第八届“泰迪杯”全国数据挖掘挑战赛

“智慧政务”中的文本挖掘应用

摘要

在信息化快速发展的现代社会，数字空间、网络社会逐渐成为人类活动的重要场景。而以“智慧政务”为代表的具有决策科学化、办公快速化、治理高效化与服务便捷化特征的智慧政府建设已成为社会治理体系和治理能力现代化的重要支撑和关键环节。其强调尊重事实、追求精确、推崇理性和逻辑、注重满足多元化服务，是精细化、智能化、人本化、平等化的深度融合。特别是在舆情监控上，“智慧政务”能够利用大数据和智能决策系统为科学决策提供动态的数据支撑，在确保决策的客观性、真实性的同时准确把握社会事件的演进规律和发展趋势，避免传统科层制政府在有限理性基础上的模糊决策、主观决策和经验决策等弊端。

本文基于自然语言处理技术，对给定附件中的群众问政留言记录，及相关部门对部分群众留言的答复意见，对留言进行分类，并挖掘出其中的热点问题，同时根据答复的相关性、完整性、可解释性等角度，制定出针对答复意见的评价方案。

针对问题一群众留言的分类问题，采用 LDA 模型进行文本的聚类，具体的步骤是首先是去重，结巴分词后进行洗词操作，去掉停用词，之后对文本中的词汇通过 TFIdf 构建矩阵，得出词在不同内容中的权重，之后输入模型，进行训练，在对其中的参数绘制折线图，并通过计算 F-Score 进行评价。

针对问题二热点问题挖掘问题，核心问题在于判断是否为焦点问题。通过以下两个维度来对此题进行解决：①问题本身的热度，通过留言问题的评论或点赞数来对热量进行量化，但是需要与时间挂钩，让热度根据时间进行衰减，只有这样才可以把真正的热点问题暴露出来；②其次，找到社会热点问题，对留言进行主题词的提取，然后在所有的留言中针对抽取的关键词进行关联，把关联起来的留言同样量化热度到具体的某个相关里面，这样就形成了同个主题的焦点问题协同加权，从而把热点问题给暴露出来。

针对问题三答复意见的评价方案问题，分别从相关性、完整性和可解释性对留言与答复意见进行评价。其中相关性分析使用了 BM25 算法，可以分析答复意见中高频词与留言文本的相关程度；而可解释性作为模型可信度的重要参考依据，可以使用 ROC、PR、KS 曲线进行分析。

本题中使用的数据分析工具：pycharm64

关键词：LDA 模型 文本聚类 分词洗词 Gibbs 采样 Word2vec 词向量训练

Text Mining Application in "Smart Government Affairs"

Abstract

In the modern society with rapid development of information technology, digital space and network society have gradually become important scenes of human activities. "Smart government" can use big data and intelligent decision-making systems to provide dynamic data support for scientific decision-making, while ensuring the objectivity and authenticity of decision-making while accurately grasping the evolution laws and development trends of social events, avoiding traditional bureaucratic government The disadvantages of fuzzy decision, subjective decision and empirical decision based on limited rationality.

This article is based on natural language processing technology, to the public's questionnaire message record in the given attachment, and the relevant departments' reply to some of the people's messages, classify the messages, and dig out the hot issues among them. From the perspective of completeness and interpretability, formulate an evaluation plan for the responses.

Aiming at the classification problem of mass messages in question 1, the LDA model is used for text clustering. The specific steps are deduplication, word washing after word segmentation, and removal of stop words. Then, the matrix of words in the text is constructed by TFIdf , Get the weight of words in different content, then input the model, train, draw a line chart of the parameters, and evaluate by calculating F-Score.

Focusing on the second hot issue of problem mining, the core problem is to determine whether it is the focus problem. Solve this problem through the following two dimensions: ① The heat of the problem itself, quantify the heat by leaving comments or likes on the question, and link it with time to let the heat decay according to time; Problem, extract the subject words of the message, and then correlate the extracted keywords in all the messages, and quantify the related messages to a specific correlation, thus forming the focus of the same topic. Collaborative weighting.

Regarding the question of the evaluation plan for the answer to question three, the comments and comments were evaluated from the perspective of relevance, completeness and interpretability. Among them, the correlation analysis uses the BM25 algorithm, which can analyze the relevance of high-frequency words in the reply opinion and the message text; and interpretability, as an important reference basis for the credibility of the model, can be analyzed using ROC, PR, and KS curves.

目录

1 群众留言分类

1.1 模型说明

本题使用的是 LDA（Latent Dirichlet Allocation）模型来进行文本的聚类。LDA 是一种非监督机器学习技术，可以用来识别大规模文档集或语料库中潜藏的主题信息。它也被称为是 PLSA 的贝叶斯版本，这是由于 LDA 模型在 PLSA 的基础上，为主题分布和词分布分别加了两个 Dirichlet 先验，使得主题分布和词分布变为不确定的随机变量。在进行参数估计时，LDA 采用贝叶斯估计，把待估计的参数看作是服从先验分布的随机变量。同时，利用 Gibbs 采样估计未知参数，再通过计算得到的联合分布，求出 Dirichlet 期望和后验分布。

1.2 数据预处理

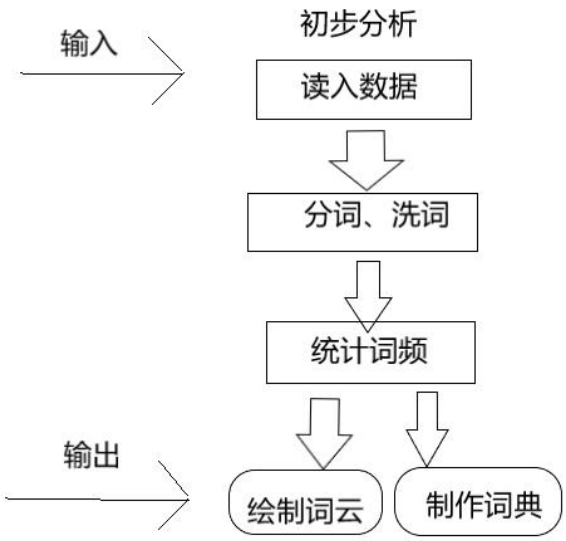

图 1 数据预处理流程示意图

针对附件 2 中给出的数据，首先要对留言详情中的语句进行分词操作，即将句子中的组成词语分割出来。具体的操作步骤是：去除换行符、制表符等无意义符号，利用 jieba 库中的分词函数 jieba.lcut 对句子进行分词。Jieba 库是第三方提供的中文分词库，主要原理是利用中文词库，确定汉字间的关联概率，并形成分词结果。本题中使用到的 lcut(s)函数就是通过精确模式的分词，返回一个列表类型的分词结果，同时去除冗余单词。

为了节省储存空间，同时提高搜索效率，要对分词结果中语气助词、副词、介词、连接词等对分类无实际意义的词语进行过滤。所以接下来利用常用的停词表，依次遍历去除分词结果中的无分类意义的词语，对结果进行洗词。

在进行完分词、洗词的操作后，定义一个字典对现有词的词频进行统计，并将结果按照词频进行排序。为了能更加直观的展示预处理的结果，可以根据词频生成图云，以及将结果保存在分类词典的文件夹中，以.txt 形势保存。同时，对于词频非常低的部分词语，可以更新加入停词表中，用于正式的数据处理中，以提高文本的处理效率。

图 2 图云-城乡建设

图 3 图云-环境保护

图 4 图云-交通运输

图 5 图云-教育文体

图 6 图云-劳动和社会保障

图 7 图云-商贸旅游

图 8 图云-卫生计生

1.3 LDA 模型实现

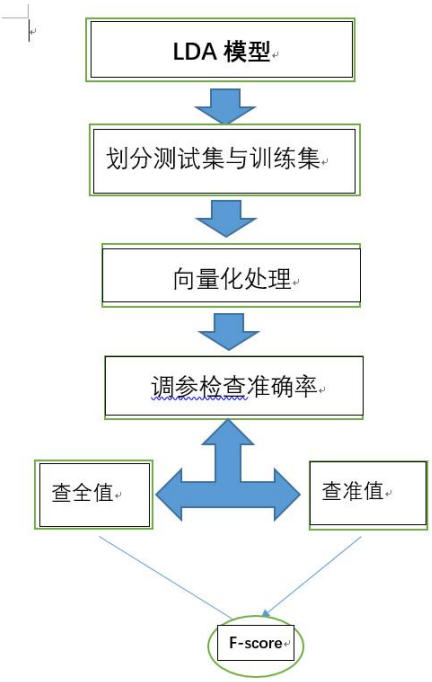

图 9　LDA 模型实现流程示意图

经过了数据预处理后，对数据集的基本情况有了一定的了解，接下来就通过 LDA 模型对数据实现聚类。

首先还是基本的数据处理：去重，提取留言详情，进行分词操作，然后利用 1.2 中数据预处理的结果更新后的停词表，对词语进行洗词，将需要的词语放入词袋中。词袋是 LDA 模型的基本框架。同时使用 LDA 的多项分布。多项分布，是二项分布扩展到多维的情况。它是指单次试验中的随机变量的取值不再是 0-1 的，而是有多种离散值可能（1,2,3…,k）。比如投掷 6 个面的骰子实验，N 次实验结果服从 K=6 的多项分布。其中：

$$\sum_{i=1}^{k} p_i = 1, p_i > 0$$

多项分布的概率密度函数为：

$$P(\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_k; n, \mathrm{p}_1, \mathrm{p}_2, ..., \mathrm{p}_k) = \frac{n!}{\mathrm{x}_1! ... \mathrm{x}_k!} p_1^{\mathrm{x}_1} ... p_k^{\mathrm{x}_k}$$

接下来构建语料库，并对已有的留言类别进行数值转换。再文本中的词汇通过 TFIdf 构建矩阵，得出词在不同内容中的权重。采用随机划分的方法，划分训练集以及测试集。而后进行向量化处理，这步中我们可以通过分析最大词频使用量来动态的分析准确率通过调整参数 max_features 观察测试结果的准确度，在经过训练集的训练时，我们需要进行模型拟合，这里要防止出现模型的过拟合，之后进行输出测试集测试，我们需要得到如下参数折线图：

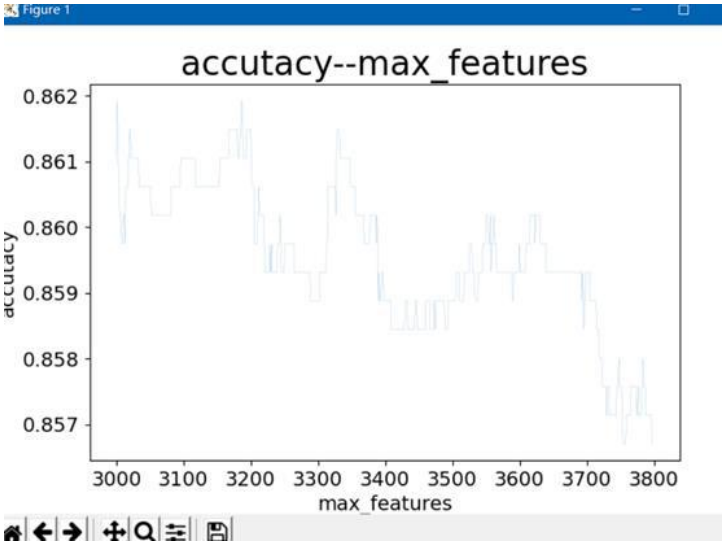

图 10 调参准确度变化曲线

接下来通过 F-Score 对 LDA 模型进行评价。

F-Score 是一种衡量机器学习模型性能的指标，计算方法为：$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i}$ 其中，P 为精确率，R 为召回率。对于普适的 F-Score 计算公式 $F = (1+\beta^2)\cdot\frac{2P_iR_i}{\beta^2\cdot P_i + R_i}$，当精确率 P 更重要时，就调整 β 的值小于 1；如果召回率 R 更重要时，就调整 β 的值大于 1.在本题中，给定的公式为 $F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i}$，这就代表 β 为 1，P 和 R 都很重要，此时成为 F 称为 F_1。

经过 F-Score 的计算中，得出结果约为 0.84。

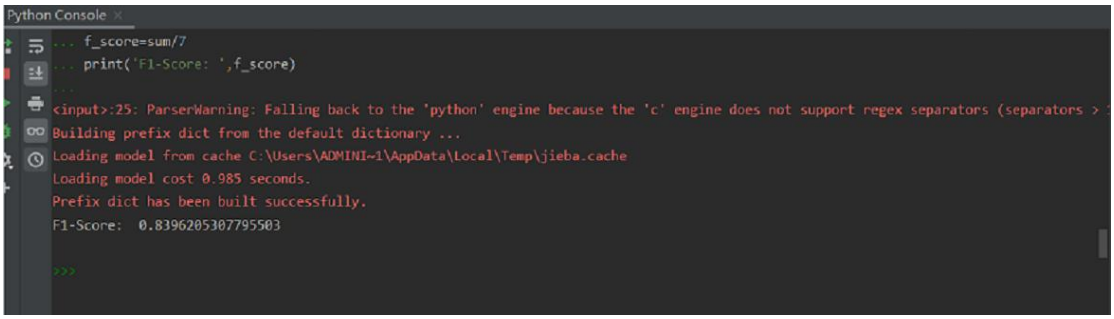

图 11 F-Score 运算结果

1.4 模型实现评价

LDA 模型具有许多优点，相较于 PLSA 模型来说它具有以下优势：

（1）采用贝叶斯框架，通过先验分布求得联合分布，再进一步求出后验分布和

期望。在训练的过程中，可以有效动态地提高准确度；
（2）采用 Gibbs 采样估计未知参数，相较于原始地变分-EM 算法，Gibbs 采样可以巧妙地求解出主题分布和此分布的后验分布，从而成功解决主题分布和词分布这两参数位置的问题。
同时，LDA 模型也具有一定的局限性，对结果会产生准确性的影响，主要是由于 LDA 采用词袋模型，一种最基础的文本表示模型。这种方式不考虑文法以及词的顺序，无法计算词的相似度，同时含有的信息量较为稀疏，与采用词向量模型的准确度会有一定的差距，因此会对留言的分析产生一定的影响。

2 热点问题挖掘

2.1 数据预处理

与上一题群众留言的分类相似，本题也要进行数据的清洗：去重、去掉标点符号、去掉一些无关词语，以及数据中大量存在的地区词语（如 A 市、A5 区）。要将这些词语提取出来，以免在分词以及后面的主题词提取中出现问题，降低搜索效率。

2.2 基本步骤

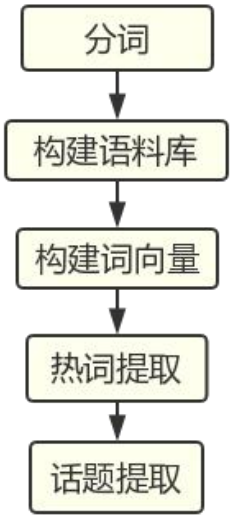

图 12 热点话题挖掘基本流程图

首先是分词操作。准备一个停用词词典，训练时要去除停用词的干扰。采用的 jieba 分词，对附件 3 里面的留言详情进行分词。接下来构建 word2vec 语料库，并将分好的词通过 word2vec 语料库进行构建数组，将文本表示成向量的形式，构成词向量。然后对热词进行提取。最终实现对话题进行提取。

2.3 Word2vec 词向量训练

最初，我们小组使用了在网上下载的百度百科词料库，但是效果不是特别的好，有效率比较低，所以我们改用自己建立语料库。

在 word2vec 中采用的是分布式表征的词向量，在向量维数比较大的情况下，每一个词都可以用元素的分布式权重来表示，因此，向量的每一维都表示一个特征向量，作用于所有的单词，而不是简单的元素和值之间的一一映射。这种方式抽象的表示了一个词的“意义”。实际上，被学习的词向量表示是用一种非常简单的方式捕捉有意义的语法和语义规律。具体来说，对于一个特定关系的词组，语法规律可以看作固定的向量偏移。

2.4 热词提取

对于一个词语来说，可能会受到多方面的影响：

① 词间影响：也许某个时间某个时间比较火，会导致一些平时关系不大的词语，一下子全部成为热词；

② 周期影响：二十四小时、星期、月份等周期性的变化，常常会使得“周日”、“三月”等事件意义性不强的词语成为热词；

③ 时间影响：白天与凌晨、工作日与节假日，留言消息的数量可能都会有一个较大的波动；

④ 自身趋势：这个就是我们所关心的热度信息了。这些由于某一事件导致某些词语的突发性、递增性等的增长，就是我们想要识别和分析出来的。

所以，要针对以上可能存在的情况，减少影响热度词语的因素，使得到的数据尽可能的准确，我们要对数据进行预处理，主要包括：文本去重、人群地区的提取，以及对数据进行一些去躁工作，并进行分词。

用梯度来作为词频增量的主要衡量指标：$S(w_i)=\dfrac{F(w_i,T_j)}{F(w_i,T_1,T_2,...,T_j)}$

其中，w_i 表示某个词语，T_j 表示时间窗口，$F(w_i,T_j)$ 表示词语在 w_i 在时间窗口 T_j 出现的频数。$S(w_i)$ 表示某个词语目前的热度分数。

而对于热度分数的计算，利用贝叶斯平均对梯度分数进行修正：

$$\text{修正分数}=\text{平均分}+\frac{\text{词频}}{\text{词频}+\text{平均词频}}\times(\text{梯度分数}-\text{平均分})$$

所谓贝叶斯平均算法，以用户投票进行排名为例，用户投票评分的人很少，则算出的平均分很可能会出现不够客观的情况。此时，可以引入外部的信息，假设还有一部分人（C 人）投了票，并且都给了平均分（m 分）。把这些人的评分加入到已有用户的评分中，再进行求平均，可以对平均分进行修正，以在某种程度或角度上增加最终分数的客观性。容易得到，当投票人数少的时候，分数会趋向于平均分；投票人数越多，贝叶斯平均的结果就越接近真实投票的算术平均，

加入的参数对最终排名的影响就越小。

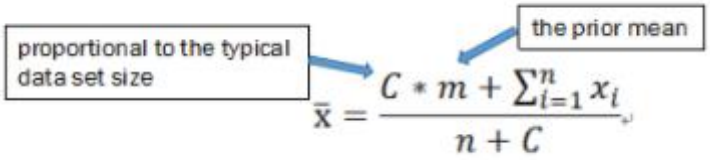

图 13 贝叶斯平均公式

而在修正公式中的平均词频就是贝叶斯公式中的 C，平均分是贝叶斯平均公式中的 m。也就是说，在热词的提取中，用梯度分数的平均分作为先验 m，用平均词频作为 C。

也可以用投票的方式对这个过程进行理解，词语每出现一次，就相当于给词的热度进行了评分。词频少，也就代表了评分的人数少，则评分的不确定性大，需要用平均分来进行修正、平滑。这里可以把一些词频很少的词语的高分数拉下来；词频大，远大于平均词频的词语，也就代表了评分的人数多。则分数会越趋向于自己的实际分数，这时平均分的影响变小。

只发现热点词语还是远远不够的，还要通过频繁项集、word2vector 等方法，发现出共现词语的关系。利用共现词语的信息，对热词进行一轮筛选，提取出最有价值的热词，避免信息冗余。

2.5 话题提取

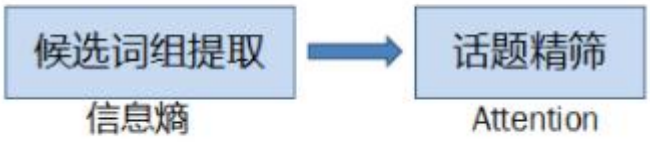

图 14 话题提取

提取工作分成了两步，第一步先找出一些候选的话题词组；第二部利用 Attention 的思想，从候选词组中找出一个包含词语的更加重要的词组，作为输出话题。

候选词组的提取主要根据信息熵的理论，而信息熵是用来衡量一个随机变量出现的期望值，一个变量的信息熵越大，表示其可能的出现的状态越多，越不确定，也即信息量越大。

$$H = -\sum_{i=1}^{n} p_i \log p_i$$

而内部的聚合度用互信息来表示，他可以说明两个随机变量之间的关系强弱，定义为：

$$I(X;Y) = \int_X \int_Y P(X,Y) \log \frac{P(X,Y)}{P(X)P(Y)}$$

对其进行变换可以得到：

$$I(X;Y)=H(Y)-H(Y\mid X)$$

该式子可表示为由 X 引入而使 Y 的不确定度减小的量越大，说明 X 出现后，Y 出现的不确定度减小，即 Y 很可能也会出现，也就是说 X、Y 关系越密切。反之亦然。而在词组中，其内部聚合度即为词语间的内部聚合度。对于一个词组，我们选取使不确定性减少的程度最多的一种词语组合，来说明词组的内部聚合度。

当然也要考虑到当一个词语在不同的语句中出现时，即它的左右信息熵，其值越大，表示词组左右的可能情况越多，与它搭配的词语越多；则说明这个词组在不同的语境里可讨论的事情越多，越可能可以独立说明一个事件或话题。

接着，要利用 Attention 模型对话题进行筛选，而 Attention 模型的要点就是加权求和，通过上面划分得到的词组，通过模型算出对应的事件或话题表示能力分数：

$$Score(s)=\frac{\sum_{i=1}^{N}\frac{Corpus(w^h,w_i)}{Corpus(w_i)}}{N}$$

其中，N 为候选词组中的词语个数，为候选词组中包含的第 i 个词语，Corpus (w)表示含有词语 w 的相关语料。当然，我们也需要考虑词组出现的频次，词组出现的次数越多，说明事件越重要。

随后筛选出排名前 5 的热点问题，并根据提取出对应话题内所包含的热点词语，将每条留言索引出来，最后根据题目所规范的要求把相应的文件输出。

	A	B	C	D	E	F
1	热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
2	1	1	50.8	2019/7/7至2019/9/1	A市伊景园滨河苑	A市伊景园滨河苑项目捆绑销售车位
3	2	1	42.5	2019/11/2至2020/1/26	A2区丽发新城小区附近	小区附近搅拌站噪音扰民
4	3	1	29.5	2019/7/21至2019/12/4	A市A5区魅力之城小区	小区临街餐饮店油烟噪音扰民
5	4	1	17.2	2019/8/1	A9市市民	关于高铁站选址
6	5	1	15.6	2017/6/8至2019/11/27	A市经济学院学生	学校强制学生去定点企业实习

图 15 热点问题表

	A	B	C	D	E	F	G	H
1	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
2	1	212323	A00020702	广铁集团要求员工购房时必须同时购买车位	2019-07-11 00:00:00	一套房子的同时也必须一对一购买售价12万的车	0	0
3	1	213584	A909172	投诉A市伊景园滨河苑定向限价商品房违规涨价	2019-07-28 13:09:08	售车位12万/户。无视法律法规，无视民生，坑	0	0
4	1	224767	A909176	伊景园滨河苑车位捆绑销售！广铁集团做个人吧！	2019-07-30 14:20:08	还不给我合同，说什么预购不用！后面就没有消	0	0
5	1	188801	A909180	投诉滨河苑针对广铁职工购房的霸王规定	2019-08-01 00:00:00	指出的问题一个接着一个，首先未取得预售资格	0	0
6	1	285897	A909191	武广新城伊景园滨河苑违法捆绑销售车位,求解决	2019-08-01 20:06:52	情况下,广铁集团就要求捆绑购买车位,还说是成	0	0
7	1	251601	A909187	A市伊景园滨河苑诈骗钱财	2019-08-01 22:42:21	者18万5千元的认购金后不与购房者签订合同，	0	0
8	1	209506	A909179	A市武广新城坑客户购房金额并且捆绑销售车位	2019-08-02 16:36:23	付款，不签合同的那种，工作人员又不说清楚，	0	0
9	1	205982	A909168	坚决反对伊景园滨河苑强制捆绑销售车位	2019-08-03 10:03:10	格，这是明显的违法捆绑销售！通过购买车位	0	2
10	1	276016	A909181	车位属于业主所有，不应该被捆绑销售！	2019-08-06 00:00:00	资格。我们查阅了《物权法》的相关规定，发	0	2
11	1	280774	A909199	反馈广铁集团铁路职工定向商品房的一些问题	2019-08-10 12:23:19	合法权益无法得到保证。今有捆绑销售车位，	0	0

图 16 热点问题留言明细表

2.6 模型实现评价

- 优点：
  ①能够通过筛选出的热词，直观地看出近期群众所关注的热点；
  ②Attention 机制能够挑重点，解决之前长距离的信息会被弱化的问题，它可以在长文本中抓住重点，不丢失重要的信息。
- 缺点：

①由于数据量较大，在进行地点和人群信息提取时，会出现提取出的信息不足等问题；

②将热度词语连接成关系时，准确率比较低，很多相关的词语不能够连接在一起；

③对于每个话题下所对应的留言，存在多个话题的问题。

3 答复意见的评价

3.1 题目分析

在文本分析中，往往趋向于选择错误率较小和准确率较高的高精度模型。但是，复杂的模型往往难以解释，无法获知所有产生模型预测结果的特征之间的关系，因此只能用准确率、错误率这样的评价标准来代替。针对第三题，评价标准则是由相关性、完整性和可解释性来构成。

3.2 相关性

针对留言与答复意见的相关性，可以通过分别对两段文本进行分词、洗词等操作，并对比词典中高词频词语的出现的数量，以及相同词语出现在词典中的占比。并根据各自的占比，计算两段文本的相关性。

使用的模型为 BM25 算法。BM25 算法是二元独立模型的拓展，是一种用于评价给定词语和文档间相关性的算法。它是基于概念检索模型提出的，考察词语在查询中的权值，拟合出综合上述考虑因素的公式，并通过引入一些经验参数。基本公式为：

$$\sum_{i\in Q}\log\frac{(ri+0.5)/(R-ri+0.5)}{(ni-ri+0.5)/(N-ni-R+ri+0.5)}\times\frac{(k1+1)fi}{K+fi}\times\frac{(k2+1)qfi}{K2+qfi}$$

其中对于查询 Q 中的每个查询此，依次计算其在文档 D 中的分支，累加后就是文档 D 与查询 Q 的相关性得分。在计算每一个查询词的权值时，公式都可以拆解为 3 个组成部分：第 1 个组成部分就是模型计算得分，第 2 个组成部分是查询词在文档 D 中的权值，K1 和 K 是经验参数，第 3 个组成部分是查询词自身的权值，如果查询较短小的话，这个值往往是 1，k2 是经验参数值。BM25 模型就是融合了这 3 个计算因子的相关性计算公式。

3.2 完整性

针对留言与答复意见的完整性，可理解为对留言的要点答复的是否全面；如果相关性较高，则回答的内容就比较全面，各个点就会都有所涉及，因此，他答复的完整性就越高。当然，还要利用自然语言处理中的 parsing tree 对句子做 semantic parsing，来消除答复意见中由于断句/同义词等原因导致的语义不通顺。

对此问题，利用 jieba.analy 分析出每条留言以及答复意见的高频词，提取每个留言详情的前 N 个关键词，之后再对比对应答复意见的前 N 个关键词，查看高频词的匹配度，从而来判断答复意见的完整性。

通过对比大量的样本之后，选取 N=10 来对答复意见的完整性做判断。

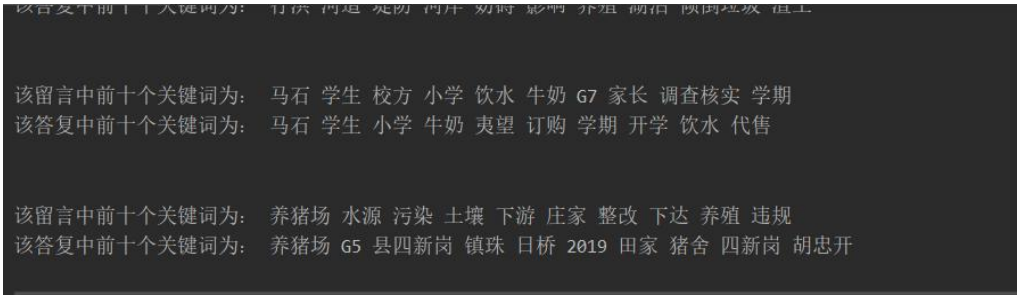

图 17 N=10 的关键词提取

3.3 可解释性

在解决问题时，往往更多地关注模型的性能指标，如准确性、精确度和召回率等等。但是指标只能说明模型预测决策的一部分，随着时间推移，由于环境中的各种因素导致的模型概念漂移，性能可能会发生变化。因此，注重模型的可解释性，可以更加透明地了解模型的决策过程，并有助于在模型上建立信任。

实际应用中，通常是先基于训练好的分类器得出测试样本的预测概率，然后将该测试样本的预测概率与给定的阈值进行比较，若该预测概率大于给定阈值，则将该测试样本划分为正类，反之则将其划分为反类。对于不同的分类任务，该分类阈值的取值也是不一样的。

ROC 曲线（The Receiver Operating Characteristic Curve）给出的是不同分类阈值情况下真正率（TPr）和假正率（FPr）的变化曲线。PR 曲线（Precision-Recall Curve）给出的是不同分类阈值情况下查准率（Precision）和查全率（Recall）的变化曲线。ROC 曲线相比 PR 曲线有一个非常好的特性：就是当正负样本分布发生变化的时候，ROC 曲线的形状能够基本保持不变，而 PR 曲线的形状会发生较剧烈的变化。为了使得 ROC 曲线之间能更好的进行比较，通常采用 AUC，即 ROC 曲线下的面积来衡量一个分类算法的性能。其中，AUC 的值越大，表明分类性能越好。

KS（Kolmogorov-Smirnov Curve）曲线横轴为不同的分类阈值，纵轴为真正率（TPr）和假正率（FPr）的变化曲线。KS 值=max|TPr-FPr|，等价于 Δ TPr= Δ FPr，这和 ROC 曲线上找最优阈值的条件一致。KS 值常在征信评分模型中用于衡量区分预测正负样本的分隔程度。一般来说，KS 值越大，表明正负样本区分的程度越好，说明模型区分度越高。

4 参考文献

[1]周志华.2015. 机器学习.清华大学出版社，4.2:75-79

[2]范淼，李超.Python 机器学习及实战.清华大学出版社，第 1 版，34-98

[3]尹裴，王洪伟.2019.基于 LDA 主题模型和领域本体的中文产品评论细粒度情感分析.同济大学出版社