摘要：在如今的大数据时代，信息越发的透明，人们所接触的信息量也越来越大，政府等相关部门也采取了网上留言等方式来获取居民的日常生活信息，以及城市中所存在的问题，而大量重复的消息往往影响了处理的速度，收到的问题，简单的问题和复杂的问题掺杂在一起，导致容易错过比较重要的问题。而如今，随着深度学习在字然语音处理这一方面的发展，智能的政务处理系统已经逐渐崭露头角。

针对赛题要求，介于自然语言的复杂，语义表达多面性等多方面的因素，我们选择用 jieba 等库对自然语言进行分词处理，之后使用 TF-IDF 算法及循环神经网络对数据进行进一步的处理和学习，以达到预期效果。

关键字：深度学习，jieba，TF-IDF，循环神经网络

Abstract: in today's big data era, information is more and more transparent, and people are exposed to more and more information. The government and other relevant departments also take the way of online message to obtain the daily life information of residents, as well as the problems existing in the city, and a large number of repeated messages often affect the processing speed, the received problems, simple problems and complex problems Mixed together, it is easy to miss more important issues. Nowadays, with the development of in-depth learning in word and natural speech processing, intelligent government processing system has gradually emerged.

According to the requirements of the competition, due to the complexity of natural language and the multi-faceted semantic expression, we choose to use the database of Jieba to segment the natural language TF-IDF algorithm and cyclic neural network are used to further process and learn the data to achieve the desired results.

Keywords: deep learning, Jieba, TF IDF, recurrent neural network

一、 引言

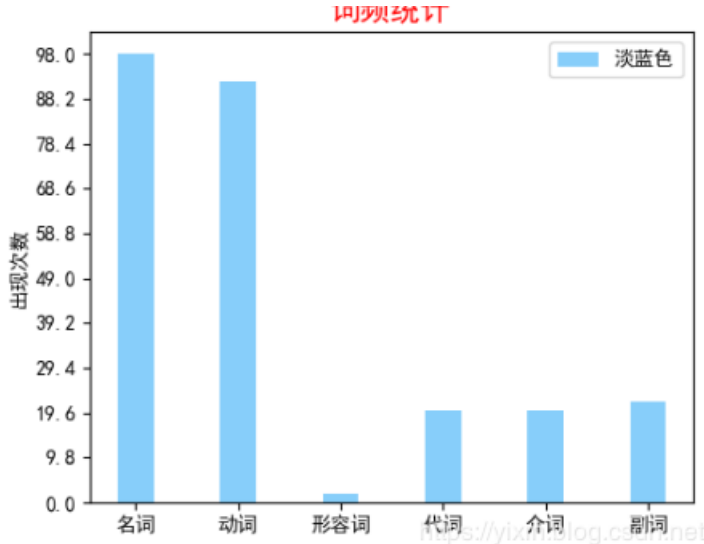
随着信息技术、信息产业的飞速发展，信息传播和更新的速度日新月异，这也使得信息的增长呈现井喷式发展，增长速度异常迅猛。如何利用好这些信息，逐渐成为企业、政府关注的焦点。传统波特价值链理论将信息作为增值过程中的支持要素，而非增值源泉，但是随着企业信息化进程的深入发展，信息有可能变成有用的资源。深层挖掘信息价值，使其与信息技术结合、与企业流程再造以及生产技术融合，就能从根本上提高企业的竞争力，实现价值增值。在政府方面，大量信息的搜集、整理，以价值为导向的数据挖掘，能为政策的制定、方案的出台提供很好的数据与理论支撑。那么，面对语义复杂，数量巨大的自然语言，如何去提取重要的信息，然后指定有效的解决档案呢？

自然语言处理的四个维度分别是：语义（Semantic）、句子结构（Syntax）、单词（Morphology）以及声音（Phonetic）。声音的处理过于复杂，而且不同地方存在方言，会大大影响辨识度，所以，主要入手点为前三个方向。获取到的数据可以通过分词技术，以及提取重

要信息，归纳相同信息，来对一大段的自然语言进行处理。通过构建语料库，来辅助解决一些简单的问题。

二、 实验方案

在对信息进行挖掘分析之前，先要把非结构化的文本信息转换为计算机能够识别的结构化信息。赛题提供的数据，以中文文本的方式给出了数据。为了便于转换，先要对这些信息进行中文分词。这里采用 python 的中文分词包 jieba 进行分词。jieba 采用了基于前缀词典实现的高效词图扫描，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)，同时采用了动态规划 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，使得能更好的实现中文分词效果。



在分词的同时，采用了 TF-IDF 算法，抽取每个问题中的前 5 个关键词，这里采用 jieba 自带的语义库。

TF-IDF 算法

在对问题描述信息分词后，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，把问题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（TermFrequency)。

词频（TF）=某个词在文本中出现的次数

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:

$$词\ 频（TF）=\frac{某个词在文本中出现的次数}{文本总词语数}$$

第二步，计算 IDF 权重，即逆文档频率（InverseDocument 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。

$$逆文档频率（IDF）=\log（\frac{语料库的文本总数}{包含该词的文本数+1}）$$

第三步，计算 TF-IDF 值（TermFrequencyDocumentFrequency)。

TF-IDF=逆文档频率（IDF）×词 频（TF）

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问题描述表中文本的关键词。

生成 TF-IDF 向量

生成 TF-IDF 向量的具体步骤如下：

（1）使用 TF-IDF 算法，找出每个问题描述的前 5 个关键词；

（2）对每个问题描述提取的 5 个关键词，合并成一个集合，计算每个问题描述对于这个集合中词的词频，如果没有则记为 0；

（3）生成各个问题描述的 TF-IDF 权重向量，计算公式如下：

TF-IDF=词频（TF）×逆文档频率（IDF）

- **词语级别 TF-IDF：**矩阵代表了每个词语在不同文档中的 TF-IDF 分数。
- **N-gram 级别 TF-IDF:** N-grams 是多个词语在一起的组合，这个矩阵代表了 N-grams 的 TF-IDF 分数。
- **词性级别 TF-IDF：**矩阵代表了语料中多个词性的 TF-IDF 分数

创建基于文本/NLP 的特征

创建许多额外基于文本的特征有时可以提升模型效果。

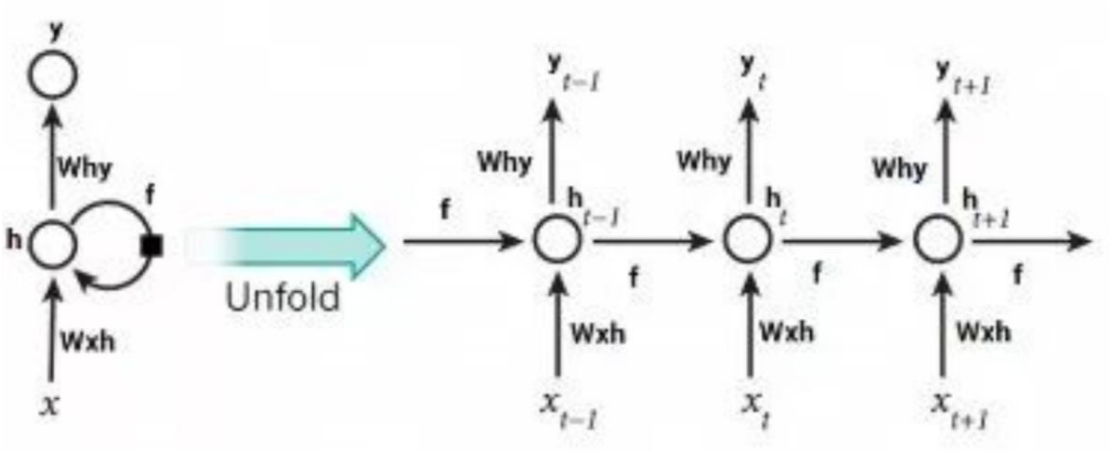
- 文档的词语计数—文档中词语的总数量
- 文档的词性计数—文档中词性的总数量
- 文档的平均字密度--文件中使用的单词的平均长度
- 完整文章中的标点符号出现次数--文档中标点符号的总数量
- 整篇文章中的大写次数—文档中大写单词的数量
- 完整文章中标题出现的次数—文档中适当的主题（标题）的总数量

- 词性标注的频率分布
  - 名词数量
  - 动词数量
  - 形容词数量
  - 副词数量
  - 代词数量

建模

- **循环神经网络-LSTM**

  与前馈神经网络不同，前馈神经网络的激活输出仅在一个方向上传播，而循环神经网络的激活输出在两个方向传播（从输入到输出，从输出到输入）。因此在神经网络架构中产生循环，充当神经元的“记忆状态”，这种状态使神经元能够记住迄今为止学到的东西。RNN 中的记忆状态优于传统的神经网络，但是被称为梯度弥散的问题也因这种架构而产生。这个问题导致当网络有很多层的时候，很难学习和调整前面网络层的参数。为了解决这个问题，开发了称为 LSTM（Long Short Term Memory）模型的新型 RNN：



三、 结果分析

在利用 sklearn 框架，在特征为多个词语的情况下，使用 TF-IDF 向量的朴素贝叶斯，结果为 0.58 左右，进行分类时，在多样本的情况下，准确率很低。

在使用**循环神经网络-LSTM 下，**学习大量样本最后所得的结果为 0.7 左右，在多样本情况下，可以保证大部分的准确率。

四、 改进

在学习过程中，并没有过的的考虑语义的问题，语义的问题会导致很多分类导致误分的情况，如果在之后分词之后，再加入情感分析，之后再进行分类，可以进一步的提升准确度。

五、 参考文献

[1] https://github.com/fxsjy/jieba

[2] https://github.com/LambdaWx/CNN_sentence_tensorflow

[3] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient Estimation of Word Representations in Vector Space[J]. arXiv preprint arXiv, 2013.

[4] 杨虎.面向海量短文文本去重技术的研究与实现.国防科学技术大学.2007

[5] MikolovT, Karafiat M,Burget L,et al.Recurrent neural network based language model[C].Interspeech.2010,2:3.

[6] Hochreiter S,Schmidhuber J.Longshort-term memory[J].Neural computation,1997,9(8):1735-1780.

[7] Yiming Cui, Zhipeng Chen, Attention-over-Attention Neural Networks for Reading Comprehension, 2016.