摘　要

随着国家对于互联网技术越来越重视，政府开始主导互联网技术更好为人民服务，更好、更快、更加真实的了解到群众的意愿，因此将互联网技术和政府行政相结合的智慧政务越来越受重视。

针对问题一，首先通过对数据的清洗处理，剔除掉了重复数据，并将重要信息“留言主题”、“留言详情”和“一级标签”提取出来；发现存在非平衡数据问题，决定通过句子打乱顺序（shuffle）方法进行数据增强；对“留言主题”和“留言详情”进行分词及剔除停用词处理，同时对于各个类别共有词也剔除，并将同行合并集合成新的文本集，训练出词袋。使用语言模型构造大量数据特征，在通过卡方检验来选取有明显区分度的特征；然后进行文本向量化，这里采用 TF-IDF 和 word2vec 两种方式对新文本集进行处理。最后借助朴素贝叶斯(NB)、支持向量机(SVM)、fasttext 等分类器进行交叉验证，调整参数选取最优模型。在对新数据进行预测时，将各个分类器预测结果进行投票选择类别。

针对问题二，将清洗后的数据提取“留言主题”、“留言详情”、“留言时间”、“支持数”、“反对数”重要信息 ，对“留言主题”、“留言详情”进行分词及去停用词处理，同时剔除在本题中没有热点划分意义的高词频，并将这两列按行合并。使用腾讯向量化库进行文本向量化，通过余弦距离计算各个文本向量的相似度，阈值取 0.85。构建‘热度指标=留言条数+支持反对数×3+峰度值×5’的热度指标，通过热度值的对比写出前五类的热点问题。

针对问题三， 为建立答复意见评价模型，选取了相关性、完整性、可解释性和时效性这四个方面作为评价方向，从数据中提取出重要数据进行研究，并对“留言主题”、“留言详情”和“答复意见”依次进行去重、分词、去停用词处理。以留言时间和回复时间的时间差作为时效性的指标，留言内容和答复意见的相似度作为完整性和可解释性的指标，将留言内容和答复意见的主题分类对比作为相关性的指标，通过等权重将这些指标合并为答复质量指标，对每条留言进行质量评价。

通过对留言文本的判别模型和评价模型进行研究，使用算法和程序构建了一套可以实时使用的系统化方法，更好的协助政府部门完成日常工作，使其更加高效的为人民服务。

关键词：分词、文本向量化、TF-IDF、fasttext、评价模型

目录

一、挖掘目标

1.1 挖掘背景

近年来，随着微信、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。基于大数据的智慧政府治理为推动政府治理模式转型和政府治理能力现代化提供了新途径。我国目前在科学化的政策制定、全程化的权力监督、网络化的协同治理、预防性的危机管理、精准化的公共服务等领域已开始推进智慧政务创新工作和实践探索，但智慧政务依然面临着制度建设不够完善、动态网络协同治理体系还未有效建构、政府数据开放与共享步伐迟滞、相关要素支撑不足以及平台建设远未跟上时代步伐等问题与挑战。为此，我国智慧政务创新发必须重视智慧政府治理的顶层设计、创新动态网络协同治理方式、推进政府数据开放共享、探索智慧公共决策的路径、加强智慧政务技术的研发与应用以及完善治理网络基础设施建设。

在大数据时代，加快推动大数据驱动的智慧政务建设已成为当前推动政府能力现代化的内在需求和必然选择。大数据与智能化有利于推动政务决策的民主化和科学化，带来更为开放、透明和负责的政府。在大数据时代，“用数据说话、用数据决策、用数据管理、用数据创新”带来了国家治理方式的根本变革。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的推动作用。

1.2 挖掘目标

根据附件 1、2 里的数据，设计出对留言主题自动分类的系统，解决之前只能使用人工分类的问题。

根据附件 3 里的数据，根据时间、地点和人群等找出留言热点问题，并建立热度值体系，找出前五个热度问题并按照格式写出来

根据附件 4 里的数据，依据相关性、完整性、可解释性等角度建立留言回复质量评价指标，并尝试实行。

二、问题分析

针对问题 1，将附件 1、2 联合分析去掉重复数据，剔除停用词和类别共有词，在本体中我们决定自己训练词袋，使用 tf-idf 和 word2vec 进行文本向量化，选取例如朴素贝叶斯、支持向量机等分类器对文本进行分类，采用交叉验证调试参数测试模型的准确度。

针对问题 2，将附件 3 中数据去掉同一用户短期重复留言数据，分词后去停用词及剔除对于本题判别意义不大的词，将分词文本向量化后使用余弦相似度算法计算同一个类别中两两之间的相似度，通过留言数量、时间间隔和支持反对数构建评价热度指标。

针对问题 3，从相关性、完整性、可解释性和时效性这四个方面构造综合评价模型，先设定为均等权重，然后通过测试调整权重。

三、数据预处理

3.1 数据清洗

3.1.1 第一题

有极少数的重复留言会有多个标签，这可能是人工误分的结果，即可以认为是噪声。这里考虑的策略是将所有重复留言去重，仅保留一个，其类别根据多数投票来决定。如果出现的不同类别次数相同，则选择留言时间最后的类别。因为当留言时间不同时，很可能是因为之前的群众留言没有得到有效的回复，可以认为是之前的分类是错误的，因而选择最后留言的类别。当留言时间相同且出现的不同类别次数相同，就随机抽选一个类别

3.1.2 第二题

将在研究“留言主题”与“留言详情”的时候，发现部分重复的留言，因此筛选出“留言用户”、“留言主题”和“留言详情”完全相同的留言：

	id	user_id	topic	time	detail	oppose_num	approve_num
550	201447	A00020543	A7县新国道107距我家仅3米，相关政府部门为何不同意拆迁?	2019/7/22 17:04:08	\n\t\t\t\t\t\n\t\t\t\t\t我是A7县青山铺镇沙塘冲组村民，新修的国道1...	0	2
1746	229903	A00020543	A7县新国道107距我家仅3米，相关政府部门为何不同意拆迁?	2019/7/22 17:05:04	\n\t\t\t\t\t\n\t\t\t\t\t我是A7县青山铺镇沙塘冲组村民，新修的国道1...	0	0
3624	273925	A00020543	A7县新国道107距我家仅3米，相关政府部门为何不同意拆迁?	2019/7/18 10:47:31	\n\t\t\t\t\t\n\t\t\t\t\t我是A7县青山铺镇沙塘冲组村民，新修的国道1...	0	3
3829	278907	A00020543	A7县新国道107距我家仅3米，相关政府部门为何不同意拆迁?	2019/7/18 10:48:28	\n\t\t\t\t\t\n\t\t\t\t\t我是A7县青山铺镇沙塘冲组村民，新修的国道1...	0	0

在讨论中认为，统一用户短期内（一天之内）连续留言多次相同主题及内容的留言为无效信息予以剔除。在不同的天数里留言完全相同的内容时，认为此事件对该用户影响较大，可以保留计算到热度值之中。对留言事件进行分析发现存在不同用户在同一时间留言完全相同的内容，认为是群众有组织性留言，予以保留：

id	user_id	topic	time	detail	oppose_num	approve_num
191474	A000110333	12123上申请驾驶证期满换证，一个星期了都无人受理	2019-04-26 15:28:42	\n\t\t\t\t\t\n\t\t\t\t\t说是推出了便民服务，可以直接网上换证，听起来...	0	0
313964	A108906	12123上申请驾驶证期满换证，一个星期了都无人受理	2019-04-26 15:28:42	\n\t\t\t\t\t\n\t\t\t\t\t说是推出了便民服务，可以直接网上换证，听起来...	0	0

3.1.3 第三题

同一时间、同一个人的同一问题，两个答复时间有间隔，属于需要转部门解决的问题（若研究政府的解决问题效率只保留时间间隔较长的；若研究政府回复问题效率应保留时间间隔叫短的）

	id	user_id	topic	lwtime	leave word	reply	replytime
1692	74211	UU0082219	投诉F9市富豪名城小区违规安装变电站	2019/9/4 8:07:25	\n\t\t\t\t\t\n\t\t\t\t\t我是F9市富豪名城6栋的业主，现将富豪名城6...	您的留言已收悉，关于您反映的问题，已转F9市调查处理。	2019/9/10 9:35:52
1774	80151	UU0082219	投诉F9市富豪名城小区违规安装变电站	2019/9/4 8:07:25	\n\t\t\t\t\t\n\t\t\t\t\t我是F9市富豪名城6栋的业主，现将富豪名城6...	9月4日下午，我公司工作人员与反映问题的客户（以下称为客户）及部分相关业主见面，并现场测量帖...	2019/12/4 16:41:57

留言内容相同但留言时间不同的（考虑可能问题没有得到有效解决）

	id	user_id	topic	lwtime	leave word	reply	replytime
637	16996	UU008757	咨询A7县体育公园建设情况	2017/11/1 11:29:30	\n\t\t\t\t\t\n\t\t\t\t\t 以前电视、报纸报道的A7县体育公园建设项...	网友"UU008757" 您好！来信收悉。现回复如下： 您所述的"A7县体育公园"已更名...	2017/11/3 15:44:16
678	17356	UU008757	咨询A7县体育公园建设情况	2017/6/5 11:51:08	\n\t\t\t\t\t\n\t\t\t\t\t 据星沙时报报道，A7县体育中心已启动建设...	网友"UU008757" 您好！来信收悉。现回复如下： 您所述的"A7县体育公园"因相关...	2017/6/6 15:44:10

完全相同的 （可删除）

	id	user_id	topic	lwtime	leave word	reply	replytime
2372	121215	UU0081001	请求合理设置L市城区的28路公交车终点站	2018/4/16 8:20:19	\n\t\t\t\t\t\n\t\t\t\t\tL市城区的28路公交车是贯穿L市南到L市北的...	"UU0081001"您好！感谢您对公交的关注和监督，您所反映的"请求合理设置L市城区的28...	2018/4/23 8:37:38
2374	121229	UU0081001	请求合理设置L市城区的28路公交车终点站	2018/4/6 21:40:29	\n\t\t\t\t\t\n\t\t\t\t\tL市城区的28路公交车是贯穿L市南到L市北的...	"UU0081001"您好！感谢您对公交的关注和监督，您所反映的"L1区居民建议增加开往盛世...	2018/4/23 9:37:38

3.2 文本分词

为了将不同内容的句子区分开来，需要找出每个句子的特征关键词，因此需要利用 jieba 分词对每句话进行分词处理。jieba 自带了一个叫做 dict.txt 的词典，里面有 2 万多条词， 包含了词条出现的次数和词性。这个第一条的 trie 树结构的词图扫描，说的就是把这 2 万多条词语，放到一个 trie 树中，而 trie 树是有名的前缀树，也就是说一个词语的前面几个字一样，就表示他们具有相同的前缀，就可以使用 trie 树来存储，具有查找速度快的优势。jieba 分词属于概率语言模型分词，在全切分所得的所有结果中求某个切分方案 S，使得 P(S)最大。例如：“小明来到月湾区”划分为“小明 来到 月湾 区”

为了使标签的判定更为准确，加入例如：经济学院，…等切分词，在进行文本划分时会将这些字不分开归为词，减小了计算量增加了预测精度。

3.3 去停用词

停用词主要包括英文字符、数字、数学字符、标点符号及使用频率特高的单汉字等。在对分词后的数据进行去停用词处理，因为停用词具有极其普遍和无法单独表达文档相关程度的特点，所以在对句子进行分类的时候这些功能词基本没有什么帮助，同时这些词扔掉减少了索引量，增加了检索效率。

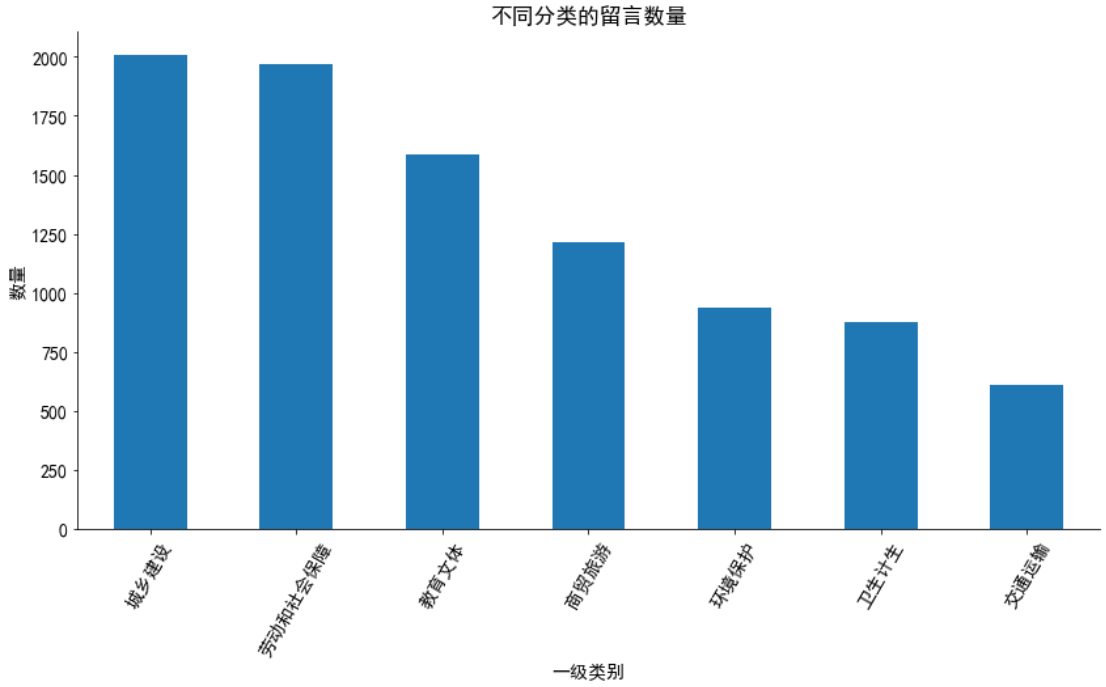
在第一题中，需要对不同留言的类别进行区分，因此每一类别共有词对于分类不会起到作用，从简化计算和消除误差的角度出发将分词中共有词剔除掉。

在第二题中，对分词后的高频词进行分析，发现存在很多在本题中没有分类意义的词语，因此将这些词也添加到本体去停用词词表中予以剔除。一定程度上简化了计算。

四、问题一

4.1 数据增强

训练数据文本类别 7 类，一级分类 15 类，对于已知的 7 类中存在一定的数据不平衡问题，因为不平衡比例大概为 1：3 左右：



选取句子打乱顺序（shuffle）的方法来进行数据增强。Shuffle 是更具索引的目标产生一个随机索引，并调换双方内容信息。

Step1：从数组末尾开始，选取最后一个元素，然后随机选择一个元素与之交换位置，并固定下最后一位。

Step2：依次向前推进，完成所有交换。

对数据中类别数量较少的数据增加数据，直到所有类别的数据大概相同。

4.2 特征的构建与选择

4.2.1 词袋模型

词袋模型(BOW)，BOW 模型假定对于一个文档，忽略它的单词顺序和语法、句法等要素，将其仅仅看作是若干个词汇的集合，文档中每个单词的出现都是独立的，不依赖于其它单词是否出现。

该方法存在两个问题，不能保留语义，不能保留词语在句子中的位置信息，“你爱我”和“我爱你”在这种方式下的向量化结果依然没有区别；维数高和稀疏性，当语料增加时，那么维数也会不可避免的增大，一个文本里不出现的词语就会增多，导致矩阵稀疏

我们通过对分词去停用词后的数据进行训练，来建立词袋为后续文本向量化做准备。

4.2.2n-gram 语言模型

n-gram 模型也称为 n-1 阶马尔科夫模型，它有一个有限假设：当前词的出现概率仅仅与前 n-1 个词相关。因此给定词语序列 S=$W_1, W_2, \dots, W_k$它出现的概率可以表示为：

$$P(S) = P(W_1, W_2, \dots, W_k) = p(W_1)P(W_2|W_1) \dots P(W_k|W_1, W_2, \dots, W_{k-1})$$

当 n 取 1、2、3 时，n-gram 模型分别 unigram、bigram 和 trigram 语言模型。每种语言模型都存在自身的弊端，为了避免遗漏重要特征，同时使用这三种语言模型来构建特征，总共构造 200000 个词作为特征然后从中剔除。

4.2.3 卡方检验

设一总体 X 服从分布：

$$H_0：P(X = a_i) = p_i(i = 1, \dots, k)$$

设定原假设H_0为：特征能够区分句子类别，通过公式计算卡方统计量

$Z = \sum_{i=1}^{a} \sum_{j=1}^{b} (nn_{ij} - n_{i.}n_{.j})^2/(nn_{i.}n_{.j})$,自由度为 k-1-r=(a-1)(b-1)

并选取卡方值最大的 20000 条特征，被舍弃的特征可以认定为对于类别划分没有明显意义。

4.2.4TF-IDF

Tf-idf 是一种用于信息检索与信息挖掘的常用加权技术，用以评估某个字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。

算法步骤：

Step1:计算词频

$$词频（TF）=\frac{某个词在句子中出现次数}{所有句子的总词数}$$

Step2:建立语料库

$$逆文档频率（IDF）=\log\left(\frac{语料库的句子总数}{包含该词的句子数+1}\right)$$

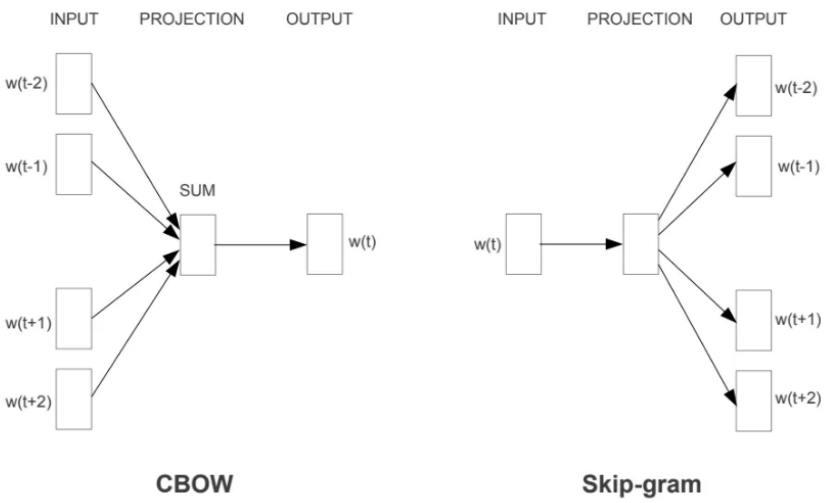
Step3:计算 TF-IDF

$$TF\text{-}IDF = 词频(TF) \times 逆文档频率(IDF)$$

TF-IDF 具有简单快捷，容易理解，适合大数据的计算的优点，将特征赋予权重，便于分类识别。

4.2.5Word2vec

在预测类别时用到深度学习，因此这里对分词进行第二种向量化。Word2Vec 是轻量级的神经网络，其模型仅仅包括输入层、隐藏层和输出层，模型框架根据输入输出的不同，主要包括 CBOW 和 Skip-gram 模型。CBOW 的方式是在知道词ω_t的上下文$\omega_{t-2},\omega_{t-1},\omega_{t+1},\omega_{t+2}$,的情况下预测当前词$\omega_t$。而 Skip-gram 是在知道了词$\omega_t$的情况下，对词的上下文$\omega_{t-2},\omega_{t-1},\omega_{t+1},\omega_{t+2}$进行预测，如下图所示：



4.3 模型的建立

4.3.1 朴素贝叶斯

利用“逆向概率”的思想，通过已知的条件概率，借助贝叶斯定理，求得需要的概率，并通过最大概率划分分类。

Step1:条件概率公式

$$P\left(X^{(j)}=a_\beta \middle| Y=c_k\right)=\frac{\sum_{i=1}^{N} I(x_i^{(j)}=a_{ji}, y_i=c_k)}{\sum_{i=1}^{N} I(y_i=c_k)}$$

$$j=1,2,\cdots,n; l=1,2,\cdots,S_j; k=1,2,\cdots,K$$

Step2:对于给定的实例$x=(x^{(1)},x^{(2)},\cdots,x^{(n)})^T$

$$P(Y=c_k)\prod_{j=1}^{n} P\left(X^{(j)}=x^{(j)}\middle| Y=c_k\right), k=1,2,\cdots,K$$

Step3:确定实例 x 的类

$$y=arg\max_{c_k} P(Y=c_k)\prod_{j=1}^{n} P(X^{(j)}=x^{(j)}|Y=c_k$$

4.3.2 支持向量机

支持向量机可以用来解决非线性问题，通过核函数与软间隔最大化，学习得到分类决策函数。

Step1：选取适当的核函数 K(x,z)和适当的参数 C,构造并求解最优化问题

$$\min_{a}\frac{1}{2}\sum_{t=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(x_i,x_j)-\sum_{t=1}^{N}\alpha_i$$

$$s.t.\quad \sum_{t=1}^{N}\alpha_i y_i=0 \qquad 0\le\alpha_i\le C, i=1,2,\dots,N$$

求得最优解$\alpha^*=(\alpha_1^*,\alpha_2^*,\dots,\alpha_N^*)^T$

Step2：选择α^*的一个正分量$0<\alpha_j^*<C$，计算

$$b^*=y_j-\sum_{i=1}^{N}\alpha_i^* y_i K(x_i\cdot x_j)$$

Step3:构造决策函数：

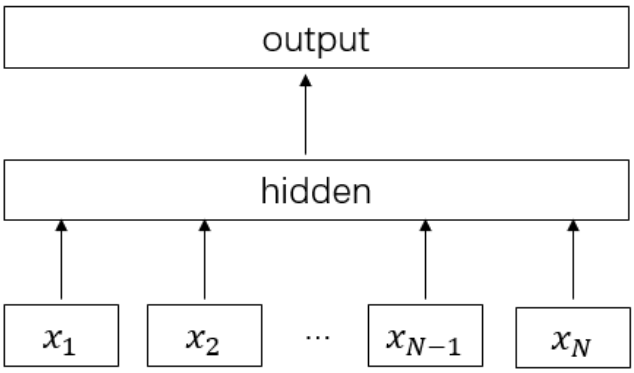
$$f(x)=sign(\sum_{t=1}^{N}\alpha_i^* y_i K(x\cdot x_i)+b^*)$$

通过支持向量机可以求出向量化后的句子类型，使用交叉验证判断它的分类准确率。

最终通过 stacking 模型求出每个句子的分类。

4.3.3 快速文本（fast Text）

Fasttext 是一个资料库，能针对文本表达和分类帮助建立量化的解决方案，该模型使用词袋以及 n-gram 袋表征语句，还有使用子字信息，并通过隐藏表征在类别间共享信息。我们另外采用了一个 softmax 层级来加速运算过程。



Step1：在文本 word2vec 后，输入向量化后的词袋

Step2：就是将第一步中输入的向量相加再求平均，得到一个新的向量 w，然后将这个向量输入到输出层。

Step3：采用了层次 softmax 的方法，根据 label 的频次建立哈夫曼树，每个 label 对应一个哈夫曼编码，预测的时候隐层输出与每个哈夫曼树节点向量做点乘，根据结果决定向左右哪个方向移动，最终落到某个 label 对应的节点上。

4.3.4 交叉验证

在给定的模型样本中，拿出大部分样本进行建模，留小部分样本用刚建立的模型进行预报，记录预报误差的和，可用来解决非平衡数据造成的过拟合。

Step1:从全部训练数据 data 中随机选择部分数据 data_tr 作为训练集，剩余部分数据 data_te 作为测试集

Step2:通过训练出的模型，对测试集进行测试，求出分类正确率

Step3:选择具有最大分类率的模型

通过交叉验证我们可以测得各个模型的准确率，所以我们可以不断调整各个模型的参数，在确保计算适度的情况下获得较高的准确率。最终，我们将模型参数定为：

文本特征的清洗：全类词典；

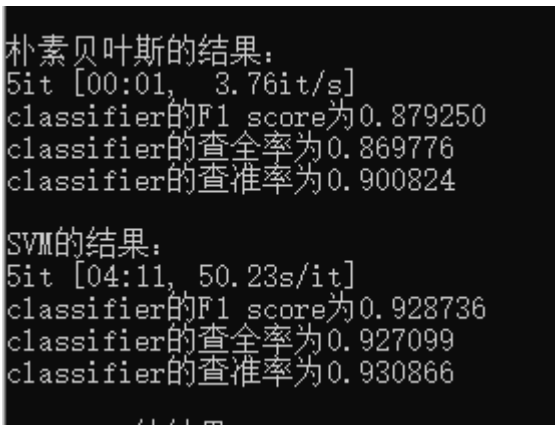
分类词典：全类；

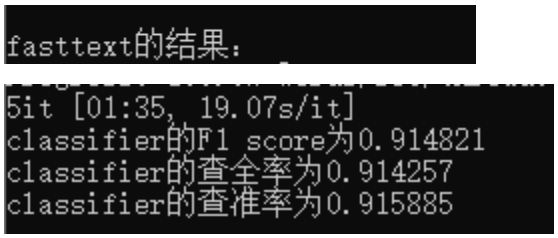
剔除低频词阈值：0；

卡方检验抽取的特征数：70000

词频标准差抽取的特征数：20000

TF-IDF 抽取的特征数：20000





到这里对于每个文本我们已经获得了三个预测值，对于三个预测结果进行投票选取，每个文本所投票数最高的类别作为该文本最终预测文本。

五、问题二

在对热点问题的文本进行分词和向量化，我们发现自己建立词库来向量化的问题突出，使用 simhash 时发现该方法对于小文本求相似度及其不友好。因此在本题中我们选择开源的腾讯分词与腾讯词向量。将“留言主题”和“留言详情”按行合并起来使其向量化，我们使用了 7 万个词的词库，存在部分词语未能向量化，我们自己将这些词建立了三维以上的向量化。

5.1 余弦相似度算法

借助余弦定理公式计算两个向量之间的夹角间的余弦值作为衡量个体之间的差异，余弦值接近 1，夹角趋于 0，表明两个向量越相似，余弦值接近于 0，夹角趋于 90 度，表明两个向量越不相似。

Step1:写出需要计算的向量

Step2:借助公式计算，并判断相似性

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

因为本体属于无监督学习，所以在对求解余弦相似度后的各个向量与原始数据进行对照发现，阈值取 0.85 是一个比较好的分界点。

5.2 综合热度指标

根据问题二的需要，建立一段时间内反应特定地点或特定人群的留言热度评价指标，通过每一类留言数、时间——数量峰度和支持反对数来综合评价，可以认为这三个方面是相同重要的。

对于某一类留言来说，留言数和支持反对数越高热度值越高，时间——数量峰度越大热度值越高。总留言条数和支持反对数的比例为 1：3，因此可以将这个比例作为权重指标。同时几乎所有留言时间集中于在一年的时间内，因此将计算时间——数量的峰度值，最大峰度值与总留言条数的比例为 1：a。因此可以写出热度值的评价公式：

C 类热度指标=C 类留言条数+C 类支持反对数×3+C 类峰度值×a

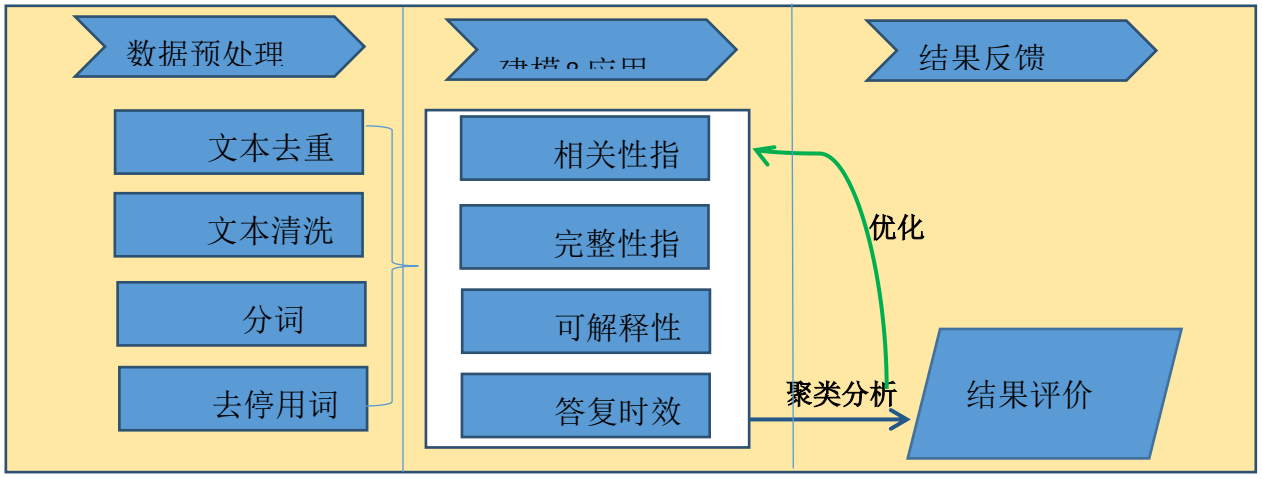
六、问题三

6.1 研究目标:

针对附件 4 相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对
答复意见的质量给出一套评价方案，并尝试实现。

6.2 分析方法与过程

6.2.1 总体流程



主要步骤描述：
步骤一、相同的留言内容的分类处理。
步骤二、分词及去停用词

步骤三、文本向量化

（1）One-hot 编码

one-hot 编码是常用的方法，我们可以用 one-hot 编码的方式将句子向量化，大致步骤为：

a.用构造文本分词后的字典

b.对词语进行 One-hot 编码

one-hot 词向量构造起来简单，但通常不是一个好的选择，它有明显的缺点：

维数过高: 短短的 2 句话，每个词语已经是一个 10 维的向量了，随着语料的增加，维数会越来越大，导致维数灾难

b. 矩阵稀疏: 利用 One-hot 编码的另一个问题就是矩阵稀疏，从上面也可以看到，每一个词向量只有 1 维是有数值的，其他维上的数值都为 0

c. 不能保留语义: 用这种方式得到的结果不能保留词语在句子中的位置信息，“我爱你”和“你爱我”的向量化结果并没有什么不同。

（2）词袋（BOW）模型

词袋模型(BOW)，BOW 模型假定对于一个文档，忽略它的单词顺序和语法、句法等要素，将其仅仅看作是若干个词汇的集合，文档中每个单词的出现都是独立的，不依赖于其它单词是否出现。

这种方式不像 one-hot 编码那样导致维数非常大，但也有自己的缺点

不能保留语义：不能保留词语在句子中的位置信息，“你爱我”和“我爱你”在这种方式下的向量化结果依然没有区别。“我喜欢北京”和“我不喜欢北京”这两个文本语义相反，利用这个模型得到的结果却能认为它们是相似的文本。

b. 维数高和稀疏性：当语料增加时，那么维数也会不可避免的增大，一个文本里不出现的词语就会增多，导致矩阵稀疏

步骤四、计算留言和回复的相似度

步骤五、把群众的留言和政府的回复做一个主题/聚类分析（看它们是不是同一主题或者类别）

步骤六、计算留言时间和回复时间之间的间隔，

6.2.2 具体步骤

6.2.3 过滤器

步骤六、计算留言时间和回复时间之间的间隔

	reply	replytime	difftime
监督与支持！	2015年1月8日	2016/1/8 15:22:06	0
		2018/5/24 16:12:33	0
祝好！	2017年11月3日	2017/10/30 15:54:34	0

最快 0 天，最慢的 1161 天，平均在 20 天左右。

6.3 建立指标体系

- 及时性：算一下间隔（做一个时间间隔的表---升序），间隔越少回复就越及时
- 相关性/完整性/解释性：先计算相似度来评判，相似度越大，这几个指标就越好（对政府和群众留言的文本分词，清洗---> 分别一一计算两段文本的相似度）
    - 把群众的留言和政府的留言做一个主题/聚类分析（看它们是不是同一主题或者类别）

七、总结

在第一问和第二问中，我们采取了多种方法进行建模与试验，在第一问中我们取得了较好的模型，能够对之后的预测给出合理结果。对于第二问我们能够得出热点问题并给出热度值。