个量的定义：

TP: True Positive, 被判定为正样本，事实上也是正样本。

FP: False Positive, 被判定为正样本，但事实上是负样本。

TN: True Negative, 被判定为负样本，事实上也是负样本。

FN: False Negative, 被判定为负样本，但事实上是正样本。

Accuracy: 表示预测结果的精确度，预测正确的样本数除以总样本数。

precision: 准确率，又称为查准率，表示预测结果中，预测为正样本的样本中，正确预测为正样本的概率；

recall: 召回率，又称为查全率，表示在原始样本的正样本中，最后被正确预测为正样本的概率；

精确率(Precision)和召回率(Recall)评估指标，理想情况下做到两个指标都高当然最好，但一般情况下，Precision 高，Recall 就低，Recall 高，Precision 就低。所以在实际中常常需要根据具体情况做出取舍，例如一般的搜索情况，在保证召回率的条件下，尽量提升精确率。而像癌症检测、地震检测、金融欺诈等，则在保证精确率的条件下，尽量提升召回率。

下面引出一个新的指标 F-score, 综合考虑 Precision 和 Recall 的调和值

$$F - Score = (1 + \beta^2) \cdot \frac{Precision \cdot Recall}{\beta^2 \cdot Precision + Recall}$$

当 $\beta=1$ 时,称为 F1-score 或者 F1-Measure,这时,精确率和召回率都很重要,权重相同。

当有些情况下,我们认为精确率更重要些,那就调整 β 的值小于 1,

如果我们认为召回率更重要些,那就调整 β 的值大于 1。

F1 指标 (F1-score):F1-score 表示的是 precision 和 recall 的调和平均评估

指标。
$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

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

在本模型中由于有 n 个分类,则 $F_1 = \frac{1}{n} \sum_{i=1}^n \frac{2P_iR_i}{P_i + R_i}$, 其中 P_i 为第 i 类的精确率, R_i 为第 i 类的召回率,结果取值区间为 0 到 1,越趋向于一说明分类效果越好。

使用第二次下发的全部数据,我们按照“8: 2”的比例将全部数据切分为训练集与测试集;而后分别使用多项式朴素贝叶斯、朴素贝叶斯、随机梯度下降三种分类算法带入运行。

按照 F-score 模型对分类结果评价得到如下结果:

	precision	recall	f1-score	support
0	0.61	0.97	0.75	386
3	0.99	0.71	0.82	190
5	0.95	0.27	0.42	142
10	0.94	0.53	0.68	253
11	1.00	0.47	0.64	179
12	0.95	0.86	0.90	329
13	0.68	0.98	0.80	363
accuracy			0.76	1842
macro avg	0.87	0.68	0.71	1842
weighted avg	0.83	0.76	0.75	1842

图 1-4-1 多项式朴素贝叶斯算法运行结果

	precision	recall	f1-score	support
0	0.84	0.94	0.88	401
3	0.91	0.95	0.93	182
5	0.92	0.68	0.78	102
10	0.91	0.79	0.85	258
11	0.96	0.87	0.91	172
12	0.92	0.94	0.93	338
13	0.91	0.95	0.93	389
accuracy			0.90	1842
macro avg	0.91	0.87	0.89	1842
weighted avg	0.90	0.90	0.90	1842

图 1-4-2 朴素贝叶斯算法运行结果

	precision	recall	f1-score	support
0	0.86	0.92	0.89	401
3	0.93	0.95	0.94	182
5	0.87	0.88	0.87	102
10	0.91	0.83	0.87	258
11	0.95	0.91	0.93	172
12	0.95	0.96	0.96	338
13	0.95	0.95	0.95	389
accuracy			0.92	1842
macro avg	0.92	0.91	0.92	1842
weighted avg	0.92	0.92	0.92	1842

图 1-4-3 随机梯度下降算法运行结果

注：图中序号列（0、3、5、10、11、12、13）是一级标签的数字标签，分别代表：“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”。

1.5 优化处理分析

先对数据进行分析如下：

由于机器学习的本质是通过留言中某些关键词的词频特征对此条留言进行分类，即每一类留言都存在一批高概率出现的关键词。以在“城乡建设”中出现的高频词为例得到一下统计结果：

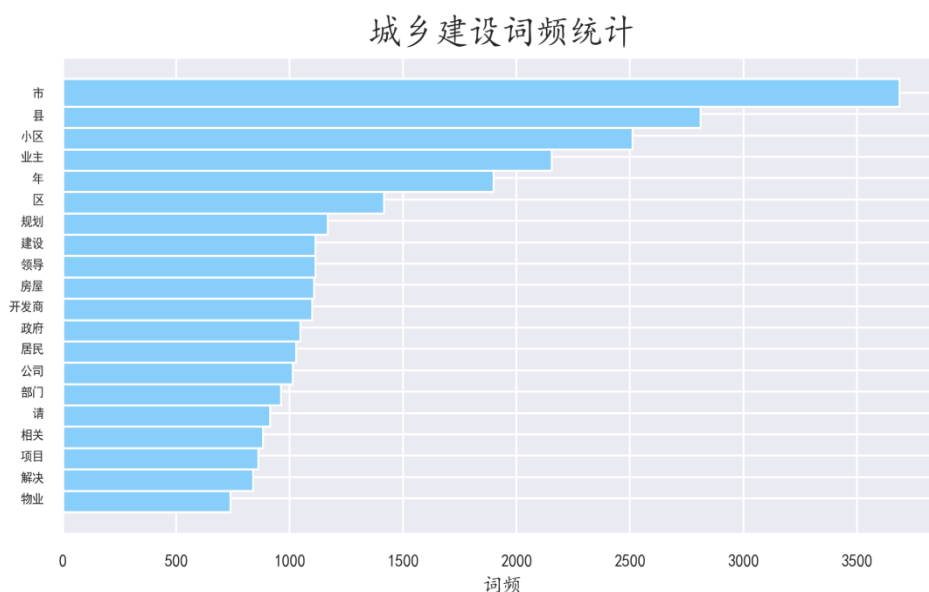


表 1-5-1 城乡建设类留言词频统计（示例）

因此，统计出每种一级标签下出现频率最高的关键词有利于从根本上找到限制准确率 and 精确率的因素。将在各个一级标签中出现词频前 1000 的词汇汇总如下：（截取前 20 行）

	城乡 建设	环境 保护	交通 运输	商贸 旅游	卫生 计生	教育 文体	劳动 和社 会保 障	稀疏 度	峰均比
市	3689	1375	1039	2320	1204	4188	3580	7	1.685312
县	2811	1510	1108	1757	1142	3486	2867	7	1.662148
小区	2511	393	47	659	0	232	0	5	3.267829
业主	2156	105	11	635	0	87	0	5	3.600534
年	1900	549	244	798	905	1959	4797	7	3.011029
区	1417	698	129	746	240	1296	1063	7	1.774736
规划	1168	69	85	81	0	131	0	5	3.80704
建设	1115	275	98	213	40	434	164	7	3.336896
领导	1114	555	221	474	517	1373	1600	7	1.913222
房屋	1108	122	11	116	0	0	0	4	3.266028
开发商	1099	0	0	305	0	72	0	3	2.23374
政府	1048	451	250	380	246	513	981	7	1.896097
居民	1028	1047	56	217	37	106	226	7	2.69746
公司	1014	533	415	900	70	161	2105	7	2.834744
部门	963	484	219	646	247	479	597	7	1.85447
请	915	422	207	569	444	1087	1020	7	1.631432
相关	884	388	130	537	255	504	766	7	1.786374
项目	863	165	20	143	50	246	135	7	3.724414
解决	840	321	114	214	146	625	1242	7	2.482581
物业	739	0	0	308	0	0	66	3	1.991914

表 1-5-2 一级标签中词频前 1000 的词汇总（截取前 20 行）

由于高频词经常会“跨类出现”，因此通过在七大一级分类中汇总前 1000 个高频词是最终得到 3481 个高频词。

1.6 优化处理

1.6.1 定义“高性能词”与“模糊词”

为便于分析高频关键词的“分类性能”，现定义两种关键词：

高性能词：仅仅在某一类标签中高频出现，具有优良的指向性，十分有利于准确分类。示例如下：

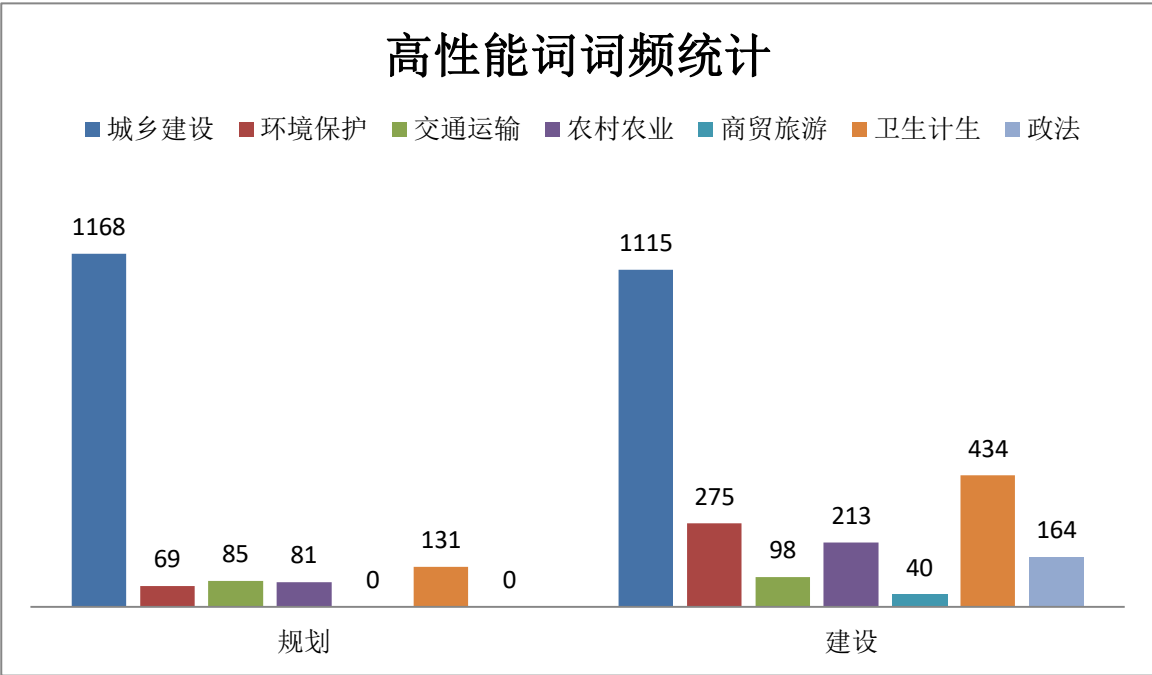


图 1-6-1 高性能词词频统计（以“规划”和“建设”为示例）

在极限情况下，假设某一篇文档中仅包含“规划、建设”词样。当算法“遇到”高性能词“规划、建设...”时，由于“规划、建设”唯一出现且大量出现在一级分类“城乡建设”之中，故可以准确判断分类为“城乡建设”。

模糊词：在两个及以上标签中有出现频率较接近，指向性较差，不利于实现准确分类。如图 1-6-2 中的标红词汇：“每人、周岁”。

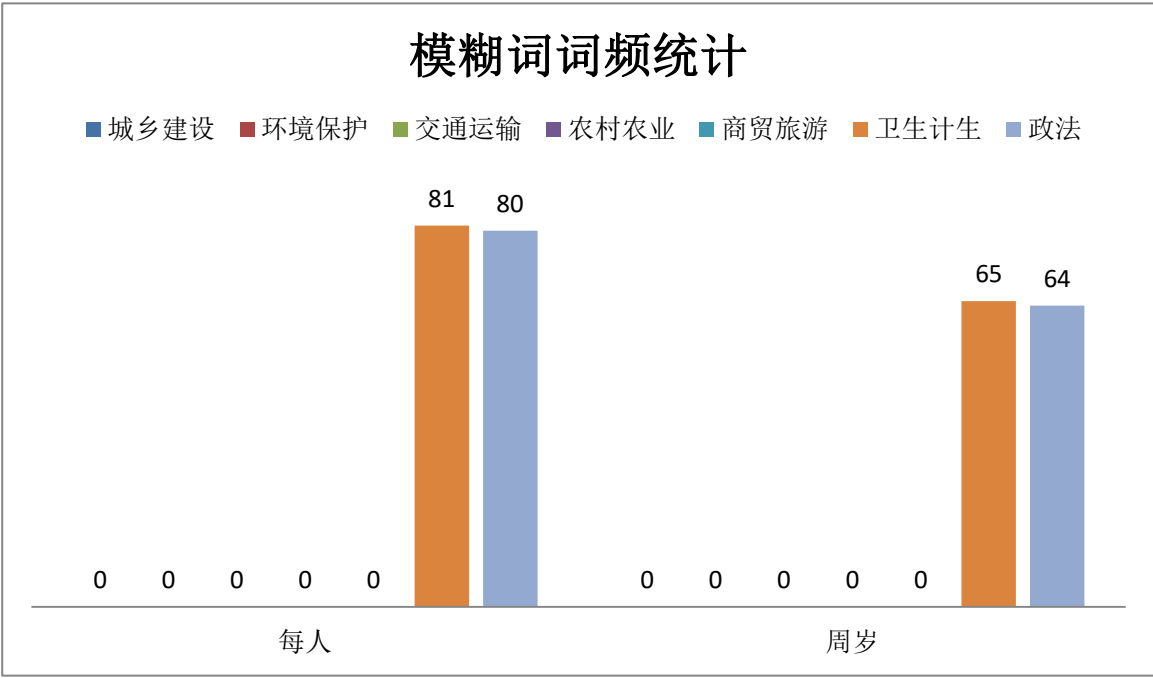


图 1-6-2 模糊词词频统计（以“每人”和“周岁”为示例）

同理，当算法“遇到”模糊词“每人、周岁”时，由于它在“卫生计生”、“政法”中出现频率十分接近，很容易发生“误判”。

1.6.2 定义模糊词评判指标

a). 稀疏度-峰均比

为进一步量化区分高性能词和模糊词，现引入两个在信号分析与处理领域十分重要的数学参数：

稀疏度：信号中非零元素的个数。在本问题中即为某一个关键词在各类一级标签中出现频率不为零的种类数。

峰均比：一种对波形的测量参数,等于波形的振幅除以有效值(RMS)所得到的一个比值。在本问题中即为一组数据中的最大值比均值。

稀疏度-峰均比：即在求解峰均比时，只考虑非零项，而忽略为零项。

$$\text{稀疏度 - 峰均比} = \frac{\text{MAX}(X_i)}{\sum(X_i)/\text{sparsity}}$$

(sparsity: 稀疏度)

主要基于以下考虑：

- 1.稀疏度等于一的词汇属于最有利于分类的词汇，即高性能词。随着稀疏度增大，非零项目增多，分类性能逐渐下降，
- 2.考虑分类性能最差的情况，即若一个词汇在七大类中完全均匀分布，算法会“纠结”于将该词最终划归到哪一类中。示例如下：

城乡	环境	交通	商贸	卫生	教育	劳动	稀疏	峰均比
----	----	----	----	----	----	----	----	-----

	建设	保护	运输	旅游	计生	文体	和社会保障	度	
XX	1	1	1	1	1	1	1	7	1

图 1-6-3 分类性能最差示例

此时该词汇分类性能最差。故峰均比指标有利于衡量该词汇在各个一级分类中偏离均匀分布的程度。峰均比越高，则该词汇在某一个一级分类中分布的更集中，考虑到极端情况如下：

	城乡 建设	环境 保护	交通 运输	商贸 旅游	卫生 计生	教育 文体	劳动 和社会 保障	稀疏 度	峰均比
XY	1	0	0	0	0	0	0	1	7

图 1-6-4 分类性能最好示例

此时峰均比取得最大值，该词汇分类性能也最好。

3. 稀疏度-峰均比相比于峰均比的优势主要是只考虑在非零项目中求局部的峰均比，避免了为零项的干扰。因为算法在“决策”的时候根本不会考虑认为这个词对于为零项有贡献。


b). 命中率（分类准确率）

将算法的“决策”过程（根据词汇在各个一级分类中的词频对词汇进行分类）类比为简单的以古典概型为基础的“摸球-投球”游戏。具体介绍如下：

将算法中遇到的一个高频词看作一个球：

词语  球

将七大一级分类看作七个要投掷的目标袋子：

一级分类{城乡建设、环境保~}  七个篮子

现假设高频词通过大样本训练后得到如下概率分布：

	城乡建 设	环境保 护	交通运 输	商贸旅 游	卫生计 生	教育文 体	劳动和 社会保 障
word	30	60	0	0	0	0	0

图 1-6-5 高频词通过大样本训练后概率

第一步：摸球

用于摸球的袋子里有好多个球，摸出球后发现这个球名为“word”，由于预先根据大样本训练得到的先验概率，可以预估袋子中名称均为“word”的球属性为“城乡

建设”的占比 1/3，“环境保护”的占比 2/3。（运用先验概率预估计位置袋子中的球）

第二步：投球

显然算法只会考虑往“城乡建设”和“环境保护”两个篮子里去投，但是算法不会因为这个词属性为“环境保护”的概率更大而将名为“word”的词全部投进一个篮子里，而是按照“经验”选择篮子。于是，1/3 的“word”被投进“城乡建设”的篮子，2/3 的“word”被投进“环境保护”的篮子。

则最终投篮的命中率为 $1/3 * 1/3 + 2/3 * 2/3 = 5/9$

命中率计算公式为：

$$Accuracy = \sum (X_i / \sum(X_i))^2$$

1.6.3 使用稀疏度-峰均比与命中率优化结果对比

使用稀疏度-峰均比与命中率两种评判模糊词的指标求解结果示例如下：

	城 乡 建 设	环 境 保 护	交 通 运 输	商 贸 旅 游	卫 生 计 生	教 育 文 体	劳 动 和 社 会 保 障	稀 疏 度	峰 均 比	命 中 率
强行	60	31	16	50	30	43	70	7	1.633333	0.166289
一事	78	39	12	63	32	42	57	7	1.690402	0.170183
规范	108	27	54	51	39	81	45	7	1.866667	0.170809
楚	43	29	51	46	0	47	55	6	1.217712	0.172125
时	399	203	82	252	274	427	593	7	1.861435	0.176897
10	554	271	168	365	269	705	767	7	1.732494	0.176959
听说	47	21	12	29	30	68	53	7	1.830769	0.177041
外	72	32	12	34	38	61	82	7	1.734139	0.177043
民生	105	51	35	34	36	60	109	7	1.774419	0.177307
举报	372	258	17	318	150	196	219	7	1.701961	0.177384

图 1-6-6 求解结果

为衡量两种指标识别模糊词的性能好坏，分别根据两种指标排序，得到性能最差的前 100 个模糊词，加入“stopword.txt”当中，得到两组结果：

	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.90	0.93	0.91	399	0	0.88	0.92	0.90	399
3	0.92	0.97	0.94	197	3	0.91	0.97	0.94	197
5	0.93	0.86	0.90	109	5	0.93	0.83	0.88	109
10	0.94	0.89	0.92	237	10	0.93	0.88	0.91	237
11	0.94	0.91	0.92	161	11	0.94	0.91	0.92	161
12	0.95	0.94	0.95	332	12	0.95	0.93	0.94	332
13	0.94	0.96	0.95	407	13	0.94	0.95	0.94	407
accuracy			0.93	1842	accuracy			0.92	1842
macro avg	0.93	0.92	0.93	1842	macro avg	0.93	0.91	0.92	1842
weighted avg	0.93	0.93	0.93	1842	weighted avg	0.92	0.92	0.92	1842

图 1-6-7 按命中率去除模糊词

图 1-6-8 按峰均比去除模糊词

因此，我们可得出结论：使用**命中率**作为判别模糊词的指标更为准确。

1.6.4 使用命中率指标去除停用词及效果展示

优化后，重新运行程序，得到结果如下：

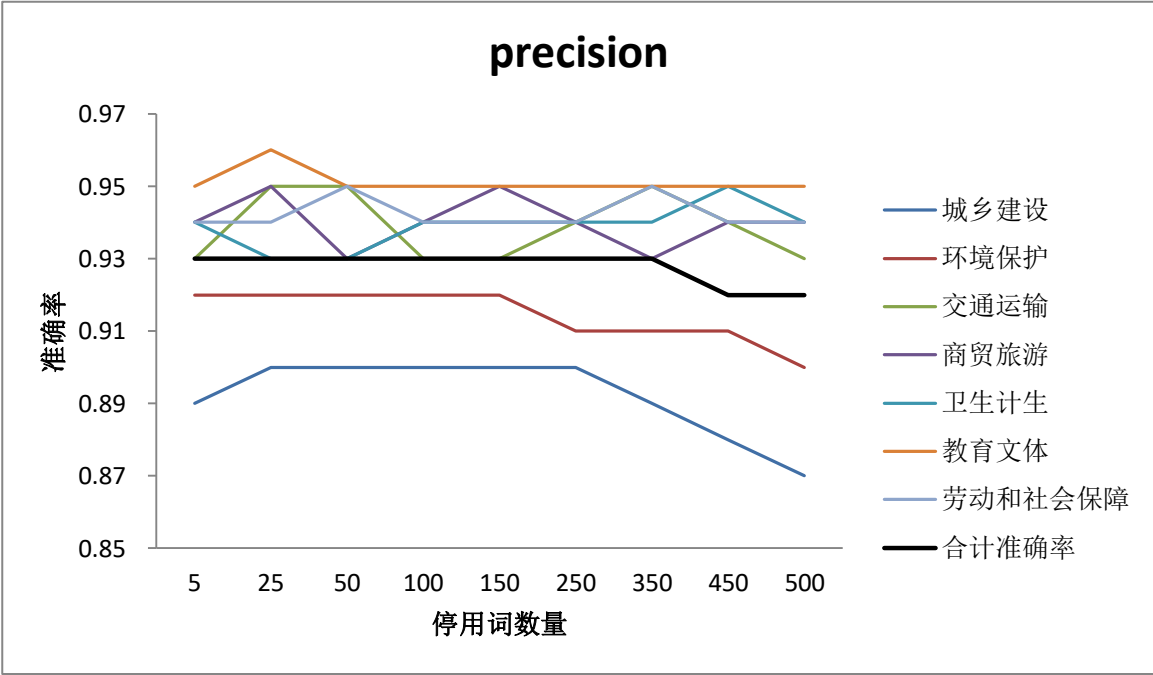


图 1-6-9 准确率结果

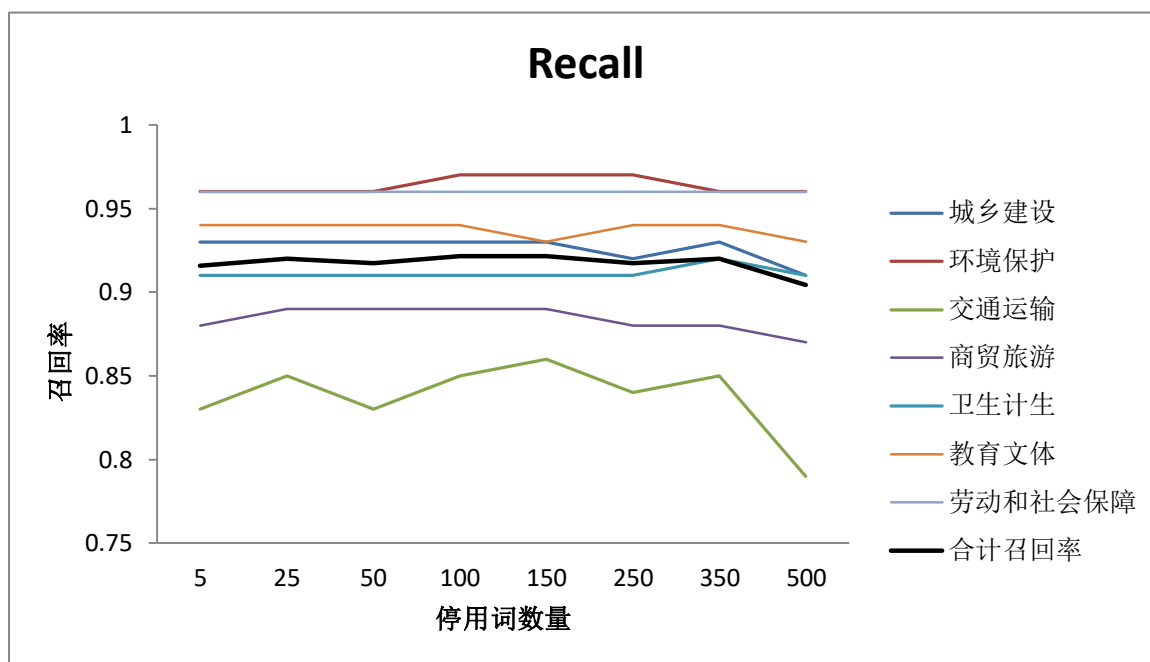


图 1-6-10 召回率结果

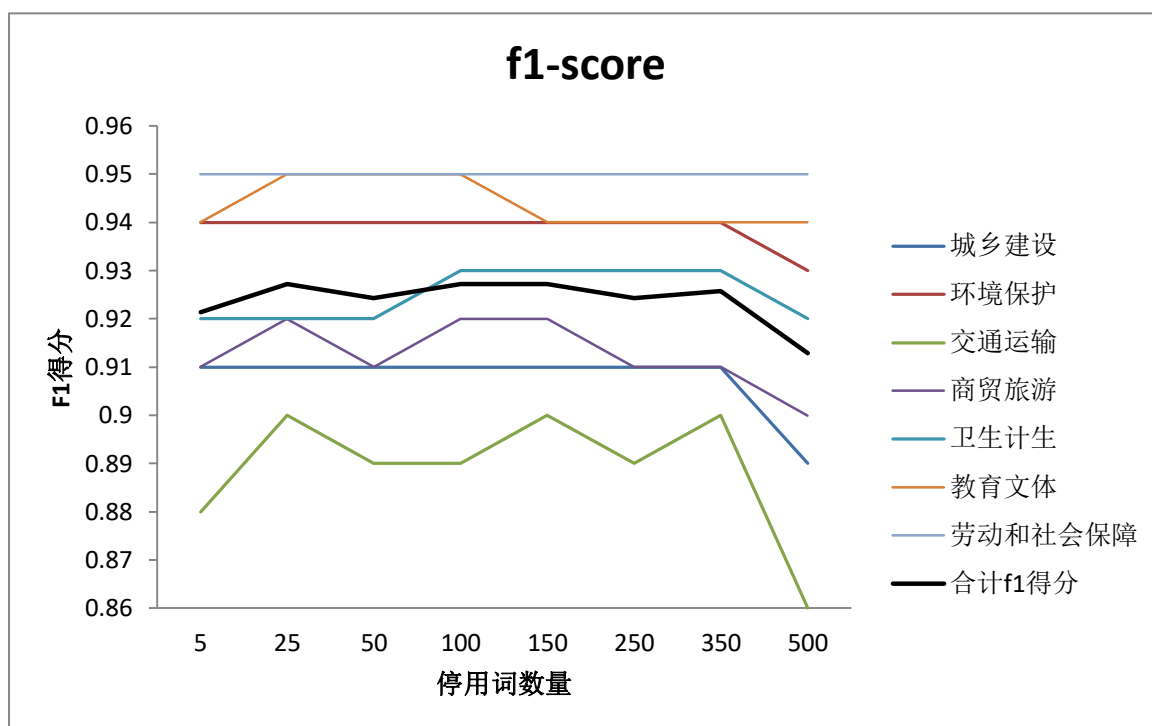


图 1-6-11 F1 得分结果

由图可知：当停用词数量在 25 附近以及 100 到 150 之间时，F1-score 达到最大值 0.927。由 IDE 中输出参数只能达到小数点后两位，所以汇总后的结果存在一定误差。

差，不能明显地看出准确率和召回率一升一降的变化关系。但可以使用这种思路在无需确定命中率阈值的情况下找到最佳的平衡点。若去除模糊词数量过少，影响机器学习的准确性，若去除模糊词数量过多，有失算法的泛化能力。

综上所述，利用命中率排名去除掉性能最差的前 25 个模糊词，最终结果如下：

	precision	recall	f1-score	support
0	0.87	0.94	0.90	399
3	0.94	0.95	0.94	197
5	0.95	0.83	0.89	109
10	0.94	0.87	0.91	237
11	0.94	0.91	0.92	161
12	0.95	0.93	0.94	332
13	0.94	0.96	0.95	407
accuracy			0.93	1842
macro avg	0.93	0.91	0.92	1842
weighted avg	0.93	0.93	0.93	1842

图 1-6-12 利用命中率排名去除掉性能最差的前 25 个模糊词后结果

第二章 热点问题挖掘

2.1 问题描述与分析

第一步，对于处理后的留言主题、留言详情数据使用 Tf-idf 向量模型完成特征提取。

第二步，利用使用门槛最低相似度比较聚类方法做为基础对照算法，同时使用高斯混合模型以及 K-means 的方法使用留言主题对留言进行聚类操作，通过这种无监督的方法实现对“无标签数据”的聚类。

第三步，也是最为棘手的一步。热度的评价并没有先例可循，我们充分关注点赞数与反对数，尽可能合理地定义热度指数评价指标，并建立热度随时间的衰减模型。按照最终你所得热度指数进行排名，名次即为题目要求中的“问题 ID”，且同一类问题的 ID 相同。至此，得到数据：“表 2-热点问题留言明细表”。

第四步，以哈工大中文处理工具包 PyLtp 为有力工具，附加多样灵活的字符串操作，实现在表二的基础上进行提取“人物/地点”和形成“问题描述”。得到数据：“表 1-热点问题表”。在对本题的论述中，首先根据聚类得到的“热点问题留言明细表”进行分析优化，优化完成后再给出“热点问题表”。

2.2 数据预处理

2.2.1 pyltp 基本介绍：

LTP (Language Technology Platform) 为中文语言技术平台，是哈工大社会计

算与信息检索研究中心开发的一整套中文语言处理系统。LTP 制定了基于 XML 的语言处理结果表示,并在此基础上提供了一整套自底向上的丰富而且高效的中文语言处理模块(包括词法、句法、语义等 6 项中文处理核心技术)。

通过和网络上其他很多可开源使用且对 python 友好的自然语言处理工具包(中科院 ICTCLAS、SnowNlp、Hanlp、清华 THULAC)对比,我们发现 pyltp 更为准确高效且功能较强大。尤其是在词性标注和命名实体识别以及句法分析方面使用简洁方便。

在本论文中,主要使用了免费开源的四大功能:分词,词性标注,命名实体识别,依存句法分析。

以某一条留言主题为例做在线测试:

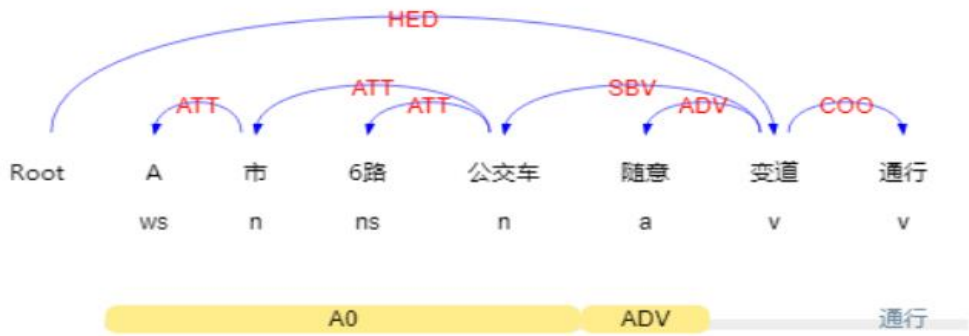


图 2-2-1 测试示例

根据定中关系“ATT”,即可获得“A、市、6路、公交车”四个单词。

根据题意:“请根据附件 3 将某一时段内反映特定地点或特定人群问题的留言进行归类”,因此,依靠提取涉及定中关系的单词可以准确获取特定地点或特定人群的特征信息,依此来实现将地点和人群作为主要区别的聚类。

通过依靠定中关系可以挖掘出最有价值信息,为之后的处理打下良好基础。

2.2.2 预处理

第一步:分词,去停用词。

第二步,使用哈工大中文处理工具包 pyltp 挖掘出留言主题中的定中关系作为预处理数据,而过滤掉其他无关词汇。

2.3 问题建模

2.3.1 特征提取

继续使用 TF-IDF 模型将留言中的文本数据转化为词向量数据。

2.3.2 留言聚类:

同时使用相似度比较、SVD、K-means、GMM 模型聚类:

a). 相似度比较聚类模型

模型框架如下:

定义：两条留言属于同一类的相似度阈值为 $\text{sim}0$ ；

说明：判断为同一类的条件是在第 j 类留言中， $\exists j$ 使得 $\text{sim}(i,j)$ 低于阈值。阈值根据聚类效果调试得到。

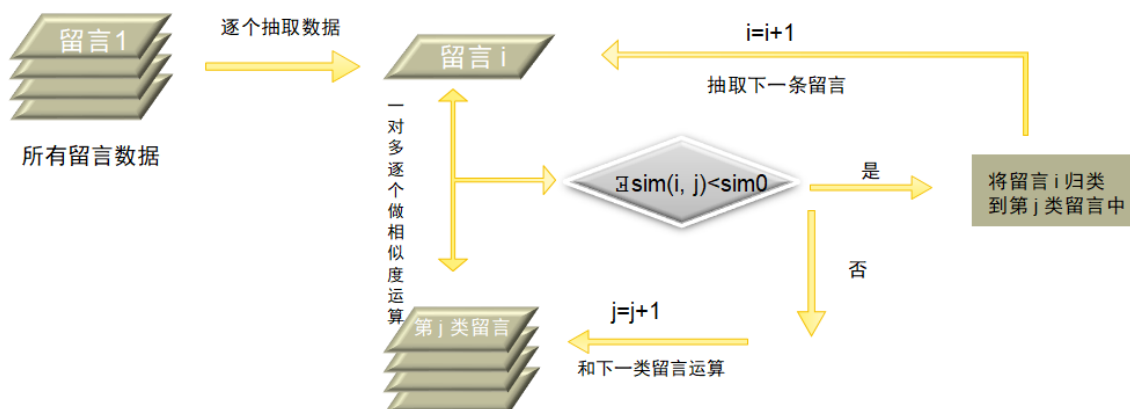


图 2-3-1 相似度比较聚类模型框架

b). SVD

c). K-means 聚类模型

K-means 算法是很典型的基于距离的聚类算法，采用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。该算法认为簇是由距离靠近的对象组成的，因此把得到紧凑且独立的簇作为最终目标。k-means 算法特点在于：同一聚类的簇内的对象相似度较高；而不同聚类的簇内的对象相似度较小。

k-means 是用来解决著名的聚类问题的最简单的非监督学习算法之一。

该过程由以下几个步骤：

将一组数据划分为预先设定好的 k 个簇。其主要思想是为每个簇定义一个质心。设置这些质心需要一些技巧，因为不同的位置会产生不同的聚类结果。因此，较好的选择是使它们互相之间尽可能远。

接下来将数据中的每个点归类为距它最近的质心，距离的计算可以是欧式距离、曼哈顿距离、切比雪夫距离等。如果所有的数据点都归类完毕，那么第一步就结束了，早期的聚合过程也相应完成。

然后，我们根据上一步所产生的结果重新计算 k 个质心作为各个簇的质心。一旦获得 k 个新的质心，我们需要重新将数据集中的点与距它最近的新质心进行绑定。一个循环就此产生。作为循环的结果，我们发现 k 个质心逐步改变它们的位置，直至位置不再发生变化为止。

d). 高斯混合模型 (Gaussian mixture model)

核心步骤如下：

1. 通过观察采样的概率值和模型概率值的接近程度，来判断一个模型是否拟合良好；
2. 通过模型来计算数据的期望值，通过更新分布的均值和标准差(参数 μ 和 σ)来让期望值最大化；
3. 反复迭代这个过程很多次，直到两个概率值非常接近时；

4. 停止更新并完成模型训练。

高斯混合模型(Gaussian Mixture Model)聚类算法作为 K-means 算法的进化，混合高斯模型聚类算法相比于 K-means 模型的核心优势如下：

1. 计算伸缩性：使用多个高斯分布的组合来刻画数据分布，计算伸缩性好；
2. 参数依赖性：可调整参数为分布的均值和标准差(参数 μ 和 σ)；
3. 普适性能力：描述能力和泛化能力优于 K 均值聚类(K-Means)；
4. 抗噪音能力：由于 K 均值聚类(K-Means)；
5. 结果解释性：模型和结果均具有解释性。

高斯混合模型是概率模型，其假设所有数据点是从具有未知参数的有限数量的高斯分布的混合生成的。高斯混合模型可以视为 K 均值聚类，协方差结构和样本中心的混合体。GaussianMixture 用于拟合高斯模型混合的期望最大化(EM)模型，并计算出置信椭圆体，准确率和聚类数量等关键参数。GaussianMixture 有不同的选项来约束估计的差分类的协方差：球形，对角线，并列或完全协方差，如下图所示：

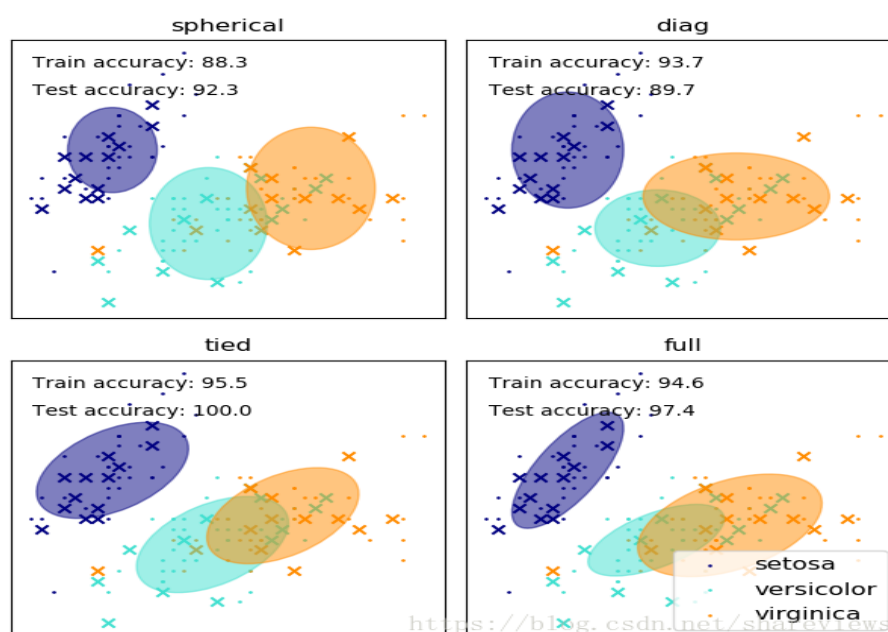


图 2-3-2 高斯混合模型(GaussianMixture)示例

2.3.3 热度计算

为了正确评价相关留言的重要程度（或者称为热度），首先要定义出一个恰当的热度指数。

1. 定义当时热度

经统计，留言点赞数目统计情况如下：

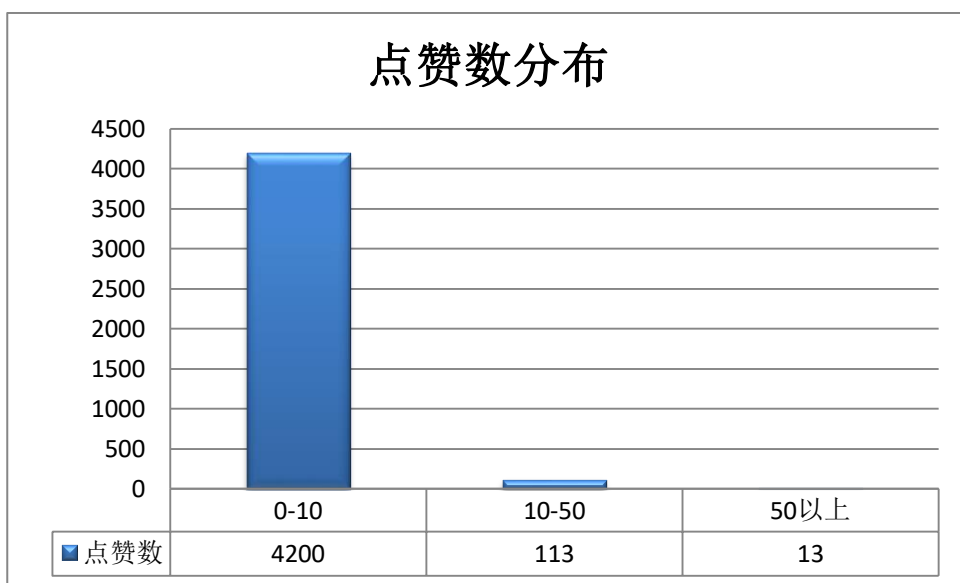


图 2-3-3 点赞数分布

在留言刚刚出现时的热度我们称之为留言的当时热度，我们对其的定义如下：

假设： N_1 =点赞数

N_2 =反对数

$$\text{Heat(当时热度)} = \begin{cases} N_1 + N_2, & N_1 \leq N_2 \\ \sum_{i=1}^{N_1-N_2} (1 * 1.5^{i-1}) + 2N_2, & 0 < N_1 - N_2 \leq 10 \\ \sum_{i=1}^{10} (1 * 1.5^{i-1}) + N_1 + N_2, & N_1 - N_2 > 10 \end{cases}$$

将 N_1 看作自变量，热度变化图如下：

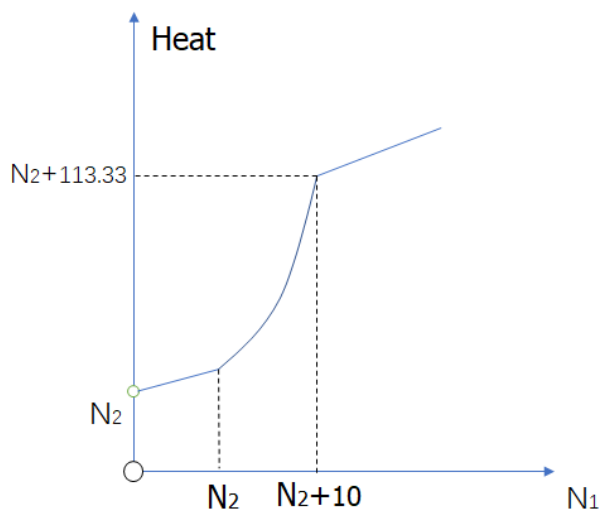


图 2-3-4 热度变化图

下面对这一定义作出数学解释：

- 1) 点赞数和反对数是反应热度的核心。如果仅有留言而没有点赞或反对，即认为没有得到相应或反馈， N_1 和 N_2 均为零，属于无效留言，热度为零。

- 2) 热度评价的本质或最终目的是为政府得到群众最为关心，亟待解决的民生或其他各个方面的问题，同时还要考虑到留言的参考价值，即最基础的真实性。
- 3) 当 $N_1 - N_2 > 10$ 时，表示反映的问题具有一定真实性，且值越大真实性越高。在增长初期，每个点赞的热度基础得分为1分，考虑到热度随点赞数目的非线性增长，从第二个点赞开始权重增大1.5倍，这样可以尽量拉开点赞数在0到10之间的留言的热度的差异。同时，当 N_1 继续增大时，为了防止指数爆炸，将热度增长模型回归到普通的线性模型上来。
- 4) 当 $N_1 \leq N_2$ 时，表示该留言的真实性或者客观性差强人意，但为了和“冷留言”相区分开，直接用点赞数和反对数之和表示出的“人气数”来代表热度。
- 5) 当 $0 < N_1 - N_2 \leq 10$ 时，表示 N_1 本身较小或者 N_1 与 N_2 均较大，公式中“ $+2N_2$ ”为调整系数，保证整个函数连续。若是第一种情况，热度随点赞数目的非线性增长，可以拉开点赞数在0到10之间的留言的热度的差距。若是第二种情况，表明该留言仍然具有一定的争议性，但此时依然保证了它的热度。
- 6) 考虑到特殊情况：即同一用户可能存在对同一问题多次留言以引起注意的行为，按照现有模型，我们并不关注他的留言次数，而关注于有没有得到其他网友的反馈，即点赞或反对。因此，现有模型不会受到此类用户留言的影响。

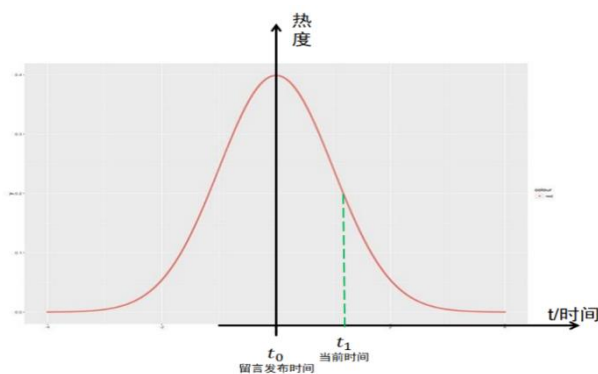
2. 定义当前热度

上面定义出了当时热度，意为留言当时其所代表的重要程度，但考虑到随着时间的推移，此留言所反映的问题会逐渐淡出公众视野，因此其热度也必然会随着时间的推移而减少，其衰减速度应为先小后大再趋于缓缓下降。具体表达为

当前热度=当时热度*高斯时间衰减

即：Heat(t)=Heat(0)*a(t)

所谓高斯时间衰减函数，即为二维空间上的正态分布函数，如下所示：



传播时间： $\Delta t = t_1 - t_0$

图 2-3-5 正态分布函数（时间衰减函数）

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

特别地，在当前模型下 $\mu = 0$ ，综上，得到最终的高斯时间衰减函数：

$$\text{Heat}(t) = \sum_{\text{所有同类留言}} \left[\text{Heat}(0) * \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{t^2}{2\sigma^2}\right) \right]$$

注：Heat(0)表示当时热度。
经调试，高斯衰减速度如下：

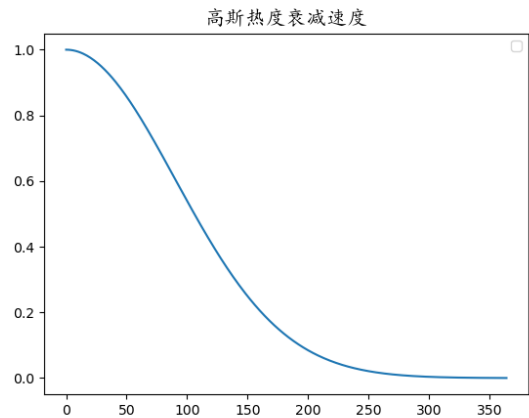


图 2-3-6 高斯衰减速度

注：20 天后大约衰减为当时热度的 0.97，360 天后衰减为 0.0003。

如此定义，主要基于以下几点考虑：

1. 热度随时间的衰减并不是简单的线性模型。根据网络舆情的发展规律，时间相近的留言热度相对接近，因为该留言中提到的事件刚刚开始发酵，大众的注意力开始慢慢聚焦到该事件上。经过一段时间的冷却之后，大众的注意力开始转移，热度下降加快。当留言淡出人们的视野较长一段时间之后，“关注量”下降减缓，热度下降也随之减缓。
2. 考虑高斯衰减：当 Δt 很小时（1 到 10 天），热度衰减较小；当 Δt 很大时，“冷留言”的热度会比较接近，不至于相差太大。当 Δt 处于中间区间时，可以将不同类型留言的热度较好地区分开。另外，通过调整方差 σ ，就可以得到一个比较理想的衰减速度。

2.3.4 挖掘地点/人群，并形成问题描述

通过上述步骤，我们已经获取到了“表 2-热点问题留言明细表”，若要制作“表 1-热点问题表”，需要得到地点/人群，以及问题描述。

通过类似于数据预处理中挖掘“ATT”定中关系的办法，选用组成定中关系的单词，再配合语义分析作为补充，在直接提取遇到困难时，提取出粗略地点。

直接使用聚类结果中同一类问题中热度指数最高的那条留言的主题，进行简化描述，得到最终的问题描述。我们默认热度最高的这一条留言对问题的描述最为清晰客观。

2.4 聚类结果展示及分析

由于聚类步骤十分重要，我们先对聚类结果进行分析，最后在优化处理中得到最终结果，包括“热点问题表”和“热点问题留言明细表”。

鉴于留言详情，篇幅过长，论文中仅展示前十五条留言。

1) 相似度比较模型聚类

在本题中,进行聚类的直接对象是数据预处理得到的留言主题中的定中关系内容

问题 ID	留言编号	留言用户	留言主题
1	236798	A00039089	A5 区劳动东路魅力之城小区油烟扰民
1	246598	A00054842	A5 区劳动东路魅力之城小区临街门面烧烤夜宵摊
1	268914	A0006238	A5 区劳动东路魅力之城小区底层餐馆油烟扰民
1	272122	A909113	A5 区劳动东路魅力之城小区一楼的夜宵摊严重污染附近的空气,急需处理!
1	284147	A909113	A5 区劳动东路魅力之城小区一楼的夜宵摊严重污染附近的空气
1	360101	A324156	A5 区劳动东路魅力之城小区油烟扰民
1	360102	A1234140	A5 区劳动东路魅力之城小区底层餐馆油烟扰民
1	360103	A0012425	A5 区劳动东路魅力之城小区临街门面烧烤夜宵摊
1	360107	A0283523	A5 区劳动东路魅力之城小区一楼的夜宵摊严重污染附近的空气
1	360108	A0283523	A5 区劳动东路魅力之城小区一楼的夜宵摊严重污染附近的空气,急需处理!
2	203187	A00024716	咨询 A9 市高铁站选址的问题
2	213099	A00038372	请彻查 A9 市北盛镇卓然村的账目!
2	223301	A00014535	A1 区望龙路南延线会不会跨过 A9 市河
2	224894	A00098248	A 市四方坪商贸城 A7 到 A9 栋卖小菜商家每天大早就试用高音喇叭卖菜
2	225342	A0008076	请落实潜江至韶关输气管道 A9 市段农田临时租赁后的恢复工作

表 2-4-1 相似度比较聚类结果

最终得到 2121 类留言,即最大“问题 ID”为 2121。

2) K-means 模型聚类

问题 ID	留言编号	留言用户	留言主题
1	188895	A00063289	票牛 A 市分公司不肯退我草莓音乐节的票钱怎么办
1	189856	A00073717)
1	197206	A00059278	举报 A 市振业城三期 A 组团房屋开裂、电表外置在过道、绿化不达标等问题
1	198393	A00046173	A 市 X115 等待时间太久了
1	199416	A00029157	请速为西地省未管所家属楼的供电安装保护接地线
1	200667	A00079480	请问 A 市为什么要把和包支付作为任务而不让市场正当竞争?
1	202604	A00051608	A 市白云路成断头路两年了
1	202744	A00078937	咨询 A 市电工证一事
1	205649	A0003502	望多增加几趟 A 市 260 路车

1	206749	A00031455	投诉 A 市捞刀河刀剪厂重复收取办证手续费
1	207161	A00026895	我觉得 A 市橘子洲的烟火蛮好的啊，请继续保持！
1	207432	A00042955	A 市燃星奥体健身房非法诈骗集资跑路
1	207702	A00060082	三问 A 市音乐厅
1	208026	A00049470	为什么要取消田东心站早上 8 点到 A 市的城铁？
1	208336	A00042963	请问曾经有网上信访经历者能不能入党？

表 2-4-2 K-means 聚类结果

3) GMM 模型聚类

问题 ID	留言编号	留言用户	留言主题
1	188895	A00063289	票牛 A 市分公司不肯退我草莓音乐节的票钱怎么办
1	189856	A00073717)
1	197206	A00059278	举报 A 市振业城三期 A 组团房屋开裂、电表外置在过道、绿化不达标等问题
1	198393	A00046173	A 市 X115 等待时间太久了
1	199416	A00029157	请速为西地省未管所家属楼的供电安装保护接地线
1	200667	A00079480	请问 A 市为什么要把和包支付作为任务而不让市场正当竞争？
1	202604	A00051608	A 市白云路成断头路两年了
1	202744	A00078937	咨询 A 市电工证一事
1	205649	A0003502	望多增加几趟 A 市 260 路车
1	206749	A00031455	投诉 A 市捞刀河刀剪厂重复收取办证手续费
1	207161	A00026895	我觉得 A 市橘子洲的烟火蛮好的啊，请继续保持！
1	207432	A00042955	A 市燃星奥体健身房非法诈骗集资跑路
1	207702	A00060082	三问 A 市音乐厅
1	208026	A00049470	为什么要取消田东心站早上 8 点到 A 市的城铁？
1	208336	A00042963	请问曾经有网上信访经历者能不能入党？

表 2-4-3 GMM 聚类结果

根据以上三个模型的运行结果中，可以得出以下结论：

- 1) 应用提取定中关系加 TF-IDF 的方式，保证了算法对人群或地点在词频上的敏感，可以尽最大可能将按照“反映同一人群或地点”的聚类标准实现聚合同类留言问题。
- 2) 在实际的程序运行过程中，我们发现最简单原始的相似度比较算法运算量并不会太大，相比之下，K-means 和 GMM 的运行更加迟钝。由于特征提取结束后得到

的词向量达到 5000 维，并且 GMM 需要 100 次的迭代，给程序运行带来了巨大的计算量。因此，在 GMM 运行前使用 PCA 降维，但 PCA 在达到降维的目的同时也带来了降低聚类精度的副作用。

- 3) 相似度比较模型聚类效果相对理想稳定，但用来衡量聚类的相似度阈值难以准确界定，这是此算法最大的缺点，优点是无需确定聚类数目。
- 4) K-means 和 GMM 的聚类操作可以看作是映射到几何空间上的聚类，这两种算法的最大缺陷在于当向量维数过大，并且点在多维空间上的分布过于密集，不同类之间相互交织，而又没有固定的协方差矩阵，给聚类带来了巨大困难。对比发现，K-means 的效果并不会优于 GMM，因此，优化处理中我们只考虑使用相似度比较模型和 GMM。

2.5 优化处理

2.5.1 使用峰均比优化相似度比较模型

在之前的模型中，将新留言归类到现有分类中的标准是在现有分类中存在一个相似度高于阈值即可。而这种方法判别过于鲁莽，因此我们引入峰均比的指标来对新留言就行归类，通过多次调试峰均比得到如下结果：

热点问题留言明细表：

问题 ID	留言编号	留言用户	留言主题	留言时间	留言详情	点赞数	反对数
1	208636	A00077171	A 市 A5 区汇金路五矿万境 K9 县存在一系列问题	2019/8/19	我是 A 市 A5 区汇金路五矿万境 K9 县 24 栋的一名业主.....	2097	0
1	241460	A00012655	A 市 A1 区华天楚水源足浴涉黄经营按摩服务	2019/10/26		0	4
1	276613	A00094265	对 A 市 A4 区委员会办公室回复不属实的质疑	2019/7/23	原帖：A4 区长房雍景湾项目.....	5	0
1	269045	A00094197	A 市 A1 区北路由捞刀河大桥南到金霞苑红灯太	2019/12/3	A1 区北路由捞刀河大.....	0	0

			久了				
1	220711	A00031682	请书记关注 A 市 A4 区 58 车贷案	2019/2/ 21	尊敬的胡书记: 您好!A4 区 p2p 公司 58 车 贷,	821	0
1	238226	A00031741	A 市 139 区 间线怎么还 在停运	2019/11 /18	因 A 市 139 公交线路突 然停运,	0	0
1	242543	A00099220	A 市区发小 卡片行为盼 查	2019/11 /14	A 市城区红绿灯、或者 各个停车场经常有人 发小卡片,	0	0
1	227451	A00025180	A 市 A1 区万 国城 3 期有 业主未经同 意住改商	2019/10 /11	A1 区万国城 3 期 7-8-9 栋 1 层住户未经楼上业 主同意	0	0
1	202514	A00056808	A 市 A6 区坡 地铁站入口 处疑似有人 发涉黄卡片	2019/10 /2	10 月 2 日下午 2: 30 左 右, 本人通过 A6 区坡 地铁站搭	0	0

表 2-5-1 留言明细表

热点问题表:

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	468.0317	2019/1/1 至 2019/9/6	A 市 A5 区汇金 路五矿万境 K9 县	县存在一系列问题
2	2	427.6953	2019/08/15 至 2020/1/7	A 市	对禁摩限令处罚解
3	3	222.6849	2019/8/23 至 2019/9/6	A4 区绿地海外 滩小区	小区距赣高铁最近有 30 米到,
4	4	188.626	2019/10/6 至	A7 县星沙派出	请民警执法

			2019/7/10	所民警	
5	5	171.6322	2019/1/17 至 2019/6/5	A4 区教育局	请教育局落实发放原 A 市七中 01 年后退休教师单位奖

表 2-5-2 热点问题表

2.5.2 使用 SVD 奇异值分解获得聚类数目优化 GMM

奇异值分解 (Singular Value Decomposition, SVD) 是一种重要的矩阵分解 (Matrix Decomposition) 方法, 可以看做对称方正在任意矩阵上的一种推广, 该方法在机器学习的中占有重要地位。SVD 具体理论如下:

设有 A 是一个 $m \times n$ 的实矩阵, 则存在一个分解使得:

$$A_{m \times n} = U_{m \times m} \Sigma_{m \times n} V_{n \times n}^T$$

其中 U 和 V 都是正交矩阵。

Σ 是一个非负实对角矩阵, U 和 V 的列分别叫做 A 的左奇异向量和右奇异向量, Σ 的对角线上的值叫做 A 的奇异值。关于这三个矩阵的求解如下:

- 1) U 的列由 $A A^T$ 的特征向量构成, 且特征向量为单位列向量。
- 2) V 的列由 $A^T A$ 的特征向量构成, 且特征向量为单位列向量。
- 3) Σ 的对角元素来源于 $A A^T$ 或 $A^T A$ 的特征值的平方根, 并且是按从大到小的顺序排列的。值越大可以理解为越重要。

根据“3)”中的结论, 通过设置奇异值的门限值, 即可获得最佳聚类数目。尔后在使用相似度比较模型、GMM 模型将聚类数目设置为通过 SVD 获得的最佳数目。

(经多次调试, 我们将奇异值的阈值设置为 0.01, 此时得到最佳聚类数目为 3807 类)

优化后结果

热点问题留言明细表:

问题 ID	留言编号	留言用户	留言主题	留言时间	留言详情	点赞数	反对数
1	208636	A00077171	A 市 A5 区汇金路五矿万境 K9 县存在一系列问题	2019/8/19	我是 A 市 A5 区汇金路五矿万境 K9 县 24 栋的一名业主,	2097	0
2	263672	A00041448	A4 区绿地海	2019/9/		669	0

			外滩小区距长赣高铁最近只有 30 米不到，合理吗？	5	您好，近日看到了渝长厦高铁最新的红线征地范围以及走向经过，.....		
3	192702	A000179	请敦促 A4 区教育局尽快落实发放原 A 市七中 01 年后退休教师的文明单位奖！	2019/10/30	尊敬的卢局长：您好！我们是原 A 市七中 01 年后退休的教师	17	0
3	205217	A00040562	请 A4 区教育局尽快落实发放原 A 市七中 01 年后退休教师的文明单位奖！	2019/10/29	您好！我们是原 A 市七中 01 年后退休的教师	33	0
4	208285	A909205	投诉小区附近搅拌站噪音扰民	2019/12/15	尊敬的领导，我是 A 市暮云街道丽发新城的一名业主，.....	24	0
4	255008	A909208	投诉小区附近搅拌站噪音扰民	2019/11/18	暮云街道丽发新城边上在建大型搅拌站，.....	0	0
4	261072	A909207	投诉小区附近搅拌站噪音扰民	2019/11/23	投诉 A 市暮云街道丽发新城附近大型搅拌站水泥厂噪音严重扰	9	2
4	266665	A00096279	投诉小区附近搅拌站噪音扰民	2019/12/04	开发商把特大型搅拌站，水泥厂从绿心范围内搬迁	0	0
5	285552	A00072434	建议 A7 县将猎豹汽车所出让工业用地转商业用地规划建设大型购物中心	2019/12/26	前些日子，我泉塘有网友建议在恒天九五那里收回用地改	22	1

表 2-5-3 优化后留言明细

热点问题表:

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	455.1675	2019/8/19	A 市 A5 区汇金路五矿万境 K9 县	县存在一系列问题
2	2	221.4082	2019/9/5	A4 区绿地海外滩小区	小区距赣高铁最近有 30 米到,
3	3	170.5479	2019/10/29 至 2019/10/30	A4 区教育局	请教育局落实发放原 A 市七中 01 年后退休教师单位奖
4	4	166.1101	2019/11/18 至 2019/12/15	小区附近搅拌站噪音	投诉噪音扰民
5	5	128.48	2019/12/26	A7 县	建议将猎豹汽车出让工业用地转商业用地规划建设大型购物中心

表 2-5-4 优化后热点问题表

第二问完整数据见附件。

第三章 留言回复评价模型

3.1 word2vec-CBOW 模型简介

作为当下普遍使用且流行的自然语言处理工具，本文作以简单介绍。

Word2Vec 是基于神经网络算法用来产生词向量的相关模型，训练完成之后，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。Word2Vec 采用 CBOW 和 Skip-Gram 来建立神经网络词嵌入。CBOW 是已知当前词的上下文，预测当前词。而 Skip-Gram 相反，是在已知当前词，预测当前词的上下文。

CBOW (Continuous Bag-of-Words Model) 和 Skip-gram (Continuous Skip-gram Model)，是 word2vec 的两种训练模式。

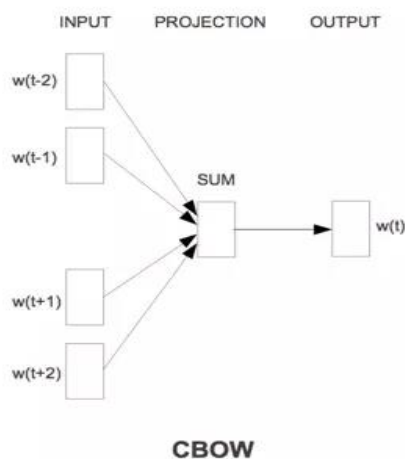


图 3-1-1 CBOW 框架

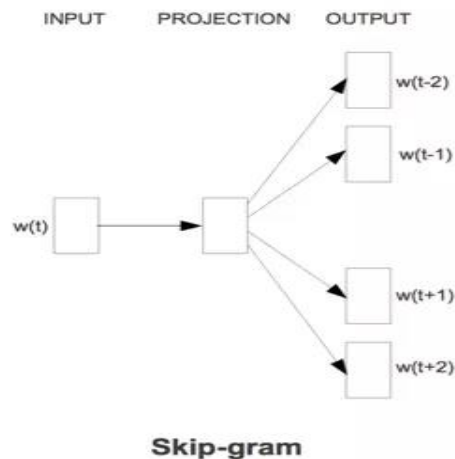


图 3-1-2 Skip-gram 框架

通过此模型,可直接得到带有语义特征的词向量,实现文本数据到词向量的转换。

后续内容中在求单词与单词之间的相似度时,直接利用映射得到的词向量求余弦相似度作为最小语义单元(单词)之间的相似度得分。

本文采用了网络上开源的 **word2vec-CBOW** 模型文件,此模型文件的训练语料库参数如下:

- 搜狐新闻 400w+条, 12G+
- 百度百科 600w+条, 20G+
- 小说: 90G+

最终训练得到的模型文件大小 1.5G, 因此没有将其加入附件。

3.2 数据预处理

3.2.1 数据分析

阅读附件四后,我们发现样本有以下几个特点:

- i) 留言详情情况较为复杂,群众留言十分多样化,“干扰性”语言,即冗余信息随机分布。理想的留言应当是简明扼要,在“留言详情”中阐述出有关于“留言主题”的 n 个方面的具体问题。而大多数群众会自我“创作”出表达自己情感色彩、对于此问题的看法或是对事件的无关描述上,举例如下:

留言主题	留言详情
c1 区国土和拆迁事务所拆迁不合理	c1 区万楼 c1 区 国土和拆迁事务所不合理拆迁 百姓苦,苦于苦中苦,难于难中难, 欺负百姓,欺负残疾人,不算好政府。我不讲了,我头痛了,你们的无能,显示了百姓的苦难。

A3 区教师村小区盼望早日安装电梯	尊敬的胡书记： 您好！过去在小区买房是为了自己，买的便宜的 7 楼，现在接了 80 多岁的老母亲来住，上楼下楼十分不方便，……现在又要 恳求党和政府领导，想群众之所想，急群众之所急 ，解决新的装电梯问题。期待回复。 谢谢尊敬的胡书记！
--------------------------	--

图 3-2-1 留言无关描述示例

这些因素对求解相关性、完整性、可解释性均十分不利。例如在求解完整性时，需要得到能够详实反应问题方面数的 n 个“子问题”，而加粗部分语句会严重影响计算问题方面数的准确性。

ii) 答复意见

留言主题	答复详情
请问 B 市水竹湖公园什么时候开工？	网友你好！ 据有关单位反馈，水竹湖公园 2019 年暂无建设计划， 感谢对我们工作的关心！ 2019 年 1 月 21 日网友您好：您反映的问题已收悉，已交相关部门调查核处，如有相关情况将及时反馈，谢谢！ 2018 年 1 月 21 日

图 3-2-2 答复意见无关信息示例

答复意见的情况相对于群众的留言更加规范一些，但依然存在干扰性语句，即它的句子本身与留言问题无关，暂且将“**已交相关部门调查核处**”这一类语言称为垃圾信息。

iii) 留言主题相对理想，属于留言的“核心”，可以大胆充分使用。

3.2.2 数据预处理

在大量浏览和研究附件四中的样本后，鉴于数据分析中所提到的因素，做如下处理：

- 1) 将“尊敬、您好、你好、谢谢、希望”一类尊称和礼貌性词语加入停用词列表，尔后分词。
- 2) 为了更为高效地评估完整性和可解释性，依据留言详情对留言分为四大类：

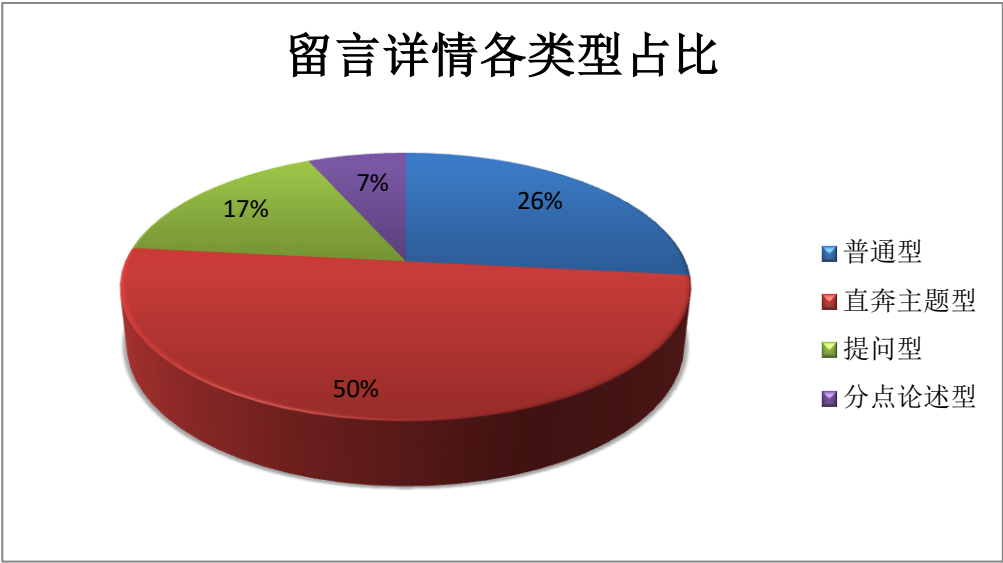


图 3-2-3 留言详情各分类占比

具体介绍如下：

留言详情类型	占比	特点
直奔主题型	50%	文本简短，与留言主题有高度重复，冗余信息少
普通型	26%	文本较长，冗余表达多，难以挖掘有用信息
提问型	17%	可通过“？”的位置直接准确定位留言主题所反映的子问题
分点论述型	7%	可通过“1. 2. 3.”的位置直接准确定位留言主题所反映的子问题，并且获得子问题的数目和内容

表 3-2-4 留言详情各分类具体介绍

举例如下：

留言详情类型	留言主题	举例
直奔主题型	请问 B 市西环线两侧辅道何时能全线贯通	请问西环线两侧辅道何时能全线贯通，城市不断发展，快速路两侧的辅路十分重要。
普通型	A3 区教师村小区盼望早日安装电梯	尊敬的胡书记：您好！过去在小区买房是为了自己，买的便宜的 7 楼，现在接了 80 多岁的老母亲来住，..... 期待 A 市住建、财政、国土、安全等部门尽早研究出台为老社区惠民装电梯的规范性文 现在又要恳求党和政府领导，想群众之所想，急群众之所急，解决新的装电梯问题。期待回复。谢谢尊敬的

		胡书记！
提问型	残疾人咨询购买电动轮椅事宜	我是一名一级肢体残疾，靠低保度日，去年 10 月在残联买了台电动轮椅代步，想咨询下现在还能购买吗？残联是不是能给相应的补助呢？
分点论述型	B 市 601 小区钻石新村 20 栋楼下快乐休闲网吧扰民	601 小区钻石新村 20 栋楼下快乐休闲网吧扰民：1、私自凿开楼道墙壁开门。2、网吧空调污水直接排放到过道。3、网吧 24 小时营业，对外排放噪音高达 80 分贝严重影响居民生活和休息。

表 3-2-5 留言详情各分类举例

后文中，将对不同类型文本做针对性处理，从而达到去除冗余信息，提取关键信息的目的。

3.3 问题建模

3.3.1 特征提取

使用 word2vec 模型自带函数将单词映射为词向量。

3.3.2 去除冗余信息，获取子问题

1) 模型准备：“贪吃蛇”算法，介绍如下：

针对“普通型”留言，为找到其中的子问题，根据大众的语言习惯：一个完整的句意总是通过紧密相邻的一个或几个句子共同表达，而不可能是第一句话和第三句话说同一件事情，而第二句话在说另外一件事情。（以“。”、“！”、“？”、“...”作为句子结束的标志，编程实现中使用正则表达式判别）

因此，具体操作方法如下：从第一句话开始，使用前后两句话做相似度判别，若相似度低于阈值，则将后一句话判为和上一句同意，否则为新建句意。另外，进行判断上下两句话的相关性操作的前提条件是这句话和留言主题相似度高于另一个阈值。

可将整个算法类比于贪吃蛇进食，即遇到“食物”后，通过嗅觉（与主题的相似度）判断吃不吃它，如果吃掉，贪吃蛇长度增长到第一个句子的长度；如果不吃，长度即为 0。遇到下一条句子时，除过判断吃不吃外，还要判断和上一个已经吃掉的食物关系。由于整个过程中表达同一句意的段落长度不断变化，故命名为“贪吃蛇”算法。

2) 针对性处理

根据数据分析的结果，鉴于留言详情中成规模出现的表达情感倾向性的语言，使用和留言主题的平均词向量做相似度筛选的办法将其过滤掉。

并针对预处理中得到的四大分类，为获取留言详情中的子问题，处理如下：

留言详情类型	处理办法
直奔主题型	用留言主题替换留言详情，得到一个子问题

普通型	运用贪吃蛇算法找出子问题
提问型	可通过“？”的位置直接准确定位留言详情所反映的子问题，为防止问句只是在发泄不满情绪，增加情感分析的操作过滤掉不反应客观问题的问句
分点论述型	可通过“1. 2. 3.”等表示分点的列表符号的位置直接准确定位留言主题所反映的子问题，同时还可以获得子问题的数目和内容

表 3-3-1 处理办法分类

处理效果如下：

留言详情类型	留言详情	提取部分
直奔主题型	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。	[['毕业', '人才', '新政', '落户', 'A 市', '买', '套', '公寓', '购买', '公寓', '享受', '万元', '购房', '补贴']]
普通型	正常工作上班时，拨打 14 县医保所信息股电话 8881528；财务股电话 8885867 十余次，且每次都间隔了一段时间再打的，一直无人接听。请相关部门重视，不然就不要公布联系电话，公布了就应该受理接听，让我们小老板信不知道找谁啊，这不是失信于民吗。	[['工作', '上班时间', '拨打', '14', '医保', '信息', '电话', '8881528', '财务股', '电话', '8885867', '每次', '接听', '部门', '联系电话', '受理', '接听', '老板', '信', '找']]
提问型	本人户籍 E 市，在杭州工作并在杭州购买了一辆雅阁混动小车，希望能够回 C 市上牌，目前开具购车发票，上了临时牌照、未缴纳购置税，请问回来上牌需要准备哪些材料，预计多久能完成上牌流程？车牌制作如需较长时间，是否可以邮寄至杭州？望予以解答为盼，谢谢！	[['户籍', 'E', '杭州', '工作', '杭州', '购买', '一辆', '雅阁', '小车', '回', 'C', '市上', '牌', '开具', '购车', '发票', '临时', '牌照', '缴纳', '购置税', '请问', '回来', '牌', '材料', '预计', '牌', '流程', '车牌'], ['制作', '如需', '邮寄', '杭州']]
分点论述型	601 小区钻石新村 20 栋楼下快乐休闲网吧扰民：1、私自	[['私自', '楼道', '墙壁', '开门', '网吧', '空调', '污水'], ['网吧', '空调', '污水'], ['网吧', '24', '小时', '营']]

	凿开楼道墙壁开门。2、网吧空调污水直接排放到过道。3、网吧 24 小时营业，对外排放噪音高达 80 分贝严重影响居民生活和休息。	业', '噪音', '80', '居民', '生活', '休息']]
--	--	------------------------------------

表 3-3-2 处理效果

3.3.3 相关性评估

- 1) 鉴于留言详情中杂质信息较多，情况相对复杂，为准确衡量答复意见与留言的相关性，以留言主题作为参考分别跟详情和答复中的分词做相似度运算，将相似度结果作为详情和答复中单词的权重。求单词间的相似度均采用 word2vec 模型的自带求相似度函数，即可判断语义上的相似程度。
求权重框架如下：

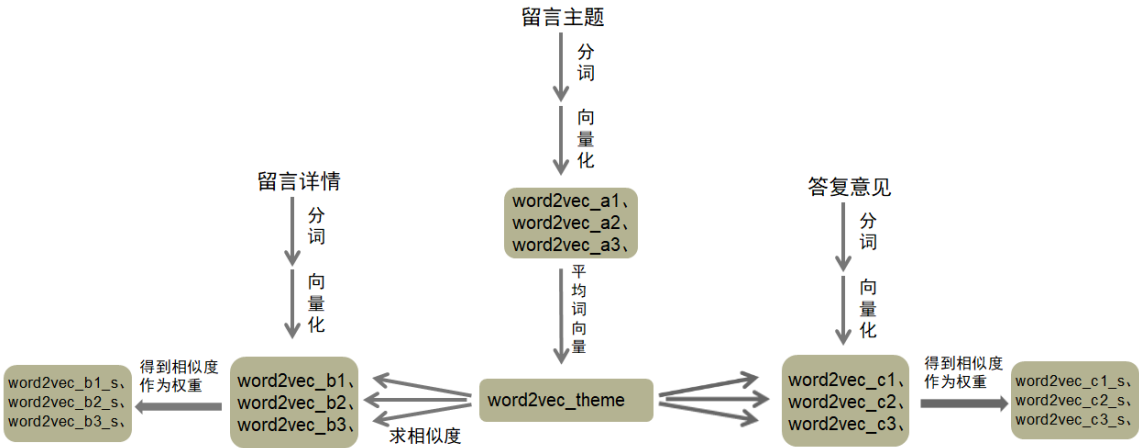


图 3-3-3 求权重框架

- 2) 利用详情和答复中的单词的权重求最终相似度
详情和答复中单词权重已知，接下来求出答复意见和详情的相似度，再累乘以它们自身的权重即可。求最终相关性框架如下：

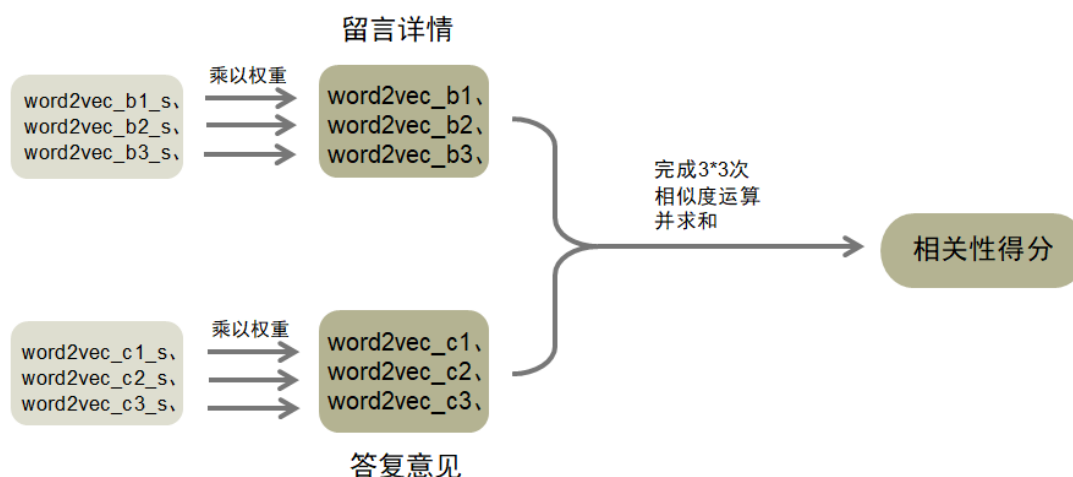


图 3-3-4 求最终相关性框架

3.3.4 完整性评估

模型准备：正交匹配追踪算法介绍及用法

1、算法思想

在正交匹配追踪 OMP 中，残差是总与已经选择过的原子正交的。这意味着一个原子不会被选择两次，结果会在有限的几步收敛。

2、算法流程

1. 用 x 表示初始样本，初始化残差 $e_0 = x$;
2. 在剩余的样本中选择与 e_0 内积绝对值最大的原子，表示为 ϕ_1 ;
3. 将选择的原子作为列组成矩阵 Φ_t ，定义 Φ_t 列空间的正交投影算子为：

$$P = \Phi_t (\Phi_t^T \Phi_t)^{-1} \Phi_t^T$$

通过从 e_0 减去其在 Φ_t 所张成空间上的正交投影得到残差 e_1 :

$$e_1 = e_0 - P e_0 = (I - P) e_0$$

4. 对残差迭代执行 (2)、(3) 步

$$e_{m+1} = e_m - P e_m = (I - P) e_m$$

其中 I 为单位阵。需要注意的是在迭代过程中 Φ_t 为所有被选择过的原子组成的矩阵，因此每次都是不同的，所以由它生成的正交投影算子矩阵 P 每次都是不同的。

5. 直到达到某个指定的停止准则后停止算法

OMP 减去的 $P e_m$ 是 e_m 在所有被选择过的原子组成的矩阵 Φ_t 所张成空间上的正交投影，而 MP 减去的 $P e_m$ 是 e_m 在本次被选择的原子 ϕ_m 所张成空间上的正交投影。

3、算法运用

由于之前的步骤“针对性处理”已经准确挖掘出了需要进行完整性评估的素材：子问题内容与问题方面数。因此通过使用正交匹配追踪，即可判断每一个子问题是否在答复意见中得到了回答或反馈，即为响应。

通过使用 OMP，可以极大地简化我们判断一个子问题是否在答复意见中得到“响应”的过程。

1) 判断某一个子问题是否在答复意见中得到了响应。

以此条留言为例：

留言详情类型	留言详情	提取部分
分点论述型	601 小区钻石新村 20 栋楼下快乐休闲网吧扰民：1、私自凿开楼道墙壁开门。2、网吧空调污水直接排放到过道。3、网吧 24 小时营业，对外排放噪音高达 80 分贝严重影响居民生活和休息。	[['私自', '楼道', '墙壁', '开门', '网吧', '空调', '污水'], ['网吧', '空调', '污水'], ['网吧', '24', '小时', '营业', '噪音', '80', '居民', '生活', '休息']]

表 3-3-5 留言示例

OMP 运算框架如下：

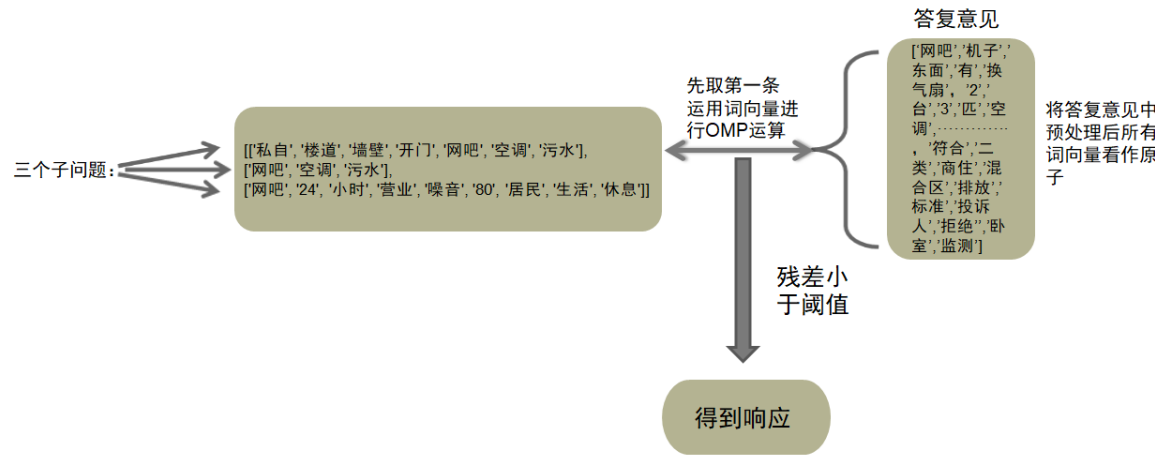


图 3-3-6 OMP 运算框架

2) 利用 omp 算法检查留言详情中的问题是否在答复意见中得到响应，即判断答复意见回复了那几个子问题。

$$\text{Score_完整性} = \frac{\text{响应数}}{\text{问题方面数}}$$

3. 3. 5 可解释性评估

广义上的可解释性指在我们需要了解或解决一件事情的时候，我们可以获得我们所需要的足够的可以理解的信息。

为了和前期在相关性和完整性评估中大量使用的相似度运算保持良好的区分度，通过建立可解释性词表，然后根据答复意见中高可解释性词的数量对其进行打分。通过大量阅读可解释性优异的答复意见，统计其中的高可解释性词汇，并给出相应赋分。

在统计高可解释性的词汇以及得分时，使用 **word2vec** 求出近似的高价值词语，并且根据相似度大小相应赋分。

3.3.6 时效性评估

为反映政府部门对群众留言热点进行回复的及时性，我们新增时效性评估。

鉴于附件四没有点赞数等数据，无法得到热度评估，因此，只考虑该条留言在留言时间到答复时间之间的时间段上的衰减。

假设留言后立即回复得到满分 100 分，则衰减模型如下：

$$\text{Score}_{\text{时效性}} = 100 * 1.036^{-t}$$

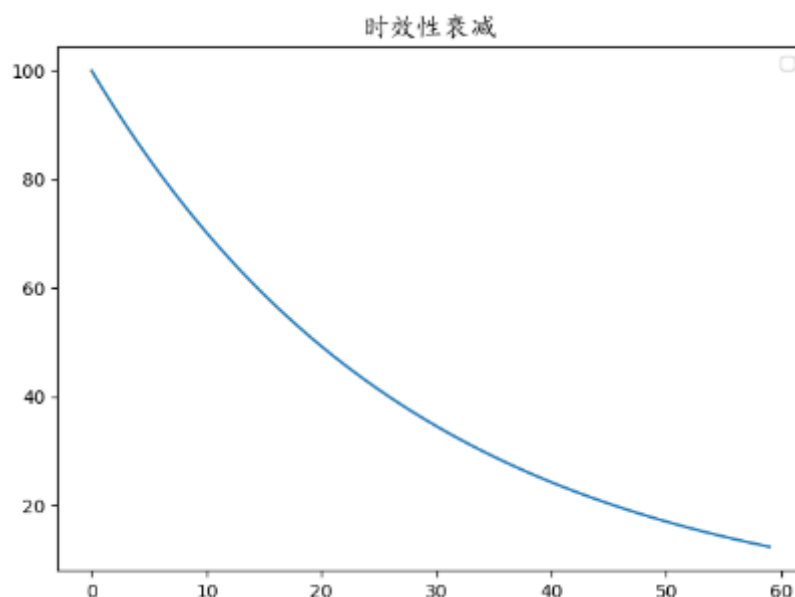


图 3-3-7 时效性衰减模型

3.4 形成综合评价方案

最终评价方案中我们对四个评价指标均进行了归一化，转化为百分制，使得各个得分均匀分布到 0 到 100 的区间内。见附件中的答案。

第四章 总结与展望

4.1 总结

在前两问中，本文通过在基于传统机器学习算法的情况下最大限度实现准确地留言分类和聚类。在完成这两个最基础操作之后，充分从群众的现实心理和情况出发，定义热度评价指标，并创新型地选择高斯函数作为热度衰减模型，有良好的区分度，尽最大努力将群众最为关切的问题反馈给政府。同时也为热度评价提供了新思路。

在对政府部门给出的答复意见进行评价时，我们以 **word2vec** 作为最基本工具，以映射到的词向量为特征进行了大量灵活丰富的操作，包括运用“贪吃蛇”算法进行最小句意单元地划分，过滤留言详情中的无效信息等等。另外，我们通过时效性指标来反映政府部门对群众留言回复的速度。

整体而言，我们基本实现了通过以机器学习进行分类聚类，运用 **word2vec** 判断语义上的相关性为基本手段对群众留言和政府答复意见的闭环处理，基本达到本赛题设立的基本目标。

4.2 展望

1) 本文中使用的 **Word2vec** 训练语料相对偏小，如果使用更有针对性的预料，比如使用大样本的具有优秀的相关性、完整性、可解释性的模板来训练，再使用关于民生、社会舆论进行增量训练，会取得更好效果。

2) 第一问中对于模糊词的识别上，我们设置了命中率指标进行判别，试图最精确地找出影响分类性能的模糊词。但在模糊词去除数量的设置上，没有量化地计算出分类准确率与泛化能力之间的平衡点。

机器学习的好坏往往高度依赖于样本集的质量和规模，只要保证质量和规模，理论上就可以保证传统机器学习算法优异的泛化能力，并且稳定可靠地运行。

但是我们在从对示例数据到对全部数据处理的过程中，也深刻体会到了机器学习的一些短板，譬如通过 **TF-IDF** 特征提取过程中随着文档的增加，词向量维数会发生指数型增长，同时给后续处理带来巨大运算量。

3) 通过经验学习也可获得用来判断答复意见质量好坏的词表，比如在评价可解释性时，完全可以通过从大众的语言习惯中总结出具有衡量可解释性优良程度的代表性词语，只要保证大量和得分梯度配置合理，便可以达到不错的效果。

参考文献

- [1]涂铭,刘祥,刘树春.Python 自然语言处理实战.杭州:机械工业出版社,2019.
- [2]Wanxiang Che, Zhenghua Li, Ting Liu. LTP: A Chinese Language Technology Platform. In Proceedings of the Coling 2010:Demonstrations. 2010.08, pp13-16, Beijing, China.
- [3]Benjamin Bengfort, Rebecca Bilbro, Tony Ojeda 著.陈光[译].基于 Python 的只能文本

分析[M]. 北京：中国电力出版社，2020. 1.

[3] https://blog.csdn.net/qc_29110265/article/details/90769363

[4] https://blog.csdn.net/weixin_44243926/article/details/90263582?depth_1-utm_source=distribute.pc_relevant.none-task&utm_source=distribute.pc_relevant.none-task

[5] https://blog.csdn.net/github_38486975/article/details/88960629

[6] https://blog.csdn.net/weixin_33743880/article/details/91438343

[7]Nitin Hardeniya, Jacob Perkins 等著. 林赐[译]. Python 和 NLTK 自然语言处理. 北京：人民邮电出版社，2019. 4.

[8]Dccpti Chopra 等著. 王威[译]. 精通 Python 自然语言处理. 北京：人民邮电出版社，2017. 8.

[9]Dipanjan Sarkar 著. 闫龙川等[译]. Python 文本分析. 北京：机械工业出版社，2018. 3.

[10]宗成庆，夏睿，张家俊著. 文本数据挖掘. 北京：清华大学出版社，2019.

附录

（一）实验环境：

运行平台： Windows

Python 版本: Python3.x
IDE: PyCharm

（二）答案数据：

由于文中不便于体现全部数据，三问的所有答案均保存在附件中相应文件夹下，有 excel 格式文件：