

“智慧政务”中的文本挖掘应用

摘 要

近年来，随着互联网的广泛应用和网上政务的发展，微信、微博、市长信箱、阳光热线等网络问政平台逐步成为政府了解民意、汇聚民智、凝聚民气的重要渠道，各类社情民意相关的文本数据量不断攀升，给以往主要依靠人工来进行留言划分和热点整理的相关部门的工作带来了极大挑战。因此对“智慧政务”进行文本分析与数据挖掘具有重要意义。一方面可以造福百姓，便民利民。另一方面，可以推动提升政府的管理水平和施政效率。

对于问题 1：本文首先将附件 2 中的文本数据进行去重去空、中文分词、停用词过滤等文本数据预处理；然后分别对每个标签下的内容统计词频并绘制词云图；最后利用 TF-IDF 以及逻辑斯蒂回归方法构建一级分类模型，并计算模型的 $f1_score$ 值。

对于问题 2：本文首先将附件 3 的文本数据进行去重去空、crf 分词等文本数据预处理；然后对已经分词的文本进行“BMES0”词性标注；最后对标注好的文本利用 CRF 进行命名实体识别提取地名；以频率为热点判断的指标，统计各地区出现的次数得出热点地区，而后从热点地区中提取出热点问题，写入“热点问题表”和“热点问题留言明细表”。

对于问题 3：本文从附件 4 中提取出“留言详情”和“答复意见”中的部分文本数据进行研究，首先分别对文本进行去重去空、中文分词、去停用词等文本数据预处理；然后基于词频（TF）对文本分别进行向量化处理；最后基于余弦相关度计算出“留言详情”和“答复意见”的相关性，并定义答复质量评价标准。

关键词： TF-IDF；逻辑斯蒂回归；命名实体识别；余弦相似度

Application of text mining in "intelligent government affairs"

Abstract

In recent years, with the wide application of Internet and the development of online government affairs, WeChat, such as weibo, mayor mailbox, sun hotline network asked ZhengPing stage gradually become the government understand the importance of public opinion, gathering intelligence, condensed bull channels, all kinds of public opinion related text data quantity rising, leave a message to past mainly rely on artificial to divide and hot spots of relevant departments work has brought great challenge. Therefore, it is of great significance to carry out text analysis and data mining on "intelligent government affairs". On the one hand, it can benefit the people and benefit the people. On the other hand, it can promote the improvement of the government's management level and administrative efficiency.

For question 1: in this paper, the text data in annex 2 is firstly preprocessed by deduplication and nullification, Chinese word segmentation, stop word filtering and other text data. Then, the word frequency of content under each label is counted and the word cloud map is drawn. Finally, tf-idf and logistic regression method are used to construct the first-level classification model and calculate the f1_score value of the model.

For question 2: in this paper, the text data of attachment 3 is firstly preprocessed to deduplicate, devoid, CRF segmentation and other text data. Then, "BMESO" part-of-speech annotation is carried out on the text with participles. Finally, CRF is used to identify the annotated text and extract the place names. Taking frequency as the indicator of hot spot judgment, the frequency of occurrence in each region is counted to get the hot spot, and then the hot spot issues are extracted from the hot spot and written into the "hot spot issues table" and "hot spot issues message list".

For question 3: this paper extracts part of the text data in "message details" and "reply comments" from appendix 4 for research. Firstly, it carries out text data preprocessing such as reduplication and nullization, Chinese word segmentation, stop words and so on. Then the text is vectorized based on word frequency (TF). Finally, the correlation between "message details" and "reply comments" is calculated based on the cosine correlation degree, and the response quality evaluation standard is defined.

Key words: tf-idf; logistic regression; named entity identification; cosine similarity

目 录

1. 挖掘目标.....	1
2. 总体流程.....	1
3. 群众留言分类.....	3
3.1 数据预处理.....	3
3.2 绘制词云图.....	4
3.3 构建分类模型.....	6
3.4 模型评分.....	7
4. 热点问题挖掘.....	7
4.1 CRF 分词.....	8
4.2 词性标注.....	8
4.3 提取并统计地名频数.....	9
4.4 获取热点问题.....	11
5. 答复意见的评价.....	13
5.1 数据预处理.....	14
5.2 文本向量化.....	16
5.3 计算文本相似度.....	17
5.4 制定评价方案.....	18
6. 结论及存在的问题.....	19
参考文献:	20

1. 挖掘目标

在当今的互联网的高速发展中，我们已经生活在大数据时代里。近年来，各种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，给以往主要依靠人工来进行留言划分和热点整理的相关部门的工作带来了极大挑战。同时，随着大数据、云计算、人工智能等技术的发展，建立基于自然语言处理技术的智慧政务系统已经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具有极大地推动作用。

本文建模目标是对网络平台的留言数据进行基本的预处理, 中文分词, 停用词过滤后, 一方面为留言主题建立一级标签分类模型, 采用基于 TF-IDF 权重及 sklearn 框架下 LogisticRegression 分类算法, 对留言主题进行分类, 之后绘制词云图统计每个标签下的词频。第二方面是对留言主题进行分析, 得出群众反映问题最多的地区, 而后根据地名搜索出时间段以及反应问题。第三方面是统计留言详情和答复意见信息含量大的词频, 将其向量化后之后, 利用余弦定理计算两个向量的夹角大小, 判定文本相关性。

2. 总体流程

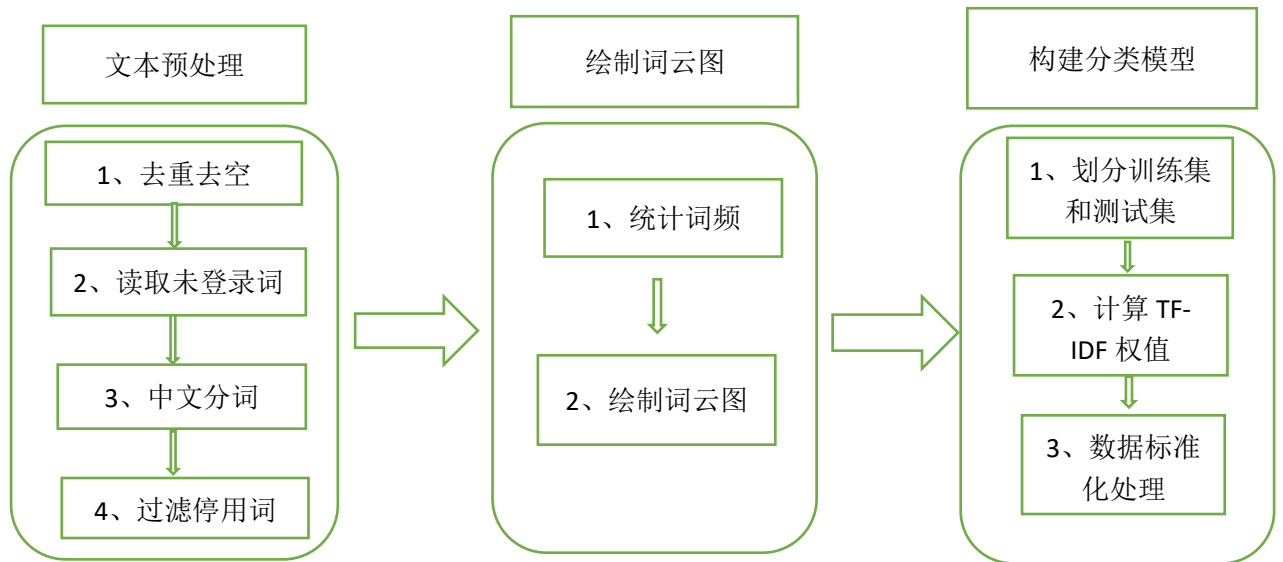


图 2-1 问题一流程图

针对问题一：第一步是文本预处理，首先对附件 2 中的“留言主题”进行去重去空，然后利用 jieba 库进行中文分词。分词过程中发现大量的小区名字被分开，通过搜查发现这些小区位于长沙，为提高分词的质量，本文从网站“安居客”中爬取了长沙的所有小区名，并写入未登录词文本文件，分词前先进行读取，能有效提高分词质量。最后过滤停用词。第二步是绘制词云图，统计各一级标签下内容的词频，利用 wordcloud 和 matplotlib.pyplot 绘制词云图。第三步是构建分类模型，首先对已分词的文本数据划分训练集和测试集，然后利用 TF-IDF 权值对文本进行向量化处理，进而对数据进行标准化处理，最后构建逻辑斯蒂回归分类模型并测试其 f1_score 值。

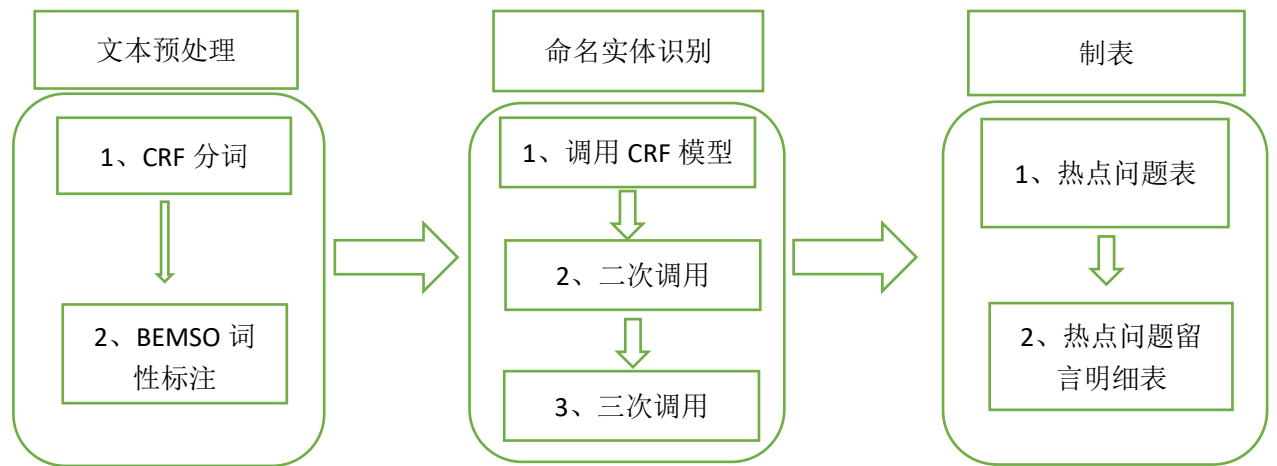


图 2-2 问题二流程图

针对问题二：第一步是文本预处理，本文对附件 3 中的“留言主题”利用 Hanlp 进行 CRF 分词，然后对已分词的文本通过“B”“E”“M”“S”“O”体系进行词性标注。第二步是命名实体识别，本文利用 CRF 模型进行命名实体识别，通过三次调用 CRF 模型，逐步缩小地区范围，最终得出热点地区。第三步是制表，获取热点地区后，可以从附件 4 中搜索出该热点地区的所有问题，然后通过统计归纳出出现频数较高的热点问题，并制作“热点问题表”和“热点问题留言明细表”。

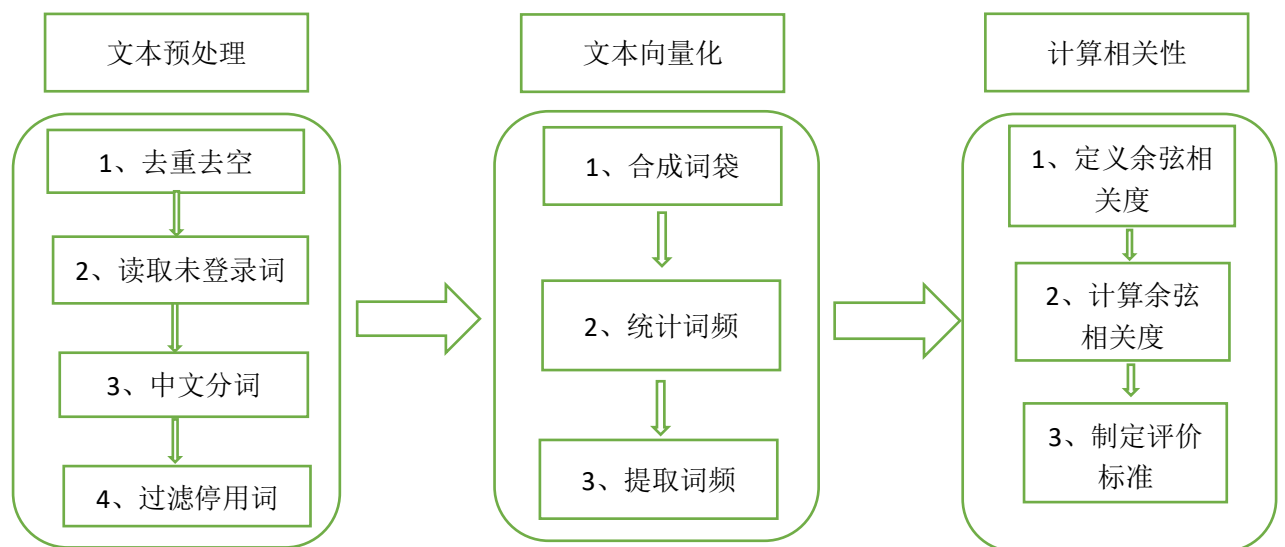


图 2-3 问题三流程图

针对问题三：本文从附件 4 的“留言详情”和“答复意见”中提取部分文本数据作为样本。第一步是文本预处理，对样本操作同问题一的第一步。第二步是文本向量化，首先对样本的分词结果进行合并得出词袋，然后分别统计出“留言详情”和“答复意见”样本在该词

袋中相应的词频，最后提取出各自的词频作为向量化处理后的数据。第三步是计算相关性，首先定义余弦相关度，然后利用向量化处理后的数据计算余弦相关度，多次抽取样本重复试验后，归纳总结出评价标准。

3. 群众留言分类

首先对文本进行数据预处理，因为文本中存在大量重复以及空格，因为对文本进行去重去空，由于文本词与词之间没有明显的界线，因而对附件 2 的留言主题进行中文分词，此后在文本中仍存在大量标点符号以及无意义信息，接着对文本进行过滤得到少量且高效有用信息；之后绘制词云图；最终构建分类模型对留言主题进行分类。

3.1 数据预处理

1. 去重去空

文本中有大量重复内容以及空白信息，利用 `data.drop_duplicates()` 去重，以及遍历判断的方法去空。

2. 中文分词

利用 python 中的中文分词模块——jieba 分词。

jieba 分词的算法：

（1）利用 Trie 树结构实现高效的词图扫描，生成语句中所有有可能构成词语情况所构成的有向无环图（DAG）。

（2）查找树结构中的最大概率路径，找出基于词频的最大切分组合

（3）对于未登录词，采用手动补充文本，读取文本的方法。

特别的，在本文中，附件 2，3，4 中存在大量小区名等未登录词，通过查询后发现其所在的城市为长沙，于是利用爬虫技术爬取网站“安居客”中长沙的所有小区名并写入 `newdict.txt` 文件。

0	[A, 市, 西湖建筑集团, 占道, 施工, 有, 安全隐患]
1	[A, 市, 在水一方大厦, 人为, 烂尾, 多年, , , 安全隐患, 严重]
2	[投诉, A, 市, A1, 区苑, 物业, 违规, 收, 停车费]
3	[A1, 区, 蔡锷, 南路, A2, 区华庭, 楼顶, 水箱, 长年, 不洗]
4	[A1, 区, A2, 区华庭, 自来水, 好大, 一股, 霉味]
5	[投诉, A, 市, 盛世, 耀凯, 小区, 物业, 无故, 停水]
6	[咨询, A, 市, 楼盘, 集中, 供暖, 一事]
7	[A3, 区, 桐梓, 坡, 西路, 可可, 小城, 长期, 停水, 得不到, 解决]
8	[反映, C4, 市, 收取, 城市, 垃圾处理, 费, 不, 平等, 的, 问题]
9	[A3, 区, 魏家坡, 小区, 脏乱差]
10	[A, 市, 魏家坡, 小区, 脏乱差]
11	[A2, 区, 泰华, 一村, 小区, 第四届, 非法, 业委会, 涉嫌, 侵占, 小区业主...]
12	[A3, 区, 梅溪湖, 壹号, 御湾, 业主, 用水, 难]
13	[A4, 区, 鸿涛翡翠湾, 强行, 对, 入住, 的, 业主, 关水, 限电]
14	[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...]
15	[A6, 区, 润和紫郡, 用电, 的, 问题, 能, 不能, 解决]
16	[A, 市, 锦楚国际新城, 从, 6, 月份, 开始, 停电, 好, 多次, 了]
17	[给, A9, 市, 城区, 南西, 片区, 城铁, 站, 设立, 的, 建议]
18	[请, A6, 区政府, 加大, 对滨水, 新城, 的, 绿化, 建设]
19	[A5区, 楚府, 线, 几个, 小区, 经常, 停电]
20	[请, 调查, 西地省, 建望, 集团, 及, 西地省, 辉, 东安, 建, 工程, 有限公...]
21	[A2, 区, 山水嘉园, 1, 栋, 三, 单元, 群, 租房, 扰民]

图 3-1-1 部分分词结果

从图 3.2 中可以看到其中含有大量标点符号以及表达无意义的词,给后续统计分析带来干扰,因而利用停用表过滤停用词。在选用停用表应选取合适的停用词,停用词有两个特征:

(2) 包含信息量低, 对文本识别无意义, 如助词, 形容词, 连接词等。

0 [占道，施工，安全隐患]
1 [烂尾，安全隐患]
2 [物业，停车费]
3 [蔡锷，区华庭，楼顶，水箱，不洗]
4 [区华庭，自来水，好大，霉味]
5 [物业，停水]
6 [楼盘，供暖]
7 [桐梓，小城，停水]
8 [收取，城市，垃圾处理，平等]
9 [魏家坡，脏乱差]
10 [魏家坡，脏乱差]
11 [业委会，小区业主，资金]
12 [梅溪湖，御湾，业主，用水]
13 [鸿涛翡翠湾，入住，业主，关水，限电]
14 [地铁，施工，停电]
15 [用电]
16 [停电]
17 [城区，南西，城铁]
18 [对滨水，绿化]
19 [楚府，线，停电]
20 [建望，工程，有限公司，违法行为]
21 [山水嘉园，租房，扰民]

3.2 绘制词云图

首先统计每个一级标签对应内容的词频，然后利用 WordCloud 库和 matplotlib.pyplot 库绘制出词云图，将每个一级标签下的内容词频可视化，词云图实例如下图所示：



4



图 3-2-2 环境保护词云图



图 3-2-3 交通运输词云图



图 3-2-4 教育文体词云图



图 3-2-5 劳动和社会保障词云图



图 3-2-6 商贸旅游词云图



图 3-2-7 卫生计生词云图

3.3 构建分类模型

1. 利用 CountVectorizer 模块将文本向量化处理，然后调用 TfidfTransformer 进行预处理。

(1) CountVectorizer 用于将文本中的词语转换为词频矩阵,通过 fit_transform 函数计算各个词语出现的次数。

(2) TfidfTransformer 用于统计每个词语的 TF-IDF 值。

词频 (TF): 用于衡量一个词在文档中出现频率。

TF=某词出现次数/总次数

逆向文档频率 (IDF): 用于衡量一个词的重要性。

IDF = $\log(\text{语料库的文档总数} / (\text{含有某词的文档数} + 1))$

TF-IDF 权值 = TF * IDF

2. 利用 StandardScale 对特征数据标准化处理。

$$Z_{ij} = \frac{X_{ij} - X_i}{S_i}$$

其中: Z_{ij} 为标准化后的变量值; X_{ij} 为实际变量值; X_i 为原始数据的均值; S_i 为原始数据的标准差。

3. 利用逻辑斯蒂回归 (LogisticRegression) 构建分类模型。

逻辑回归模型通过构造一个二分类模型逻辑斯蒂回归方程对未知的请求进行分类,其中包含取样,特征选取,参数拟合,测试。

3.4 模型评分

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

其中 P_i 为第 i 类的精确率, R_i 为第 i 类的召回率。

模型预测结果如图 3.4.1 所示:

	精确率	召回率	f1_score	score
结果	0.804056	0.801303	0.801546	0.801303

图 3.4.1 模型 f1_score 评分

4. 热点问题挖掘

某一时段内群众集中反映的某一问题可称为热点问题,本文以出现的频数作为热度指数,为了找出排名前 5 的热点问题,需要统计出附件 3 中哪个时间段以及地点出现的频率最高,则需要命名实体识别。首先是对附件 3 中的留言主题进行 Hanlp-crf 分词,之后将分词得出的结果进行“BEMSO”标注,再用过调用 CRF 模型提取地名并统计频数。最后获取出现高频

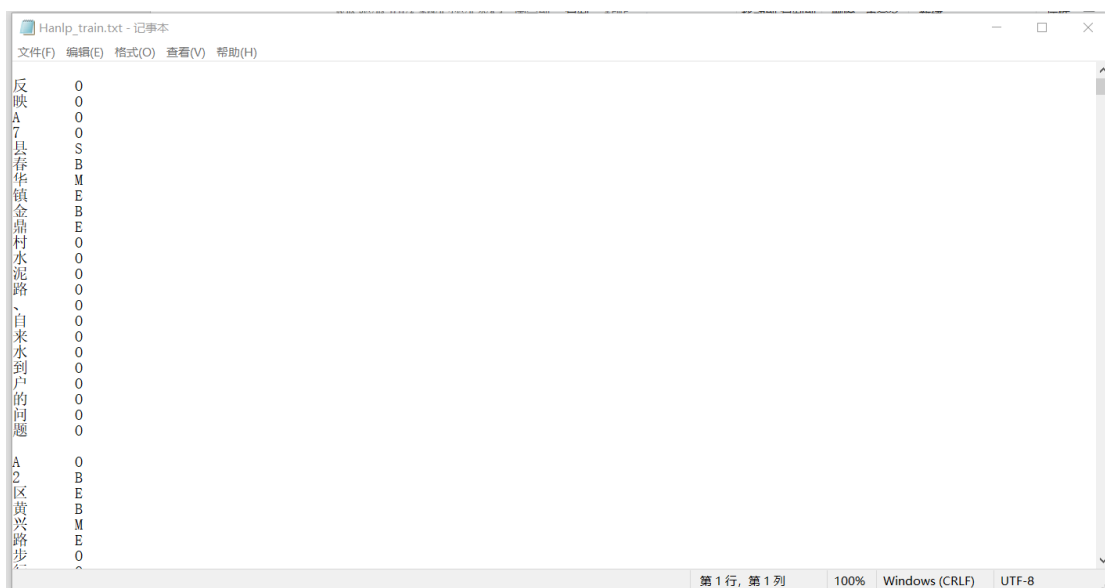


图 4-2-1 Hanlp_train.txt 中的部分标注结果

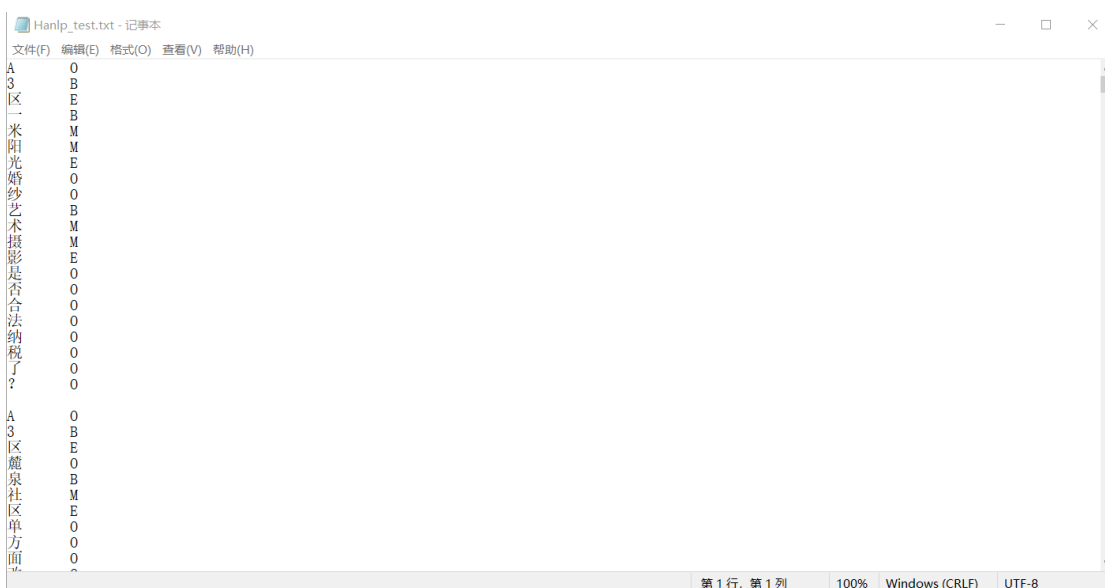


图 4-2-2 Hanlp_test.txt 中的部分标注结果

4.3 提取并统计地名频数

1. 构建 CRF 模型

CRF 模型的原理:

$$P(\theta | O | W) = \frac{1}{Z_W} \left\{ \sum_{n=1}^N \sum_k f_k(O_{n-1}, O_n, W, n) \right\}$$

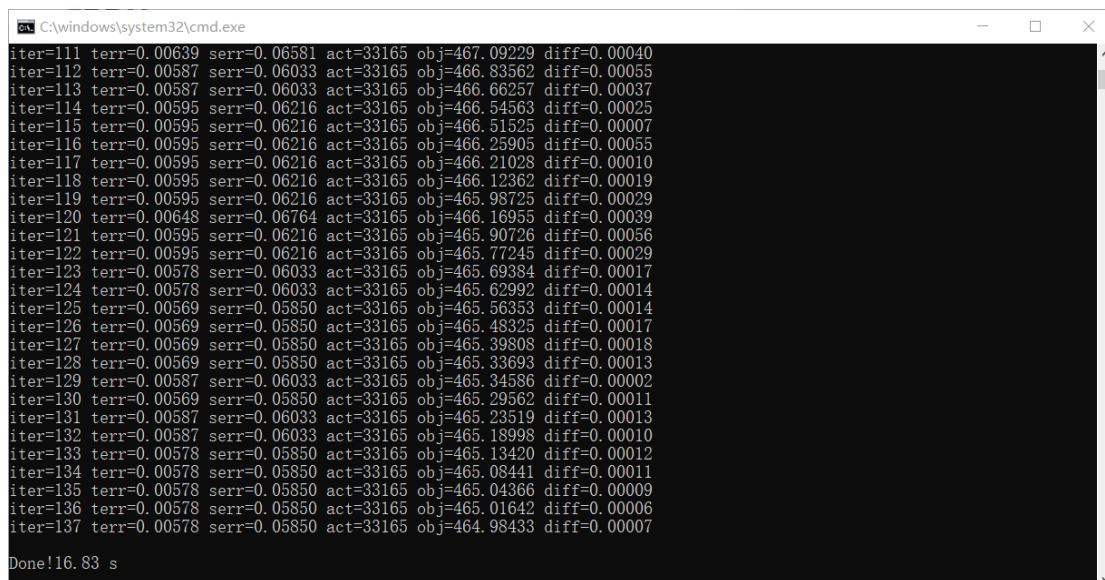
$$Z_W = \sum_o \exp \left\{ \sum_{n=1}^N \sum_k f_k(O_{n-1}, O_n, W, n) \right\}$$

其中, Z_W 是归一化参数, 它使得给定输入的所有可能状态序列的概率之和为 1。 $f_k(O_{n-1}, O_n, W, n)$ 是对于整个观察序列, 标记位于 N 和 $N-1$ 之间的特征函数, 特征函数可以是 0,1 值, 也可以是任意实数 $\theta = \theta_1 \theta_2 \dots \theta_k$ 是特征函数对应的权重。对与 W 来说,

目标是搜索概率最大的 $O^* = \text{argmax}_O P(W|O)$ 。

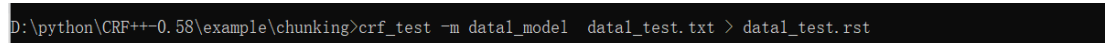
事先安装好 CRF++-0.58 的相关组件后，利用命令行构建 CRF 模型。

CRF 模型的训练结果以及测试结果如图 4-2-4 训练模型，图 4-2-5 测试模型所示



```
C:\windows\system32\cmd.exe
iter=111 terr=0.00639 serr=0.06581 act=33165 obj=467.09229 diff=0.00040
iter=112 terr=0.00587 serr=0.06033 act=33165 obj=466.83562 diff=0.00055
iter=113 terr=0.00587 serr=0.06033 act=33165 obj=466.66257 diff=0.00037
iter=114 terr=0.00595 serr=0.06216 act=33165 obj=466.54563 diff=0.00025
iter=115 terr=0.00595 serr=0.06216 act=33165 obj=466.51525 diff=0.00007
iter=116 terr=0.00595 serr=0.06216 act=33165 obj=466.25905 diff=0.00055
iter=117 terr=0.00595 serr=0.06216 act=33165 obj=466.21028 diff=0.00010
iter=118 terr=0.00595 serr=0.06216 act=33165 obj=466.12362 diff=0.00019
iter=119 terr=0.00595 serr=0.06216 act=33165 obj=465.98725 diff=0.00029
iter=120 terr=0.00648 serr=0.06764 act=33165 obj=466.16955 diff=0.00039
iter=121 terr=0.00595 serr=0.06216 act=33165 obj=465.90726 diff=0.00056
iter=122 terr=0.00595 serr=0.06216 act=33165 obj=465.77245 diff=0.00029
iter=123 terr=0.00578 serr=0.06033 act=33165 obj=465.69384 diff=0.00017
iter=124 terr=0.00578 serr=0.06033 act=33165 obj=465.62992 diff=0.00014
iter=125 terr=0.00569 serr=0.05850 act=33165 obj=465.56353 diff=0.00014
iter=126 terr=0.00569 serr=0.05850 act=33165 obj=465.48325 diff=0.00017
iter=127 terr=0.00569 serr=0.05850 act=33165 obj=465.39808 diff=0.00018
iter=128 terr=0.00569 serr=0.05850 act=33165 obj=465.33693 diff=0.00013
iter=129 terr=0.00587 serr=0.06033 act=33165 obj=465.34586 diff=0.00002
iter=130 terr=0.00569 serr=0.05850 act=33165 obj=465.29562 diff=0.00011
iter=131 terr=0.00587 serr=0.06033 act=33165 obj=465.23519 diff=0.00013
iter=132 terr=0.00587 serr=0.06033 act=33165 obj=465.18998 diff=0.00010
iter=133 terr=0.00578 serr=0.05850 act=33165 obj=465.13420 diff=0.00012
iter=134 terr=0.00578 serr=0.05850 act=33165 obj=465.08441 diff=0.00011
iter=135 terr=0.00578 serr=0.05850 act=33165 obj=465.04366 diff=0.00009
iter=136 terr=0.00578 serr=0.05850 act=33165 obj=465.01642 diff=0.00006
iter=137 terr=0.00578 serr=0.05850 act=33165 obj=464.98433 diff=0.00007
Done!16.83 s
```

图 4-3-1 训练模型



```
D:\python\CRF++-0.58\example\chunking>crf_test -m datal_model datal_test.txt > datal_test.rst
```

图 4-3-2 测试模型

CRF 模型的原理：

$$P_{\theta}(O|W) = \frac{1}{Z_W} \left\{ \sum_{n=1}^N \sum_k f_k(O_{n-1}, O_n, W, n) \right\}$$
$$Z_W = \sum_o \exp \left\{ \sum_{n=1}^N \sum_k f_k(O_{n-1}, O_n, W, n) \right\}$$

其中, Z_W 是归一化参数，它使得给定输入的所有可能状态序列的概率之和为 1。 $f_k(O_{n-1}, O_n, W, n)$ 是针对整个观察序列，标记位于 N 和 $N-1$ 之间的特征函数，特征函数可以是 0,1 值，也可以是任意实数 $\theta = \theta_1 \theta_2 \dots \theta_k$ 是特征函数对应的权重。对与 W 来说，目标是搜索概率最大的 $O^* = \text{argmax}_O P(W|O)$ 。

2. 调用模型

通过 python 代码调用 CRF 模型，将附件 3 所有数据放入模型，提取地名并统计频数。

7县	575
3区	267
4区	230
2区	221
西地	210
1区	196
5区	160
星沙	120
6区	106
县	102
8县	70
公交车	68
大道	60
安全隐患	57
街道	50
9市	47
不合理	45
泉塘	41
社区	39
楚龙	26

图 4-3-3 地名统计频数

不难发现，地名频数较大的关键字分别为“7 县，3 区，4 区，2 区，西地，1 区，5 区，星沙，6 区，8 县”。

4.4 获取热点问题

提取出附件 3 中含有以上地名的数据并写成 txt 文件，部分结果如图 4.3.1 所示：



图 4-4-1 “7 县”部分统计结果

图 4.3.1 是经过第一次提取的结果，不难发现在此次提取出的地域范围太大，所以重复步骤（一）到（三）进行第二次提取，得到频数较大的地名关键字为“星沙，梅溪湖，泉塘，丽发新城”等；因为该阶段提取到的地理位置还不够准确，进而再重复（一）到（三）的步骤进行第三次提取，最终得到出现频数前五的地区关键字为“丽发新城、魅力之城、旧城、月亮岛、松雅”。下图所示的是“7 县”→“星沙”→“旧城”的地

名识别过程，其他地区做法相同。

7县	575
星沙	100
县	94
泉塘	33
楚龙	26
街道	21
泉塘街道	19
大道	18
松雅湖	15
公交车	13
北山镇	12
黄兴镇	11
星沙镇	11
春华镇	11
凉塘路	11
东六路	10
华庭	10
旧城	10
星沙街道	10
星城	9
安全隐患	9
六路	9

图 4-4-2 对“7 县”第二次提取部分结果

星沙	120
7县	110
星沙镇	11
星沙街道	10
旧城	9
凉塘路	9
大道	8
县	6
中茂城	4
四区	4
星城	3
文化公园	3
联络线	3
恒大	3
华庭	3
9县	3
公交车	2
星湖	2
圣力华苑	2
摆摊设点	2

图 4-4-3 对“星沙”第三次提取部分结果

A市星沙城区旧城区棚户改造项目范围是什么？
咨询A7县星沙旧城改造项目问题
反映A7县星沙一区14栋旧城改造之痛
请问A7县星沙凉塘路的旧城改造要拖到何时何月何时才能再次启动？
A7县星沙四区凉塘路的旧城改造要拖到何时？
A7县星沙凉塘路旧城改造究竟要拖到何年何月才能开始？
A7县星沙街道凉塘路的旧城改造什么时候会启动？
A7县星沙街道凉塘路旧城改造什么时候可以进行
A7县星沙四区凉塘路旧城改造要待何时
A7县星沙四区凉塘路旧城改造要拖到何年何月才能动工

图 4-4-4 “旧城”中的部分问题

分别获取附件 3 中以上地区的所有留言主题，由于数量较少，所以进行人工统计，得出热度指数前五的热点问题，并从附件 3 中搜索排除得出时间范围以及其他数据，写入“热点问题表.xls”。

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
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	53	2019/11/02至2020/01/26	A市A2区丽发新城	丽发新城小区附近搅拌站噪音扰民和污染环境									
2	2	21	2019/07/21至2019/09/25	A市A5区魅力之城	魅力之城小区一楼教育楼扰民和污染空气									
3	3	9	2019/02/14至2019/09/19	A7县里沙西区本德路	A7县里沙西区本德路旧城改造尚未进行									
4	4	8	2019/03/26至2019/04/15	A市A6区月亮岛路	关于A6区月亮岛路110m高压线的投诉建议									
5	5	6	2019/04/17至2019/09/06	A市地铁3号线松雅西地省站	A市地铁3号线松雅西地省站出入口设计不合理									

图 4-4-5 热点问题表

最终通过关键字搜索在附件 3 中通过找出具体数据并写入“热点问题留言明细表.xls”

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述	回复内容	回复时间	回复状态	回复类型	回复来源	回复渠道	回复平台	回复部门	回复人员	回复电话	回复邮箱	回复地址	回复邮编	回复城市	回复国家	回复备注
1	203393	A0005306	A市丽发新城	2019/11/19 14:51	尊敬的领导:	0	2														
2	206285	A909205	投诉小区	2019/12/15 12:32	尊敬的领导:	0	24														
3	208714	A0004201	A2区丽发新城	2020/1/2 0:00	尊敬的领导:	0	4														
4	213464	A909233	投诉丽发新城	2019/12/10 12:34	我是暮云	0	0														
5	213930	A909218	A2区丽发新城	2019/12/27 23:34	A市	0	0														
6	214282	A909209	A市丽发新城	2020/1/25 9:07	你们管不管?	0	0														
7	215563	A909231	A2区丽发新城	2019/12/6 12:21	领导,您好!	0	0														
8	215842	A909210	A2区丽发新城	2020/1/26 19:47	请领导领导	0	0														
9	216824	A909214	投诉丽发新城	2019/12/25 12:15	最近一段	0	0														
10	217700	A909239	丽发新城	2019/12/21 2:33	开发商把!	0	1														
11	222831	A909228	噪音、灰	2019/12/22 10:23	A2区丽发新城	0	0														
12	225217	A909223	A2区丽发新城	2019/11/15 9:17	我已经好	0	0														
13	231136	A909204	投诉A2区	2019/12/2 11:20	尊敬的领导:	0	0														
14	232158	A909242	丽发新城	2019/12/5 8:46	本人是暮云	0	0														
15	235362	A909215	暮云街道	2020/1/6 20:45	暮云街道	0	0														
16	238212	A909203	丽发新城	2019/12/12 10:23	请问在居民	0	0														
17	238336	A909213	A市A2区	2019/12/11 11:44	尊敬的领导:	0	0														
18	239648	A909211	A市A2区	2020/1/6 22:41	丽发新城	0	0														
19	243692	A909201	丽发新城	2019/11/15 11:23	领导您好!	0	2														
20	244335	A909135	A市暮云街	2019/12/2 12:11		0	0														
21	244512	A0009470	投诉丽发新城	2019/12/25 20:57	我是暮云	0	1														
22	247160	A0001041	A市丽发新城	2019/11/25 19:08		0	0														
23	253040	A909202	投诉A2区	2019/12/4 12:10	投诉A2区	0	0														
24	255008	A909208	投诉小区	2019/11/18 12:23	暮云街道	0	0														
25	255276	A909219	再次希望!	2019/12/11 0:00	尊敬的领导:	0	0														
26	257091	A0007843	反对搅拌站	2019/12/13 0:00	尊敬的领导:	0	2														
27	258242	A909220	A市暮云街	2019/12/2 12:23	A市暮云街	0	0														
28	258378	A0008422	丽发新城	2019/11/23 0:00	开发商在!	0	0														
29	259788	A909221	A市暮云街	2019/12/7 0:00	A市暮云街	0	0														
30	260979	A909229	反应A市暮云街	2019/12/4 14:21	A市A2区	0	0														

图 4-4-6 热点问题留言明细表部分数据

5. 答复意见的评价

本文抽取附件 4 留言详情和答复意见中的部分数据记作 text1 和 text2 进行研究。判断留言详情与答复意见的相关性，可以将两个文本数据 text1,text2 向量化处理之后，通过其夹

角判断相关性。利用 jieba 分词分别对 text1 和 text2 进行分词，去除停用词，再将去除停用词之后的两个列表 text1_after_stop, text2_after_stop 合并成一个词袋 all_word, 再进行去重；接着再将去重后的词袋转换成初始字典 all_word_dict, 从而利用初始字典分别统计去除停用词之后两个文本的词频 text1_dict, text2_dict, 将得到词频值分别提取出来合成 ndarray 类型，得到两个向量 vec_a, vec_b, 根据余弦定理求取夹角，可得出文本相似度。

本文将利用两个文本数据解释说明：

留言详情：text1 = '2019 年 4 月以来，位于 A 市 A2 区桂花坪街道的 A2 区公安分局宿舍区（景蓉华苑）出现了一番乱象，该小区的物业公司美顺物业扬言要退出小区，因为小区水电改造造成物业公司的高昂水电费收取不了（原水电在小区买，水 4.23 一吨，电 0.64 一度）所以要通过征收小区停车费增加收入，小区业委会不知处于何种理由对该物业公司一再挽留，而对业主提出的新应聘的物业公司却以交 20 万保证金，不能提高收费的苛刻条件拒之门外，业委会在未召开全体业主大会的情况下，制定了一高昂收费方案要各业主投票，而投票不采用投票箱只制定表格要物业公司人员这一利害关系机构负责组织，对投票业主隐私权没有任何保护，还对投反对票的业主以领导做工作等方式要求改变为同意票，这种投票何来公平公正公开，面对公安干警采用这种方式投票合法性在哪？'

答复意见：text2 = '现将网友在平台《问政西地省》栏目向胡华衡书记留言反映“A2 区景蓉花苑物业管理有问题”的调查核实情况向该网友答复如下：您好，首先感谢您对我们工作的信任和支持，关于您在平台栏目给胡华衡书记留言，反映“A2 区景蓉花苑物业管理有问题”的情况已收悉。现将我们调查处理情况答复如下：经调查了解，针对来信所反映的“小区停车收费问题”，景蓉华苑业委会于 2019 年 4 月 10 日至 4 月 27 日以“意见收集方式”召开了业主大会，经业委会统计，超过三分之二的业主同意收取停车管理费，在业主大会结束后业委会也对业主提出的意见和建议进行了认真梳理归纳并进行了反馈，业委会制定的停车收费标准不高于周边小区价格。针对来信所反映的“物业公司去留问题”，5 月 5 日下午，辖区桂花坪街道牵头组织社区、物业公司、业主委员会、业主代表的会议，区住房和城乡建设局也参加了会议。在综合各方面的意见后，辖区桂花坪街道、区住房和城乡建设局已要求业委会依法依规召开业主大会，根据业主大会的表决结果再执行相应的程序。再次感谢您对我区工作的理解和关心。2019 年 5 月 9 日'

（即为“附件 4”中除标题外的第一行数据。）

5.1 数据预处理

1. 分词，去除停用词

利用 jieba 分词工具对 text1, text2 分词，而后对分词结果进行去除停用词处理，保存为 text1_after_stop, text2_after_stop 部分分词结果如下图所示：

```

['桂花',
'街道',
'公安分局',
'宿舍区',
'景蓉华苑',
'物业公司',
'美顺',
'物业',
'扬言',
'退出',
'水电',
'改造',
'物业公司',
'高昂',
'水电费',
'收取',
'水电',
'买水',
'4.23',
'一吨',
'电',
'0.64',

```

图 5-1-1 留言详情 text1_after_stop 部分分词结果

```

000[200].
['网友',
'平台',
'问政',
'栏目',
'胡华衡',
'留言',
'区景蓉',
'花苑',
'物业管理',
'调查核实',
'网友',
'答复',
'感谢您',
'信任',
'支持',
'平台',
'栏目',
'胡华衡',
'留言',
'区景蓉',
'花苑',
'物业管理',

```

图 5-1-2 答复意见 text2_after_stop 部分分词结果

2. 形成词袋

将把去除停用词后的两个列表 text1_after_stop,text2_after_stop 合并成一个词袋 all_word,对 all_word 去重，all_word 部分数据如下图所示：

```
[
    '高于',
    '表决',
    '收费',
    '信任',
    '机构',
    '业委会',
    '胡华衡',
    '社区',
    '代表',
    '平台',
    '一吨',
    '水电',
    '制定',
    '征收',
    '日至',
    '收集',
    '调查核实',
    '电',
    '日以',
    '停车',
    '坦中'
]
```

图 5-1-3 all_word 中的部分词语

5.2 文本向量化

1. 词袋转为字典

将 all_word 转换成初始字典类型 text_dict（即所有值为 0），某一留言详情与其答复意见组成的词袋的初始字典如图所示：

```
{
    '高于': 0, '表决': 0, '收费': 0, '信任': 0, '机构': 0, '业委会': 0, '胡华衡': 0, '社区': 0,
    '代表': 0, '平台': 0, '一吨': 0, '水电': 0, '制定': 0, '征收': 0, '日至': 0, '收集': 0, '调查核实': 0,
    '电': 0, '日以': 0, '停车': 0, '提出': 0, '票': 0, '住房': 0, '改造': 0, '美顺': 0,
    '结束': 0, '桂花': 0, '价格': 0, '挽留': 0, '网友': 0, '一度': 0, '归纳': 0, '业主': 0, '超过': 0,
    '增加收入': 0, '退出': 0, '问政': 0, '买水': 0, '反对票': 0, '公安干警': 0, '疏理': 0,
    '牵头': 0, '公正': 0, '理解': 0, '物业': 0, '方式': 0, '宿舍区': 0, '何种': 0, '何来': 0, '水电费': 0,
    '苛刻': 0, '管理费': 0, '拒之门外': 0, '物业公司': 0, '停车费': 0, '扬言': 0, '关心': 0,
    '收取': 0, '保证金': 0, '依规': 0, '采用': 0, '应聘': 0, '花苑': 0, '委员会': 0, '我区': 0,
    '收悉': 0, '业主大会': 0, '去留': 0, '栏目': 0, '以交': 0, '0.64': 0, '表格': 0, '物业管理': 0,
    '辖区': 0, '感谢您': 0, '合法性': 0, '会议': 0, '统计': 0, '城乡': 0, '区景蓉': 0, '街道': 0,
    '利害关系': 0, '建设局': 0, '4.23': 0, '答复': 0, '留言': 0, '隐私权': 0, '公安分局': 0,
    '高昂': 0, '反馈': 0, '面对': 0, '周边': 0, '投票箱': 0, '理由': 0, '投': 0, '综合': 0,
    '投票': 0, '景蓉华苑': 0, '支持': 0, '来信': 0, '意见': 0
}
```

图 5-2-1 某一词袋的初始字典

2. 统计词频

根据 text_dict 统计 text1_after_stop 和 text2_after_stop 的词频，记为 text1_dict, text2_dict，词频统计结果如下图所示：

```
{'高于': 0, '表决': 0, '收费': 2, '信任': 0, '机构': 1, '业委会': 2, '胡华衡': 0, '社区': 0,
'代表': 0, '平台': 0, '一吨': 1, '水电': 2, '制定': 2, '征收': 1, '日至': 0, '收集': 0, '调查
核实': 0, '电': 1, '日以': 0, '停车': 0, '提出': 1, '票': 1, '住房': 0, '改造': 1, '美顺': 1,
'结束': 0, '桂花': 1, '价格': 0, '挽留': 1, '网友': 0, '一度': 1, '归纳': 0, '业主': 4, '超
过': 0, '增加收入': 1, '退出': 1, '问政': 0, '买水': 1, '反对票': 1, '公安干警': 1, '疏理': 0,
'牵头': 0, '公正': 1, '理解': 0, '物业': 1, '方式': 2, '宿舍区': 1, '何种': 1, '何来': 1, '水
电费': 1, '苛刻': 1, '管理费': 0, '拒之门外': 1, '物业公司': 5, '停车费': 1, '扬言': 1, '关
心': 0, '收取': 1, '保证金': 1, '依规': 0, '采用': 2, '应聘': 1, '花苑': 0, '委员会': 0, '我
区': 0, '收悉': 0, '业主大会': 1, '去留': 0, '栏目': 0, '以交': 1, '0.64': 1, '表格': 1, '物
业管理': 0, '辖区': 0, '感谢您': 0, '合法性': 1, '会议': 0, '统计': 0, '城乡': 0, '区景蓉': 0,
'街道': 1, '利害关系': 1, '建设局': 0, '4.23': 1, '答复': 0, '留言': 0, '隐私权': 1, '公安分
局': 1, '高昂': 2, '反馈': 0, '面对': 1, '周边': 0, '投票箱': 1, '理由': 1, '投': 1, '综合':
0, '投票': 5, '景蓉华苑': 1, '支持': 0, '来信': 0, '意见': 0}
```

图 5-2-2 text1_dict 词频统计结果

```
{'高于': 1, '表决': 1, '收费': 2, '信任': 1, '机构': 0, '业委会': 5, '胡华衡': 2, '社区': 1,
'代表': 1, '平台': 2, '一吨': 0, '水电': 0, '制定': 1, '征收': 0, '日至': 1, '收集': 1, '调查
核实': 1, '电': 0, '日以': 1, '停车': 3, '提出': 1, '票': 0, '住房': 2, '改造': 0, '美顺': 0,
'结束': 1, '桂花': 2, '价格': 1, '挽留': 0, '网友': 2, '一度': 0, '归纳': 1, '业主': 4, '超
过': 1, '增加收入': 0, '退出': 0, '问政': 1, '买水': 0, '反对票': 0, '公安干警': 0, '疏理': 1,
'牵头': 1, '公正': 0, '理解': 1, '物业': 0, '方式': 1, '宿舍区': 0, '何种': 0, '何来': 0, '水
电费': 0, '苛刻': 0, '管理费': 1, '拒之门外': 0, '物业公司': 2, '停车费': 0, '扬言': 0, '关
心': 1, '收取': 1, '保证金': 0, '依规': 1, '采用': 0, '应聘': 0, '花苑': 2, '委员会': 1, '我
区': 1, '收悉': 1, '业主大会': 4, '去留': 1, '栏目': 2, '以交': 0, '0.64': 0, '表格': 0, '物
业管理': 2, '辖区': 2, '感谢您': 2, '合法性': 0, '会议': 2, '统计': 1, '城乡': 2, '区景蓉': 2,
'街道': 2, '利害关系': 0, '建设局': 2, '4.23': 0, '答复': 2, '留言': 2, '隐私权': 0, '公安分
局': 0, '高昂': 0, '反馈': 1, '面对': 0, '周边': 1, '投票箱': 0, '理由': 0, '投': 0, '综合':
1, '投票': 0, '景蓉华苑': 1, '支持': 1, '来信': 2, '意见': 3}
```

图 5-2-3 text2_dict 词频统计结果

3. 词频转换为数组类型

根据 2 中的结果可以将得到的词频的数值提取出合并成 ndarray 类型，vec_a 和 vec_b 分别为 text1_dict 和 text2_dict 的转换结果，即为文本向量化的结果。

```
[0 0 2 0 1 2 0 0 0 0 1 2 2 1 0 0 0 1 0 0 1 1 0 1 1 0 1 0 1 0 4 0 1 1 0
1 1 1 0 0 1 0 1 2 1 1 1 1 1 0 1 5 1 1 0 1 1 0 2 1 0 0 0 0 1 0 0 1 1 1 0 0
0 1 0 0 0 0 1 1 0 1 0 0 1 1 2 0 1 0 1 1 1 0 5 1 0 0 0]
```

图 5-2-4 vec_a

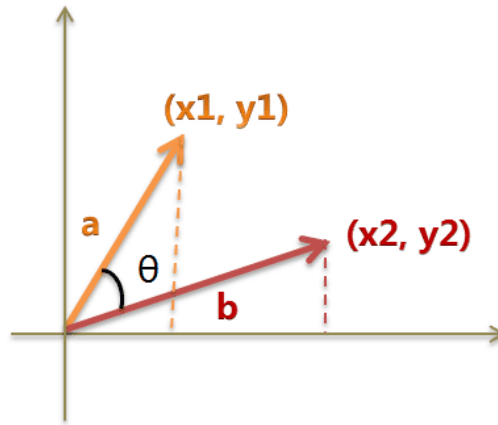
```
[1 1 2 1 0 5 2 1 1 2 0 0 1 0 1 1 1 0 1 3 1 0 2 0 0 1 2 1 0 2 0 1 4 1 0 0 1
0 0 0 1 1 0 1 0 1 0 0 0 0 0 1 0 2 0 0 1 1 0 1 0 0 2 1 1 1 4 1 2 0 0 0 2 2
2 0 2 1 2 2 2 0 2 0 2 2 0 0 0 1 0 1 0 0 0 1 0 1 1 2 3]
```

图 5-2-5 vec_b

5.3 计算文本相似度

自定义函数求余弦相似度，把 vec_a, vec_b 想象成空间中的两条线段，都是从原点 ([0, 0, ...]) 出发，指向不同的方向。两条线段之间形成一个夹角，如果夹角为 0 度，意味着方向相同、线段重合；如果夹角为 90 度，意味着形成直角，方向完全不相似；如果夹角为 180 度，意味着方向正好相反。因此，可以通过夹角的大小，来判断向量的相似程度。夹角越小，就代表越相似。假定 vec_a 向量是 [x1, y1]，vec_b 向量是 [x2, y2]，根据余弦定理可得：

$$\cos\theta = \frac{x_1x_2 + y_1y_2}{\sqrt{x_1^2 + y_1^2} \times \sqrt{x_2^2 + y_2^2}}$$



数学家已经证明，余弦的这种计算方法对 n 维向量也成立。假定 A 和 B 是两个 n 维向量，A 是 [A1, A2, ..., An]，B 是 [B1, B2, ..., Bn]，则 A 与 B 的夹角 θ 的余弦等于：

$$\begin{aligned}\cos\theta &= \frac{\sum_{i=1}^n (A_i \times B_i)}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}} \\ &= \frac{A \cdot B}{|A| \times |B|}\end{aligned}$$

5.4 制定评价方案

上文列举例子最后的计算结果为：

```
: getCos(vec_a,vec_b)
: 0.34
```

图 5-4-1 示例文本的余弦相似度

由“附件 4”中随机抽取系列数据进行试验，得出最终结果分别为 0.54、0.11、0.34、0.26、0.57 等，通过文本内容比对，最终可将相关度区间划分为：①>0.5，②0.3~0.5，③0.1~0.3，④<0.1 其中分别对应：①相关②相关性较好③相关性一般④不相关；其中特别注意，当答复意见为“2016 年 6 月 12 日”或者“网友：您好！您反映的问题已转市客运办调查核处。2019 年 2 月 13 日”之类时，余弦相似度为 0 或者“nan”，记为“完全无关”。

6. 结论及存在的问题

随着互联网的广泛应用和网上政务，政府也充分利用互联网了解民意、汇聚民智、凝聚民气，因此各类社情民意相关的文本数据量不断攀升，传统的一看人工分类已经不能满足数据量庞大的信息，本文对”智慧政务”进行文本分析与数据挖掘具有重要意义。

本文通过利用逻辑斯蒂回归方程对留言数据进行分类，刚开始采用高斯朴素贝斯模型的准确率较低，在后续探索中优化了模型，采用逻辑斯蒂回归模型，从而提高了准确率，实现了群众留言分类。在建立 CRF 模型提取地名时，第一次的提取范围较广，因而采用了三重 CRF 提取地名，成功挖掘到热点问题。在第三个评价答复意见，则是通过自定义余弦函数计算文本向量，测试数据给出数值用来判断留言详情与答复意见的相关性，本文所求出的相关度系数不高，通过验证排查后发现原因为：jieba 分词效果并不理想。解决方法为：完善停用词表。

参考文献:

- [1] 陶伟. 警务应用中基于双向最大匹配法的中文分词算法实现[J]. 电子技术与软件工程, 2016(04):153-155.
- [2] 张波, 黄晓芳. 基于 TF-IDF 的卷积神经网络新闻文本分类优化[J]. 西南科技大学学报, 2020, 35(01):64-69.
- [3] 陈春玲, 吴凡, 余瀚. 基于逻辑斯蒂回归的恶意请求分类识别模型[J]. 计算机技术与发展, 2019, 29(2):124-128.
- [4] 马孟铖, 艾斯卡尔·艾木都拉, 吐尔地·托合提. 基于条件随机场多特征融合的中文地名、机构名实体识别[J]. 现代计算机, 2019(12):13-17.
- [5] 王志刚, 谢恺, 朱慧. 降成本政策的文本分析——基于文本相似度计算原理[J]. 地方财政研究, 2020(03):90-97.