

# “智慧政务”中的文本挖掘应用

## 摘 要

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。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的推动作用。附件给出了收集自互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见。

对于问题 1 群众留言分类，通过 jieba 对预处理之后的留言详情进行分词。利用 TF-IDF 算法将文本转化为权值向量。然后把向量，和类别导入贝叶斯多项式分类器进行建模，再讲需要预测的数据导入模型进行预测，并评价预测结果。

对于问题 2 热点问题挖掘，仍然是使用 jieba 进行数据预处理。问题热度的参数分为问题出现的次数和点赞数，我们通过命名实体识别提取出各个留言主题的实体，来统计同一问题出现的次数，再统计点赞总数。将二者以不同的权重构建评分模型，得出结果并排序。

对于问题 3 的答复评价指标，我们采用 TF-IDF 算法查看留言关键词在答复中的重要程度以评价相关性，通过对文章结构的分析评价其完整性和可解释性，采用 ahp 层次分析法

**关键词：**sklearn, 命名实体识别, 多项式朴素贝叶斯, TF-IDF 算法

## Abstract

In recent years, with the online inquiry platforms such as WeChat, Weibo, Mayor's Mailbox, and Sunshine Hotline, the government has gradually become aware of public opinion. It is an important channel for gathering people's wisdom and condensing people's popularity. The amount of text data related to various social conditions and public opinions has continued to rise, giving the past major work of related departments that rely on manual to divide messages and organize hot spots has brought great challenges. At the same time, with big data, the development of cloud computing, artificial intelligence and other technologies, and the establishment of smart government affairs systems based on natural language processing technology are already social governance. The new trend of the development of scientific innovation has greatly promoted the improvement of the government's management level and governance efficiency. Attached is a record of the messages of the masses collected from open sources on the Internet, and the relevant departments' comments on some of the masses. Reply to comments..

For question 1, the message classification of the masses, through jieba, the high-frequency word statistics on the subject and details of the message, get the key information of the high-frequency words to classify. Use the jieba Chinese word segmentation tool to segment the problem description information, and use the jieba.analyse.extract\_tags method to extract the first 5 keywords of each problem description. Then use the word2vec module to process the corpus into a vector (Skip-Gram algorithm), and then import the vector into the svm classifier for classification. The TF-IDF algorithm obtains the TF-IDF weight vector for each job description. K-means is used to cluster the TF-IDF weight vector to obtain 7 centroids. The work of related departments that rely on manual to divide messages and organize hot spots has brought great challenges. At the same time, with big data, the development of cloud computing, artificial intelligence and other technologies, and the establishment of smart government affairs systems based on natural language processing technology are already social governance. The new trend of the development of scientific innovation has greatly promoted the improvement of the government's management level and governance efficiency. Attached is a record of the messages of the masses collected from open sources on the Internet, and the relevant departments' comments on some of the masses. Reply to comments.

For problem 2 hotspot problem mining, it is actually similar to problem 1. You need to classify and filter the problem. We first read the excel data content through pandas, and then introduced the jieba method for Chinese segmentation. The word2vec module processes the corpus into a vector (Skip-Gram algorithm), and then imports the vector into the svm classifier for classification. The TF-IDF algorithm obtains the TF-IDF weight vector for each job description. K-means is used to cluster the TF-IDF weight vector to obtain 7 centroids. The work of related departments that rely on manual to divide messages and organize hot spots has brought great challenges. At the same time, with big data, The development of cloud computing, artificial intelligence and other technologies, and the establishment of smart government affairs systems based on natural language processing technology are already social governance. The new trend of the development of scientific innovation has greatly promoted the improvement of the government's management level and governance efficiency. Attached is a record of the messages of the masses collected from open sources on the Internet, and the relevant departments' comments on some of the masses. Reply to comments.

**Key words:** sklearn, named entity recognition, polynomial naive bayes, TF-IDF algorithm

目 录

第一章 问题概述 ..... 5

1.1 本文结构和流程图..... 7

1.2 引言..... 7

第二章数据探索 .....8

2.1 数据预处理 .....8

2.1.1 挖掘目标.....8

2.1.2 数据去重，分词..... 10

2.1.3 数据清洗.....10

2.2 sklearn 算法 .....12

2.3 TF-IDF 算法 .....13

第三章 群众留言分类..... 13

3.1 多项式朴素贝叶斯..... 13

3.1.1 基本原理..... 14

3.1.2 训练阶段..... 14

第四章 构建评价模型..... 16

4.1 精准率..... 14

4.2 召回率..... 14

4.3 命名体识别..... 15

第五章模型评价.....19

# 第一章 问题概述

## 1.1 本文结构和总体流程图

本文共分为五章，各章内容安排如下：

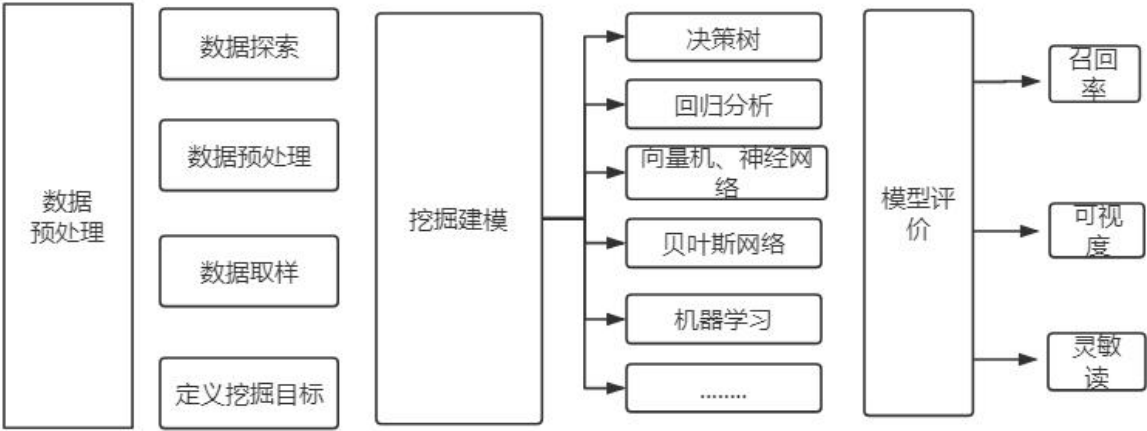
第一章，对论文需要解决的问题进行描述，并简单介绍整篇论文的结构安排。

第二章，对数据进行分析，对数据进行预处理，并从处理后的数据中提取出有用的信息，包括关键词、高频事件。

第三章，进行分类，热度评价，构建判断模型，并通过结果分析验证判断方法的有效性。

第四章，运用 F-score 评价第一问的分类模型，搭建结构输出问题 2 要求的表格，最后使用决策树算法构建安全性评判模型的评价准则。

第五章，综合考虑、效率、高频的因素，通过网络层次分析法建立建立行车安全的综合评价指标体系与综合评价模型。



流程图3

## 1.2 引言

### 1.2.1 问题描述

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。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的推动作用。收集自互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见。

在处理网络问政平台的群众留言时，工作人员首先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，以便后续将群众留言分派至相应的职能部门处理。目前，大部分电子政务系统还是依靠人工根据经验处理，存在工作量大、效率低，且差错率高等问题。

请根据给出的数据，建立关于留言内容的一级标签分类模型。某一时段内群众集中反映的某一问题可称为热点问题，如“XXX 小区多位业主多次反映入夏以来小区楼下烧烤店深夜经营导致噪音和油烟扰民”。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。

## 第二章 数据探索

在分析留言详情时，一些信息的提取和分割是必要的，比如留言的主题和详情，留言的类别，留言的时间和所提及的地点等。所有其中有包括等级的划分，时间段的划分、地区的划分、不同类划分。本章将主要就上述几个方面的问题，对数据进行初步的探索分析。

### 2.1 数据预处理

#### 2.1.1 挖掘目标

在数据挖掘过程中，数据预处理是第一步，同时也是很重要的一步，数据预处理的好坏直接决定着之后特征提取、分类预测等步骤能否顺利进行。在对数据进行分析之前和分析的过程中，我们逐渐对题目所给的数的认识逐步加深，并最终得出了一套较为完整的数据预处理流程。

本次建模目标是利用网络信息平台系统发布的信息数据，利用 jieba 中文分词工具对留言描述进行分词，并使用 TF-IDF 算法、贝叶斯多项式算法等分析方法，达到以下目标：

问题 1：在网上爬取语料训练词向量，用逻辑回归分类器对群众留言进行分类，并用 F-score 对分类方法进行评价。

问题 2：利用文本分词和命名实体识别方法对非结构化的数据进行文本挖掘，通过提取留言的实体（地点），结合问题出现频率，问题点赞数分析其热度并排序。

问题 3：利用 TF-IDF 算法分析留言关键词在答复中的重要程度，以及其他参数评价答复的完整性，相关性，可解释性。

留言编号	留言用户	留言主题	留言时间	留言详情	一级分类
24	A00074011	A 市西湖建筑集团占道施工有安全隐患	2020/1/6 12:09:38	A3 区大道西行便道.....	城乡建设



181619	A00019042	I3 县南洲镇 鑫顺广场 A 栋楼下开烧 烤摊,业主该 如何维权	2018/6/6 19:51:27	I3 县南洲镇鑫 顺广场 A 栋...	环境保护
185537	A00035431	运输车辆重 复办证的行 为要制止	2014/5/29 15:40:21	尊敬的厅长: 你好,对予运 输车辆,名...	交通运输
1957	A00012120	A3 区含浦镇 玉江村高考 美术培训机 构是否合 法?	2019/6/20 10:37:18	A3 区含 浦镇玉江村 欧公组有一 个叫同行添 艺 .....	教育文体

部分数据表 1

一级分类	二级分类	三级分类
城乡建设	安全生产	事故处理
城乡建设	安全生产	安全生产管理
城乡建设	安全生产	安全隐患

部分数据表 2

热度	问题 ID	热度指数	时间范围	地点/人群	问题
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

部分数据表 3

### 2.1.2 数据去重，分词

首先我们用 pandas 库自带的 drop—duplicates 方法对所有数据进行去重，再调用 jieba 类库进行分词。

**jieba 系统简介：**

“结巴”中文分词：Python 中文分词组件。

**特点：**

1. 支持三种分词模式
2. 支持繁体分词
3. 支持自定义词典
4. MIT 授权协议

**涉及算法：**

1. 基于前缀词典实现词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图（DAG），采用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；
2. 对于未登录词，采用了基于汉字成词能力的 HMM 模型，采用 Viterbi 算法进行计算；
3. 基于 Viterbi 算法的词性标注；
4. 分别基于 tfidf 和 textrank 模型抽取关键词；

### 2.1.3 数据清洗

文本中存在很多的无意义的字符，停用词以及乱码，这些都会对分类模型的性能有一定影响，在这里我们用 split（）指令将无法识别的字符串转换为可识别的空格字符，并导入停用词表，通过循环语句删除文本中的停用词和字符，得到一个由中文词汇构成的列表。

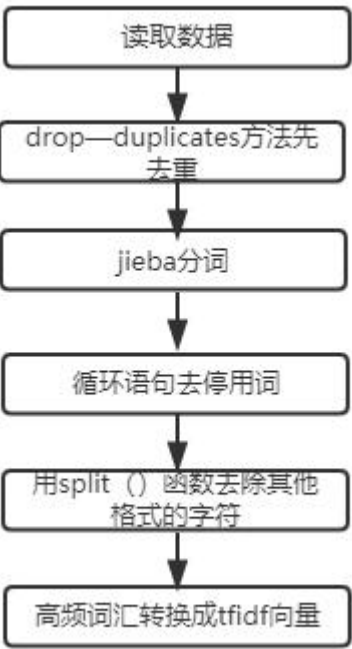
由数据图 3 可知，留言详情的每个句子都被分为词或字，概率越高，出现的频率越高，表示该词语与句子的相关性较大，可以视为高频词汇。

```
0.34681364566105266
0.31485698129192985
0.277341247837193
0.6596071597535544
0.4783893441800948
0.19763996661611374
0.1932339919772512
0.16997299767156399
2453183e-04 1.3894194e-03 1.7576506e-04 2.4997428e-04
772339e-04 -1.0981711e-03 -9.6409692e-04 -1.4905281e-03
ODO Terminal Python Console
2833:23 CRLF UTF-8 4 spaces P
```

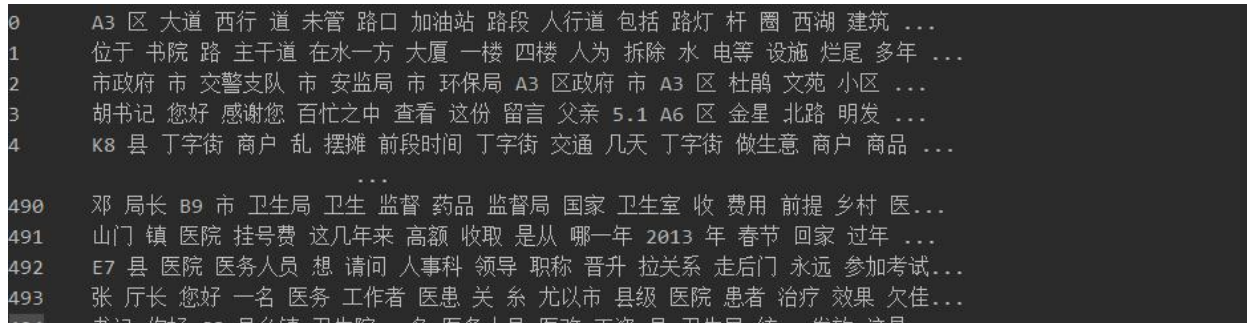
预处理数据图 3

我们根据给定的数据可以看出，会出现很多冗余且不需要的数据，每个留言都会出现一大串问候语，以及没有意义的词，比如“的”、“你”、“吧”等等。而在其后才是重要的想要反应的事件，这给我们在计算其准确率上有着很大的干扰，那么我们该如何解决此类问题呢。

首先我们解决此类问题的思路是这样：首先我们用 drop—duplicates 方法先去重，再 jieba 分词，然后用循环语句去停用词，将列表转换为字符串之后用 split 函数去除其他格式的字符，之后将每个句子的高频词汇转换成 tfidf 向量



思路图 4



分词结果图 6

## 2.2 sklearn

sklearn 库提供很多可以进行数据预处理，数学建模的函数：

`train_test_split` 用于划分训练集和测试集

`Labelencoder` 用于将不同的文本类别分配不同的标签

```
from sklearn.feature_ext\fraction.text import CountVectorizer
```

```
vec = CountVectorizer() #实例化向量化对象
```

```
X_train_count_vectorizer = vec.fit_transform(X_train) #将训练集中的新闻向量化
```

```
X_test_count_vectorizer = vec.transform(X_test) #将测试集中的新闻向量化
```

## 2.3 TF-IDF 算法

在对职位描述信息分词后，需要把这些词语转换为向量，以供挖掘分析使用。这里采用 TF-IDF 算法，把职位描述信息转换为权重向量。TF-IDF 算法的具体原理如下：

第一步，计算词频，即 TF 权重（Term Frequency）。

$$\text{词频}(TF) = \text{某个词在文本中出现的次数} \quad (1)$$

考虑到文章有长短之分，为了便于不同文章的比较，进行“词频”标准化，除以文本的总词数或者除以该文本中出现次数最多的词的出现次数即：

$$\text{词频}(TF) = \frac{\text{某个词在文本中的出现次数}}{\text{文本的总词数}} \quad (2)$$

或

$$\text{词频}(TF) = \frac{\text{某个词个词在本中出现次数}}{\text{该文本出现次数最多的词的出现次数}} \quad (3)$$

第二步，计算 IDF 权重，即逆文档频率（Inverse Document Frequency），需要建立一个语料库（corpus），用来模拟语言的使用环境。IDF 越大，此特征性在文本中的分布越集中，说明该分词在区分该文本内容属性能力越强。

$$\text{逆文档频率}(IDF) = \log\left(\frac{\text{语料库的文本总数}}{\text{包含该词文本数} + 1}\right) \quad (4)$$

第三步，计算 TF-IDF 值（Term Frequency Document Frequency）。

$$TF - IDF = \text{词频}(TF) \times \text{逆文档频率}(IDF) \quad (5)$$

实际分析得出 TF-IDF 值与一个词在职位描述表中文本出现的次数成正比，某个词文本的重要性越高，TF-IDF 值越大。计算文本中每个词的 TF-IDF 值，进行排序，次数最多的即为要提取的职位描述表中文本的关键词。

## 第三章 群众留言分类

### 3.1 多项式朴素贝叶斯

当特征是离散的时候，使用多项式模型。多项式模型在计算先验概率  $P(y_k)$  和条件概率  $P(x_i | y_k)$  时，会做一些平滑处理，具体公式为：

$$P(y_k) = \frac{N_{y_k} + \alpha}{N + K\alpha}$$

N 是总的样本个数，k 是总的类别个数，是类别为  $y_k$  的样本个数， $\alpha$  是平滑值

$$P(x_i | y_k) = \frac{N_{y_k, x_i} + \alpha}{N_{y_k} + n\alpha}$$

$N_{yk}$  是类别为  $y_k$  的样本个数,  $n$  是特征的维数,  $N_{yk, x_i}$  是类别为  $y_k$  的样本中, 第  $i$  维特征的值是  $x_i$  的样本个数,  $\alpha$  是平滑值

### 3.1.1 基本原理

朴素贝叶斯法利用贝叶斯定理首先求出联合概率分布, 再求出条件概率分布。这里的朴素是指在计算似然估计时假定了条件独立。基本原理可以用下面的公式给出:

$$P(Y|X) = \frac{P(Y)P(X|Y)}{P(X)}$$

其中,  $P(X|Y) = P(X_1, X_2, \dots, X_n|Y) = P(X_1|Y)P(X_2|Y)\dots P(X_n|Y)$ ,  $P(Y|X)$  叫做后验概率,  $P(Y)$  叫做先验概率,  $P(X|Y)$  叫做似然概率,  $P(X)$  叫做证据。

### 3.1.2 训练阶段

先验概率

$$P(C=c) = \frac{\text{属于类 } c \text{ 的文档数}}{\text{训练集文档总数}}$$

条件概率

$$P(w_i|c) = \frac{\text{词 } w_i \text{ 在属于类 } c \text{ 的所有文档中出现次数}}{\text{属于类 } c \text{ 的所有文档中的词语总数}}$$

注:

- (1) 条件概率  $P(w_i|c)$  表示的是词  $w_i$  在类别  $c$  中的权重

(2) 条件概率独立性假设，丢失了词语的位置信息，在文本表示上来说，就是它失去了语义信息。当然可以通过 ngram 的特征来减少损失，但是也不能有效解决语义上的损失。

(3) 先验概率和条件概率的计算都利用了最大似然估计。它们实际算出的是相对频率值，这些值能使训练数据的出现概率最大。

拉普拉斯平滑（加 1 平滑）

$$P(w_i|c) = \frac{\text{词 } w_i \text{ 在属于类 } c \text{ 的所有文档中出现次数} + 1}{\text{属于类 } c \text{ 的所有文档中的词语总数}}$$

加 1 平滑可以认为是采用均匀分布作为先验分布，即每个词项在每个类别中出现一次，然后根据训练数据对得到的结果进行更新。也就是说未登录词的估计值为 1/词汇表长度

## 第四章构建评价模型

### 4.1 精准率

精准率 (Precision) 指的是模型预测当前类别时的准确度，其计算公式如下：

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

假如癌症检测系统的混淆矩阵如下：

真实\预测	0	1
0	9978	12
1	2	8

则对类 1， 精准率=8/(8+12)=0.4。

0.4 表示如果有 100 个人被预测患有癌症，其中有 40 人是真的患有癌症。也就是说，精准率越高。

### 4.2 召回率

召回率 (Recall) 指的是当前类别中被正确预测的比率，其计算公式如下：

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

例如对上面的混淆矩阵，对类 1，召回率 =8/(8+2)=0.8 。

0.8 表示如果有 100 个患有癌症的病人使用这个系统进行检测，系统能够正确检测出其中的 80 人。也就是说，召回率越高，漏掉的可能性越低。

精准率与召回率之间的关系

假设有这么一组数据，菱形代表 Positive，圆形代表 Negative 。

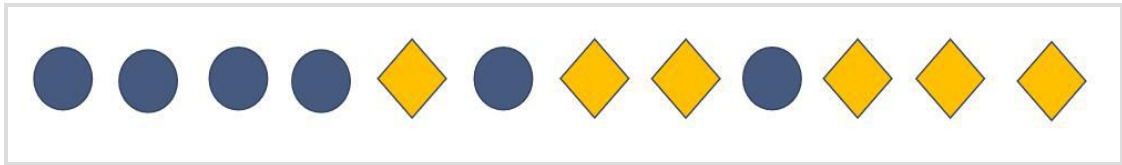


图 7

现在需要训练一个模型对数据进行分类，假如该模型非常简单，就是在数据上画一条线作为分类边界。模型认为边界的左边是 Negative，右边是 Positive。如果



该模型的分类边界向左或者向右移动的话，模型所对应的精准率和召回率如下图所示：



图 8

从上图可知，模型的精准率变高，召回率会变低，精准率变低，召回率会变高。

应该选精准率还是召回率作为性能指标？

到底应该使用精准率还是召回率作为性能指标，其实是根据具体业务来决定的。

比如，要训练一个模型来预测股票涨 ( Positive ) 还是跌 ( Negtive )，那么我们应该主要使用精准率作为性能指标。因为精准率高，模型预测该股票要涨的可信度就高。

如果要训练一个模型来预测人是 ( Positive ) 否 ( Negtive ) 患有艾滋病，那么我们应该主要使用召回率作为性能指标。因为召回率太低，很可能导致漏检，这样可能导致病人错过最佳的治疗时间。

### 4.3 命名实体识别 (NER)

是在自然语言处理中的一个经典问题，其应用也极为广泛。比如从一句话中识别出人名、地名，从电商的搜索中识别出产品的名字，识别药物名称等等。传统的公认比较好的处理算法是条件随机场 (CRF)，它是一种判别式概率模型，是随机场的一种，常用于标注或分析序列资料，如自然语言文字或是生物序列。简单是说在 NER 中应用是，给定一系列的特征去预测每个词的标签。

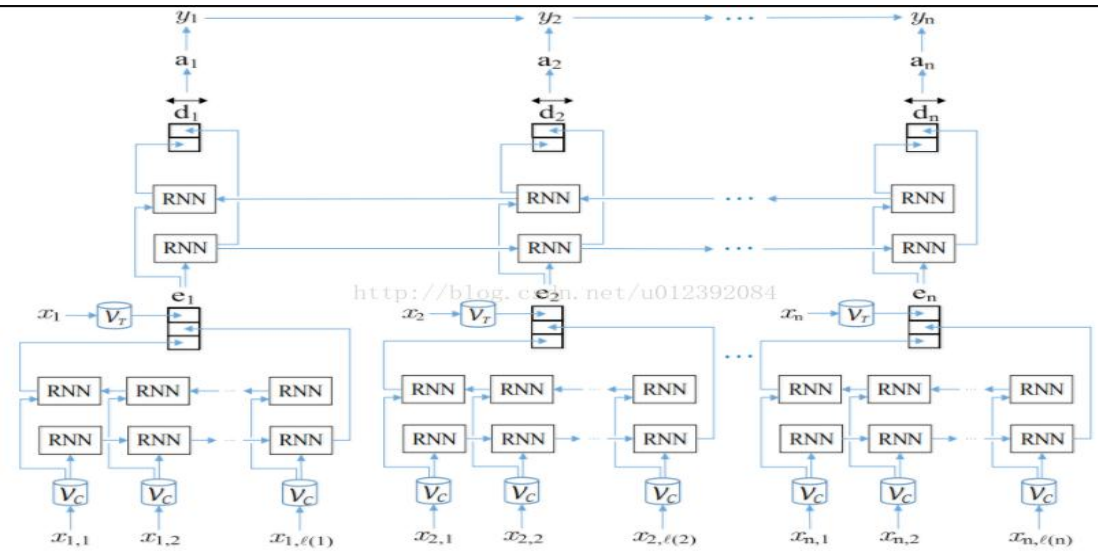
如何把词转换成神经网络能接受的数据？

神经网络只接受数字，不接受字符串，所以我们需要用工具把词转换成为词 向量，这个工具可以是 gensim word2vec、glove 等等。训练的数据最好是要 有

个庞大的数据集，比如从网上爬取下来的新闻然后用这些数据来训练词向量。

如何处理训练数据中没有见过的词？

我们之所以需要用庞大的新闻数据来预训练词向量的原因就是为了解决训练数据量小的问题。因为人力有限，我们不可能有很大的标记过的数据，如果在测试样例中出现了我们没有标记过的词，那么显然会影响 NER 结果。然后 word2vec 的用处就是将相似的词的“距离”拉的很近，这样可以一定程度上减少未出现词的影响。



流程图 9

## 第五章模型评价

### 1. 预测结果

由图 10 知，随着生活水平的提高，在交通运输、卫生计生、环境保护、商贸旅游这块出现的问题相对较少，说明城乡建设、教育问题、劳动和社会保障方面反应的问题次数较多。

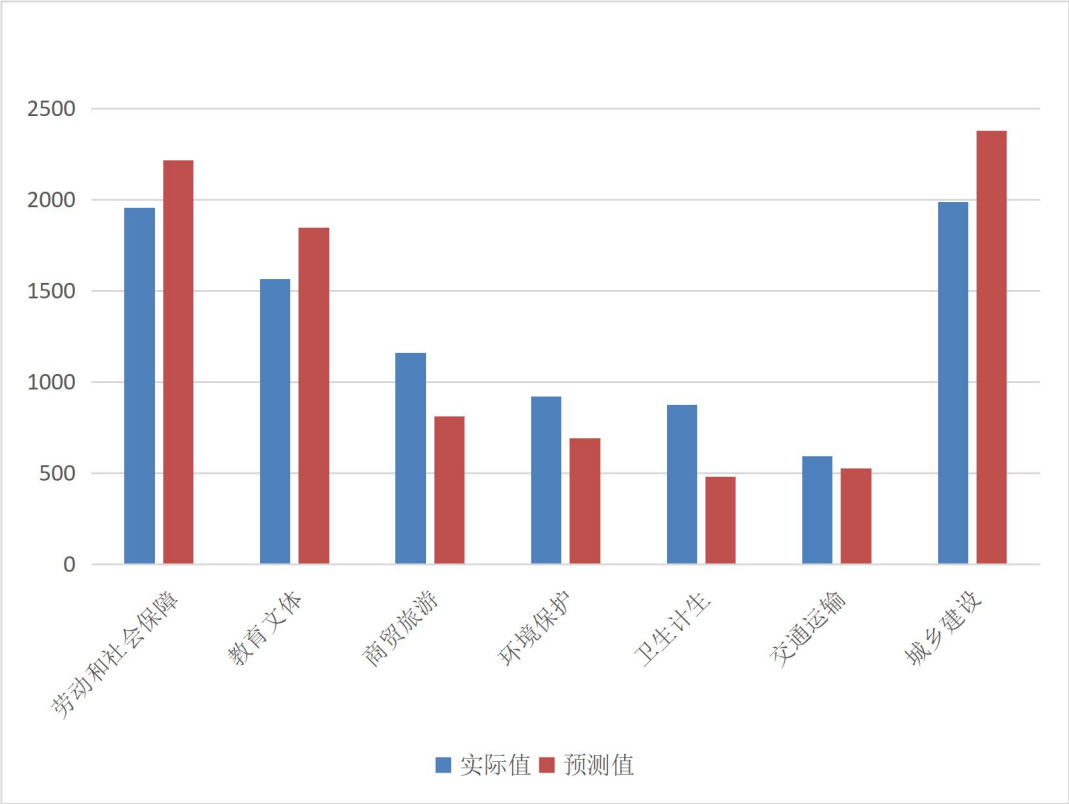


图 10

## 总结

在这次泰迪杯比赛里，我们学到了不少算法，也学会了运用。非常复杂，需要很多细节，和编程功底，对算法的要求也是相对比较高。通过看 c 题讲解，我们的思路是首先定义挖掘目标，再进行数据探索。

首先识别相似留言，再问题归类，相似的归为同一类，然后进行热度评价。我们在处理问题期间也遇到很多问题，比如地点，人群的识别，表达有多样化。相似的计算复杂度等等，通过网上搜索文献，以及视频观看和老师指导，我们逐渐解决了一些问题。

## 参考文献

【1】<https://blog.csdn.net/appleyuchi/article/details/79057307> 随机森林针对中文文本分类

【2】<https://blog.csdn.net/xiexf189/article/details/79092629> 用 Python 进行简单的文本相似度分析

【3】何小东,刘卫国.数据挖掘中关联规则挖掘算法比较研究[J].计算机工程与设计,2005,26(5):1265-1268.

【4】Kaur P,Attwal KS.Data Mining:Review[J].International Journal of Computer Science & Information Technolo,2014.

【5】Python 数据分析与应用

【6】《Python 编程基础》——<https://edu.tipdm.org/course/96/task/1836/show>

【7】《Python 与数据挖掘》——<https://edu.tipdm.org/course/96/task/1863/show>

【8】《Python 数据分析与挖掘实战》——<https://edu.tipdm.org/course/96/task/1866/show>

【9】《地球科学大数据挖掘与机器学习》——<https://edu.tipdm.org/course/96/task/2143/show>

【10】《Python3 智能数据分析快速入门》——<https://edu.tipdm.org/course/96/task/8250/show>

【11】《数据挖掘：实用案例分析》——<https://edu.tipdm.org/course/96/task/10792/show>