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。

在本次数据挖掘过程中，我们首先对获取得到的评论数据利用 python 进行数据预处理、分词以及停用词过滤操作，实现了对评论数据的优化，并提升了其可建模度。

接着，采用多种方法来进行数据挖掘模型的构建，为后面的评论分析构建分析的基础。为此我们先利用深度学习的方法，通过多种工具构建栈式自编码神经网络:其次，我们运用武汉大学的 ROSTCM6 为所搜取来的各类社情民意相关的文本数据文本构建语义网络:再有，利用 LDA 主题模型的思想，结合统计学的角度实现评论主题模型的构建。

最后，运用构造出来的多种数据挖掘模型的结果，对这些评论数据进行多方面多角度的评论文本分析，以提取评论中隐藏的信息。栈式自编码神经网络被用以进行情感倾向性分析:语义网络重建了有价值高频词之间的关系，在共词矩阵以及评论定向筛选回查的帮助下，一定程度上得到了各类社情民意相关的文本数据包括特有建议信息:LDA 主题模型则滤取出了从统计学角度上得出人们觉得政府应该作出调整的关注点，方便政府对此作出对策。

目录

1.挖掘目标

本次建模针对京东电商平台海尔、美的、万和三种品牌型号的热水器的消费者的文本评论数据，在对文本进行基本的机器预处理、中文分词、停用词过滤后，通过建立包括栈式自编码深度学习、语义网络与 LDA 主题模型等多种数据挖掘模型，实现对文本评论数据的倾向性判断以及所隐藏的信息的挖掘并分析，以期望得到有价值的内在内容。

2.分析方法与过程

2.1.总体流程

本论文的分析流程可大致分为以下四步:

第一步:获取分析用的原始数据(文本评论语料),部分数据自行爬取:

第二步:对获取的数据进行基本的处理操作，包括数据预处理、中文分词、停用词过滤等操作:

第三步:文本评论数据经过处理后，运用多种手段对评论数据进行多方面的分析:

第四步:从对应结果的分析中获取文本评论数据中有价值的内容。

2.2.具体步骤

2.2.1 数据介绍

2.2.2 文本评论预处理

取到文本后，我们首先要进行文本评论数据的预处理。文本评论数据里面存在大量价值含量很低甚至没有价值含量的条目，如果将这些评论数据也引入进行分词、词频统计乃至情感分析等，则必然会对分析造成很大的影响，得到的结果的质量也必然是存在问题的。那么在利用到这些文本评论数据之前就必须要先进行文本预处理，把大量的这些无价值含量的评论去除。论数据之前就必须要先进行文本预处理，把大量的这些无价值含量的评论去除。

我们运用 Python2.7 对这些文本评论数据的预处理主要由二个部分组成:文本去重、以短句删除。按照各自处理的特性，我们依照这个顺序进行文本评论数据的预处理。

2.2.2.1 文本去重

(1)文本去重的基本解释

文本去重，顾名思义，就是去除文本评论数据中重复的部分。无论获取到什么样的文本评论数据，首先要进行的预处理应当都是文本去重。

2.2.2.2 短句删除

(1)短句删除的原因及思想

我们进行最后的预处理步骤:短句删除。我们知道，虽然精简的辞藻在很多时候是一种比较良好的习惯，但是由语言的特点我们知道，从根本上说，字数越少所能够表达的意思是越少的，要想表达一些相关的意思就- -定 要有相应量的字数，过少的字数的评论必然是没有任何意义的评论，比如三个字，就只能表达诸如“很不错”、“质量差”等等。为此，我们就要删除掉过短的评论文本数据，以去除掉没有意义的评论，包括: ①原本就过短的评论文本，如“很不错”。②过短的评论文本,即原本为存在连续重复的且无意义的长文本，如“好好好好好好好好好好好好好好好好好好”。

(2) 保留的评论的字数下限的确定

显然，短句删除最重要的环节就是保留的评论的字数下限的确定，这个没有精确的标准，可以结合特定语料来确定，- -般 6 到 10 个国际字符都是较为合理的下限，在此处我们设定下限为 7 个国际字符，即经过前两步预处理后得到的语料若小于等于 6 个国际字符，则将该语料删去。

2.2.3 情感倾向性分析

为了得到人们的总体情感倾向，我们可以对该人们的评论集做情感倾向分析，以得到对某政策的总体印象。传统的情感分析是基于情感词典的方法，对每条评论中的情感词做加权，来得到每条评论的情感倾向值，进而获得整个评论集的情感倾向。这种方法直观，容易理解:此外，运用机器学习的方法进行情感二极分类也是当前的流行方法:对每条评论抽取特征(tf, t-idfF 值等)构成特征向量，然后采用朴素贝叶斯法或者 SVM 进行分类。如今，这种方法的运用也是比较成熟。

本文抛弃这些传统的方法，大胆尝试采用新的方法:基于词向量和深度学习方法对评论集做情感倾向性分析。

2.2.3.1 训练生成词向量

我们首先训练以得到词向量，为了将文本情感分析(情感分类)转化为机器学习问题，首先就是需要将符号数学化。在 NLP 中，最常见的词表示方法就是 Onc-hot Representatin:将--一个词映射成一个很长的单位向量，向量的长度就是词表的大小，如:“学习”表示成[000 10000 00000000...“复习”表示成[0000 000010000000...这样就完成了词语的数学化表示。

我们在这里使用最后一种词向量模型。

word2vec 采用神经网络语言模型 NNLM 和 N-gram 语言模型，每个词都可以表示成-

一个实数向量。

2.2.3.2 评论集子集的人工标注与映射

利用词向量构建的结果,我们进行评论集子集的人工标注，正面评论标为(,负面评论标记为 2(或者采用 python 的 NLP 包 snownlp 的 sentiment 功能做简单的机器标注，减少人为工作量)，然后我们将每条评论映射为一个向量，将分词后评论中的所有词语对应的词向量相加做平均，使得一条评论对应一个向量。

2.2.3.3 训练栈式自编码网络

自编码网络是由原始的 BP 神经网络演化而来。在原始的 BP 神经网络中我们从特征空间输入到神经网络中，并用类别标签与输出空间来衡量误差，用最优化理论不断求得极小值，从而得到一个与类别标签相近的输出。但是在编码网络并不是如此，我们并不用类别标签来衡量与输出空间的误差，而是用从特征空间的输入来衡量与输出空间的误差。

完成评论映射后，我们将标注的评论划分为训练集和测试集，在 MATLAB 下，利用标注好的训练集(标注值和向量)训练栈式自编码网络(SAE),对原始向量做深度学习提取特征，并后接 Softmax 分类器做分类，并用测试集测试训练好的模型的正确率。

2.2.3.4 情感分析

当 SAE 模型训练好后，我们便可以对整个评论集进行情感倾向分析。

2.2.4 基于 LDA 模型的主题分析

基于语义网络的评论分析进行初步数据感知后，我们从统计学习的角度，对主题的特征词出现频率进行量化表示。本文运用 LDA 主题模型，用以挖掘人们评论中更多的信息。

主题模型在机器学习和自然语言处理等领域是用来在一- 系列文档中发现抽象主题的一种统计模型。直观上来说，传统判断两个文档相似性的方法是通过查看两个文档共同出现的单词的多少，如 TF. TF-IDF 等，这种方法没有考虑到文字背后的语义关联，可能在两个文档共同出现的单词很少甚至没有，但两个文档是相似的，因此在判断文档相似性时，应进行语义挖掘，而语义挖掘的有效工具即为主题模型。

2.2.4.1 LDA 主题模型介绍

潜在狄利克雷分配(Latent Dirichlet Allocation, LDA) 是由 Blei 等人在 2003 年提出的生成式主题模型。生成模型，即认为每一篇文档的每一个词都是通过“一定的概率选择了某个主题，并从这个主题中以一-定的概率选择了某个词语”。LDA 模型也被称为三层贝叶斯概率模型，包含文档(d)、主题(z)、词(w)三层结构，能够有效对文本进行建模，和传统的空间向量模型(VSM) 相比，增加了概率的信息。通过 LDA 主题模型，能够挖掘数据集中的潜在主题，

进而分析数据集的集中关注点及其相关特征词。

2.2.4.2 运用 LDA 模型进行主题分析的实现过程

在本文人们评论关注点的研究中，即对评论中的潜在主题进行挖掘，评论中的特征词是模型中的可观测变量。- -般来说，每则评论中都存在一个中心思想，即主题。如果某个潜在主题同时是多则评论中的主题，则这一潜在主题很可能是整个评论语料集的热门关注点。在这个潜在主题上越高频的特征词将越可能成为热门关注点中的评论词。

2.3.1 情感倾向性分析结果

2.3.2 语义网络的结果与分析

2.3.3 LDA 模型构造结果与分析

3.结论

但是从我们的分析结果当中也可以看出总体来讲效果还不是特别的好，比如我们的情感倾向性分析结果就会与真实结果有一-定程度上的出入，这里面既涉及到中文语言结构所必然导致的文本评论分析的缺陷的问题，也涉及到当今中文文本挖掘模型的不足以及评论数据本身所具有的问题，这也是我们在后期进-一步的对中文文本数据的研究过程中可以继续深入探讨的地方。