第八届“泰迪杯”数据挖掘挑战赛

C

题

论

文

参赛题目：“智慧政务”中的文本挖掘应用

“智慧政务”中的文本挖掘应用

【摘要】

随着互联网的发展，从大量的文本数据中挖掘出简要的文本信息，把群众问政留言进行准确的提取、分析、划分、整理，对政府高效的管理和施政有着重要的意义，对社会治理的创新发展起着重要的推动作用。

针对问题一，文本多分类问题，对每种标签进行独立处理，把分类任务当做对多个二分类任务，以此提高分类的正确率，并且需要考虑到标签之间的依赖关系，利用二级三级标签反推一级标签。对文本数据进行预处理，将分词得到的单词集合作为文本特征，并对文本特征进行筛选，将文本用特征空间向量表示，用 TF-IDF 的计算结果表示文档权值，TF-IDF 能同时考虑到在文本中出现的频率和文本中出现特征词的频率。构建分类模型，在得到文本特征向量后，利用深度学习的方法生成分类模型。

针对问题二，热点问题的挖掘，对群众的留言根据特定时间、特定地点、发生的问题进行识别，在众多留言中识别相似的留言，用命名实体识别对地点进行识别。对识别出的相似的留言进行相似度的计算，把特定地点或人群的数据归并，把相似的留言归为同一问题 。相关的问题归并后，量化评价指标，确定指标的定义和计算方法，对指标要考虑到群众反映的次数、群众的关注度、反映人群的多少、群众对留言点赞的次数、相似留言在有一段时间内出现的频率。

针对问题三，答复意见的评价，量化指标，相关部门对于群众留言的反馈是否是相关的，给予群众的答复是否能解决群众留言的问题，答复意见是否满足相关部门的规范性，并且对群众的疑问进行解释。将相关性、完整性、可解释性等描述量化，对答复意见进行计算和评价。

通过解决以上三个问题，解决将群众留言划分给相应的部分，提高解决群众问题的效率，及时了解当下群众最为关心的热点问题，解决问题，体察民情，有望提升政府的管理水平和施政效率。

关键词：多标签文本分类、文本表示、深度学习、指标量化

Abstract

With the development of the Internet, it is of great significance for the efficient management and governance of the government to extract, analyze, classify and sort out the message of the people from a large number of text data and play an important role in promoting the innovative development of social governance.

For problem 1, text multi-classification problem, each label is treated independently, and the classification task is treated as multiple binary classification task, so as to improve the accuracy of classification. Moreover, the dependency relationship between labels should be considered, a nd the second-level and third-level labels should be used to push the first-level labels backward. The text data is preprocessed, the word set obtained by word segmentation is taken as the text feature, and the text feature is screened. The text is represented by the feature space vector, and the document weight is represented by the calculation result of TF-IDF. The classification model is constructed, and after the text feature vector is obtained, the classification model is generated by the method of deep learning.

Aiming at problem two, the mining of hot issues, the message of

the masses according to a specific time, a specific place, the occurrence of the problem to identify, in many messages to identify similar message s, with the name of the entity to identify the location. The similarity of identified similar messages is calculated, and the data of specific places or people are merged, and similar messages are classified as the same problem. After the relevant questions are merged, the evaluation index is quantified, and the definition and calculation method of the index are determined. For the index, the number of the reflection of the masses, the attention of the masses, the number of the reflection of the masses, the number of thumb up comments from the masses, and the frequency of similar comments in a period of time should be considered.

In view of question 3, the evaluation of the reply, relevant departments to the public message feedback is relevant, whether the reply given to the masses can solve the problem of the comments of the masses, and whether the reply meets the standards of the relevant departments.

Key words: multi-label text classification, text representation, deep learning, index quantification

目录

0 引言

0.1 问题重述

问题一：群众留言分类

按照一定的划分体系 （参考附件 1 提供的内容分类三级标签体系）对留言进行分类，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。

问题二：热点问题挖掘

请根据附件 3 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，按表 1 的格式给出排名前 5 的热点问题，并保存为文件“热点问题表.xls″。按表 2 的格式给出相应热点问题对应的留言信息，并保存为“热点问题留言明细表.xls″。

表 1-热点问题表

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	…	2019/08/18 至 2019/09/04	A 市 A5 区魅力之城小区	小区临街餐饮店油烟噪音扰民
2	2	…	2017/06/08 至 2019/11/22	A 市经济学院学生	学校强制学生去定点企业实习
…	…	…	…	…	…

表 2-热点问题留言明细表

问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	点赞数	反对数
1	360104	A012417	A 市魅力之城商铺无排烟管道，小区内到处油烟味	2019/08/18 14:44:00	A 市魅力之城小区自打交房入住后，底层商铺无排烟管道，经营餐馆导致大量油烟排入小区内，每天到凌晨还在营业……	0	0
1	360105	A120356	A5 区魅力之城小区一楼被搞成商业门面，噪音扰民严重	2019/08/26 08:33:03	我们是魅力之城小区居民，小区朝北大门两侧的楼栋下面一楼，本来应是架空层，现搞成商业门面，噪声严重扰民，有很大的油烟味往楼上窜，没办法居住……	1	0
1	360106	A235367	A 市魅力之城小区底层商铺营业到凌晨，各种噪音好痛苦	2019/08/26 01:50:38	2019 年 5 月起，小区楼下商铺越发猖张，不仅营业到凌晨不休息，各种烧烤、喝酒的噪音严重影响了小区居民休息……	0	0
…	…	…	…	…	…	…	…
1	360109	A0080252	魅力之城小区底层门店深夜经营，各种噪音扰民	2019/09/04 21:00:18	您好：我是魅力之城小区的业主，小区临街的一楼是商铺，尤其是餐馆夜宵摊等，每到凌晨都还在营业，每到晚上睡觉耳边都充斥着吆喝……	0	0
2	360110	A110021	A 市经济学院寒假过年期间组织学生去工厂工作	2019/11/22 14:42:14	西地省 A 市经济学院寒假过年期间组织学生去工厂工作，过年本该是家人团聚的时光，很多家长一年回来一次，也就过年和自己孩子见一次面，可是这样搞……	0	0
2	360111	A1204455	A 市经济学院组织学生外出打工合理吗？	2019/11/5 10:31:38	学校组织我们学生在外边打工，在东莞做流水线工作，还要倒白夜班。本来都在学校好好上课，十月底突然说组织到外省打工……	1	0
…	…	…	…	…	…	…	…
2	360114	A0182491	A 市经济学院变相强制实习	2017/06/08 17:31:20	系里要求我们在实习前分别去指定的不同公司实训，我这的公司的工作内容和老师之前介绍以及我们专业几乎不对口，不做满 6 个月不给实训分，不能毕业……	9	0

问题三：答复意见的评价

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。

1 挖掘目标

1.1 挖掘背景

随着互联网的发展和计算机技术的不断进步，微信、微博、市长信箱、阳光热线等网络问政平台成为人们生活的一部分，传递和表达人们内心的想法，同时也产生了大量的文本数据。若是仅仅依靠人工来对这些大量的文本数据进行整理，不仅需要消耗大量的人力、物力、时间，还容易在对文本数据进行整理时出现错误。

因此，如何在文本数据中高效且快速的挖掘出有用的信息，对文本数据进行分类，帮助工作人员更好的对群众的观点看法、留言进行处理、解决，提高政府的管理水平和施政效率就显得尤为重要了。这时就要合理的应用自然语言处理和文本挖掘的方法对文本数据进行分类、分析与处理。

1.2 挖掘目标

按照附件一的三级分类标签体系，根据附件 2 给出的数据，建立关于群众留言内容的一级分类标签模型。

结合涉及的人群、反映人数的多少、关注度、出现的次数、点赞数定义合理的热度评价指标，根据附件三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，评价出排名前 5 的热点问题。

量化指标对相关部门对留言的答复意见进行计算和评价。

2 问题分析

2.1 问题一的分析

首先了解附件一中给出的三级标签体系的关系，三级标签之间是从属和包含的关系，并且给出的标签是完整的，但数据中并不一定包含全部数据。对数据进行预处理，注意文本语义带来的词语交叉。对于文本多分类问题一般常用的是 word2vec，将自然语言中的转为 计算机可以理解的稠密向量（Dense Vector）。在建立一级分类标签模型方面，利用文本循环神经网络 TextRNN 的深度学习的方法构建模型。

2.2 问题二的分析

对群众的留言根据特定时间、特定地点、发生的问题进行识别，在众多留言中识别相似的留言，用命名实体识别对地点进行识别。对识别出的相似的留言进行相似度的计算，把特定地点或人群的数据归并，把相似的留言归为同一问题 。相关的问题归并后，量化评价指标，确定指标的定义和计算方法，对指标要考虑到群众反映的次数、群众的关注度、反映人群的多少、群众对留言点赞的次数、相似留言在有一段时间内出现的频率。

2.3 问题三的分析

量化指标，相关部门对于群众留言的反馈是否是相关的，给予群众的答复是否能解决群众留言的问题，答复意见是否满足相关部门的

规范性，并且对群众的疑问进行解释。将相关性、完整性、可解释性等描述量化，对答复意见进行计算和评价。

3 问题求解

3.1 问题一：群众留言分类

3.1.1 数据预处理

数据清洗，对特殊字符进行处理，用正则表达式对文本进行替换、删除、查找。其次，利用 jieba 分词，将清洗好的留言数据进行分词，利用词性标注功能，将标注标注句子分词后每个词的词性，将群众留言抽象出来。

```
8532    [\n,  ,  ,  ,  ,  , \n,  ,  ,  ,  ,  ,  ,  , 患...
8533    [\n, \t, \t, \t, \t, \t, \n, \t, \t, \t, \t, \...
8534    [\n, \t, \t, \t, \t, \t, \n, \t, \t, \t, \t, \...
8535    [\n,  ,  ,  ,  ,  , \n,  ,  ,  ,  ,  ,  ,  , 市...
8536    [\n,  ,  ,  ,  ,  , \n,  ,  ,  ,  ,  , 厅长, :, ...
8537    [\n,  ,  ,  ,  ,  , \n,  ,  ,  ,  ,  , 尊敬, 的, ...
8538    [\n,  ,  ,  ,  ,  , \n,  ,  ,  ,  ,  , 尊敬, 的, ...
8539    [\n, \t, \t, \t, \t, \t, \n, \t, \t, \t, \t, \...
8540    [\n,  ,  ,  ,  ,  , \n,  ,  ,  ,  ,  ,  , 守, 在...
8541    [\n, \t, \t, \t, \t, \t, \n, \t, \t, \t, \t, \...
```

去停用词，过滤掉无意义的词，文档中含有大量的代词、介词等没有实际含义的词语，这些单词不携带语义信息，如果保留保留这些单词作为特征会干扰关键词的分类效果，去掉这些单词可降低计算的复杂度，同时使分类更加精确。

```
4548    [小孩, 岁, 转进, 县城, 读, 二年级, 按着, B7, 县, 教育局, 统一, 登记...
4549    [M3, 县, 水车, 镇, 古城村, 建立, 光大, 希望, 小学, 校内, 新建, 大型...
4550    [尊敬, 市委书记, 您好, 现向, A3, 区, 莲花, 镇, 双枫, 中学, 学校食堂,...
4551    [领导, 您好, 工作, 原因, 透漏, 姓名, 工作, 请, 谅解, 所说, 信息, 皆,...
4552    [领导, 您好, 工作, 原因, 透漏, 姓名, 工作, 请, 谅解, 所说, 信息, 皆,...
4553    [您好, 西地省, 艺术, 研究院, 研究员, 国家一级, 编剧, 孙文辉, 单位, 地处,...
4554    [您好, 西地省, 艺术, 研究院, 研究员, 国家一级, 编剧, 孙文辉, 单位, 地处,...
4555    [D6, 县一中, 高一, 学生家长, 开学, 那天, 孩子, 说, 办, 重点班, 还, ...
4556    [L, 市, L2, 县, 人民政府, 相关, 部门, L2, 县铜湾, 镇, 黄溪村, 境...
```

3.1.2 文本特征选择

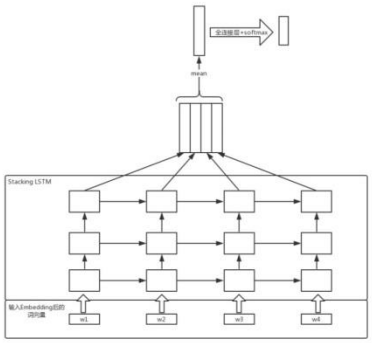
进行词频统计，绘制词云图，用数学方法选择最具分类信息的特征，一个单词出现的文本频数越小，它区别不同类别文本能力就越大。

先转化为词频向量，再转化成 TF-IDF 权重矩阵，以 TF 和 IDF 的乘积作为特征空间坐标系的取值测度，调整权值突出重要单词，抑制次要单词。词频权值用特征词在文本中出现的次数作为贡献度，单词出现的次数组成文档的特征向量。

3.1.3 构建分类模型

得到文档特征向量之后，根据文本循环神经网络 TextRNN 构建分类模型。模型结构包括：Embedding 层，获得词的分布式表示；Stacking LSTM 层，堆叠多个 LSTM，并对 LSTM 的输出在句子的维度取平均值，

将平均后的向量作为包含整个句子信息的向量；Dropout+全链接层。



对输入到模型的句子进行 Word Embedding 将每个词表示成一个数值型的词向量。在经过 embedding 之后，进入 LSTM 层，经过时间序列得到 n 个隐藏 LSTM 神经单元的向量，这些向量经过 mean pooling 层之后，得到向量 h 然后到 Softmax 层，得到一个类别分布概率向量，取概率值最大的类别作为最终预测结果。

3.1.4 分类与评估

要全面评估模型的有效性，必须同时检查精确率 Precision 和召回率 Recall，但是精确率 Precision 和召回率 Recall 往往是一种此消彼长的关系，精确率 Precision 通常会降低召回率 Recall，利用 F-score 综合考虑 Precision 和 Recall 的调和值。

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Precision}=\frac{TP}{TP+FP}$$

$$\text{Recall}=\frac{TP}{TP+FN}$$

$$\text{F1}=\frac{2PR}{P+R}=\frac{2TP}{2TP+FP+FN}$$

TP: True Positive，被判定为正样本，事实上也是正样本。

FP: False Positive，被判定为正样本，但事实上是负样本。

TN: True Negative，被判定为负样本，事实上也是负样本。.

FN: False Negative，被判定为负样本，但事实上是正样本。

引入 F1-Score 作为综合指标，F1-Score 是精确率和召回率的调和平均。

3.2 问题二：热点问题挖掘

3.2.1 相似问题的识别

我们需要利用命名实体识别的方法对留言中的日期、地点等特定的实体进行识别，以便归拢反映相同问题的留言。命名实体识别（NER）系统的目标是识别所有文字提及的命名实体。可以分解成两个子任务：确定 NE 的边界和确定其类型。 命名实体识别非常适用于基于分类器类型的方法来处理的任务。 NLTK 有一个已经训练好的可以识别命名实体的分类器，可以使用函数 nltk.ne_chunk()进行访问。如果我们设置参数 binary=True，那么命名实体只被标注为 NE。否则分类器会添加类型标签，如 PERSON，ORGANIZATION 以及 GPE。如下：

```
# 取出语料库中的一个句子
sent = nltk.corpus.treebank.tagged_sents()[22]

# 使用NE分块器进行命名实体识别，返回Tree对象，
# Tree对象的label()方法可以查看命名实体的标签
for tree in nltk.ne_chunk(sent, binary=True).subtrees():
    # 过滤根树
    if tree.label() == 'S':
        continue

    print(tree)
```

以上代码我们设置参数 binary=True 得到如下结果：

> *(NE U.S./NNP)*
>
> *(NE Brooke/NNP)*
>
> *(NE University/NNP)*
>
> *(NE Vermont/NNP College/NNP)*
>
> *(NE Medicine/NNP)*

如果设置参数 binary=False 则可以得到如下结果：

> *(GPE U.S./NNP)*
>
> *(PERSON Brooke/NNP T./NNP Mossman/NNP)*
>
> *(ORGANIZATION University/NNP)*
>
> *(PERSON Vermont/NNP College/NNP)*
>
> *(GPE Medicine/NNP)*

一旦文本中的命名实体已被识别，我们就可以提取它们之间的关系。我们通常会寻找指定类型的命名实体之间的关系。我们可以找出所有(X, a, Y)形式的三元组。X，Y 是指定类型的命名实体，a 表示 X 和 Y 之间关系的字符串。

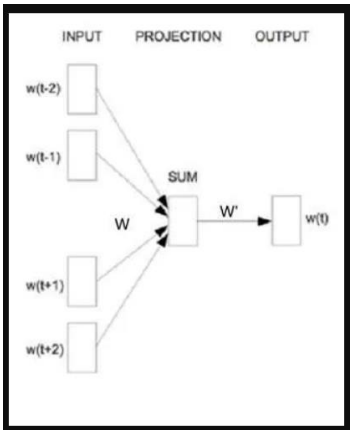
nltk.sem.extract_rels 函数接受的参数主要是主实体名(ORGANIZATION)，宾实体名(LOCATION)，以及代表它们之间关系的正则表达式，该函数返回关系的列表，关系是字典结构。关系中的键 subjtext 可以取出主实体对象，关系中的键 objtext 可以取出宾实体对象。

3.2.2 问题的归类

在归拢出反映相似问题的留言之后，我们需要计算这些不同类留

言之间的相似度，以达到将问题归成大类的目的。首先我们需要先将文本转化为可计算和比较的数学表示的形式。由于第一问已经计算过词频和权重，我们现在可以直接构造 Bag of words(词袋)。但由于词袋模型并不能保留原本语句的语序信息，所以我们现在需要继续用 Word2Vec 训练出词向量来表示一个词。

Word2Vec 有两种训练方法，这里我们运用通过上下文预测中间这个词被称为 CBOW（Continuous Bag-of-Words Model）的方法。它其实是一个三层神经网络模型。如图所示：



以图中的神经网络为例，它 Input Layer 是一个词上下文各两个词的 one-hot 编码 4 个 10 维的向量：w(t-2),...w(t+2)，实际的 Output 是一个概率分布，训练目标是让这个概率分布尽可能接近所要预测词的 one-hot 编码 w(t)

具体过程如下：

1.这里 one-hot 编码是 10 维向量，假设需要得到的是 8 维词向量。

2.通过训练得到两个权值矩阵 W（10×8）和 W’（8 ×10）。

3.从 Input Layer 到 Hidden Layer，w(t-2),...,w(t+2)分别乘上 W 得到 4 个 8 维的向量 v(t-2),...v(t+2)，再取平均得到一个 8 维向量 v’；

4.从 Hidden Layer 到 Output Layer,v’又与 W’相乘，再用 Softmax 处理得到一个概率分布。

5.这个概率分布与要预测的词的 one-hot 编码越接近越好。

6.事实上，v(t-2),...,(t+2)就是我们需要的 8 维词向量；

有了每个词对应的词向量后，我们可以将句子转化为一组向量表示。例如词向量为 8 维，片段一中的每句话有 6 个词，就将每个句子转化为 6 个 8 维的向量作为一组。

然后我们就可以进行相似度的计算了。通常比较两个长度相等的向量(N 维）可以计算余弦距离或欧式距离：

- 余弦相似度：

$$\cos\theta = \frac{\sum_{i=1}^{N} x_i y_i}{\sqrt{\sum_{i=1}^{N} x_i^2}\sqrt{\sum_{i=1}^{N} y_i^2}}$$

- 欧氏距离：

$$dist(X,Y) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2}$$

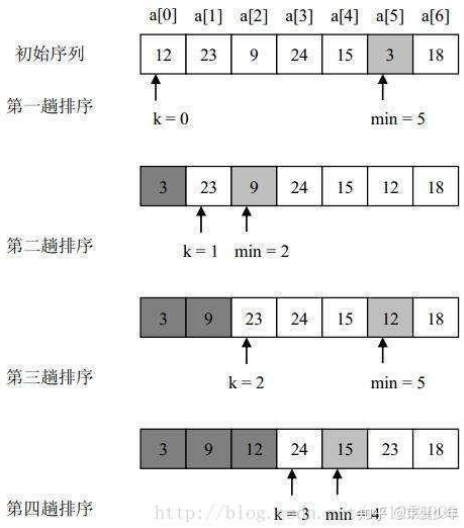
而 Word2Vec 将每个句子转化为了一组 N 维向量，要比较两组 N 维向量，我们需要再将每组 N 维向量转化为一个 N 维向量再计算余弦相似度或欧氏距离。这个过程称为 Sentence embedding。

至此，我们就可以完成问题的归并了。

3.2.3 热度评价

相关的问题归并后，我们要量化评价指标，并给出评价结果，找出排名前五的热点问题。根据附件三中的点赞数和反对数和的大小作为评价标准来判断是否为热点问题。将归并好的问题大类中所有的留言的点赞数和反对数相加，得到的值作为“热点值”，然后将这些热点值按降序排列并选出排在前五的问题作为热点问题。用选择排序法在 Python 中实现“热点值”的降序排列，具体过程如下：

首先在未排序序列中找到最小（大）元素，存放到排序序列的起始位置，然后，再从剩余未排序元素中继续寻找最小（大）元素，然后放到已排序序列的末尾。以此类推，直到所有元素均排序完毕。即选择数组中最小的数字放到 index=0 的位置，数组会越来越小，数组为 0 时，排序完毕。



n 个记录的直接选择排序可经过 n-1 趟直接选择排序得到有序结果。具体算法描述如下：

1.初始状态：无序区为 R[1..n]，有序区为空；

2.第 i 趟排序(i=1,2,3…n-1)开始时，当前有序区和无序区分别为 R[1..i-1]和 R(i..n)。该趟排序从当前无序区中-选出关键字最小的记录 R[k]，将它与无序区的第 1 个记录 R 交换，使 R[1..i]和 R[i+1..n)分别变为记录个数增加 1 个的新有序区和记录个数减少 1 个的新无序区；

3.n-1 趟结束，数组有序化了。

3.3 问题三：答复意见的评价

3.3.1 将相关性、完整性、可解释性等描述量化

对于相关性的指标，给出的答复的内容是否与群众留言中所提出的问题相关，是否正确的对群众问题提出合理有效的解决方法。

对于完整性，答复的格式是否满足规范，使群众群众能够从答复中获得是哪个部门机构在何时对问题进行答复，何时解决问题，若还有疑问时的咨询方法等。

对于可解释性，对于答复意见中相关部门是否给予群众了相关的解释。

3.3.2 构建指标进行计算和评价

4 总结

本文的主要目的利用数据挖掘和建模技术对文本数据建立分类

模型，对热点问题进行数据挖掘，以及对文本内容答复建议的相关性进行评价。

首先对文本数据进行预处理，分词，去停用词，进行词频统计后绘制词云图，先转化为词频向量，再转化成 TF-IDF 权重矩阵，以 TF 和 IDF 的乘积作为特征空间坐标系的取值测度，调整权值突出重要单词，抑制次要单词。词频权值用特征词在文本中出现的次数作为贡献度，单词出现的次数组成文档的特征向量。得到文档特征向量之后，根据文本循环神经网络 TextRNN 构建分类模型。然后对都建好的模型使用 F-Score 的方法进行评价。

5 参考文献

[1]、谢晨阳，武汉大学，基于层次监督的多标签文档分类问题的研究，2018-05-01.

[2]、罗峰，山西大学，基于深度学习的文本情绪多标签分类方法研究，2019-06-01.

[3]、闫琰，北京科技大学，基于深度学习的文本表示与分类方法研究，2016-06-06.

[4]. 《用 Python 进行自然语言处理》宾州大学计算机与信息科学系教材

[5]. Efficient Estimation of Word Representation in Vector Space: https://arxiv.org/pdf/1301.3781v3.pdf

[6].GENSIM:https://radimrehurek.com/gensim/models/word2vec.html