基于文本挖掘技术的“智慧政务”系统

摘要：随着互联网和大数据的发展，各大网络政务平台留言文本量不断攀升，给人工划分和整理带来较大挑战。为了帮助提高政务处理效率，我们建立了基于文本挖掘技术的智慧政务系统。本次任务分为三大部分：群众留言分类，热点问题挖掘和答复意见评价方案的建立。

对于留言分类问题，任务要求建立关于留言内容的一级标签分类模型。我们首先选择了 $BERT$（Bidirectional Encoder Representation from Transformers）模型，并看到了 $UER-py$ 的重现性，模块化，高效率等优点，使用其进行预训练模型的加载，完成文本分类任务。我们进行了三轮实验，从实验的损失值趋势可见损失值逐渐稳定，分类准确率在 0.92。

对于热点问题的挖掘，我们需要攻克两方面的难点。第一是命名实体的识别，需要尽可能提高识别准确率，为热点问题的提取做前期准备；第二是使用自动摘要方法提取问题，做到简练、可读。

首先，我们进行数据预处理。我们采用 python 正则表达式去除标点；$jieba$ 库进行分词处理；对于未登录词采用 HMM 模型，使用了 $Viterbi$ 算法；另外，为了使用主题模型的文档-词汇矩阵，使用 from sklearn 里面的 CountVectorizer api 进行文本特征提取。

解决第一难点选择命名实体识别模型。我们考虑了隐马尔可夫模型（HMM），最大熵马尔可夫模型（MEMM）最大熵马尔可夫模型（MEMM），条件随机场（CRF）模型，分析不同模型的优缺点和适用性。我们采用的是短文本内容，文本词汇稀疏，共现概率低，基于表 3.3-1 测验数据，选择 CRF 模型。

解决第二难点，使用 LDA（Latent Dirichlet Allocation）文档主题生成模型，包含词、主题和文档三层结构，把文档—词汇矩阵变成文档-主题矩阵和主题-词汇矩阵。在此过程中，涉及到最优主题数选取问题，我们引用困惑度（$perplexity$）作为评价指标，最终确定 140 的最优主题数。在此主题数下，我们采用的 $textrank$ 算法解决自动摘要方法提取问题，通过把文本分割成若干组成单元并建立图模型，利用投票机制对文本中的重要成分进行排序，仅利用单篇文档本身的信息即可实现关键词提取、文摘。使用相关组件 $textrank4zh$，抽取文章的关键字以及关键句作为文摘，通过 lda 模型分类后，将同一主题文本输入 $textrank4zh$ 进行主题句分析，得到句子的权重，按照权重输出。权重最大的为关键句，我们选择它作为问题描述。同时我们引入自己的热度指标算法，得到任务要求的“热点问题表.xls”和“热点问题留言明细表.xls”。

对于答复意见评价方案的建立，我们基于文献参考，根据国家政府办公规定，总结出高分和低分回复的普遍特点。对于这些特点，我们决定建立三个评价标准，对每个评价标准给出 0 到 1 分之间的分数，最后把三个得分按照一定的权重求和，得到最终得分。评价标准为：字数，关键词提及比例，是否引用法律条规。设置评价系统总得分$\mathrm{mark}_{\mathrm{total}} = 0.2\mathrm{mark1} + 0.7\mathrm{mark2} + 0.1\mathrm{mark3}$。从结果可以看出符合评判标准的合理分布，并使用人工验证了其在一定范围内的准确性。

最后我们采用 F-Score 评价和鲁棒性可拓展性，进行模型相关性验证和评价。并提出三种不同方面的模型展望。

关键词：文本挖掘技术；Bert；LDA 主题模型；命名实体识别；文本质量评价

"Intelligent government affairs" system based on text mining technology

Abstract: With the development of the Internet and big data, the amount of message texts of major online government platforms has been rising, which has brought great challenges to manual division and sorting. To help improve the efficiency of government processing, we have established a smart government system based on natural language processing technology. This task is divided into three major parts: the classification of mass messages, the mining of hot issues and the establishment of an evaluation plan for answering opinions.

For the message classification problem, the task requires the establishment of a first-level label classification model on the content of the message. We first selected the BERT (Bidirectional Encoder Representation from Transformers) pre-training model, and saw the advantages of UER-py's reproducibility, modularity, and high efficiency. We used it to load the pre-training model and complete the text classification task. We have conducted three rounds of experiments, and we can see from the trend of the loss value of the experiment that the loss value gradually stabilizes. Based on the results, we use a confusion matrix to evaluate the classification method based on F-Score. It can be seen that the model has an excellent effect on first-level label classification, and the accuracy rate is 0.92.

For the mining of hot issues, we need to overcome two difficulties. The first is the recognition of named entities. It is necessary to improve the recognition accuracy as much as possible to prepare for the extraction of hot issues. The second is to use automatic summarization to extract problems, which is concise and readable.

First, we perform data preprocessing. We use Python regular expressions to remove punctuation; $Jieba$ library performs word segmentation; for the unregistered words, the HMM model is used, and the Viterbi algorithm is used; in addition, in order to use the document-vocabulary matrix of the topic model, the $CountVectorizer\ api$ from sklearn is used for text features extract.

To solve the first difficulty, we choose a named entity recognition model. We considered hidden Markov model (HMM), maximum entropy Markov model (MEMM) maximum entropy Markov model (MEMM), conditional random field (CRF) model, and analyzed the advantages and disadvantages and applicability of different models. We used short text content, sparse text vocabulary, and low probability of co-occurrence. Based on the test data in Table 3.3-1, we chose the CRF model.

To solve the second difficulty, use the LDA (Latent Dirichlet Allocation) document topic generation model, which contains a three-layer structure of words, topics and documents, and turn the document-vocabulary matrix into a document-topic matrix and a topic-vocabulary matrix. In this process, which involves the selection of the optimal number of topics, we quote the perplexity as the evaluation index, and finally determine the optimal number of topics of 140. In this topic number, we use the $textrank$ algorithm to solve the problem of automatic summary extraction. By dividing the text into several

constituent units and establishing a graph model, the voting mechanism is used to sort the important components in the text, using only the single document itself. Information can be used to extract keywords and abstracts. Use the related component textrank4zh to extract the keywords and key sentences of the article as abstracts. After classifying by the lda model, input the same topic text into textrank4zh for topic sentence analysis to get the weight of the sentence and output it according to the weight. The key sentence is the one with the largest weight, and we choose it as the problem description. At the same time, we introduce our own heat index algorithm to get the "Hot Question List.xls" and "Hot Question Message List.xls" required by the task.

Regarding the establishment of the evaluation plan for answering opinions, we summarized the general characteristics of high- and low-score responses based on literature references and in accordance with national government office regulations. For these characteristics, we decided to establish three evaluation criteria, give each evaluation criteria a score between 0 and 1, and finally sum the three scores according to a certain weight to get the final score. The evaluation criteria are: the number of words, the proportion of keywords mentioned, and whether the laws and regulations are cited. Set the total score of the evaluation system: $mark\_total = 0.2mark1 + 0.7mark2 + 0.1mark3$. From the results, it can be seen that the reasonable distribution meets the evaluation criteria, and the accuracy within a certain range is verified manually.

Finally, we use F-Score evaluation and robust scalability to perform model correlation verification and evaluation. And put forward three different aspects of model outlook.

Key words: Text mining technique；LDA；Named Entity Recognition；Text quality evaluation

目 录

1. 问题背景与任务

1.1 问题背景

互联网和大数据的迅猛发展，对政府行政管理变革和创新提出了新的挑战，政府与民众互动关系随着智慧政府的建设逐渐走向和谐.目前,在智慧政府建设中还存在政务网站服务现状欠佳、信息共享程度不高、政府回应与民众期待衔接不到位等问题，一定程度上阻碍了政府与民众互动的良性发展[1]。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。

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，能够从各大平台的互动中高效汇集问题，方便进行统计及回顾分析，帮助提高政府的管理水平和施政效率[2]。

1.2 挖掘任务

- **群众留言分类。**按照一定的划分体系（内容分类三级标签体系）对留言进行分类，以便后续将群众留言分派至相应的职能部门处理。
- **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。及时发现热点问题，有助于相关部门进行有针对性地处理，提升服务效率。我们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，定义合理的热度评价指标，并给出评价结果。给出排名前 5 的热点问题，制作“热点问题表”和“热点问题明细表”。
- **答复意见评价方案。**针对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。

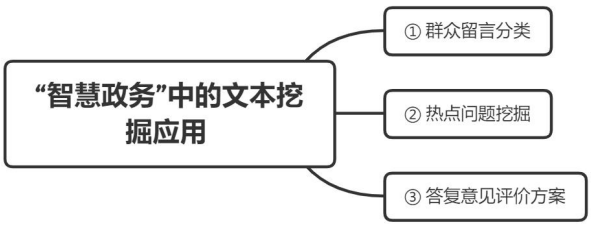


图 1-1 整个工作流程的大概框架

2. 群众留言分类

2.1 BERT 预训练模型

BERT 的全称是 Bidirectional Encoder Representation from Transformers[6]，即双向 Transformer 的 Encoder，因为 decoder 是不能获要预测的信息的。此模型的主要创新点有：

- 使用了 Transformer 作为算法的主要框架，Trabsformer 能更彻底的捕捉语句中的双向关系；
- 使用了 Mask LanguageModel(MLM)和 Next Sentence Prediction(NSP) 的多任务训练目标；
- 使用更强大的机器训练更大规模的数据，使 BERT 的结果达到了全新的高度，并且 Google 开源了 BERT 模型，用户可以直接使用 BERT 作为 Word2Vec 的转换矩阵并高效的将其应用到自己的任务中。

2.2 UER-py 进行预训练模型加载

UER-py 是一个在通用语料预训练以及对下游任务进行微调的工具包。其具有以下特点：

- 重现性。UER-py 已在多个数据集上进行了测试，应与原始实现的性能相匹配。
- 多 GPU。UER-py 支持 CPU 模式，单 GPU 模式和分布式训练模式。
- 模块化。UER-py 分为多个组件：子编码器，编码器，目标和下游任务微调。每个组件中实现了足够的模块。清晰而强大的界面使我们可以结合尽可能少的限制来组合模块。
- 高效率。UER-py 完善了其预处理，预训练和微调阶段，从而极大地提高了速度并且需要更少的内存。

预训练模型有几个关键的部分：语料、编码器、目标任务、以及微调策略。UER-py 能让用户轻易的对不同的部分进行组合，复现已有的预训练模型（比如 BERT），并为用户提供进一步扩展的接口。我们通过 UER-py 加载预训练模型，从而完成文本分类的任务。

2.2 结果

2.2.1 启动代码

$$python\ run\_classifier.py$$
$$--\ pretrained\_model\_path\ models/book\_review\_model.bin$$

$--vocab\_path\ models/google\_zh\_vocab.txt$

$--train\_path\ datasets/taidi\_second/train.tsv$
$--dev\_path\ datasets/taidi\_second/dev.tsv$
$--test\_path\ datasets/taidi\_second/test.tsv$
$--epochs\_num\ 3$
$--batch\_size\ 32$
$--encoder\ bert$
$--report\_steps\ 5$

参数说明:
-- pretrained_model_path：预训练模型的位置。这里使用的是豆瓣书评的预训练模型。
$--vocab\_path$：词汇表的位置。这里使用的是谷歌的中文词汇表。

$--train\_path --dev\_path --test\_path$

分别是训练集，校验集和测试集的位置。这里三个数据集取自题目的附件 2。首先对附件 2 进行随机排序，再通过 8:1:1 的比例分别提取到训练集，校验集和测试集中。
$--epochs\_num$：训练轮数
在这次训练中我们选取了 3 轮。
$--batch\_size$：缓存大小
在这次训练中缓存大小为 32
$--encoder$：编码器
使用 bert 对文本进行编码
$--report\_steps$：报告步数，仅为了在训练过程中监视损失值，不影响训练过程。

2.2.2 实验结果

我们进行了以下测试:

亚马逊 lstm 预训练模型:
3 轮 学习率 $2e-5$
测试集准确率 0.8843（780/882）

人民日报预训练模型:
3 轮 学习率 $2e-5$
测试集准确率 0.8854（781/882）

豆瓣预训练模型:

3 轮 学习率 $1e-3$ 第一轮 100 步以内损失函数不下降 未拟合

3 轮 学习率 $1e-4$ 第一轮 100 步以内损失函数不下降 未拟合

3 轮 学习率 $1e-5$ 0.9116 (804/882)

最后选择准确度最高的超参数（即 2.2.1 中的超参数）作为我们的模型。

三轮实验的损失值趋势如图 2.2-1，2.2-2，2.2-3。可见损失值逐渐稳定。

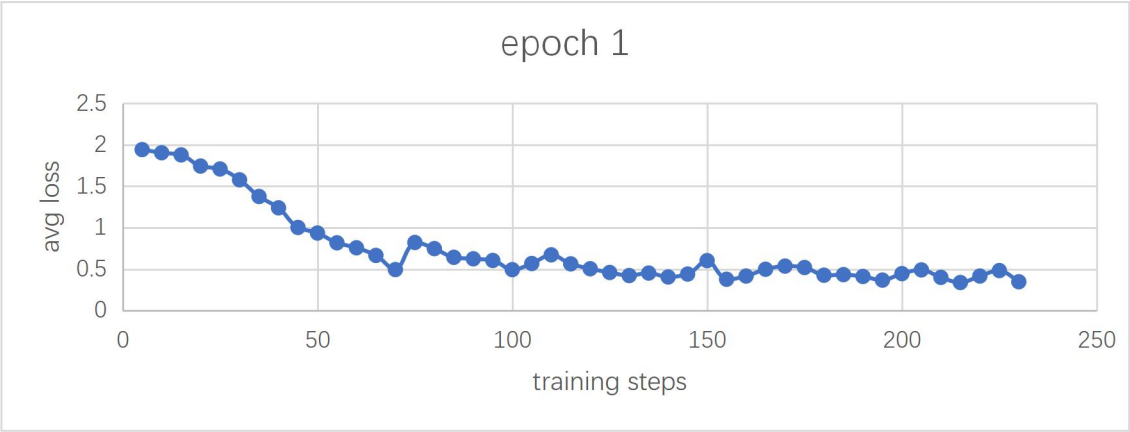


图 2.2-1 第一轮实验结果

图 2.2-2 第二轮实验结果

图 2.2-3 第三轮实验结果

测试集的输出混淆矩阵:

```
tensor( [ [197,  11,   1,   1,   1,   8,   0],
          [  4,  77,   0,   0,   0,   1,   0],
          [  0,   0,  56,   0,   0,   0,   0],
```

```
[  5,   1,   0, 145,   3,   3,   0],
[  3,   2,   0,   1, 169,   1,   5],
[  3,   1,   0,   3,   0,  92,   0],
[  0,   0,   0,   0,   7,   2,  79] ] )
```

测试集 7 个标签的查准率，查全率和 F1：

```
Report precision, recall, and f1:
Label 0: 0.900, 0.929, 0.914
Label 1: 0.939, 0.837, 0.885
Label 2: 1.000, 0.982, 0.991
Label 3: 0.924, 0.967, 0.945
Label 4: 0.934, 0.939, 0.936
Label 5: 0.929, 0.860, 0.893
Label 6: 0.898, 0.940, 0.919
```

测试集的准确率：Acc. (Correct/Total): 0.9240 (815/882)

可以看到该模型对一级标签的分类效果相当好，准确率在 0.92。

3. 热点问题挖掘

3.1 任务的分析

对于热点问题的挖掘，我们需要解决攻克下面几个难点：

第一，在自然语言处理应用领域中，命名实体识别是信息检索、机器翻译、问答系统等多项自然语言处理应用的基础任务，其识别的准确率将会直接影响到后续的一系列工作。这方面的难点有：

- 命名实体在不同领域或不同场景下的识别具有较大的差异。目前已标注的语料通常局限于某些领域，难以适用于其他语料，例如：基于新闻语料进行训练，然后在社交语料进行测试，测试的结果往往难以达到理想的效果，因为社交语料中存在大量非规范的用语。
- 命名实体识别标注成本较大，目前命名实体识别标注语料较少，如何从较少的语料中学习到较好的模型，或者借助其他相似任务语料以及大量未标记的语料进行学习。这些问题都是命名实体识别的挑战。
- 中文命名实体识别中“字”的边界是确定的，但是“词”的边界是模糊的，因此通常会出现一些语义理解歧义的情况，不同分词方案的语句意思完全不一样。中文命名实体识别通常要与中文分词、浅层语法分析等过程相结合，而分词、语法分析的准确率直接影响了命名实体识别的效果。

待识别的文本中存在着大量新的实体词，随着文本出现时间的增加，难以维护这些新词。

第二，提取问题需要使用自动摘要方法。这方面的难点有：

- 理解文档。和人类阅读一篇文章一样，可以说明白文档的中心思想，涉及到的话题等等。
- 可读性强。可读性是指生成的摘要要能够连贯（Coherence）与衔接（Cohesion），通过图灵测试。
- 简练总结。在理解了文档意思的基础上，提炼出最核心的部分，用最短的话讲明白全文的意思。

3.2 数据预处理

- getdata：我们取留言主题和留言详情组合作为我们的数据集。
- 删除标点符号：通过 python 正则表达式去除标点符号。
- jieba 分词：

  英语文本中，单词之间采用空格作为强制分隔符，而中文文本没有这种空格区域，为了解决中文分词问题，我们采用由中国程序员开发的 jieba 库进行分词处理。

  jieba 分词算法使用了基于前缀词典实现高效的词图扫描，生成句子中汉字所有可能生成词情况所构成的有向无环图(DAG)，再采用了动态规划查找最大概率路径，找出基于词频的最大切分组合，对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法。
- 文档—词汇矩阵：

  由于 lda 主题模型需要输入文档-词汇矩阵，所以需要提前把数据预处理。我们使用 from sklearn 里面的 CountVectorizer *api* 进行文本特征提取，它只考虑每种词汇在该训练文本中出现的频率，CountVectorizer 会将文本中的词语转换为词频矩阵，它通过 fit_transform 函数计算各个词语出现的次数。

3.3 命名实体识别模型的选择

3.3.1 基于统计模型的命名实体识别方法归纳

- 隐马尔可夫模型（HMM）

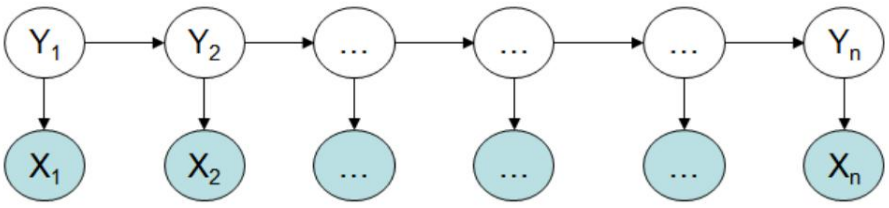
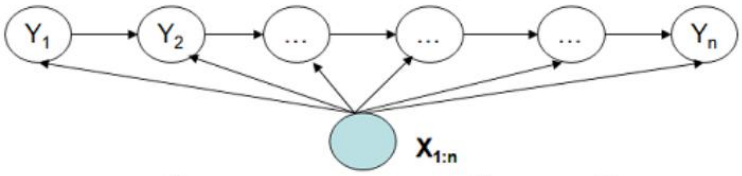


图 3.3-1 HMM

HMM 具有强大的统计基础，具有高效的学习算法，可以直接从原始序列数据进行学习。它允许以本地可学习的方法形式对插入和删除罚分进行一致的处理，并且可以处理可变长度

的输入。它们是序列配置文件的最灵活的概括。它还可以执行多种操作，包括多重对齐，数据挖掘和分类，结构分析和模式发现。合并到库中也很容易。

- 最大熵马尔可夫模型（MEMM）



$$P(\mathbf{y}_{1:n}|\mathbf{x}_{1:n}) = \prod_{i=1}^{n} P(y_i|y_{i-1},\mathbf{x}_{1:n}) = \prod_{i=1}^{n} \frac{\exp(\mathbf{w}^T\mathbf{f}(y_i,y_{i-1},\mathbf{x}_{1:n}))}{Z(y_{i-1},\mathbf{x}_{1:n})}$$

图 3.3-2 MEMM

MEMM 考虑到相邻状态与整个观察序列之间的依赖性，因此具有更好的表达能力。MEMM 不考虑 P（X），因为 P（X）减少了建模工作量，并且了解了目标函数和估计函数之间的一致性。

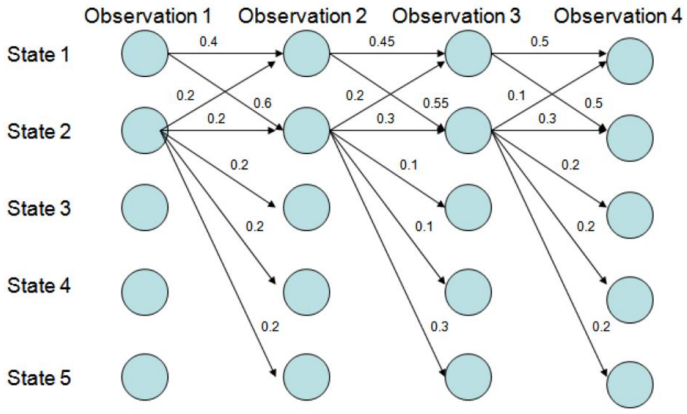


图 3.3-3 MEMM 的 Viterbi 算法解码

在图 3.3-3 中，状态 1 倾向于转换为状态 2，而状态 2 倾向于同时停留在状态 2。因为状态 2 比状态 1 具有更多的可转换状态，因此降低了转换概率。MEMM 倾向于选择具有较少可转换状态的状态。这种选择称为标签偏差问题。CRF 很好地解决了标签偏差问题。

- 条件随机场（CRF 模型）

条件随机场（Conditional Random Fields，简称 CRF）[8]是给定一组输入序列条件下另一组输出序列的条件概率分布模型，CRF 是马尔科夫随机场的特例，它假设马尔科夫随机场中只有 X 和 Y 两种变量，X 一般是给定的，而 Y 一般是在给定 X 的条件下我们的输出。这样马尔科夫随机场就特化成了条件随机场。

对于 CRF，我们给出准确的数学语言描述：设 X 与 Y 是随机变量，P(Y|X)是给定 X 时 Y 的条件概率分布，若随机变量 Y 构成的是一个马尔科夫随机场，则称条件概率分布 P(Y|X) 是条件随机场。

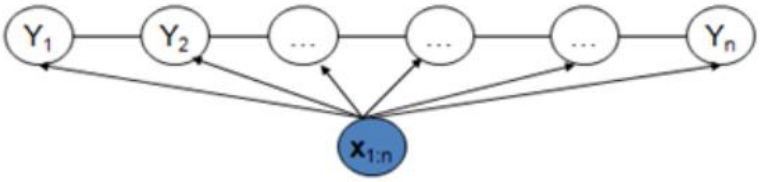


图 3.3-4 CRF

3.3.2 模型的选择

HMM 对转移概率和表现概率直接建模，统计共现概率。HMM 算法默认只考虑前一个状态（词）的影响，忽略了更多上下文信息（特征）。后来的 MEMM、CRF，都是循序渐进的改进方法。

MEMM 是对转移概率和表现概率建立联合概率，统计时统计的是条件概率，由于其只在局部做归一化，所以容易陷入局部最优。

CRF 是在全局范围内统计归一化的概率，而不像是 MEMM 在局部统计归一化概率。是全局最优的解。解决了 MEMM 中标注偏置的问题。

CRF 与其他模型比较：

- 与 HMM 比较。CRF 没有 HMM 那样严格的独立性假设条件，因而可以容纳任意的上下文信息。特征设计灵活（与 ME 一样），适合我们的短文本内容。
- 与 MEMM 比较。由于 CRF 计算全局最优输出节点的条件概率，它还克服了最大熵马尔可夫模型标记偏置（Label-bias）的缺点。对于全局求解有更好的下过。
- 与 ME 比较。CRF 是在给定需要标记的观察序列的条件下，计算整个标记序列的联合概率分布，而不是在给定当前状态条件下，定义下一个状态的状态分布。

	precision	recall	f1-score	support
B_LOC	0.819	0.806	0.813	253
I_LOC	0.841	0.809	0.825	1134
B_ORG	0.881	0.94	0.909	764
I_ORG	0.925	0.935	0.93	1215
B_PER	0.958	0.944	0.951	432
I_PER	0.957	0.964	0.961	808
B_T	0.993	0.67	0.8	415
I_T	0.995	0.727	0.84	1046
micro avg	0.918	0.857	0.886	6067
macro avg	0.921	0.849	0.879	6067
weighted avg	0.923	0.857	0.884	6067

表 3.3-1 CRF 测验数据

我们采用的是短文本内容，文本词汇稀疏，共现概率低，所以选择 CRF 模型。

在进行模型训练时，选用人民日报语料+MSRA 语料进行训练，我们选择训练的实体词依次为 nr（人名）、ns（机构名）、nt（地点）。

对语料需要做以下处理：

① 将语料全角字符统一转为半角；

② 合并语料库分开标注的姓和名，例如：温/nr 家宝/nr；

③ 合并语料库中括号中的大粒度词，例如：[国家/n 环保局/n]nt；

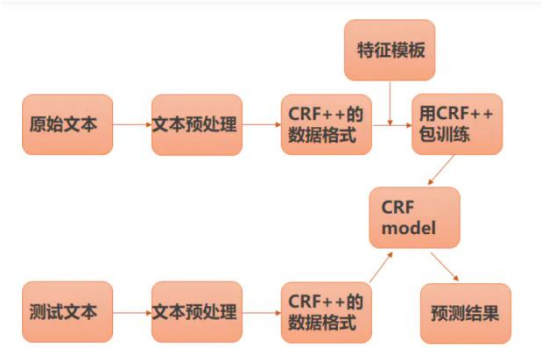


图 3.3-5 使用 CRF 模型处理思路

3.4 主题提取

3.4.1 提取问题描述：

自动摘要（Automatic Summarization）的方法主要有两种：Extraction 和 Abstraction。其中 Extraction 是抽取式自动文摘方法，通过提取文档中已存在的关键词，句子形成摘要；Abstraction 是生成式自动文摘方法，通过建立抽象的语意表示，使用自然语言生成技术，形成摘要。由于自动摘要方法需要复杂的自然语言理解和生成技术支持，应用领域受限。抽取式摘要成为现阶段主流，它也能在很大程度上满足人们对摘要的需求。

目前抽取式的主要方法：

- 基于统计：统计词频，位置等信息，计算句子权值，再简选取权值高的句子作为文摘，特点：简单易用，但对词句的使用大多仅停留在表面信息。
- 基于图模型：构建拓扑结构图，对词句进行排序。例如，$TextRank/LexRank$
- 基于潜在语义：使用主题模型，挖掘词句隐藏信息。例如，采用 LDA，HMM
- 基于线路规划：将摘要问题转为线路规划，求全局最优解。[5]

3.4.2 TextRank 算法

我们采用的 $textrank$ 算法，$TextRank$ 是受到 Google 的 $PageRank$ 的启发，通过把文本分割成若干组成单元(单词、句子)并建立图模型，利用投票机制对文本中的重要成分进行排序，仅利用单篇文档本身的信息即可实现关键词提取、文摘。和 LDA、HMM 等模型不同，$TextRank$ 不需要事先对多篇文档进行学习训练，因其简洁有效而得到广泛应用。

TextRank 一般模型可以表示为一个有向有权图 $G = (V, E)$，由点集合 V 和边集合 E 组成，E 是 $V \times V$ 的子集。图中任两点 V_i, V_j之间边的权重为w_{ji}，对于一个给定的点 V_i， 为指向该点的点集合，$\mathrm{Out}(V_i)$为点V_i指向的点集合。点V_i的得分定义如下:

$$WS(V_i) = (1-d) + d * \sum_{V_i \in \mathrm{In}\ (V_i)} \frac{w_{ji}}{\sum_{V_i \in \mathrm{Out}(V_i)} w_{jk}} WS(V_j)$$

其中, d 为阻尼系数，取值范围为 0 到 1，代表从图中某一特定点指向其他任意点的概率，一般取值为 0.85。

使用 TextRank 算法计算图中各点的得分时，需要给图中的点指定任意的初值，并递归计算直到收敛，即图中任意一点的误差率小于给定的极限值时就可以达到收敛，一般该极限值取 0.0001。

3.4.3 LDA 文档主题生成模型

LDA（Latent Dirichlet Allocation）是一种文档主题生成模型[7]，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，就是指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。

LDA 模型可以把文档—词汇矩阵变成文档-主题矩阵（分布）和主题-词汇矩阵（分布）。

我们选取了 python 里面的 lda 库进行主题分类，根据识别出来不同的命名实体对应的语料输入模型，进行分类。

这里涉及一个最优主题数问题，lda 主题模型需要在训练前输入一个最优主题数，为了确定最优主题数，我们使用困惑度进行判断。

其中困惑度可以理解为对于一篇文档 D，所训练出来的模型对文档 D 属于哪个主题有多不确定，这个不确定成都就是困惑度。困惑度越低，说明聚类的效果越好。

计算公式:

$$perplexity(D) = e^{-\frac{\sum logp(w)}{\sum_{d=1}^{M} N_d}}$$

其中:

$\sum_{d=1}^{M} N_d$是测试集中所有单词之和，即测试集的总长度。

$p(w)$指的是测试集中每个单词出现的概率。

用于 lda 主题模型计算公式:

$$p(w) = p(z|d) * p(w|z)$$

其中:

$p(z|d)$表示的是一个文档中每个主题出现的概率。

$p(w|z)$表示的是词典中的每一个单词在某个主题下出现的概率。

我们通过范围（1,20）内，以 1 为步长选取主题数，计算该主题数下的困惑度，最后选取困惑度最小的为最优主题数，确认分类。

操作流程：相关组件为 textrank4zh，其用于抽取中文文章的关键字以及关键句（作为文摘）。

通过 LDA 模型分类后，将同一主题文本输入 $textrank4zh$ 进行主题句分析，得到句子的权重，按照权重输出，权重最大的为关键句，我们选择它作为问题描述。

3.5 热点问题挖掘与分类结果

热点指标是评价某主题受关注流量大小的指标，我们根据数据集特征建立起自己的热点指标。我们取同一主题下的文本数为 $text\_num$，取同一主题下的文本的总支持度为 $support$，取同一主题下的文本的总反对数为 $oppose$，给出一个热度指标计算

$$hotIndex = \text{text}_{\text{num}} * 0.5 + (\text{support} * 0.7 + \text{oppose} * 0.3) * 0.75$$

基于上面建立的模型，训练生成“热点问题表.xls”和“热点问题留言明细表.xls”，见附件。困惑度-主题数曲线见表 3.5-1，可见在主题数确定为 140 时困惑度最小。图 3.5-2 展示了第 n（n=10,700,1500,2500,4000）文本出现在各主题的概率。

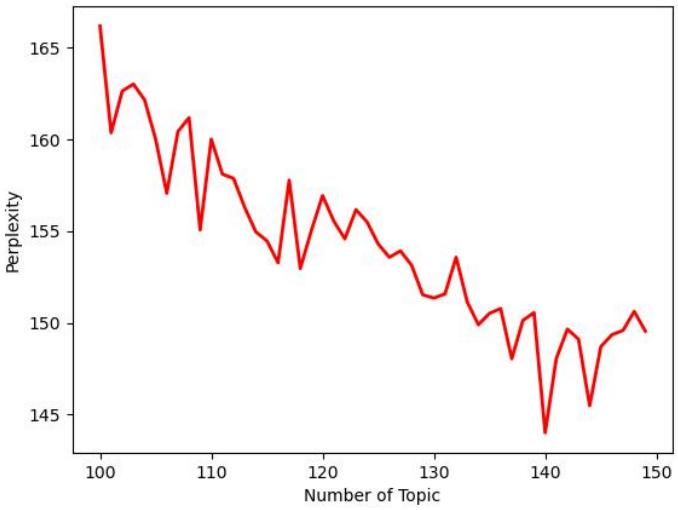


图 3.5-1 困惑度-主题数曲线

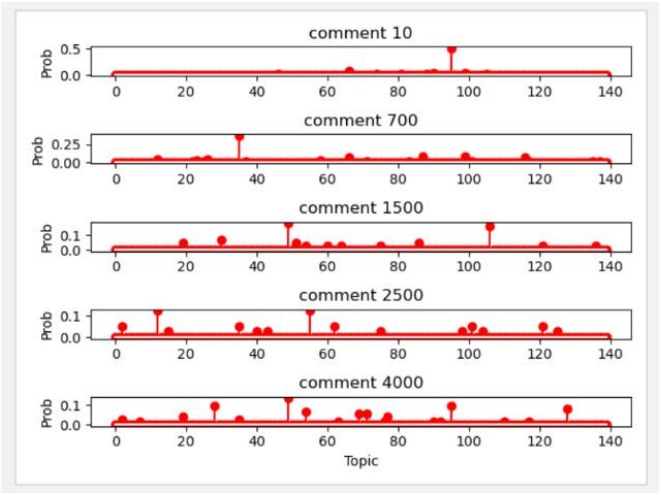


图 3.5-2 文本在每个主题的概率

4. 答复意见评价方案的建立

4.1 任务的分析

通过阅读答复意见的文本集可以看到，大部分的回复与留言关联紧密，对留言中的问题进行了详细的回答。这种回复应该属于优秀回复，在评价系统中给分要高，如图 4.1-1。

	留言编号	留言用户	留言主题	言时	留言详情	答复意见
2	2549	A00045581	A2区景蓉华苑物业管理有问题	25	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？	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疏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

图 4.1-1 高分回复样例

而有些回复的文本与留言毫无关系，根本没有解决留言的提问，甚至有些回复只有一个日期。这种回复应该属于较差回复，在评价系统中给分要低，如图 4.1-2 和图 4.1-3。

23	37459	A00039732	请问B9市带小孩打疫苗要带什么证件	13	请问，带小孩去打疫苗要带什么证件呢？是所有医院的都可以打吗？	2019年1月14日

图 4.1-2 低分回复样例 1

77	179880	A00091124	H市学院在军训期间无理收费	24	1	我是2018级新生家长，H市学院在军训期间无理收费，两双军训袜子，1双鞋垫，1个本子和一支笔共计45元，45元是我孩子3天的生活费，学院如此这般威逼，希望有关部门能处理。	您好，您所反映的问题，已转交相关部门调查处置。

4.1-3 低分回复样例 2

4.2 基于数据特征性的评价标准的建立

关于如何识别出不同质量的回复并分类，是评价系统的重点。基于文献参考[10]，根据国家政府办公规定：坚持以人为本的原则，留言回复必须坚持对组织负责和对人民群众负责相统一，既要坚持政策，又要体现人文关怀；处处事事从当事人的角度去思考，尽可能为当事人提供便捷、实用的回复；回复文风要质朴，方式方法要得当。

我们通过人工理解题目数据集的回复文本，结合有关规定，归纳出高分和低分回复的特点如下：

- 高分回复的特点：
  a) 与留言内容关联紧密
  b) 内容较为详细和丰富
  c) 会引用法律条规
- 低分回复的特点：
  a) 与留言内容毫无关联
  b) 字数很少
  c) 使用套话来回复

对于这些特点，我们决定建立三个评价标准，对每个评价标准给出 0 到 1 分之间的分数，最后把三个得分按照一定的权重求和，得到最终得分。评价标准为：a）字数，b）关键词提及比例，c）是否引用法律条规。

4.3 评价分数量化标准

4.3.1 字数

虽然字数多少不能直接体现回复质量的高低，但是根据对数据集的分析归纳发现，较差的回复普

遍字数较少，而优秀的回复通常字数较多。因此把字数作为其中一个评价标准是简单而又有效的。

在这里我们设计了最小字数 min 和最大字数 max，当回复字数小于最小字数 min 时，该回复在这个标准得分为 0。最大字数时为了避免字数多得分就很高这种情况。当回复字数超过最小字数时，将会给出一个得分

$$mark1 = (\log_m n - move)/(\log_m max - move))$$

m 底数根据得分情况给出。$move$ 是为了将最小字数时的得分和 0 分之间的距离拉近。

在示例集数据中测试，我们变量的设计为 $min = 30$，$\max = 1000$，$m = 2$，$move = \log_2 30$。

得到部分高分回复和低分回复如图 4.3-1 和图 4.3-2。可以看到通过该标准来评价回复质量效果很好。

mark:0.8159237965872557 reply:尊敬的网友：网友反映的问题已收悉，现就有关情况回复如下：根据省厅文件精神，城乡居民养老保险可以进行一次性补缴，因城居保政策从2011年开始实行，最多可以从2011年开始补缴，也就是可以补缴8年（2011年-2018年）。目前缴费标准有每年100元、200元、300元、400元、500元、600元、700元、800元、900元、1000元、1500元、2000元、2500元、3000元共14个档次。2011年-2013年最高缴费标准为1000元，2014年至今，每年的缴费标准最高为3000元。参保人员可以自主选择档次缴费。但是根据《西地省人力资源和社会保障厅关于暂停灵活就业人员和无雇工的个体工商户职工基本养老保险补建补缴的通知》（楚人社函[2016]96号）规定，企业养老保险中，个人缴费从2016年4月1日起已经不能一次性补建补缴了，只能逐年缴纳。2018年12月29日网友您好：您反映的问题已收悉并交有关单位调查核处，如有情况将及时回复！2018年12月27日
mark:0.8632384369615925 reply:尊敬的网友：经查，欧洲城小区一、二、三号栋由西地省富东集团B9市新兴房地产开发有限公司开发，地址位于B9市五里牌村。该项目现处于地下室浇筑建设阶段，针对市民反映夜间施工噪声扰民一事，我局环境执法人员约谈施工方B9市远大建设工程有限公司负责人，该公司负责人表示之前确有夜间施工超时现象。根据《中华人民共和国噪声污染防治法》，第四章第三十条规定：“在城市市区噪声敏感建筑物集中区域内，禁止夜间进行产生环境噪声污染的建筑施工作业，但抢修、抢险作业和因生产工艺上要求或者特殊需要必须连续作业的除外，因特殊需要必须连续作业的，必须有县级以上人民政府或者其有关部门的证明”，该工地夜间施工并未向B9市环境保护局报备审批。执法人员要求项目负责人立即改正违法行为、如需特殊工艺连续作业，需向环保部门报备，并在附近张贴公示方可施工，如若有再次发生类似事件，将按照相关法律从严处理。欧洲城项目负责人承诺，今后将会严格按照规定要求落实，避免施工扰民。同时，我局将加大监管力度，确保市民良好的生活环境。2019年1月4日网友您好：您反映的问题已收悉并交有关单位调查核处，如有情况将及时回复！2018年12月4日

4.3-1 字数标准高分回复示例

mark:0.05788201005952963 reply:网民您好：您反映的问题建议您拨打12345市长热线，可以就有关问题进咨询，谢谢！
mark:0 reply:2019年1月14日
mark:0 reply:2018年12月12日
mark:0 reply:网友：您好！您所反映的问题，已进行过回复。
mark:0 reply:您好，你所反映的问题已转交相关单位调查处置。
mark:0 reply:您好，你所反映的问题已转交相关单位调查处置。
mark:0 reply:您好，你所反映的问题已转交相关单位调查处置。
mark:0 reply:您好，您所反映的问题，已转交相关部门调查处置。
mark:0 reply:您好，您所反映的问题，已转交相关部门调查处置。

4.3-2 字数标准低分回复示例

4.3.2 关键词提及比例

回复中如果跟留言内容关联紧密，必然会提及留言中的关键词。因此把关键词提及比例作为回复评判的标准之一。

我们通过 $jieba$ 库的 $analyse$ 将每个留言内容的关键词提出，提取数量 topK。然后计算关键词在回复中出现的数量 mark_keyword。由于考虑到回复中不需要出现所有的关键词，只需要提及部分就可以给高分，因此给出 max_keyword 作为 mark_keyword 的最高得分。该项目的得分为

$$\text{mark2} = \text{mark\_keyword}/\text{max\_keyword}$$

得到部分高分和低分的回复如图 4.3-3 和图 4.3-4。

问题：关于F9市龙源水库移民后扶政策咨询

颜局长：　　你好！我原是西地省F9市龙源水库移民户口，由于2005年户口婚迁至湖北省后，就取消了移民户待遇。请问我能否享受国家关于水库移民后扶政策？

['移民', 'F9', '市龙源', '水库', '户口', '我原', '西地省', '2005', '关于', '政策', '迁至', '你好', '湖北省', '待遇', '请问', '局长', '咨询', '取消', '享受', '能否']
回答：网名“A00094791”：　　从你的来信中得知，你虽是西地省F9市龙源水库的移民，但由于你的户口已于2005年迁入湖北省，因此，根据国务院2006年17号文件的规定，你应在湖北省享受移民后扶政策。有关具体情况请咨询当地移民部门。 西地省水库移民开发管理局　2013年8月28日
问题：咨询森林公安领导年龄问题

尊敬的领导　　您好！我是西地省某县的一名基层干部，今年48岁，有多年的纪检和政法工作经验，县里打算将我交流到森林公安担任领导职务，经向市局请示，市局以我年龄已超过45岁，省局不会批准授警衔为由予以拒绝。请问省里是否有此项规定，省里是否有相关政策放宽政法，纪检干部从警授衔，如果有，授衔的最大放宽年龄是多少？

['授衔', '市局', '纪检', '政法', '年龄', '省里', '公安', '放宽', '从警', '森林', '西地省', '48', '经向', '45', '某县', '省局', '警衔', '领导', '是否', '基层干部']
回答：基层干部：　　你好！首先感谢你对森林公安事业的热爱。　　根据国家录警政策，科级干部年龄不得超过35周岁，特殊情况不得超过40周岁。从县以上地方政法机关、纪检机关、组织部门交流到公安机关任职的，适当放宽录警授衔年龄，一般不得超过40周岁，特殊情况不得超过45周岁。由于你已经48周岁，不能录警授衔。　2016年4月26日

4.3-3 关键词提及比例高分回复示例

问题：B2区泉中路车辆乱停乱放，占用人行横道

敬爱的阳市长：您好！我是辉煌时代的业主，小区门口的泉中路一直没有相关部门来管理整改，道路两边车子扎堆，乱停乱放，占用人行横道，占用消防通道，小孩出门回家安全问题堪忧，之前投诉到B2区交警执法部门，交警部门给的答复是该道路他们没权执法，说这条路没有路标路牌，没有划线，后期投诉到B2区建设局，申请来整改，但一直未给予答复，没有过来处理。恳请市长督促相关职能部门抓紧落实处理，给我们广大业主一个安稳安全的居住环境。

['B2', '占用', '人行横道', '乱放', '整改', '中路', '投诉', '答复', '业主', '执法', '没有', '市长', '道路', '区泉', '条路', '消防通道', '安全', '划线', '路牌', '处理']
回答：尊敬的网友：您好，来信收悉！经区住建局核实，我局已经做了过渡方案，图纸已经出来了。区住建局审核盖了章，经市交警支队审批后就可以实施了。感谢来信！
问题：请问B9市带小孩打疫苗要带什么证件

请问，带小孩去打疫苗要带什么证件呢？是所有医院的都可以打吗？

['证件', '请问', '小孩', '疫苗', 'B9', '市带', '什么', '医院', '所有', '可以']
回答：2019年1月14日

4.3-4 关键词体积比例低分回复示例

4.3.3 是否引用法律条规

优秀的回复应当适当引用法律条规和说明来回复群众的问题，但是由于存在一些回复可能不需要添加法律条规，因此这一项标准的得分占比不会很高。

我们通过添加一些引用法律条规时的关键词，如“出台”,“文件精神”,“通知”和“下发”等，来识别回复中是否引用了法律条规。如果引用了法律条规给 1 分，如果没有引用给 0 分。因此得分表示为

$$\text{mark3} = \text{is\_cited}$$

得到引用了法律条规和未引用的示例如图 4.3-5 和图 4.3-6。

市民同志：你好！您反映的“请加快提高民营幼儿园教师的待遇”的来信已收悉。现回复如下：为了改善和提高民办幼儿园教师待遇，根据2019年1月8日出台的《中共A市委A市人民政府关于学前教育深化改革规范发展的实施意见》长发〔2019〕2号文件精神，对于学前教育教师的培养和待遇问题做出了明确要求。一是在提高教师待遇方面，依法保障民办幼儿园教职工待遇，民办幼儿园聘任教职工要依法签订劳动合同，依法缴纳城镇企业职工养老保险、医疗保险、生育保险、工伤保险、失业保险和住房公积金，民办园要参照当地公办园教师工资收入水平，合理确定相应教师的工资收入。二是加强监管协同推进，加强对民办幼儿园的日常监管和质量管理，保障民办幼儿园教师待遇，在完善人事（劳动）、工资待遇、社会保障和职称评聘等方面继续推进。谢谢您对我市学前教育的关注和支持！

网友“A000110735”：您好！您在平台《问政西地省》上的留言已收悉，市住建局及时将您反映的问题交由市房屋交易管理中心办理。现将相关情况回复如下：按照《A市人才购房及购房补贴实施办法（试行）》第七条规定：新落户并在A市域内工作的全日制博士、硕士毕业生（不含机关事业单位在编人员），年龄35周岁以下（含），首次购房后，可分别申请6万元、3万元的购房补贴。“首次购房”是指在A市限购区域内首次购买商品住房（含住宅类公寓）。因此，如购买商业性质公寓（非商品住房），则不可申领购房补贴。以上情况，望您知晓和理解。如您还有疑问，建议可拨打市房屋交易管理中心咨询电话0000-00000000详询。特此回复！2019年4月30日

4.3-5 引用了法律条规示例

网友“A0009233”，您好，您的留言已收悉，现将具体内容答复如下：关于来信人建议“白竹坡路口”更名为“马坡岭小学”，原“马坡岭小学”取消，保留“马坡岭”的问题。公交站点的设置需要方便周边的市民出行，现有公交线路均使用该三处公交站站名，市民均已熟知，因此不宜变更。感谢来信人对我市公共交通的支持与关心。2019年5月5日

网友“A00077538”：您好！针对您反映A3区含浦镇马路卫生很差的问题,A3区学士街道、含浦街道高度重视，现回复如下：您留言中反映的含浦镇在2013年已经析出两个街道，分别是学士街道和含浦街道，鉴于您问题中没有说明卫生较差的具体路段，也没有相应的参照物，同时您也未留下联系方式，请您看到回复后，致电学士街道0731-0000-00000000或者含浦街道0731-0000-00000000反映相关问题。感谢您对我们工作的关心、监督与支持。2019年4月24日

4.3-6 未引用法律条规示例

4.3.4 评价整合

1) 遍历$Q^{'}$中的所有元素，根据上述的三种标准给出该回答的各项得分，表示为向量$M_k = (F_1,F_2,F_3)$。
2) 设置权重向量为$\lambda = (\lambda_1,\lambda_2,\lambda_3)$，计算 $\mathrm{F(K)} = M_k\lambda^T$作为政府此条回答的最终回复质量评价函数 F(K)。将每项回复内容的三个标准得分整合，得到该回复的总评价。每个标准的得分都是 0-1 分，通过权重和为 1 的三个权重对标准得分进行求和，得到总评价也是 0-1 分。
3) 这里我们字数，关键词体积比例，是否引用法律条规的三个权重分别为 0.2，0.7 和 0.1。

结合上面的评价标准整合，可以得到该评价系统的总得分为

$$\mathrm{mark_{total}} = 0.2\mathrm{mark1} + 0.7\mathrm{mark2} + 0.1\mathrm{mark3}$$

测试中参数设定分别为 $min = 30$，$\max = 1000$，$m = 2$，$move = \log_2 30$, $\mathrm{topK} = 15$，得到部分高分回复和低分回复如图 4.3-7 和图 4.2-8。通过结果可以发现，该评价系统能有效区分优秀回复和较差回复，效果很好。得到部分高分回复和低分回复如图 4.3-7 和图 4.2-8。通过结果可以发现，该评价系统能有效区分优秀回复和较差回复，效果很好。

mark = 0.9521334529365574
尊敬的网友：经查，欧洲城小区一、二、三号栋由西地省富东集团B9市新兴房地产开发有限公司开发，地址位于B9市五里牌村。该项目现处于地下室浇筑建设阶段，针对市民反映夜间施工噪声扰民一事，我局环境执法人员约谈施工方B9市远大建设工程有限公司负责人，该公司负责人表示之前确有夜间施工超时现象。根据《中华人民共和国噪声污染防治法》，第四章第三十条规定：“在城市市区噪声敏感建筑物集中区域内，禁止夜间进行产生环境噪声污染的建筑施工作业，但抢修、抢险作业和因生产工艺上要求或者特殊需要必须连续作业的除外，因特殊需要必须连续作业的，必须有县级以上人民政府或者其有关部门的证明”，该工地夜间施工并未向B9市环境保护局报备审批。执法人员要求项目负责人立即改正违法行为、如需特殊工艺连续作业，需向环保部门报备，并在附近张贴公示方可施工，如若有再次发生类似事件，将按照相关法律从严处理。欧洲城项目负责人承诺，今后将会严格按照规定要求落实，避免施工扰民。同时，我局将加大监管力度，确保市民良好的生活环境。2019年1月4日网友您好：您反映的问题已收悉并交有关单位调查核处，如有情况将及时回复！2018年12月4日
****
mark = 0.9300527436572376
网友：您好！您反映的问题已收悉，现将具体情况回复如下：根据《B市B2区2019年中小学生招生方案》安排，我区于6月2日—6日开展B2区常住居民子弟及符合试点条件的外来经商、务工随迁子弟入学报名工作。B2区常住居民子弟依据《B2区2019年义务教育阶段公办学校招生范围（常住居民子弟）》进行报名登记。报名登记时须提供户口本、出生证、预防接种证（含查验证明）、房产证（无房产者提供无房证明）或购房合同等有效证件或材料。针对您反映问题，您的孩子属于B2区城区户籍，购买的翡翠公园房屋目前尚未办理房产证。您可以按照上述要求提供材料申请登记，并按相关提示办理入学手续。如果您孩子错过或因其他原因没有及时登记办理，您还可以在7月25日—8月8日通过网上预约的方式登记报名。如果您仍存在疑问，欢迎到局基教股咨询，地址：B市B2区枫溪大道668号；联系电话：0000-00000000（基教股）。感谢您对我局工作的支持和理解，祝您生活和工作愉快。
****

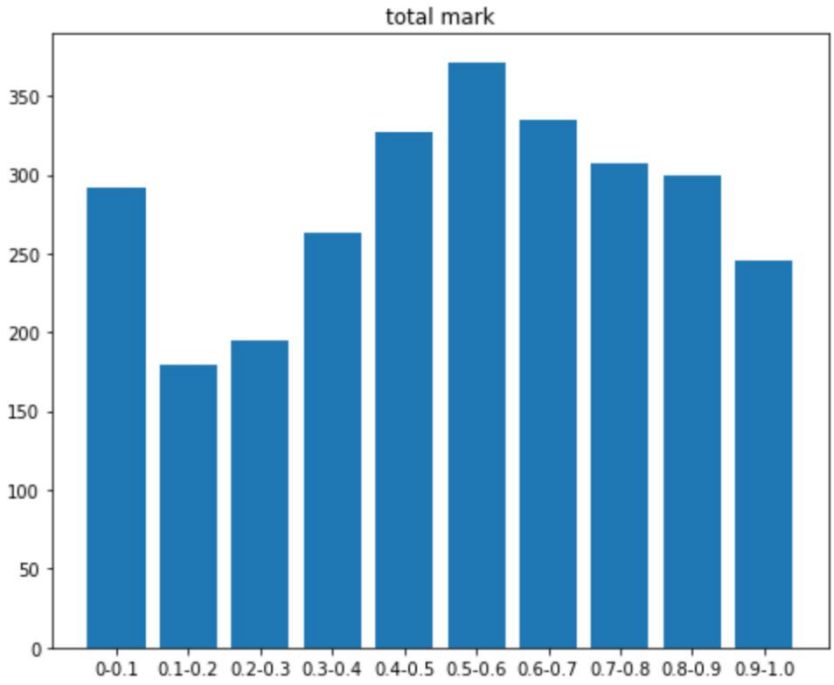
4.3-7 高分回复示例

mark = 0.08136981463194648
网民您好：您反映的问题建议您拨打12345市长热线，可以就有关问题进咨询，谢谢！
****
mark = 0.1935419707605571
尊敬的网友：您好，来信收悉！经核实，西环线西辅道剩余部分及西环线东辅道目前正在开展征地拆迁工作，预计明年底可完工通车。感谢来信！
****
mark = 0.08244054604522051
尊敬的网友：您好，来信收悉！经区住建局核实，我局已经做了过渡方案，图纸已经出来了。区住建局审核盖了章，经市交警支队审批后就可以实施了。感谢来信！
****
mark = 0.0
2019年1月14日
****
mark = 0.0
2018年12月12日
****

4.3-8 低分回复示例

4.4 总分数的分布情况

总分数的分布情况如下图所示。可以看到除了 0-0.1 的分数以外，其他的分数分布中间分数多两边分数低。这个分布符合评判标准的合理分布。而由于不符合本评判系统的三个标准的答复比较多，因此使得获得 0-0.1 评分的回复相对较多。



4.4-1 总分数的分布直方图

5.相关性验证

5.1 F-Score 评价

机器学习中分类模型的精确率(Precision)和召回率(Recall)评估指标。对于 Precision 和 Recall，虽然从计算公式来看，并没有什么必然的相关性关系，但是，在大规模数据集合中，这 2 个指标往往是相互制约的。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。

所以，很多时候我们需要综合权衡这 2 个指标，这就引出了一个新的指标 F-score。这是综合考虑 Precision 和 Recall 的调和值。

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

根据题目要求β=1，我们得出的准确达到 92%以上，在作为测试集的 882 数据中正确分类达到 815 个，比起其他的模型，bert 模型在文本分类的准确度上效率更高，准确度更高。

5.2 鲁棒性和可拓展性

我们基于 $bert$ 模型实现的文本分类模型，通过双头注意力机制实现文本多段理解分类，在分类收敛速度较快，基本在 7 次迭代可以得到较好结果，而且对于缺失值和异常值会采用一个一般值取代，保证了鲁棒性。

而我们的 $LDA + CRF$ 文本分拣模型，采用由 c 语言作为的 lda 库进行迭代处理，比起其他方法实现的模型，效率更高，速度更快，对于长度大于 10 的文本，分类效果较为理想。而 CRF 模型仅需要少量已经标注的语料训练，就可以实现命名实体识别，比起其他的命名实体识别方法，可以极大减轻人工标注的压力。

6. 展望

6.1 嵌入层的参数化方法-- 矩阵分解

把模型变小可以提高训练速度，$ALBERT$ 的作者提出了一个很漂亮的解构嵌入层的方法。在 $BERT$ 和 $XLNet$ 中，$embedding\ size\ E$ 和 $hidden\ size\ H$ 大小一样。从模型的角度来看，嵌入层是用来学习独立于环境的表征(context-independent representations)，而隐藏层的 $embeddings$ 是用来学习与环境有关(想一想 Transformer 的 attention module)的表征(context-independent representations)。这二者没有必要保持 $embedding$ 大小一致，完全可以解耦。

从实用的角度来看，嵌入层的 $vocabulary\ size\ V$ 一般都是 $10K$ 的量级。为了模型的精度，H 一般很大。如果 $E = H$，则嵌入层的参数大小 $V \times \backslash times \times H$。这很容易就导致模型有超多参数。

矩阵分解的做法是，把原来的嵌入层矩阵(大小 $V \times \backslash times \times H$)投影到一个低维的空间(维度为 E, $E << H$)，然后再从这个空间投影到隐藏层的空间(维度 H)。算一笔账:原来的嵌入层大小是 $O(V \times \backslash times \times H)$，经过矩阵分解，嵌入层的参数大小合计为 $O(V \times \backslash times \times E + E \times \backslash times \times H)$。

当 H 远大于 E 时，这种参数化方法效果显著。

6.2 动态掩码

$BERT$ 在预训练中随机遮掩 $tokens$，然后保持这些被遮掩的 $tokens$ 不变。这被称为静态掩码。$RoBERTa$ 提出了动态掩码，具体是每一次都对同样的 $sentence$ 产生不同的掩码。在对更大的数据进行更长时间的预训练时，动态掩码比静态掩码效果更好。

6.3 优化 lda 模型主题数确定方法

LDA 主题模型属于最基础的主题模型，采用 $perplexity$ 来选择最优主题数，容易受到词频影响，在短文本分类效率不高。不少研究表明，基于 $perplexity$ 选择的主题，语义上与人工的判别有一定

差距。如果不用 LDA，而是用 HDP，划分出来的主题数多而细，难以满足群众问政留言记录数据挖掘的需求。国外 2016 年以来的研究表明，采用 $semantic\ coherence$ 和 $topic\ exclusivity$ 指标来评价，能较好地解决这一问题。

7. 参考文献

[1]. 雷瑞萍,王晋玲.智慧政府建设中政府与民众互动关系探究[J].中共山西省委党校学报,2019,42(6):72-75.

[2]. 翁士洪. 参与-回应模型：网络参与下政府决策回应的一个分析模型--以公共工程项目为例[J]. 公共行政评论,2014,(5):109-130.doi:10.3969/j.issn.1674-2486.2014.05.007.

[3]. 李炜程. 地方政府回应网络民意问题研究[D].南京工业大学,2019.

[4]. 杨开平,李明奇,覃思义.基于网络回复的律师评价方法[J].计算机科学,2018,45(9):237-242. dol:10.11896/j.issn.1002-137X.2018.09.039.

[5]. 明拓思宇,陈鸿昶. 文本摘要研究进展与趋势[J]. 网络与信息安全学报,2018,4(06):1-10.

[6]. 王月,王孟轩,张胜，等.基于 BERT 的警情文本命名实体识别[J].计算机应用,2020,40(2):535-540. DOI:10.11772/j.issn.1001-9081.2019101717.

[7]. D. M. Blei, et al., "Latent Dirichlet allocation," Journal of Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[8]. T. L. Griffiths and M. Steyvers, "Finding scientific topics," Proceedings of the National Academy of Sciences, vol. 101, pp. 5228-5235, 2004

[9]. 周星瀚,刘宇,邱秀连.基于深度学习和 CRF 的新闻文章的观点提取[J].电子设计工程,2020,28(3):18-22. DOI:10.14022/j.issn1674-6236.2020.03.005.

[10]. 陈文权,余雅洁.网络环境下服务型政府建设的回应性及路径研究——以 2013 年五省(市)书记和省长集中回复网友留言为例[J].中国行政管理,2014(07):74-77.