“智慧政务”中的文本挖掘应用

目录

摘　要

随着近年来网络的大力发展，微信、微博、以及各类相关热线的兴起，各类企业以及政府可以合理的借助这类平台进行对客户或者民众意见的收集和整理。该途径可以更加方便快捷的获取信息，也可以更加针对性的实现对问题的解决。因此结合如今当下流行的大数据、云计算、人工智能等技术的发展。“智慧政务”的兴起可谓是十分的顺应潮流。基于已有的自然语言处理技术可以更好地适应如今大信息量，多方面问题的解决。因此合理的建立基于自然语言处理的智慧政务系统已经是如今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。

因此，针对本次“智慧政务”中的文本挖掘应用，主要包含三个解决问题：群众留言分类、热点问题挖掘、答复意见评价。首先针对群众留言分类问题，更具已有的数据，我们根据其一级标签体系，将具有相同的一级标题的评价进行汇总提取，并且运用分词去停用词的方法进行处理，之后用文本向量化的方法进行词频的统计，最后基于此向量化的结果建立模型，进行比较选择，选取最为适合的模型

针对热点问题的挖掘，我们提取某一时间断内反应特定地点或特定人群问题进行归类，并且识别相似的留言，定义合理的评价指标，最后得出评价结果。该问题的重点在于对特定人群的归并以及相似留言的识别，最后有针对性的进行指标排名得出对应的热点问题。

在答复意见评价方面，任何问题的提出，都需要有针对性的回复，因此需要很好地把握及时性，相关性以及完整性三个特点，因此句法分析是自然语言处理中的关键技术，也是该问的核心方法。针对输入的文本句子进行分析以得到句子的句法结构处理过程。本位使用依存关系分析，用于识别句子中词汇的与词汇之间的相互依存关系，运用合理的搭配以及处理方法，因而可以有效地展现出回复的及时性以及规范性。最后建立答复意见的评分要点，进行答复意见效果评分，得到改进措施。

任务一：群众留言分类

1.任务目标：

针对大量的网络问政平台的群众留言，需要按照一定的划分体系对其留言进行分类，根据已有的主题以及留言内容，进行合理的划分，以便后续系统性的分配给相应的职能部门。

2.问题分析方法、流程与具体步骤：

2.1 总体流程：

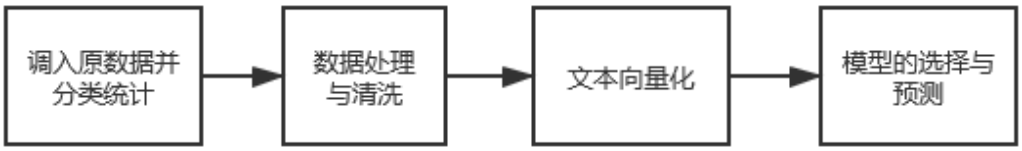

图 1 任务一总体流程

1.调入原数据并分类统计：将已有的数据进行导入，可以观察到数据已有特定的一级分类，将一级分类相同的数据进行整理汇总。

2.数据处理与清洗：对于每一类的留言主题、留言详情进行文本的分词、去停用词的处理。得到较为精准简洁的反馈。

3.文本向量化：针对每一类问题，需要将其精准化，因而能够更好的呈现给相关的管理部门，因此根据上一步的结果，进行文本的向量化从而得到词频的统计可以更加高效的反映问题。其次，计算每个单词在文本中的权重，可以很好地帮助分析统计。

4.模型的选择与预测：为了更好地建立标签分类，需要选择合适的模型，通过多个模型进行分类预测，并且针对每个模型训练出的结果，计算查准率、查全率。最终选择效果最佳的计算 F-Score 值。

2.2 **具体步骤：**

1.针对已有的数据运用 pandas 包进行导入，可以观察到数据已经将各条留言进行了一级标签的分类，因此可以根据该列进行一级标签的分类，提取相同的一级标签的信息进行汇总。

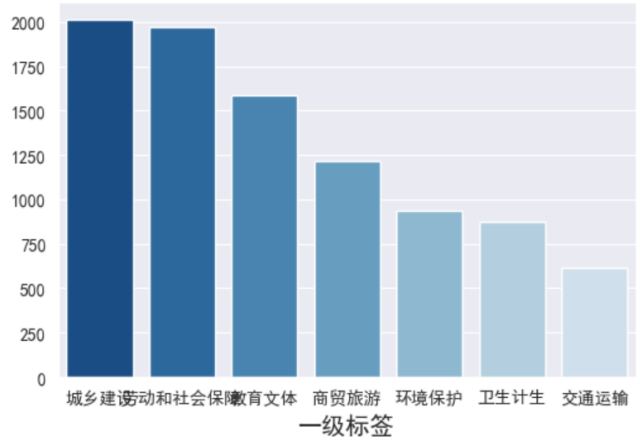

图 1 一级标签分类

2.在数据清洗一块，为了能够更加准确简介的反映问题，我们需要将一条条复杂的反馈进行分词以及去停用词的处理，利用 jieba 函数进行分词的处理，之后运用现有的停用词文本，将分词后的结果进行去停用词的处理，进而可以简化反馈，有助于后续的词频统计与模型构建。

	留言主题	一级标签	留言详情	cuted_留言主题	cuted_留言详情	cuted_l留言详情
0	A市西湖建筑集团占道施工有安全隐患	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\tA3区大道西行便道，未管所路口至加油站路段，...	A 市 西湖 建筑 集团 占 道 施工 有 安全隐患	\n \t \t \t \t \t \n \t \t \t \t \t A3 区 大道 西行...	A3 区 大道 西行 便 道 ， 未管 所 路口 至 加油站 路段 ， 人行道 包括 路灯 ...
1	A市在水一方大厦人为烂尾多年，安全隐患严重	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t位于书院路主干道的在水一方大厦一楼至四楼人为...	A 市 在水一方 大厦 人为 烂尾 多年 ， 安全隐患 严重	\n \t \t \t \t \t \n \t \t \t \t \t 位于 书院 路 主干...	位于 书院 路 主干道 的 在水一方 大厦 一楼 至 四楼 人为 拆除 水 、 电等 设施 ...
2	投诉A市A1区苑物业违规收停车费	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的领导：A1区苑小区位于A1区火炬路，小...	投诉 A 市 A1 区苑 物业 违规 收 停车费	\n \t \t \t \t \t \n \t \t \t \t \t 尊敬 的 领导 ： ...	尊敬 的 领导 ： A1 区苑 小区 位于 A1 区 火炬 路 ， 小区 物业 A 市程明 ...
3	A1区蔡锷南路A2区华庭楼顶水箱长年不洗	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\tA1区A2区华庭小区高层为二次供水，楼顶水箱...	A1 区 蔡锷 南路 A2 区华庭 楼顶 水箱 长年 不洗	\n \t \t \t \t \t \n \t \t \t \t \t A1 区 A2 区华...	A1 区 A2 区华庭 小区 高层 为 二次 供水 ， 楼顶 水箱 长年 不洗 ， 现在 自...
4	A1区A2区华庭自来水好大一股霉味	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\tA1区A2区华庭小区高层为二次供水，楼顶水箱...	A1 区 A2 区华庭 自来水 好大 一股 霉味	\n \t \t \t \t \t \n \t \t \t \t \t A1 区 A2 区华...	A1 区 A2 区华庭 小区 高层 为 二次 供水 ， 楼顶 水箱 长年 不洗 ， 现在 自...

图 2 文件的导入及按一级标签划分

	留言主题	一级标签	留言详情	cuted_留言主题	cuted_留言详情	cuted_l留言详情	labels
3567	还教育一片蓝蓝的天--在学校引入村委会的海选制度	教育文体	\n\n 教育强省，教育强市，教育强县-----这是政府大会小会，三番五次强调的。可县里教...	还 教育 一片 蓝蓝的 天 -- 在 学校 引入 村委会 的 海选 制度	\n \n 教育 强省 ， 教育 强市 ， 教育 强县 ----- 这是 政府 大会...	教育 强省 ， 教育 强市 ， 教育 强县 ----- 这是 政府 大会 小 会 ， 三番五...	3
6955	咨询离退休人员享受一次性生活补助的政策	劳动和社会保障	\n \n 近几年来，我们K市直机关和K1区、K2区两区及K4县的离退休人员...	咨询 离退休人员 享受 一次性 生活 补助 的 政策	\n \n 近几年来 ， 我们 K 市 直机关 和 K...	近几年来 ， 我们 K 市直机关 和 K1 区 、 K2 区 两区 及 K4 县 的 离退休...	4
6891	投诉C3县楚才会计培训学校拖欠工资	劳动和社会保障	\n \n 2014年进入的会计培训学校上班，一直到19年1月18号提出离职...	投诉 C3 县楚 才 会计 培训 学校 拖欠工资	\n \n 2014 年 进入 的 会计 培训 学校...	2014 年 进入 的 会计 培训 学校 上班 ， 一直 到 19 年 1 月 18 号 提...	4
1786	C市易俗河不动产登记过户从一个多月变成5个多月	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t 我按揭买的房子提前还了贷款拿了他项权证，...	C 市 易俗 河 不动产 登记 过户 从 一个多月 变成 5 个 多月	\n \t \t \t \t \t \n \t \t \t \t \t 我 按揭 买...	我 按揭 买 的 房子 提前 还 了 贷款 拿 了 他项权证 ， 6 月 15 号 之前 咨...	0

图 3 进行分次及去停用词的处理

3.文本向量化的构建中，我们调用 sklearn 包，通过 CountVectorizer 函数可以进行文本中出现的所有词语的词频统计，进而可以再通过 TfidfVectorizer 对每个单词的权重进行计算，从而可以根据权重得到反馈中用户更加注重的问题，有助于更加高效精准的解决问题。

4.模型的构建中，为了能够更好地针对今后的大量反馈进行留言的分类，需要进行模型的运用，因此我们选取逻辑回归、多项式朴素贝叶斯、线性支持向量机、随机森林四个模型分别进行训练，从而得到四个模型的预测结果，针对四个模型可以对其准确率进行计算，并且最终我们可以经过比较得到最合适的模型。

```
from sklearn.model_selection import cross_val_score
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import MultinomialNB
from sklearn.svm import LinearSVC

models = [RandomForestClassifier(n_estimators=200, max_depth=3, random_state=0),
          LinearSVC(),
          MultinomialNB(),
          LogisticRegression(random_state=0)]
CV = 7
cv_df = pd.DataFrame(index=range(CV * len(models)))
entries = []
for model in models:
    model_name = model.__class__.__name__
    accuracies = cross_val_score(model, features, labels, scoring="accuracy", cv=CV)
    for fold_idx, accuracy in enumerate(accuracies):
        entries.append((model_name, fold_idx, accuracy))
cv_df = pd.DataFrame(entries, columns=['model_name','fold_idx','accuracy'])
```

图 4 调用四个模型代码展示

5.针对以上的四种模型构建方法，我们得到最为合适的为利用支持向量机的 SVC 模型，在四个模型的比较中，我们进行准确率计算以及 ROC 曲线的绘

制，最终选择支持向量机模型。以下是模型的选择比较结果：

```
In  [12]: cv_df.groupby('model_name').accuracy.mean()

Out[12]: model_name
         LinearSVC                 0.871246
         LogisticRegression        0.795573
         MultinomialNB             0.615317
         RandomForestClassifier    0.396319
         Name: accuracy, dtype: float64
```

图 5 四个模型准确率结果

可见，支持向量机模型的准确率还是很高的，之后我们选取 SVC 模型进行题目要求的 F-Score 指数评价计算：

```
In  [ ]: #支持向量机
         print_evaluation_scores(y_test,prediction)

         accuracy:  0.9044724272687799
         f1_score_macro:  0.9018363035521347
         f1_score_micro:  0.9044724272687799
         f1_score_weighted:  0.9043083081934756
```

图 6 支持向量机评价指标

```
from sklearn.metrics import classification_report
print(classification_report(y_test,prediction))
```

	precision	recall	f1-score	support
0	0.83	0.90	0.86	498
1	0.95	0.95	0.95	229
2	0.97	0.79	0.87	177
3	0.94	0.95	0.94	397
4	0.93	0.95	0.94	513
5	0.88	0.83	0.85	296
6	0.92	0.88	0.90	193
accuracy			0.90	2303
macro avg	0.92	0.89	0.90	2303
weighted avg	0.91	0.90	0.90	2303

图 7 测试集分类结果

任务二：政务热点问题挖掘

1. 任务目标

任务二目的是需要根据给出的数据将某一时间段内反映特定地点或特定人群问题进行归类，定义合理的热度评价指标，并给出评价结果。通过对热点问题的挖掘且即使发现热点问题，有助于相关部门进行针对性地集中处理，提升政府的服务效率。

2. 问题分析方法与过程

2.1 总体流程

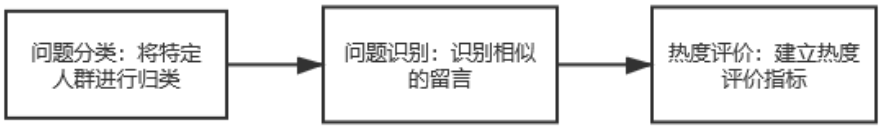

图 8 任务二流程

由于热点问题的挖掘相对复杂，所以将该任务进行任务分解，分解成三个子任务。其中：

子任务一：将特定人群或地点的数据归并；

子任务二：从众多留言中找到相似的留言，结果与特定人群或地点相对应再进行归并；

子任务三： 进行热度评价指标的定义和计算方法对指标排名之后得出对应的热点问题。

2.2 具体步骤

2.2.1 命名实体识别

针对找到特定人群或特定地点需要用到自然语言处理中的命名实体识别。即识别到特定的名词。本文使用到的是 HanLP 中的命名实体识别的方法，命名实

体识别模块的输入是单词列表，输出是命名实体的边界和类别，实现代码如下：

```
Recognizer=hanlp.load(hanlp.pretrained.ner.MSRA_NER_BERT_BASE_ZH)
recognizer([list(i.strip()) for i in data_1["留言"]])
[output]:
```

```
[[('春华镇', 'NS', 5, 8), ('金鼎村', 'NS', 8, 11)],
 [('A2', 'NS', 0, 2), ('黄兴路步行街', 'NS', 3, 9), ('大古道巷', 'NS', 9, 13)],
 [('中海国际社区三期', 'NS', 5, 13)],
 [],
 [('保利', 'NS', 3, 5),
  ('保利麓谷', 'NS', 3, 7),
  ('桐梓坡路', 'NS', 9, 13),
  ('麓松路', 'NS', 14, 17)],
 [('A7 县', 'NS', 0, 3), ('特立路', 'NS', 3, 6), ('东四', 'NS', 7, 9)],
……]
```

通过命名实体识别获取相关地点后，将地点通过列表的形式保存,对留言中含有同一地点的留言归为一类。

图 9 词云展示

通过统计 ('A7 县', 109)('A3 区', 62)('A7', 45)('滨河苑', 39)('伊景园', 35)('A', 35)('A5 区', 29)('A3', 29)('丽发新城小区', 27)('A4 区', 27)('A2', 27)('梅溪湖', 19)('魅力之城小区', 18)('A6 区', 18)('A4', 17)('丽发新城', 15) 可以知道出现次数较多的命名实体依次是 A7、A3、滨河苑、伊景园、A5、丽发新城、A2、梅溪湖、魅力之城、西地省、泉塘街道、A1、A 市、广铁集团、劳动东路、A6、泉星公园、楚龙街道、号线等。

2.2.2 文本相似度计算

对于任务二而言，根据第一步中含有相同地点的留言进行相似度计算，筛选出相似度相对较高的留言。文本相似度问题包含：词、句、段落、篇章相互之间的相似度问题，其中本文使用的是基于文本特征的相似度计算方法：

①将文本转换为 feature vectors(特征向量)：使用 TF-IDF 得到 feature vectors，向量维度为词典大小，向量的每一维度是字段中该位置的词在文本计算的 TF-IDF 值，未在文本中出现则为 0。

其中的公式说明：

$\mathrm{tf_{ij}} = \frac{\mathrm{n_{ij}}}{\sum_k \mathrm{n_{ij}}}$($\mathrm{n_{ij}}$: 在某一类中词条 w 出现的次数, $\sum_k \mathrm{n_{ij}}$: 该类中所有词条数目)

$\mathrm{idf_i} = \log \frac{|\mathrm{D}|}{\left|1 + \mathrm{j}: \mathrm{t_i} \in \mathrm{d_j}\right|}$(|D|: 语料库的文档总数，$\{\mathrm{j}: \mathrm{t_i} \in \mathrm{d_j}\}$：包含此词条 w 的文档数)

$$TF - IDF = TF * IDF$$

②利用 feature vectors 计算文本之间的相似度：可以利用余弦相似度，基于两个文本的特征向量，计算文本之间的相似度：

$$similarity = \cos(\theta) = \frac{A \cdot B}{||A||||B||} = \frac{\Sigma_{i=1}^{n} A_i B_i}{\sqrt{\Sigma_{i=1}^{n} A_i^2}\sqrt{\Sigma_{i=1}^{n} B_i^2}}$$

2.2.3 热度评价指标的建立

目前关于政务的问题其时间跨度较大，并且关于政务问题阅读的数量实际上是很少的，所以关于热度评价指标，本文将赋予点赞数和反对数较大权重，且减少时间的权重。通过设置合理的基准点，有效的数据结果：

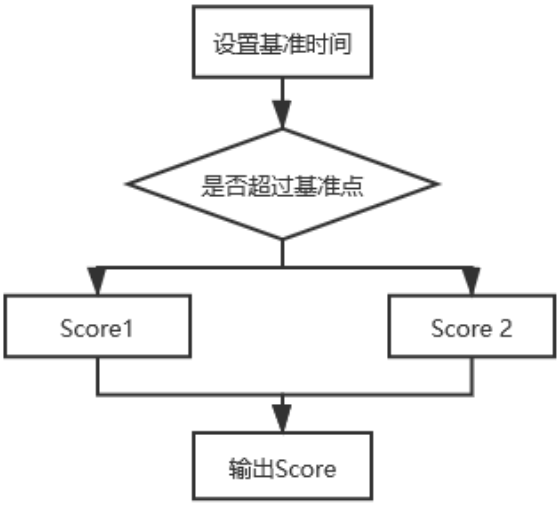

图 10 热度评价指标建立

本文将所有时间的均值设置为基准时间点，取所有时间与基准时间的差值。可视化相差的天数，观察相对时间主要集中的区域在哪一块。

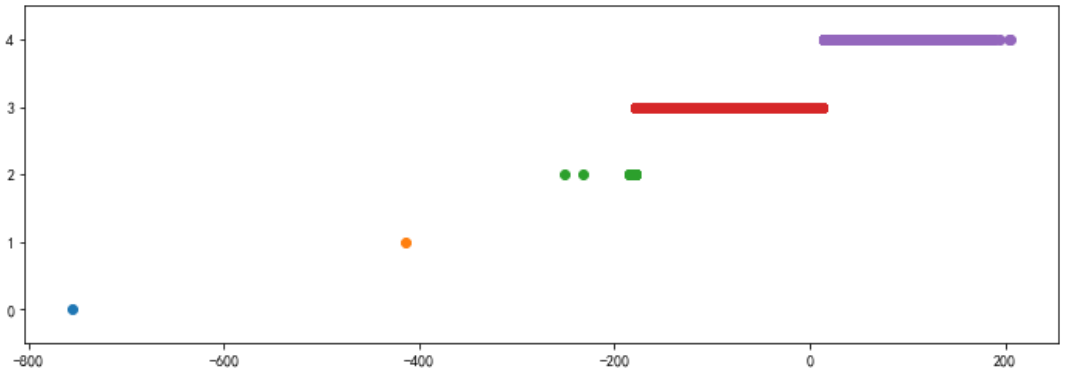

图 11 留言时间范围

由图我们可以看到留言时间基本集中在基准时间的±200day 范围内。所以我们可以集中讨论该范围内的数据。其中 Score1 表示发生在基准时间之前的热度评分，Score2 表示发生在基准时间之后的热度评分。

热度评分公式：

$$Score = S_0 + Score1 + Score2$$

$$Score2 = \Sigma_{相似内容集合} \frac{点赞数 + 反对数}{sum(该时间段内点赞数、反对数)}$$

$$Score1 = \frac{Score2}{T(Time)}$$

（$T(Time) = e^{(k*(T1-T0))}$ 其中k为自定义常数，$T(Time)$为时间衰退因子）

3 结果显示

表 1 热度评价结果

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	3.6	2019/7/10 至 2019/8/29	A 市伊景园	A 市伊景园商品房捆绑车位强制销售
2	2	3.1	2019/7/28 至 2020/1/26	A 市 A2 区丽发新城	A 市 A2 区丽发新城周围因违规修建产生噪音污染严重影响居民的正常生活起居
3	3	2.7	2019/3/24 至 2019/12/30	A 市城市地铁交通	A 市地铁交通存在各种问题：地下通道存在安全隐患、地铁周围环境脏乱差等。且规划多年的一号线、二号线等延申未开始施工
4	4	2.1	2019/3/19 至 2019/11/11	A 市 A3 区江山帝景小区	A 市 A3 区江山帝景小区物业管理不善，对居民的生活造成了严重影响
5	5	1.8	2019/1/22 至 2019/11/13	A 市幼儿园	针对 A 市幼儿园存在惠普性幼儿园还未建立且幼儿园收费较高，教育局学区划分不合理，且幼儿园难近等问题

任务三：答复意见评价

1. 答复意见评价规则建立

任何问题的提出，都必需回复。而政务部门对民众的及时且准确的回复，不仅能进行有效的政务管理，而且由于及时回复的态度会提升民众对政务部门的好感度。由此本文中设定了以下指标来判断和评价留言问题的相关回复。

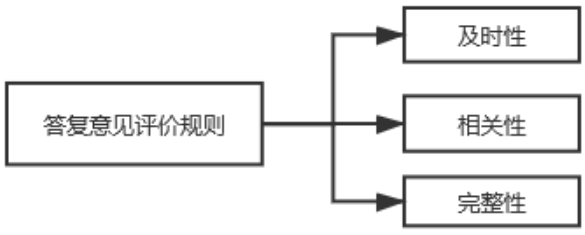

图 12 答复意见评价规则

本文将从时间维度上的及时性，内容上的相关性，以及回复内容的完整型这三个方面来对政务问题的答复进行。

从时间维度上操作，本文将运用最简单的时间差值进行判断。设定阈值为5days，如果时间差值小于 5days，记作得一分，如果差值在一周内，记 0.5 分，其余回复时间记 0 分。

内容是最重要的部分，内容回复的好，尽管回复时间迟，民众也会感受到政府对问题的重视程度。其中答复意见的相关性评价，本文将针对回复与其对应的留言进行一个相似度的计算。对于内容的完整性，本文将从句法和语义以及内容是否详实进行分析。

2. 内容指标评价原理

句法分析（syntactic parsing）是自然语言处理中的关键技术之一，它是对输入的文本句子进行分析以得到句子的句法结构的处理过程。对句法结构进行分析，一方面是语言理解的自身需求，句法分析是语言理解的重要一环，另一方面也为其它自然语言处理任务提供支持。例如句法驱动的统计机器翻译需要对源语言或目标语言（或者同时两种语言）进行句法分析。语义分析通常以句法分析的输出结果作为输入以便获得更多的指示信息。根据句法结构的表示形式不同，本文使用的分析方法是依存关系分析，又称依存句法分析（dependency syntactic parsing），简称依存分析，作用是识别句子中词汇与词汇之间的相互依存关系。以所给数据中一随机抽取留言为例，本文通过 HanLP 中的依存句法分析方法进行句法的分析，利用的是基于神经网络依存句法分析器。

3. 答复意见评价句法分析实例

句子：“业委会制定的停车收费标准不高于周边小区价格“

词法分析结果：

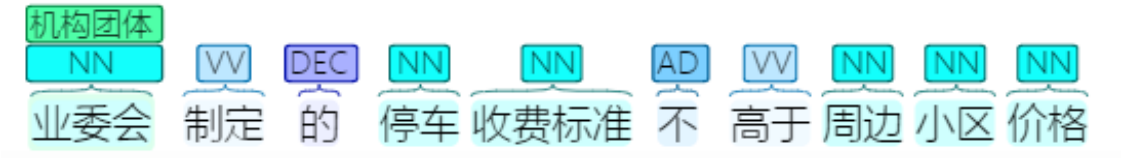

图 13 词法分析

句法分析结果：

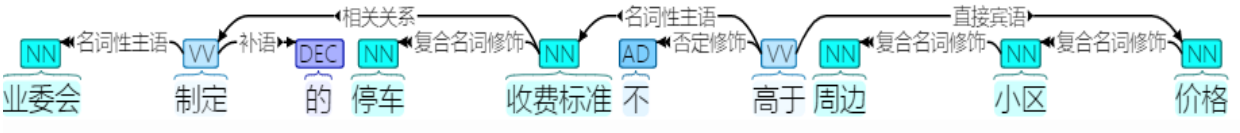

图 14 句法分析

通过句法与词法的分析，可以看出政务问题的答复是有一定的规范的，这样更提现出专业性与规范性。

4.答复意见评分细则及评分结果说明

我们针对答复意见评价规则进行了评分细则的设定，且通过评分细则的设定，计算出每个答复留言的评分，之后根据评分的分布进行回复质量的分类。我们设定了五个指标，其中包含答复意见的字数、回复时间与提问时间的间隔时长、回复文本和留言文本的相似度、答复意见是否符合语法规范以及答复过程中是否使用敬称。其中最后的指标是判断政务人员回复是否持有真诚态度，一个礼貌的称呼，不仅会让留言者感受到被尊敬，同时能让留言者明白，他所提供的意见以及相关问题不会被忽视。

表 2 评分要点

评分要点	分值分配
回复的字数	超过均值记 1′；超过均值 2/3 记 0.5′；其余字数记 0′
回复的内容	通过句法和词法分析后，若有理有据记 1′，否则记 0′
回复内容是否与留言相关	文本相似度高于 0.5 记 1′，否则记 0′
回复时间与提问时间间隔	时间差值小于 5days，记作得 1′，如果差值在一周内，记 0.5′，其余回复时间记 0′
回复中是否有敬称	存在敬称记 1′，否则记 0′

针对所有的回复意见，我们将其进行了打分，并且储存在了 Excel 中，根据所给数据本文计算了所有答复意见的分值，其分值占比如下饼图所示：

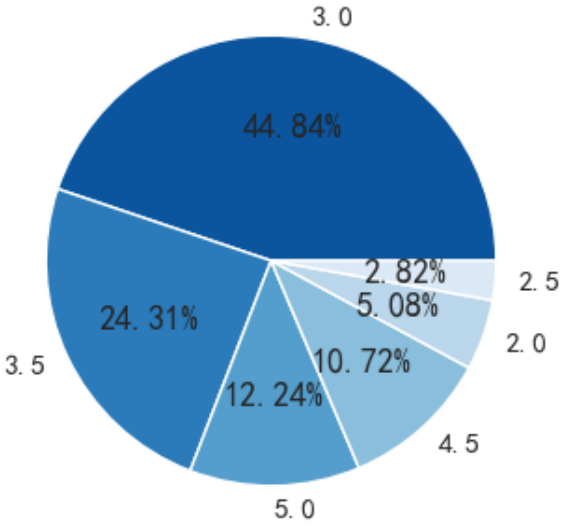

图 15 答复意见占比

下图为我们进行答复意见的两个实例，针对留言我们基本能做到回复的及时以及准确：

留言编号	68907
留言用户	UU0081190
留言主题	咨询生态安葬政策
留言时间	2016/3/16 16:54:35
留言详情	\n\t\t\t\t\t\n\t\t\t\t\t 前两天听说E市出政策说奖励生态安葬，是有...
答复意见	国家鼓励实行生态安葬，目前我市关于“生态安葬”奖励政策暂未出台，如有政策出台，会在政府门...
答复时间	2017/4/28 16:38:59
format_time_2	2017-04-28 16:38:59
format_time_1	2016-03-16 16:54:35
time_gap	407
score	2
length	61

Name: 1684, dtype: object

图 16 答复意见实例一

在第一回复实例中，我们根据留言中针对“生态安葬”的问题进行反馈，从恢复效果上可以看出，答复意见的语句上还是十分符合官方的发言，并且能有针对的提取到留言者的问题所在，但是在时间和回复能否解决的后期效果上还有所欠缺。

留言编号	17461
留言用户	UU0082272
留言主题	咨询绕城高速绿化带A7县段建设的农田租赁问题
留言时间	2017/5/5 13:26:02
留言详情	\n\t\t\t\t\t\n\t\t\t\t\t 曾书记，您好 关于A市绕城高速绿带A7...
答复意见	网友“UU0082272” 您好！来信收悉。现回复如下： 根据《中共A市委办公厅A市人民...
答复时间	2017/5/8 10:02:18
format_time_2	2017-05-08 10:02:18
format_time_1	2017-05-05 13:26:02
time_gap	2
score	5
length	363

Name: 695, dtype: object

图 17 答复意见实例二

在第二个回复实例中，我们能够做到礼貌用语，并且能够根据相关政策进行回复，字数较长，并且回复及时，所以能够获得 5 分的满分。

留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	文本相似度	相似度分值	及时度	答复评分
2549	A00045581	x区景蓉华苑物业管理有问	2019/4/25 9:32:09	业公司却以交20万保证金，对收取停车管理费，在业主大会结束后业委会		2019/5/10 14:56:53	0.052432	1.0	1.0	3.0
2554	A00023583	浦楚南路详湖段怎么还没修	2019/4/24 16:03:40	面的生意带来很大影响，里	需整体换填，且换填后还有三路雨污水管	2019/5/9 9:49:10	0.024242	0.5	1.0	2.5
2555	A00031618	关提高A市民营幼儿园老师的	2019/4/24 15:40:04	同时更是加大了教师的工作	协幼儿园聘任教职工要依法签订劳动合同，	2019/5/9 9:49:14	0.013412	0.5	0.5	3.0
2557	A00011073	公寓能享受人才新政购房补	2019/4/24 15:07:30	落户A市，想买套公寓，请问	年龄35周岁以下（含），首次购房后，可分	2019/5/9 9:49:42	0.052352	1.0	0.5	4.5
2574	A0009233	A市公交站点名称变更的建	2019/4/23 17:03:19	“马坡岭小学”，原“马坡峰	留“马坡岭”的问题。公交站点的设置需	2019/5/9 9:51:30	0.032523	0.5	0.0	3.0
2759	A00077538	A3区含浦镇马路卫生很差	2019/4/8 8:37	将把泥巴冲到右边，越是上下	您问题中没有说明卫生较差的具体路段，也	2019/5/9 10:02:08	0.034252	0.5	1.0	4.0
2849	A00010080	教师村小区盼望早日安装	2019/3/29 11:53:23	为老社区惠民装电梯的规范	A市A3区人民政府办公室下发了《关于A市A	2019/5/9 10:18:58	0.004331	0.0	1.0	3.0
3681	UU00812	区东澜湾社区居民的集体民	2018/12/31 22:21:59	好远，天寒地冻的跑好远，	修前期准备及设施设备采购等工作。下一步	2019/1/29 10:53:00	0.000322	0.0	0.5	2.0
3683	UU008792	麓阳光住宅楼无故停工以及	2018/12/31 9:55:00	也没得到相关准确开工信息。	单位落实分户检查后，西地省楚江新区建设	2019/1/16 15:29:43	0.024865	0.5	1.0	3.0
3684	UU008687	和顺路洋湖壹号小区路段公	2018/12/31 9:45:59	交桥等地方做立体绿化，取	部分也按规划要求完成了建设，其中西边绿	2019/1/16 15:31:05	0.002597	0.0	1.0	2.5
3685	UU0082204	2区大托街道大托新村违建	2018/12/30 22:30:30	规划局审批通过《温室养殖	司支付一笔耕地征收补偿款给原大托村，但	2019/3/11 16:06:33	0.004252	0.0	0.5	2.5
3692	UU008829	邵阳村D区安置房人防工程的	2018/12/29 23:27:51	区安置房地下室近两万平方	来，按长人防发[2014]7号文件要求，鄱阳村	2019/1/29 10:52:01	0.029704	1.0	1.0	4.5
3700	UU00877	区段请求修建一座人行天桥	2018/12/29 11:55:34	峰，大量从小区开车出去的	业分局配合进行具体选址，招标（邀标）进	2019/1/14 14:34:58	0.002467	0.0	0.5	3.5
3704	UU0081480	报A市芒果金融平台涉嫌诈	2018/12/28 17:18:45	复省相关政府部门的大力支	映的相关警情，已由银盆岭派出所立刑事案	2019/1/3 14:03:07	0.057976	1.0	1.0	4.0
3713	UU0081227	建议增开A市261路公交车	2018/12/28 7:53:25	时以上！天寒地冻，其他公	正常。由于驾驶员工作时间长，劳动强度大	2019/1/14 14:33:17	0.065757	1.0	1.0	4.0
3720	UU008444	路与披塘路交叉路口通行乱	2018/12/27 15:18:07	址：https://baidu.com/。出	的“披塘路路口两端各拆除20米中间花坛	2019/3/6 10:26:14	0.063463	1.0	1.0	3.5
3727	UU0081194	区桐梓坡路益丰大药房以北	2018/12/27 1:55:21	便以各种理由拒绝退货，并	根据您提供的信息进行投诉信息的登记分送	2019/1/3 14:02:47	0.098985	1.0	1.0	3.0
3733	UU008706	区在A市梅溪湖开办一个图书	2018/12/26 16:51:40	称。建议在艺术中心先期借	营业。梅溪湖二期金菊路与雪松路东南角规	2019/1/14 14:32:40	0.034325	0.5	0.5	3.0
3747	UU008201	A3区中海国际社区一期旁	2018/12/25 19:35:12	上很早就施工，严重影响居	查，施工单位由于需要夜间连续作业，已办	2019/1/8 16:19:16	0.000536	0.0	0.5	2.0

图 18 总体结果展示

5.评价指标说明

在本次答复意见分值的展示中，我们的评价指标从字数、内容及语法、回复时间及时性、文中是否包含敬称五个方面着手，建立了评价指标体系。在满分为 5 分的情况下，可以看到超过半数的指标在 3 分及以上。首先，都做到了回复的及时性，并且大部分能够针对问题给出针对性的满意回答。但少许的回复还存在文本的语法连贯、以及回复的内容相关性方面的缺陷。今后还需要在语法结构上进行改进，并且能够做到精准的回答。

总　结

本次泰迪杯竞赛，我们选择“智慧政务”中的文本挖掘应用为题，针对近年来“智慧政务”的兴起中伴随的关键性问题做出了针对性的谈论研究。在文本处理方面，自然语言处理技术首当其冲。首先是群众留言的分类，运用高效精准的分类模型进行对于群众留言进行分类，将不同类型的留言根据一级标题进行准确的分类，并且运用文本向量化的方法，可以较为高效的精准化留言评价，将大量复杂的评论转化为高频词语，从而能够方便对应的部门或政府及时的了解到所存在的问题，并进行针对性的解决。训练合适的分类模型从而有助于处理大数据量的留言。其次针对评论时间、评论问题描述进行深度研究，发掘实时热点问题，有助于及时的解决问题。针对性的热度评价指标的建立，通过对留言按照命名实体识别，建立热门评价指标体系，从而可以构建模型最终得到热门的问题表。能够及时地汇总整理出热门问题对于有关部门是至关重要的，有助于更加精准的解决存在的重大问题或者民众最为关注的焦点问题，能够很好做到为人民服务的宗旨。针对答复意见的评价，首先我们需要牢牢把握住及时性、相关性、完整性三原则，建立时间差值的阈值以及相似度的判断，从而通过句法分析技术进行有针对行的解决，最终通过建立的模型实现准确高效的回复，之后建立评分指标，对于回复的好坏进行评级，从而有利于进一步的完善。

本次泰迪杯，我们组针对所选的赛题，研究讨论了自然语言处理技术，在其中不断学习进步，最终运用相关方法解决了各个问题，可谓收获满满。也十分感谢指导老师的辛勤付出，对于我们存在问题的指点和帮助，让我们圆满的完成了本次论文。

参考文献：

[1]Nita Patil,Ajay Patil,B.V. Pawar. Named Entity Recognition using Conditional Random Fields[J]. Elsevier B.V.,2020,167.

[2]Jiuniu Wang,Wenjia Xu,Xingyu Fu,Guangluan Xu,Yirong Wu. ASTRAL: Adversarial Trained LSTM-CNN for Named Entity Recognition[J]. Elsevier B.V.,2020.

[3]杨慧,谭海波,马彦涛.“互联网+政务服务”绩效评估主体的协调机制研究——基于利益相关者的视角[J/OL].中共天津市委党校学报:1-8[2020-05-07].http://kns.cnki.net/kcms/detail/12.1285.D.20200414.1013.002.html.

[4]陈曙东,欧阳小叶.命名实体识别技术综述[J/OL].无线电通信技术:1-11[2020-05-07].http://kns.cnki.net/kcms/detail/13.1099.TN.20200414.1436.002.html.

[5]Ji Bin,Li Shasha,Yu Jie,Ma Jun,Tang Jintao,Wu Qingbo,Tan Yusong,Liu Huijun,Ji Yun. Research on Chinese medical named entity recognition based on collaborative cooperation of multiple neural network models.[J]. Pubmed,2020,104.

[6]姜涛,陆阳,张洁,洪建.无监督分词算法在新词识别中的应用[J].小型微型计算机系统,2020,41(04):888-892.

[7]李玥.机器学习的分类、聚类研究[J].电脑知识与技术,2020,16(04):161-162.

[8]韦灵,黎伟强.基于机器学习的中文文本自动分类的实践研究[J].智库时代,2019(46):265-266.

[9]韦灵,黎伟强.基于机器学习的中文文本自动分类的实践研究[J].智库时代,2019(45):233-234.

[10]刘树栋,张可.类别不均衡学习中的抽样策略研究[J].计算机工程与应用,2019,55(21):1-17.

[11]胡冠军. 中国省会城市政府门户网站政务公开评估指标体系研究[D].长春工业大学,2019.