C 题

摘要

近年来，随着微信、微博、市长信箱等网络问政平台逐步成为政府了解民意、汇聚民智、凝聚民气的重要渠道，各类社情民意相关的文本数据量不断攀升，因此以往依靠人工来进行留言划分和热点整理工作量过大且效率不高。因此通过自然语言处理等技术，来建立智慧政务系统是社会治理创新发展的新趋势，对提升政府的管理水平和施政效率也具有极大的推动作用。

首先我们所给的数据，运用 tf-idf 算法和 n-gram 模型进行提取关键词和提高文本的语义连续性表示，实现数据的预处理。

针对任务一：群众留言分类，经过数据预处理后，对比数据处理前能够比较好地提取出数据中的关键信息。随后在分成测试集和数据集导入多层感知机 MLP 中进行训练得到模型。再运用 k-fold 交叉验证模型对随得到的模型进行评估取平均值得到比较好的一级标签分类模型，并且能够把附件 2 的信息比较准确地分类到所对应的一级标签中。

针对任务二：热点问题挖掘，运用 repeat bisection 算法对处理过后的数据进行聚类，从而实现对留言进行分类，然后定义热度指标，对分尅后的问题进行热度排序，挖掘出排名前五的热点问题。最后使用 HanLP　分词工具对聚类后的数据进行关键词和留言摘要的提取。

针对任务三：建立答复意见的评价模型。主要从相关性、完整性、可解释性三个方面进行评价模型的建立。对于相关性，使用 LDA 模型分别对留言内容和答复意见进行提取，再用余弦相似度算法计算两者的相似度，从而评价其相关性；对于完整性，主要通过句子成分分析，判断答复意见中的语言是否通顺，语法是否无误，再通过词库对答复意见进行礼貌用语文本匹配，查看答复的语言是否有礼恰当，最后结合这两方面，评价其完整性；对于可解释性，运用正则表达式，进行文本匹配，查看对于涉及专业问题的答复中，是否有引用相关文献，从而进行可解释性的评价。最后综合三者的评价建立答复意见评价模型。

关键词：tf-idf 算法；n-gram 模型；多层感知机(MLP)；repeat bisection 聚类；HanLP　分词工具；LDA 模型；句子成分分析；文本匹配；答复意见评价模型

Abstract

In recent years, as WeChat, weibo, mayor's mailbox and other network political platform gradually become an important channel for the government to understand public opinion, gather people's wisdom, and condense people's spirit, the amount of text data related to various social situations and public opinions keeps increasing. Therefore, the workload of manual message division and hot spot sorting is too large and the efficiency is not high. Therefore, it is a new trend of social governance innovation and development to establish a smart government system through natural language processing and other technologies, which also plays a great role in promoting the management level and efficiency of the government.

Firstly, we used tf-idf algorithm and n-gram model to extract keywords and improve the semantic continuous representation of the text, so as to realize the data preprocessing.

For task 1: the classification of public comments, the key information in the data can be extracted well after data preprocessing and before comparison data processing. Then the model is trained in MLP with test set and data set import. Then the k-fold cross-validation model is used to evaluate the resulting models and take the average value to obtain a better first-level label classification model, and the information in attachment 2 can be classified into the corresponding first-level label more accurately.

For task 2: hot issues mining, repeat bisection algorithm is used to cluster the processed data, so as to achieve the classification of the comments, and then define the heat index, the heat of the problems after the sorting, mining the top five hot issues. Finally, HanLP word segmentation tool was used to extract keywords and message abstracts from the clustering data.

For Task 3: to establish the evaluation model of response. The evaluation model is mainly established from three aspects: relevance, completeness and interpretability. For the correlation, LDA model is used to extract the message content and the reply comments respectively, and then the cosine similarity algorithm is used to calculate the similarity between the two, so as to evaluate the correlation. As for the completeness, it mainly judges whether the language in the reply is fluent and the grammar is correct through the sentence component analysis, and then conducts the polite language text matching to the reply through the thesaurus to check whether the language of the reply is polite and appropriate. Finally, it evaluates the completeness by combining the two aspects. For interpretability, regular expressions are used to perform text matching and to check whether relevant literature is cited in the replies to professional questions, so as to evaluate the interpretability.

Key words: tf-idf algorithm; N - "gramm model; Multilayer perceptron (MLP); Repeat bisection clustering; HanLP word segmentation tool; The LDA model; Sentence component analysis; Text matching; Response comment evaluation model

目录

一、问题分析

1.1 任务一问题分析

对于任务一：群众留言分类，我们将题目做以下解读：

（1）请根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型。

根据题目可以看出本体的目的是让我们参考附件 1 对附件 2 的数据进行分类。附件 1 中给出了三级指标这里主要用一级指标的数据进行分类。

（2）其中附件 1 中一级标签包含城乡建设，党务政务，国土资源，环境保护，纪检监察，交通运输，经济管理，科技与信息产业，等 14 个标签。因此我们需要把附件二的数据分类到这些标签中。

因此，我们决定采用 tf-idf 算法提取数据中的关键词再利用 n-gram 模型提高文本语义连续性，之后再用 MLP 进行训练。

1.2 任务二问题分析

对于任务二，热点问题挖掘，任务要求我们对留言进行归类，并建立热度评价模型，并形成“热点问题表.xls”，表内包含排名前 5 的热点问题，和“热点问题留言明细表.xls”，表内包含相应热点问题对应的留言信息。

首先是

观察给出的格式表一、表二、附件 3 时，可以发现留言中，相同的地点出现的问题或事件基本是类似甚至一样的，因此将留言的出现的地点作为关键词，将相同地点的留言归为一类，计算每一类问题的留言条数、点赞数、反对数进行降序排序，就可以得出排名前五的热点问题。

1.3 任务三问题分析

对于任务三，建立答复意见评价模型，对于答复意见的评价可以从以下三方面入手：

（1）相关性：指两个变量的关联程度。一般地，从散点图上可以观察到两个变量有以下三种关系之一：两变量正相关、负相关、不相关。

而对于 NLP 任务，一般是指两段文本的内容或者语义是否相关，一般我们可以采用对文本向量的相似度计算来获得句子的相关性

（2）完整性：对于 NLP 任务，分析完整性就是分析句子的结构，语法是否完整，比如说，一个完整的句子，应该具有主谓宾结构，还需要有定语，状语等修饰成分，另外，对于不同的语法环境，例如，回复意见的完整性，我们还需要分析，句子结构中是否存在敬辞——“您好”等成分，这就是句子的完整性分析

（3）可解释性：回复意见的可解释性，就是回复内容是否有据可依，这种时候，我们需要去寻找句子的特征，比如说“根据，XX 法律”等等内容，我们可以定性的分析句子的可解释性。寻找句子的特征，采用文本匹配法+正则表达式。

二、数据预处理

2.1 分词与多余词的删除

（1）由于是文本类数据，且文本都是一整段的句子，因此需要对这些数据进行分词，就是把每个句子分成一个一个词。分词前和分词后如下图所示：

A3区大道西行便道，未管所路口至加油站路段，人行道包括路灯杆，被圈西湖建筑集团燕子山安置房项目施工围墙内。每天尤其上下班期间这条路上人流车流极多，安全隐患非常大。强烈请求文明城市A市，尽快整改这个极不文明的路段。

{"A3区"，"大道"，"西行"，便道"，"，"未管所"，"路口"，"至"，"加油站"，"路段"，"，"，"人行道"，"包括"，"路灯杆"，"，"，"被圈"，"西湖"，"建筑"，"集团"，"燕子山"
"施工"，"围墙"，"内"，"。"，"每天"，"尤其"，"上下班"，"期间"，"这条"，"路上"，"人流"，"车流"，"极多"，"，"，"安全"，"隐患"，"非常"，"大"，"。"，"强烈"，"请求
"尽快"，"整改"，"这个"，"极"，"不文明"，"的"，"路段"，"。"}

（2）因为句子中有很多的语气助词，等一些与文本内容无关的词，像“这”，“至”，“期间”等等因此我们通过去停用词的方法把这些词去掉。

——
一下
一个
一些
一何
一切
一则
一则通过
一天
一定
一方面
一旦
一时
一来
一样
一次
一片
一番
一直
一致
一般
一起
一转眼
一边
一面

最后通过上述的做法，数据便可以导入下面的算法中。

2.2 文本向量表示

2.2.1 tf-idf 算法提取关键词

由于 tf-idf 算法中，tf 也就是词频，能够算出一个词在文中出现的次数。因此，一个词在文中出现的次数越多，那么这个词肯定在文中有比较大的作用。公式如下图所示：

$$tf_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

即

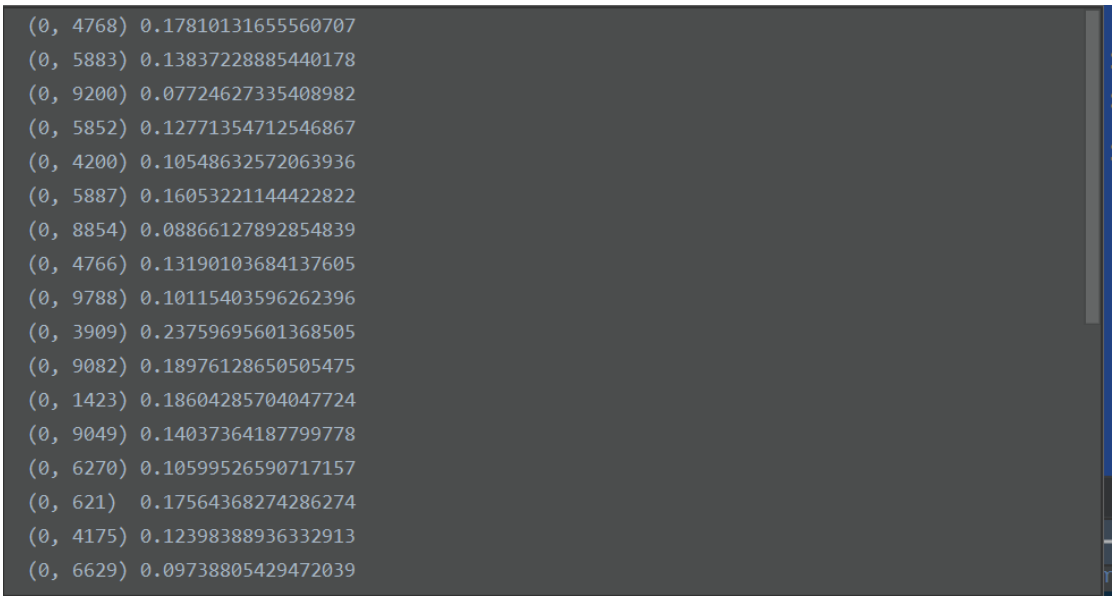
$$TF_w = \frac{\text{在某一类中词条}w\text{出现的次数}}{\text{该类中所有的词条数目}}$$

但是有时统计出来的例如“的”，“是”这样的词，出现的次数一般比较多，显然这样的词多我们的统计没有什么帮助，甚至会干扰我们的统计。所以我们需要利用 idf 对这些词进行加权，也就是在词频的基础上对每个词分配一个权重。即对一些常见词如“的”，“是”等等这些词赋予一个较低的权重，对于一些比较少见的词语赋予较高的权重，这个权重就是 idf（逆文档频率）。得到 tf 以及 idf 以后他们的乘积就是一个 tf-idf 的值。所以如果某些词的 tf-idf 的值越高，那么这个词在文本中的重要性就越大。因此，如果我们想要提取文本中的关键词，只需要 tf-idf 值排在最前面的几个就可以。这种方法在提取关键词的效率较高，而且准确性比较好。因此我们选用了 tf-idf 算法来提取。

$$idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$$

最后得到

$$TF - IDF = TF * IDF$$



if idf 算法运行结果图

2.2.2 采用 n-gram 模型提高文本语义连续性的表示

n-gram 模型简单来说，就是在得到前面词语或者句子的意思后，就能根据前面的话来推测出后面词语，即根据前(n-1)个 item 来预测第 n 个 item。由于现在许多输入法例如搜狗，讯飞，百度等等都含有 n-gram 的影子，就像有时输入完一个字后他会出现一系列的候选词来与你刚刚输入的词进行组词，一般排在比较前的都是组词频率比较高的字。这里需要通过调用 sklearn 中的 TfidfVectorizer 模块处理。

因此我们需要利用 n-gram 模型来对 tf-idf 所提取出来的关键词重新组成一个句子来提高文本语义连续性的表示。其中所利用到的原理公式如下图：

$$P(B \mid A) = \frac{P(AB)}{P(A)}$$

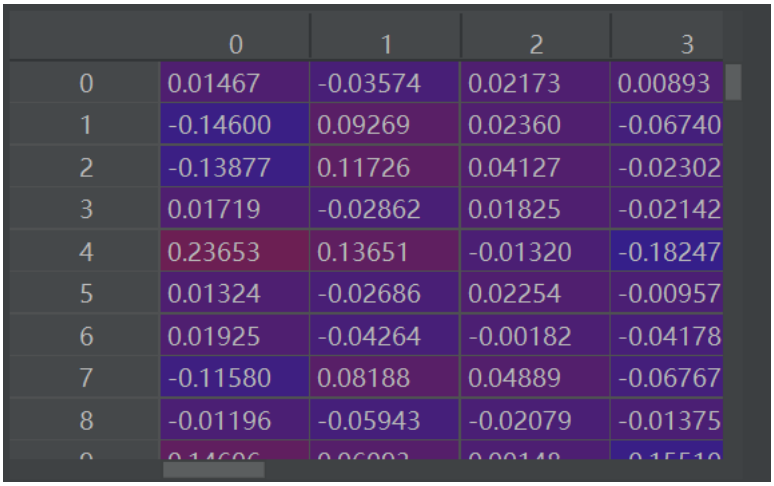
乘法公式

$$P(AB) = P(A)P(B \mid A) \quad (P(A) > 0)$$

得到

$$P(A_1A_2 \cdots A_n) = P(A_1)P(A_2 \mid A_1) \cdots P(A_n \mid A_1A_2 \cdots A_{n-1})$$
$$(P(A_1A_2 \cdots A_{n-1}) > 0)$$

即： $p(S)=p(w_1w_2\cdots w_n)=p(w_1)p(w_2 \mid w_1)\cdots p(w_n \mid w_{n-1}\cdots w_2w_1)$



	0	1	2	3
0	0.01467	-0.03574	0.02173	0.00893
1	-0.14600	0.09269	0.02360	-0.06740
2	-0.13877	0.11726	0.04127	-0.02302
3	0.01719	-0.02862	0.01825	-0.02142
4	0.23653	0.13651	-0.01320	-0.18247
5	0.01324	-0.02686	0.02254	-0.00957
6	0.01925	-0.04264	-0.00182	-0.04178
7	-0.11580	0.08188	0.04889	-0.06767
8	-0.01196	-0.05943	-0.02079	-0.01375

文本向量表示结果图

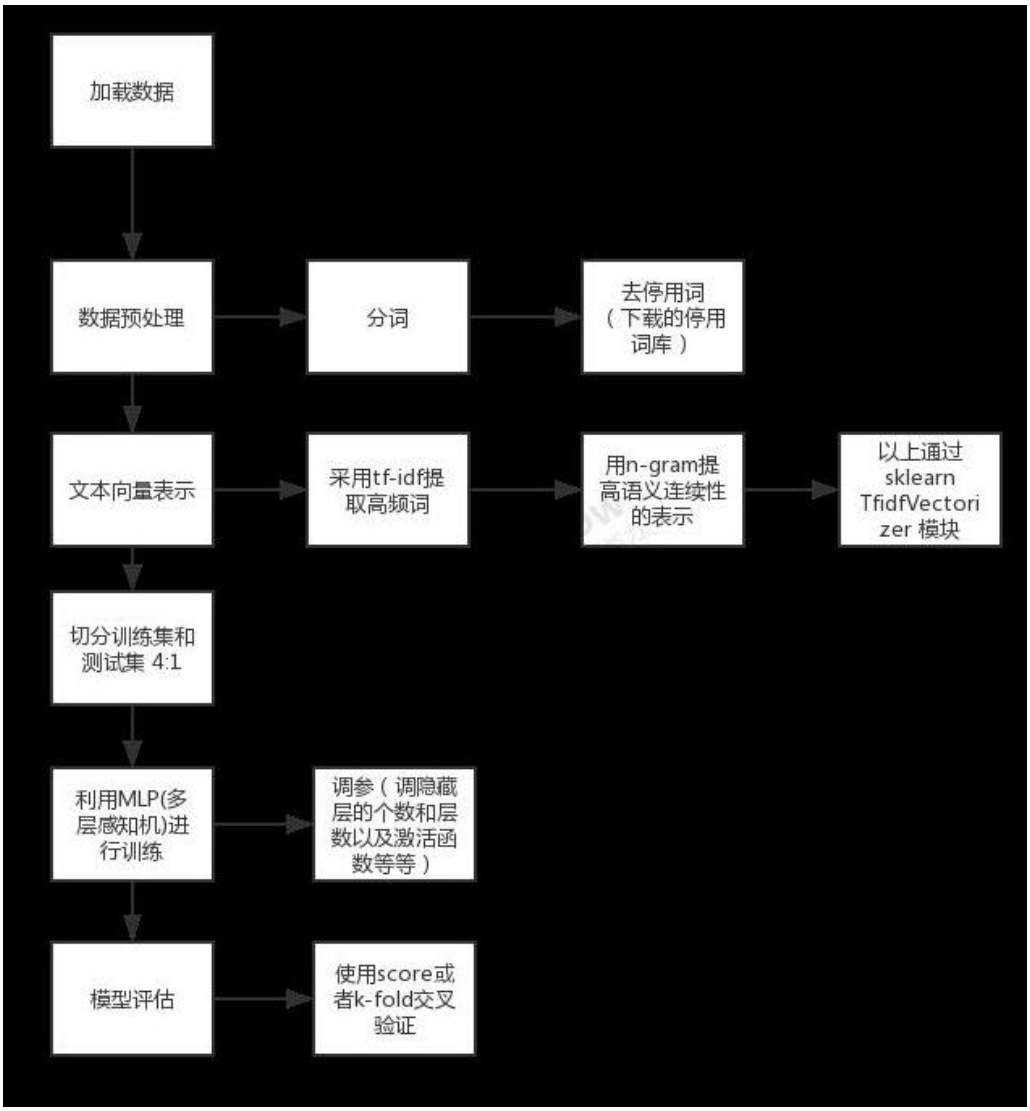
2.3 综合

对于三道题所使用的附件数据均采用上述方法进行预处理，针对问题二附件 3 的数据，在上述方法处理过后还需使用 scikit-learn 工具的类 TfidfVectorizer 进行文本特征提取。（scikit-learn 工具是一个用于数据挖掘和数据分析的简单且有效的工具，它是基于 Python 的机器学习模块，基于 BSD 开源许可证。它的基本功能主要被分为六个部分：分类(Classification)、回归(Regression)、聚类(Clustering)、数据降维(Dimensionality reduction)、模型选择(Model selection)、数据预处理(Preprocessing)。）

其他数据处理方法无添加。

三、任务一：群众留言分类

3.1 解题思路

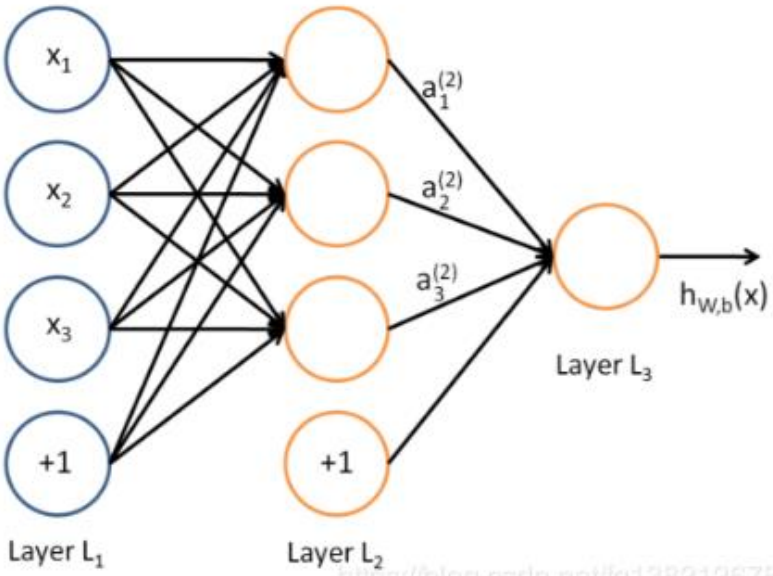


3.2 利用 MLP 对训练集进行训练

在对数据预处理，以及用文本向量表示后，我们就开始训练我们的模型。在这之前，要把数据分成训练集和测试集，在这次题目中我们把训练集和测试集划分的比例为 4:1。

把训练集导入，利用 MLP（多层感知机制）对划分出来的训练集进行训练，由于 MLP 除了输入输出层，它中间可以有多个隐层，最简单的 MLP 只含一个隐

层，即三层的结构。如下图所示。因此在这个过程中需要进行调参，即调整隐藏层的个数和层数以及各种激活函数等等。

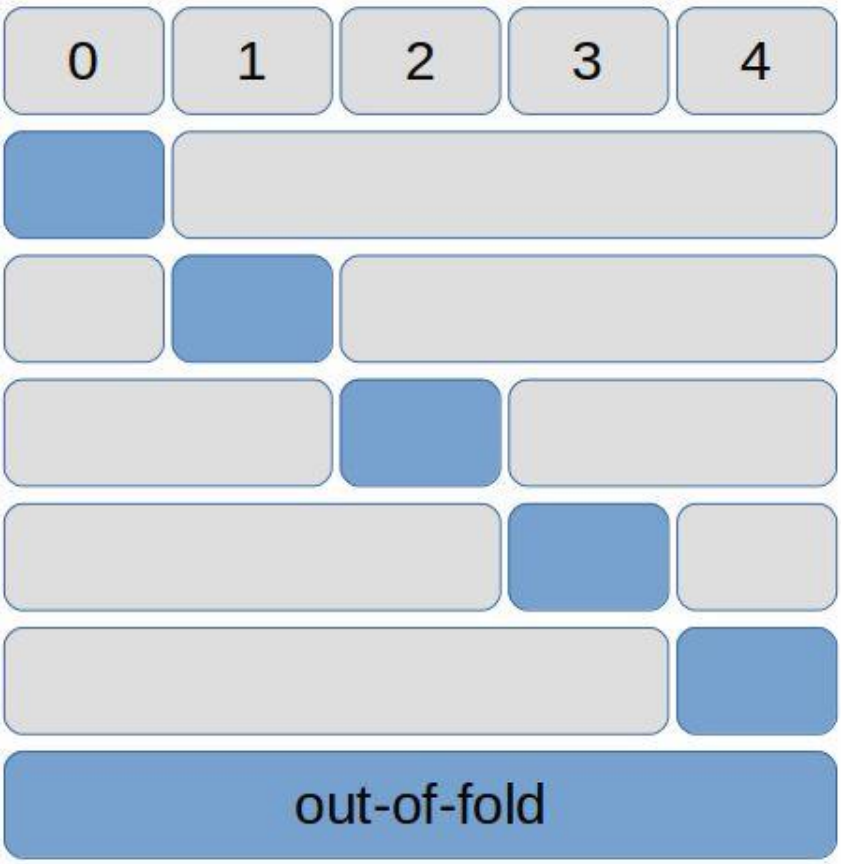


3.3 k-fold 交叉验证模型对模型进行评估

在通过 MLP 训练后得出来的模型进行评估，在这里使用 k-fold 交叉验证模型的准确率。这样就可以排除模型通常对训练数据好，但是对训练数据之外的数据拟合程度差的问题。

由于 K-fold 交叉验证模型验证数据时，取用自训练数据。k-fold 交叉验证就是把数据分成 K 组，将分成的每一组数据集都分别做一次测试集，基于的 k-1 组数据作为训练集，这样就可以得到 k 个模型，每个模型的测试集得到的结果，利用 MSE 加和平均就得到交叉验证误差。这种方法有效利用了有限的数据，并且评估结果能够尽可能接近模型在测试集上的表现。

如下列所示：把数据分成 5 组（0~4），每一次都选一组用来做一次测试集，其余做训练集。这样就得到 5 个模型的结果，再把结果进行平均就可以得到比较准确的测试结果。



3.4 结果展示

```
[[370   7   5   1   9   8   1]
 [  6 172   2   1   0   1   0]
 [ 16   0  90   0   3   8   1]
 [ 11   0   1 294   9   3   0]
 [  6   0   1   6 387   3   6]
 [ 14   4   1   3   4 190   1]
 [  3   1   1   0   8  10 174]]
```

混淆矩阵图

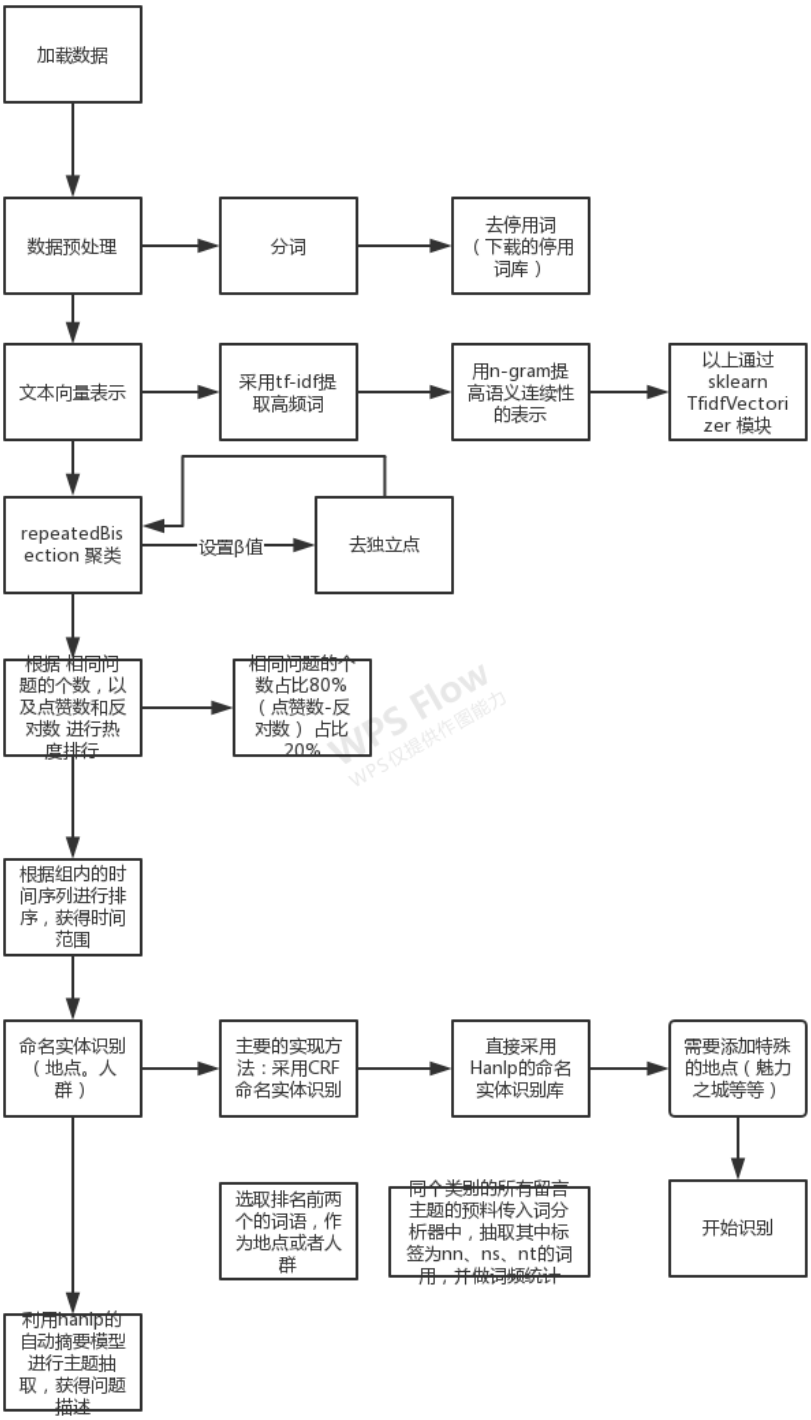
混淆矩阵是看每个类的预测结果和实际结果的情况 x 方向是预测，y 方向是实际结果

```
0.9104234527687296
```

分类方法评价 F1 值

四、任务二：热点问题挖掘

4.1 解题思路



4.2 repeat bisection 聚类算法

repeat bisection 算法即重复二分算法，repeat bisection 算法也被理解为二次 k-means 算法，但较之 k-means，速度更快，性能更优，且 repeat bisection 算法不用人工设定 k 值，而是通过给准则函数的增幅设定阈值 beta 来自动判断 k。此时算法的停机条件为，当一个簇的二分增幅小于 beta 时不再对该簇进行划分，即认为这个簇已经达到最终状态，不可再分；当所有簇都不可再分时，算法终止，此时产生的聚类数量就不再需要人工指定了。

因此我们使用 repeat bisection 算法对处理过后的附件 3 的数据进行聚类，先设置 β 值，去除独立点，有助于提高聚类的准确度，然后再进行聚类，从而可以实现对留言进行聚类分类，得到同类问题。

4.3 热度指标的定义

同类问题的留言个数占比 80%，同类问题中的点赞数和反对数的数据和占比 20%。根据上述指标进行热度统计。

	index	留言编号	留言用户	...	反对数	点赞数	分类
0	0	188006	A000102948	...	0	0	1185
1	1	188007	A00074795	...	0	1	469
2	2	188031	A00040066	...	0	1	912
3	3	188039	A00081379	...	0	1	670

4.4 热点问题留言明细表的形成

根据热度进行降序排序，将热度前五的问题的所有留言输出即得到热点问题留言明细表。

4.5 热点问题表的形成

4.5.1 时间范围的形成

将同类问题的时间序列进行排序，即可得到所需的时间范围。

4.5.2 命名实体的识别

（1）CRF 模型

CRF——条件随机场，由 Lafferty 等人于 2001 年提出，结合了最大熵模型和隐马尔可夫模型的特点，是一种无向图模型。条件随机场是一个典型的判别式模型，其联合概率可以写成若干势函数联乘的形式，

其中最常用的是线性链条件随机场。若让 x=(x1，x2，…xn)表示被观察的输入数据序列，y=(y1，y2，…yn)表示一个状态序列，在给定一个输入序列的情况下，线性链的 CRF 模型定义状态序列的联合条件概率为

$$p(y|x)=\exp\{\} \quad (2\text{-}14)$$

$$Z(x)=\{\} \quad (2\text{-}15)$$

其中:Z 是以观察序列 x 为条件的概率归一化因子;fj(yi-1，yi，x，i)是一个任意的特征函数;是每个特征函数的权值。

CRF 模型在分词、词性标注和命名实体识别等序列标注任务中，可以实现很好的运用效果。

（2）HanLP　分词工具。

HanLP　是由一系列模型与算法组成的 Java 工具包，目标是普及自然语言处理在生产环境中的应用。具备功能完善、性能高效、架构清晰、语料时新、可自定义的特点。HanLP　支持中文分词（N-最短路分词、CRF 分词、索引分词、用户自定义词典、词性标注），命名实体识别（中国人名、音译人名、日本人名、地名、实体机构名识别），关键词提取，自动摘要，短语提取，拼音转换，简繁转换，文本推荐，依存句法分析（MaxEnt 依存句法分析、CRF 依存句法分析）。

因此我们直接采用 HanLP　自带的命名实体识别库，添加所需的一些特殊词语，如地点名词等，用 CRF 进行命名实体识别。我们将之前得到同类问题的留言主题传入上述的词法分析器，抽取其中标签为 nn, nt, ns 的词用， 做词频统计，选区排名前二的词语，作为地点或人群。这样就得出我们所需的地点或人群的信息。

```
In[14]: dic
Out[14]: {'汇金路': 2, 'K9县': 2, '二房东': 1, '社区': 1, '黄谷路': 1}
```

命名识别结果图

4.5.3 问题描述的生成

我们直接采用 HanLP　工具中的自动摘要模型对同类问题中的留言主题进行提取，从而得到我们所需的问题描述。

```
qdescription = {str} '反映A7县恒基凯旋门小区配套幼儿园公办或者普惠问题'
```

提取结果图

4.5 结果展示（表格源文件件请查看作品附件）

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	7.0	2019/8/19 至 2019/9/19	K9县 五矿集团	A市五矿万境K9县交房后仍存在诸多问题
2	2	5.9	2019/2/27 至 2019/12/31	金毛湾 A市	反映A市山水湾孩子上学问题
3	3	5.1	2019/1/16 至 2019/7/8	西地省 美国	不要让A市因为58车贷案件而臭名远扬
4	4	2.8	2019/1/14 至 2019/10/22	汉回村 西地省	对A市A4区委员会办公室回复不属实的质疑
5	5	2.2	2019/8/23 至 2019/9/5	临路 绕城	A4区绿地海外滩小区距长赣高铁最近只有30米不到

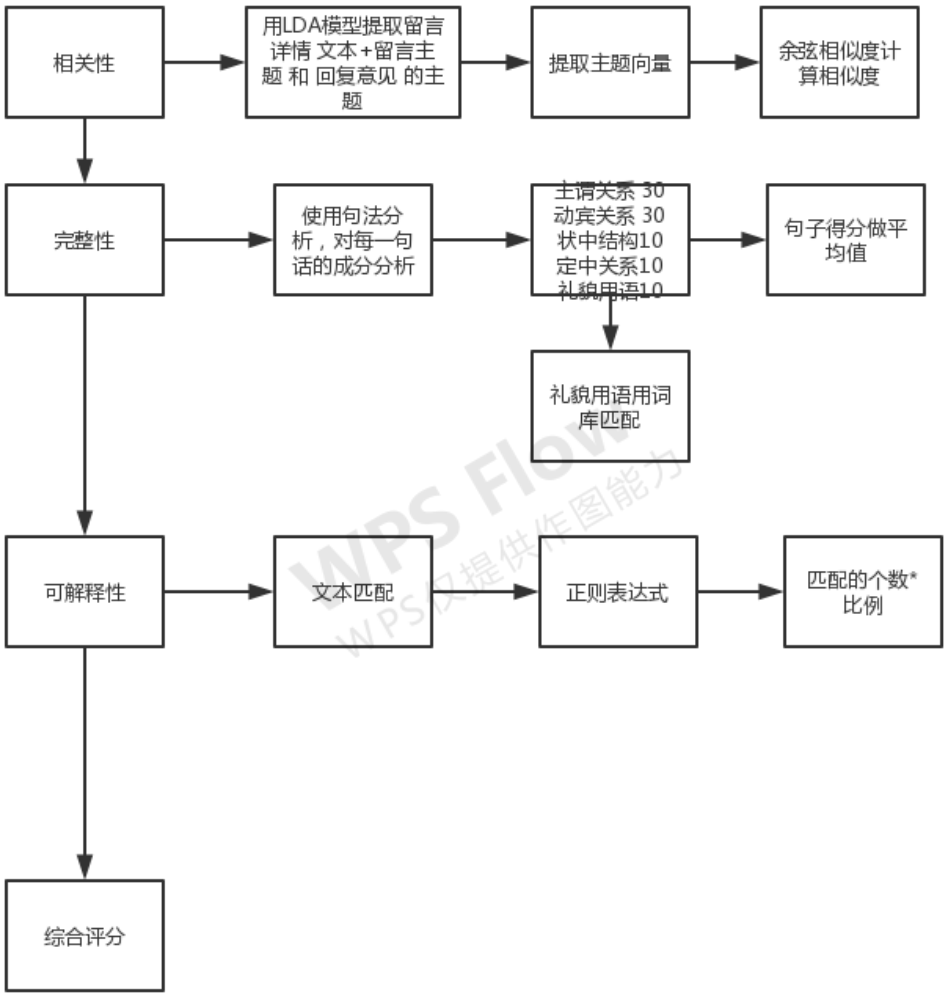
热点问题表

问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数
1	252650	A00010531	A市五矿万境K9县交房后仍存在诸多问题	2019-09-11 15:16:02	尊敬的相关部门，本人家庭	0	0
1	208636	A00077171	A市A5区汇金路五矿万境K9县存在一系列问题	2019-08-19 11:34:04	我是A市A5区汇金路五矿万	0	2097
1	262599	A000100428	A市五矿万境K9县房屋出现质量问题	2019-09-19 17:14:49	我是西地省A市五矿万境K9	0	0
2	286582	A00095839	反映A市合能枫丹丽舍精装修问题及业主子女入读小学事	2019-02-27 10:01:21	合能6、7业主要求参与毛坯	0	0
2	278895	A000103918	反映人民快速路与A市大道和劳动路衔接问题	2019-10-27 22:49:38	A市往东的放射快速路竟然	0	0
2	262494	A00024241	反映A市山水湾孩子上学问题	2019-12-31 17:06:53	山水湾住户为什么不能到西	0	0
2	223297	A00087522	反映A市金毛湾配套入学的问题	2019-04-11 21:02:44	书记先生：您好！我是梅溪	5	1762
2	202799	A000106944	反映A市商贸旅游学院白田宿舍摆摊问题	2019-10-28 23:23:54	A市商贸旅游学院白田宿舍	0	0
2	209048	A00098971	反映A市理工大学教职工宿舍的住房安全问题	2019-12-25 10:41:33	尊敬的领导：你们好！我是	0	0
2	206939	A00080330	反映A市大道沙湾路立交桥拥堵治理问题	2019-10-14 11:09:56	A市大道（沙湾路往西400米	0	2
3	234320	A000106592	不要让A市因为58车贷案件而臭名远扬	2019-07-08 17:16:57	胡书记：您好反映关于西地	0	0
3	226265	A000106448	恳请A市经侦公正办理58车贷案件，还我们受害人一个公	2019-05-28 15:08:51	唐局长，您好西地省A市58	0	3
3	272858	A00061787	A市58车贷恶性退出案件为什么不发布案情进展通报？	2019-01-16 23:21:21	唐局长，您好。我是A市58	0	0
3	217032	A00056543	严惩A市58车贷特大集资诈骗案保护伞	2019-02-25 09:58:37	胡市长：您好！西地省展星	0	790
3	218132	A000106090	再次请求过问A市58车贷案件进展情况	2019-01-29 19:15:49	尊敬的胡书记：您好！西地	0	0
3	194343	A000106161	承办A市58车贷案警官应跟进关注留言	2019-03-01 22:12:30	胡书记：您好！58车贷案发	0	733
4	252493	A00074783	A市A4区胜利村鑫利家园违规分房	2019-04-19 20:09:52	2014年底A4区四方坪胜利社	0	1
4	220711	A00031682	请书记关注A市A4区58车贷案	2019-02-21 18:45:14	尊敬的胡书记：您好!A4区	0	821
4	238241	A00096117	A4区A9市河幼儿园强制收取门禁卡费	2019-09-06 21:56:22	又到开学的时候了，小朋友	0	0
4	272413	A000106062	西地省A市58车贷恶性退出，A4区立案已近半年毫无进展	2019-01-14 20:23:57	西地省58车贷邢ze恶性退出	0	2
4	229426	A00048684	A4区A9市河隧道大量违规搅拌罐车极度扰民	2019-10-22 01:39:23	A市A4区楚江北路A9市河隧	0	0
4	280367	A00050197	举报A市A4区捞刀河街道高源村村长欺负村民	2019-07-25 18:10:54	尊敬的领导：我西地省A市	0	1
4	230533	A0005799	A市A4区沙坪街道汉回村支书纵容华颖实业集团侵占农民	2019-07-26 11:30:05	尊敬的领导：我们是A市A4	0	2
4	276613	A00094265	对A市A4区委员会办公室回复不属实的质疑	2019-07-23 14:11:00	原帖：A4区长房雍景湾项目	0	5
4	196144	A0005821	A市北二环与A4区大道交界附近有三处超高减速带	2019-07-09 10:29:17	该减速带位于A市北二环与	0	0
4	218025	A000111121	A市一中A4区中学初三国庆只放三天假	2019-09-23 11:29:46	A市初三国庆只放三天假，	0	0
5	191951	A00041448	A4区绿地海外滩小区距渝长厦高铁太近了	2019-08-23 14:21:38	尊敬的领导：你好，近日看	0	1
5	263672	A00041448	A4区绿地海外滩小区距长赣高铁最近只有30米不到，合理	2019-09-05 13:06:55	您好，近日看到了渝长厦高	0	669

热点问题明细表

五、任务三

5.1 解题思路

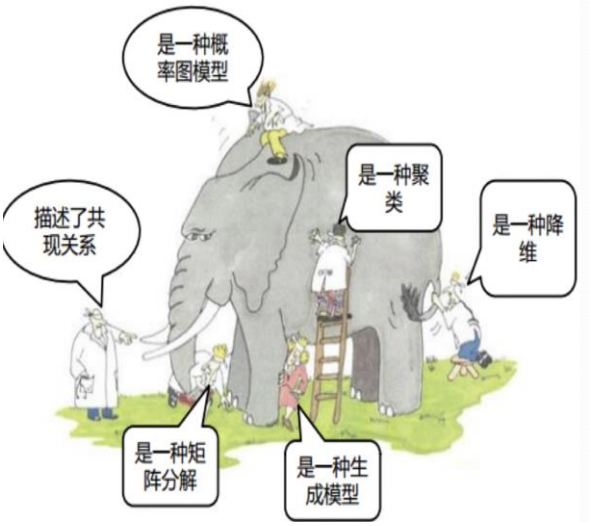


5.2 相关性

5.2.1 LDA 主题模型

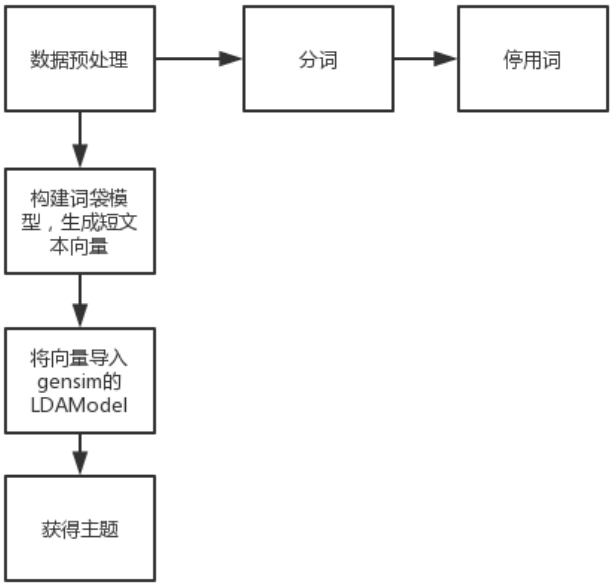
（1）LDA（Latent Dirichlet Allocation）是一种文档主题生成模型，也称为一个三层贝叶斯概率模型，包含词、主题和文档三层结构。所谓生成模型，就是说，我们认为一篇文章的每个词都是通过“以一定概率选择了某个主题，并

从这个主题中以一定概率选择某个词语”这样一个过程得到。文档到主题服从多项式分布，主题到词服从多项式分布。



从这张图片，可以了解到 LDA 主题模型的大致作用。

（2）通过 LDA 主题模型，可以提取一段文本的主题，从而可以关注到文本的关键词，筛选出当前文本重要的词句，从而有利于后面的相似度计算。



5.2.2 文本相似度（余弦相似度算法）

（1）余弦相似度：

输入：文本向量

计算公式：

① 二维坐标计算公式：

$$\cos(\theta) = \frac{a^2 + b^2 - c^2}{2ab} = \frac{x_1^2 + y_1^2 + x_2^2 + y_2^2 - (x_2 - x_1)^2 - (y_2 - y_1)^2}{2\sqrt{x_1^2 + y_1^2} * \sqrt{x_2^2 + y_2^2}}$$

$$= \frac{x_1 * x_2 + y_1 * y_2}{\sqrt{x_1^2 + y_1^2} * \sqrt{x_2^2 + y_2^2}}$$

② 多维坐标计算公式

最小化

$$\cos(\theta) = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2} \times \sqrt{\sum_{i=1}^{n}(y_i)^2}}$$

$$= \frac{a \bullet b}{||a|| \times ||b||}$$

公式(4)

输出结果：两个文本向量的余弦值

（3）利用 LDA 主题模型，可以分别提取留言主题和留言主题、答复意见的主题，这样我们可以得到每部分内容的主题，过滤重要程度不高的词语，

① 把生成的主题用词袋模型转换成词向量

② 将词向量采用多维余弦计算公式获得**余弦值**，得到的余弦值即我们需要的相似度

5.2.3 结果以及计算规则

（1）将生成的余弦值在利用极差法，是分数能够均匀得分布在（0,1）区间中

（2）结果：

H	K	L	M
留言主题	答复意见	答复时间	相关性
A2区景蓉华苑物业管理有问题	现将网友	2019/5/10 14:56:53	0.516511
A3区潇楚南路洋湖段怎么还没修好？	网友“A0	2019/5/9 9:49:10	0.099971
请加快提高A市民营幼儿园老师的待遇	市民同志	2019/5/9 9:49:14	0.509935
在A市买公寓能享受人才新政购房补贴吗？	网友“A0	2019/5/9 9:49:42	0.511261
关于A市公交站点名称变更的建议	网友“A0	2019/5/9 9:51:30	0.528398
A3区含浦镇马路卫生很差	网友“A0	2019/5/9 10:02:08	0.253703
A3区教师村小区盼望早日安装电梯	网友“A0	2019/5/9 10:18:58	0.551333
反映A5区东澜湾社区居民的集体民生诉求	网友“UU	2019/1/29 10:53:00	0.286885
反映A市美麓阳光住宅楼无故停工以及质量	网友“UU	2019/1/16 15:29:43	0.208531
反映A市洋湖新城和顺路洋湖壹号小区路段	网友“UU	2019/1/16 15:31:05	0.363868
反映A2区大托街道大托新村违建问题	网友“UU	2019/3/11 16:06:33	0.081343
A5区鄱阳村D区安置房人防工程的咨询	网友“UU	2019/1/29 10:52:01	0.513445
A4区万国城小区段请求修建一座人行天桥	网友“UU	2019/1/14 14:34:58	0.360174
举报A市芒果金融平台涉嫌诈骗	网友“UU	2019/1/3 14:03:07	0.135139
建议增开A市261路公交车	网友“UU	2019/1/14 14:33:17	0.071009
关于A市新开铺路与披塘路交叉路口通行安	网友“UU	2019/3/6 10:26:14	0.441444
投诉A3区桐梓坡路益丰大药房以次充好	网友“UU	2019/1/3 14:02:47	0.100079
建议在A市梅溪湖开办一个图书馆	网友“UU	2019/1/14 14:32:40	0.338754
希望相关部门治理A3区中海国际社区一期	网友“UU	2019/1/8 16:19:16	0.156089
希望A市社保卡、医保卡、居民健康卡尽快	网友“UU	2019/1/4 15:48:23	0.077249
希望A市潇楚一卡通尽快支持手机nfc虚拟	网友“UU	2019/1/4 15:49:46	0.267911
反映A9市北盛镇对泉水村塘下组土地征收	网友“UU	2019/1/8 16:18:00	0.429208

5.3 完整性

5.3.1 句法分析

（1）句法分析的基本任务是确定句子的语法结构或句子中词汇之间的依存关系。句法分析不是一个自然语言处理任务的最终目标，但它往往是实现最终目标的关键环节。

句法分析分为句法结构分析和依存关系分析两种。以获取整个句子的句法结构为目的的称为完全句法分析，而以获得局部成分为目的的语法分析称为局部分析，依存关系分析简称依存分析。

（2）使用句法分析可以分析句子的成分，例如主语，谓语，宾语等等，通过句子的成分是否完整，我们即可以分析句子的结构成分是否完整，因此来判断答复意见中的语言是否通顺完整。

（3）使用 HanLP 实现句法分析

HanLP 是一个强大的自然语言处理的开源框架，由于句法分析的实现比较复杂，我们可以直接使用 HanLP 的句法分析

使用示例：

```
        sentence = HanLP.parseDependency(sentence[:length])
        isPolite = False
        for word in list(sentence):
            if not isPolite:
                if word.LEMMA in polite_word:
                    isPolite = True
            if word.HEAD.DEPREL is not None:
                if word.DEPREL == "主谓关系":
                    obj = word.HEAD.LEMMA
                    result[obj] = [SentenceIngredient.PREDICATE]
                    continue
```

（4）算法流程

① 首先，我们需要找出每一个回复意见的每一个句子，因为一个句子只能由一个主谓关系，那么通过找主谓关系我们就可以找到每一个句

② 通过主谓关系我们可以得到每个句子的谓语动词；

③ 通过谓语动词我们可以查找所有他的宾语；

④ 在一句话内，我们可以寻找是否有状语，以及定语；

⑤ 通过以上流程，我们可以确定句子成分是否完全，主谓关系 30 分，动宾关系 30 分，状语和定语各有 10 分。

5.3.2 文本匹配

（1）面对不同语言场景时，有时候我们需要分析句子的特殊成分，例如对答复意见，我们需要分析其是否有敬辞（你好，您好等等），这时候我们需要使用文本匹配的方法，来分析文本的特殊成分。

（2）算法流程：

① 建立关于敬辞的词库

② 遍历所有的回复意见，然后判断是否包含词库的词库

③ 包含则总分值加 20

5.3.3 结果计算规则

将句法分析和文本匹配后得到的分值归一化处

H	K	O
留言主题	答复意见	完整性
A2区景蓉华苑物业管理有问题	现将网友在平台《问政西地省》栏目向	0.8
A3区潇楚南路洋湖段怎么还没修好？	网友“A00023583”：您好！针对您反映	0.75
请加快提高A市民营幼儿园老师的待遇	市民同志：你好！您反映的“请加快提	0.966667
在A市买公寓能享受人才新政购房补贴吗？	网友“A000110735”：您好！您在平台	0.790909
关于A市公交站点名称变更的建议	网友“A0009233”，您好，您的留言已	0.76
A3区含浦镇马路卫生很差	网友“A00077538”：您好！针对您反映	0.75
A3区教师村小区盼望早日安装电梯	网友“A000100804”：您好！针对您反	0.863636
反映A5区东澜湾社区居民的集体民生诉求	网友“UU00812”您好！您的留言已收悉	0.92
反映A市美麓阳光住宅楼无故停工以及质量	网友“UU008792”您好！您的留言已收	0.816667
反映A市洋湖新城和顺路洋湖壹号小区路段	网友“UU008687”您好！您的留言已收	0.72
反映A2区大托街道大托新村违建问题	网友“UU0082204”您好！您的留言已收	0.84
A5区鄱阳村D区安置房人防工程的咨询	网友“UU008829”您好！您的留言已收	0.85
A4区万国城小区段请求修建一座人行天桥	网友“UU00877”您好！您的留言已收悉	0.825
举报A市芒果金融平台涉嫌诈骗	网友“UU0081480”您好！您的留言已收	0.7
建议增开A市261路公交车	网友“UU0081227”您好！您的留言已收	0.778571
关于A市新开铺路与披塘路交叉路口通行安	网友“UU008444”您好！您的留言已收	0.929412
投诉A3区桐梓坡路益丰大药房以次充好	网友“UU0081194”您好！您的留言已收	0.833333
建议在A市梅溪湖开办一个图书馆	网友“UU008706”您好！您的留言已收	0.742857
希望相关部门治理A3区中海国际社区一期	网友“UU008201”您好！您的留言已收	0.84
希望A市社保卡、医保卡、居民健康卡尽快	网友“UU0081681”您好！您的留言已收	0.733333
希望A市潇楚一卡通尽快支持手机nfc虚拟	网友“UU0081681”您好！您的留言已收	0.816667
反映A9市北盛镇对泉水村塘下组土地征收	网友“UU0081500”您好！您的留言已收	0.9

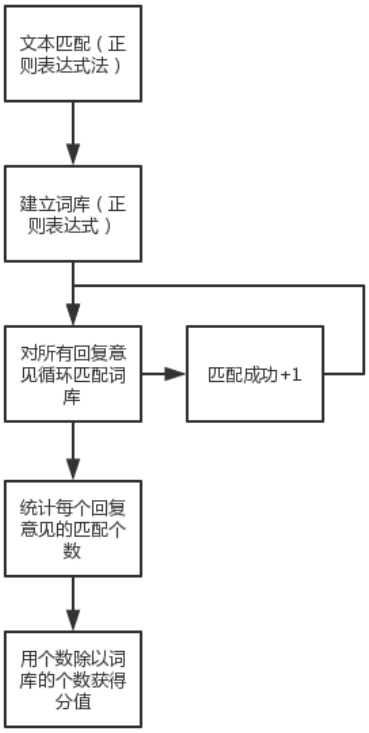
5.4 可解释性

5.4.1 文本匹配法+正则表达式

（1）正则表达式：对字符串（包括普通字符（例如，a 到 z 之间的字母）和特殊字符（称为“元字符”））操作的一种逻辑公式，就是用事先定义好的一些特定字符、及这些特定字符的组合，组成一个“规则字符串”，这个“规则字符串”用来表达对字符串的一种过滤逻辑。正则表达式是一种文本模式，该模式描述在搜索文本时要匹配的一个或多个字符串。 [1]

使用正则表达式，我们可以使用更加灵活的匹配方式，例如使用“《.+》”我们则可以匹配所有具有书名号的词语

（2）算法流程：



5.4.2 结果以及计算规则

（1）计算公式：

得分 = 匹配个数 X （1 / 词库词语总个数）

（2） 结果：

留言主题	答复意见	可解释性
A2区景蓉华苑物业管理有问题	现将网友在平台《问政西地	0.666666667
A3区潇楚南路洋湖段怎么还没修好？	网友“A00023583”：您好！	0
请加快提高A市民营幼儿园老师的待遇	市民同志：你好！您反映的	0.666666667
在A市买公寓能享受人才新政购房补贴吗？	网友“A000110735”：您好	0.333333333
关于A市公交站点名称变更的建议	网友“A0009233”，您好，	0
A3区含浦镇马路卫生很差	网友“A00077538”：您好！	0
A3区教师村小区盼望早日安装电梯	网友“A000100804”：您好	0.333333333
反映A5区东澜湾社区居民的集体民生诉求	网友“UU00812”您好！您的	0.666666667
反映A市美麓阳光住宅楼无故停工以及质量	网友“UU008792”您好！您	0
反映A市洋湖新城和顺路洋湖壹号小区路段	网友“UU008687”您好！您	0
反映A2区大托街道大托新村违建问题	网友“UU0082204”您好！您	0
A5区鄱阳村D区安置房人防工程的咨询	网友“UU008829”您好！您	0.666666667
A4区万国城小区段请求修建一座人行天桥	网友“UU00877”您好！您的	0
举报A市芒果金融平台涉嫌诈骗	网友“UU0081480”您好！您	0
建议增开A市261路公交车	网友“UU0081227”您好！您	0
关于A市新开铺路与披塘路交叉路口通行安	网友“UU008444”您好！您	0.666666667
投诉A3区桐梓坡路益丰大药房以次充好	网友“UU0081194”您好！您	0.333333333
建议在A市梅溪湖开办一个图书馆	网友“UU008706”您好！您	0
希望相关部门治理A3区中海国际社区一期	网友“UU008201”您好！您	0.333333333
希望A市社保卡、医保卡、居民健康卡尽快	网友“UU0081681”您好！您	0
希望A市潇楚一卡通尽快支持手机nfc虚拟	网友“UU0081681”您好！您	0
反映A9市北盛镇对泉水村塘下组土地征收	网友“UU0081500”您好！您	0.333333333

5.5 结果展示

将以上生成的 相关性，完整性，可解释性 相加 在除以 3 则可以得到最后的得分。

相关性	可解释性	完整性	综合得分
0.516511	0.666667	0.8	0.661059
0.099971	0	0.75	0.283324
0.509935	0.666667	0.966667	0.714423
0.511261	0.333333	0.790909	0.545168
0.528398	0	0.76	0.429466
0.253703	0	0.75	0.334568
0.551333	0.333333	0.863636	0.582768
0.286885	0.666667	0.92	0.624517
0.208531	0	0.816667	0.341732
0.363868	0	0.72	0.361289
0.081343	0	0.84	0.307114
0.513445	0.666667	0.85	0.676704
0.360174	0	0.825	0.395058
0.135139	0	0.7	0.27838
0.071009	0	0.778571	0.283193
0.441444	0.666667	0.929412	0.679174
0.100079	0.333333	0.833333	0.422248
0.338754	0	0.742857	0.360537
0.156089	0.333333	0.84	0.443141
0.077249	0	0.733333	0.270194
0.267911	0	0.816667	0.361526
0.429208	0.333333	0.9	0.554181

六、参考文献

(1) https://www.cnblogs.com/tan2810/p/11202874.html（TF-IDF 算法介绍及实现）

(2)https://blog.csdn.net/songbinxu/article/details/80209197（自然语言处理 NLP 中的 N-gram 模型）

(3)https://blog.csdn.net/fg13821267836/article/details/93405572（多层感知机（MLP）简介）

(4)https://blog.csdn.net/xiaosongshine/article/details/88658274（[深度概念]·K-Fold 交叉验证 (Cross-Validation)的理解与应用）

(5)https://blog.csdn.net/evillist/article/details/61912827（【机器学习】k-fold cross validation（k-折叠交叉验证））