“智慧政务”中的文本挖掘应用

摘要

本文对“智慧政务”中的文本挖掘应用进行了研究，分别运用朴素贝叶斯算法建立模型，引进*Tempotron*算法，采用*LIF*模型，Perplexity，分别根据给出数据，对留言内容的一级标签进行分类；定义合理的热度评价指标，给出排名前五的热点问题以及相应热点问题对应的留言信息；给出评价方案。

针对问题一：通过去掉某些文字，稀有词处理，同类词合并对文字进行预处理以后，运用贝叶斯分类算法，选取最大值所属分类对群众留言进行分类。

针对问题二：运用*Tempotron*算法，采用*LIF*模型，对脉冲序列信息分为正负两类，根据输入的脉冲序列和对应的标签来匹配进行权值调整，重复计算，直到两个模式分类正确，得出排名前五的热点问题。

针对问题三：我们通过文本预测来还原问题文本，运用Perplexity，通过词法分析，解决自动分词面临的问题，用自动分词技术方法，词性标注，词法分析评价指标来预测下一个词并给出评价方案。

最后，本文对模型的优缺点进行了评价以及将模型进行了推广。同时，本文还对模型的结果进行了误差分析，并运用拉普拉斯平滑估计和最短路径法（统计粗分模型分别进行了对建立的模型进行了改进

关键词：朴素贝叶斯算法　　*Tempotron*算法　　*LIF*模型　　Perplexity

§1 问题的重述

一、引言

1、问题背景

近年来，随着网络技术的快速发展，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。人们通过在指定网站上留言来向政府反映生活中的问题，使得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，应用现代信息技术，整合信息服务资源，通过应用各种平台，提高政府服务和管理的质量，让政府更智能，更便捷，成为“智慧政府”。

智慧政务将融合考虑建筑、环境、交通、人群、政府、公众、企业等的以管理和服务为导向，融入政务服务、智能建筑、社区服务、文体教育等多种要素属性，以“物联化、互联化、智能化”的整体智慧的体系进行设计建设，实现政务服务事项清单标准化、办事指南标准化，打造安全可靠、高效便捷、健康舒适的政务环境。

2、问题产生

网络的普及使得人民群众可以更方便的“说”出自己生活中存在的问题，各类社情民意相关的文本数据量也随之不断攀升。在处理网络问政平台的群众留言时，需要工作人员按照内容分类三级标签划分体系，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。但是目前大部分电子政务系统还是依靠人工根据经验处理，工作量大、效率低，且差错率高等问题普遍存在。

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。它可以更好地将群众问题集中反映呈现出来，帮助政府快速解决人民生活中的问题，优化服务改革措施落地，为社会公众提供便捷、优质的服务，提高人民生活幸福值；同时释放人工解放生产力，有助于社会主义和谐社会的建设。群众提出问题和政府解决问题的有机结合、线上服务和线下办事的紧密融合，对提升政府的管理水平和施政效率具有极大的推动作用。

如何为公众提供便捷、安全、高效率的政务服务，是提升政府的管理水平和施政效率的重要课题。

二、要解决的问题

1、问题一：群众留言分类

在处理网络问政平台的群众留言时，工作人员首先按照内容分类三级标签体系划分，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。建立适当的分类模型，根据题目给出的数据，对关于留言内容的一级标签进行分类。并对模型用 F-Score 方法进行评价。

2、问题二：热点问题挖掘

在某一时段内群众集中反映的某一问题可称为热点问题，如在一段时间内多名群众投诉 “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”，则此事件可称为热点问题。及时发现热点问题，有助于相

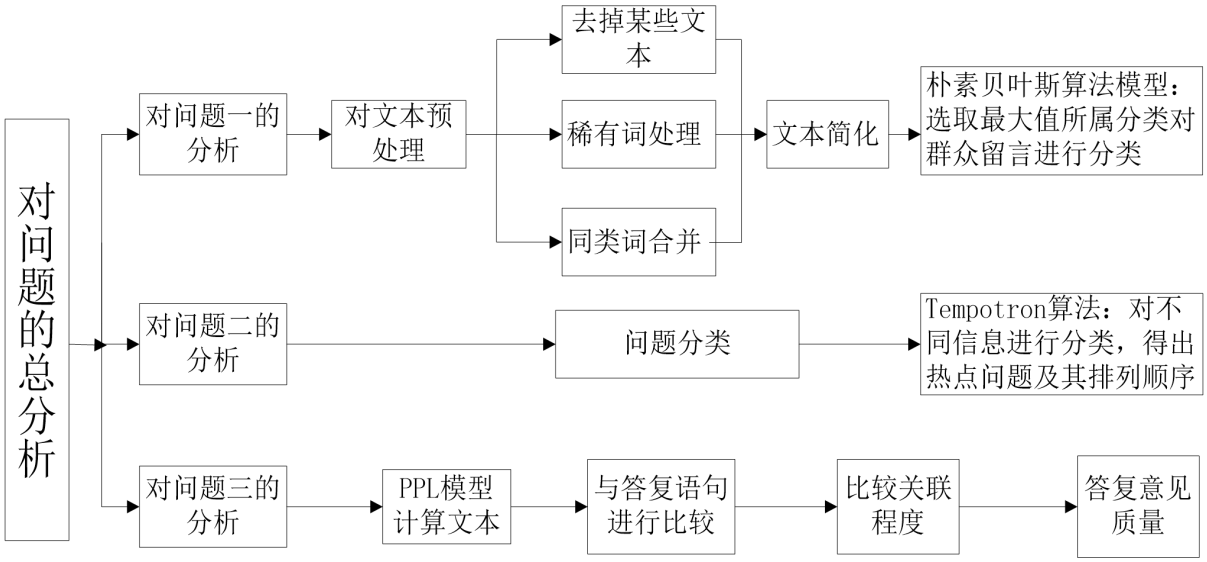
关部门进行有针对性地处理，提升服务效率。根据群众反映的文本内容将某一时段内反映特定地点或特定人群问题的留言进行留言主题、时间、详情划分归类，以及统计赞同或者有相似困扰的群众人数。建立适当的模型定义合理的热度评价指标，得出评价结果，并按给定的格式给出排名前 5 的热点问题，以“热点问题表.xls”命名文件。按题中相应的格式给出相应热点问题对应的留言信息，以“热点问题留言明细表.xls”命名文件。

3、问题三：答复意见的评价

针对相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等多个方面对答复意见的质量建立一套评价方案，并实现。

§2 问题的分析

一、问题的总分析



二、对具体问题的分析

1、对问题一的分析

在对群众的留言进行分类时，工作人员首先对留言进行三级分类，但是工作量大效率低且容易出错。首先对留言文本进行处理，简化文本，引进贝叶斯分类算法通过计算联合概率，利用参数估计原理的分类方法，从中选取最大值作为其所属于的类，即对留言进行分类。

2、对问题二的分析

在某一时段内群众集中反映的某一问题可称为热点问题。及时发现热点问题，并将问题进行主题、时间、详情划分归类，以及统计赞同或者有相似困扰群

众人数，有助于相关部门进行有针对性地处理，提升服务效率。引进*Tempotron*算法，对不同的脉冲序列信息进行分类排序，可以得到热点问题的排名及类别。

3、对问题三的分析

根据相关部门对留言的答复意见从多个方面对答复意见的质量进行评价。运用*PPL*模型进行计算，把答复的语句与提出问题的语句进行比较，看关联程度，关联程度高，则答复意见的质量较高，如果关联程度不高，则答复意见的质量较低。

§3 模型的假设

1、假设一个词语只有一个意思，如一个苹果和一个苹果手机只有一个意思。
2、假设不含情感词语，即说反话的情况。

§4 符号说明

序号	符号	符号说明
1	C	训练集中所有类标记的集合
2	a_i	测试实例 X 的第 i 个属性的属性值
3	a_{ij}	第 i 个训练实例的第 j 个属性值
4	$V(t)$	表示在时刻 t 神经元的膜电压
5	ω_i	表示各个前突触的影响权重
6	t_i	表示第 i 个输入神经元发射脉冲的时刻
7	$K(t-t_i)$	表示第 i 个神经元在 t_i 时刻对突触后膜电位的影响
8	V_{rest}	表示重置膜电压
9	$V(t_{\max}^{+})$	在 L^{+} 模式下最大膜电压
10	dt_i	施加延迟
11	$V(t_{\max}^{-})$	L^{-} 模式下最大膜电压

序号	符号	符号说明
12	S	代表句子
13	N	句子长度
14	$p(w_i)$	是第i个词的概率
15	w_0	表示句子的起始

§5 模型的建立与求解

一、对问题一的求解

问题一：群众留言分类

模型准备：

文本分类技术是通过计算机来模拟人们的分类能力，是目前数据分析、数据挖掘、模式识别、机器学习等领域研究的核心内容。面对网络上的海量信息资源，分类技术将大量的文本划分类别，使得不同类别的文本代表不同的文本将大量的文本划分类别，使得不同类别的文本代表不同的文本主题，方便政府全面地、准备地、快速地找到所需要的信息，解决信息问题。

问题一我们引进贝叶斯分类算法，贝叶斯分类是以贝叶斯统计学理论和贝叶斯网络基础的，通过计算联合概率，利用参数估计原理的分类方法。其主要方法是通过已知的先验概率，利用贝叶斯公式计算其后验概率，算其后验概率，即得到了属于所有类的概率，从中选取最大值作为其所属于的类。贝叶斯分类器由结构和选取的参数构成，决定了其分类的表现形式和参数布局，从而得到了不同的联合概率分解，即不同概率的分类器。

模型建立：

1. 进行文字预处理

（1）去掉某些文本

将一些在文本中出现频次较高但含义虚范的字词去掉，例如文本中的“的、地、得、但是、因为”等，确保去掉的字词不能作为文本特征。

（2）稀有词处理

有些词在文本中出现的频次很低，不适合作为文本特征项，通过对文本集进行词条频率的统计，设定一个词频阀值，词条频率低于该词频阀值时，可删除该词条，即该词不作为文本特征。

（3）同类词合并

为了提高文本词条的分类效果，采取单词合并和同义词合并的方式，把表达形式不同但含义相同或者含义相似的词条合并，视为同一个词条来处理。

2.运用朴素贝叶斯算法建立模型

在贝叶斯定理中，我们设X,Y是一组随机变量，组成的联合概率分布

$P(X=x,Y=y)$是指随机变量X取值x且随机变量Y取值y的概率， X和Y的联

合概率和条件概率满足：

$$P(X,Y)=P(Y|X)P(X)=P(X|Y)P(Y) \tag{1.1}$$

进而得到贝叶斯定理: $P(Y|X)=\dfrac{P(X|Y)P(Y)}{P(X)}$ （1.2）

其中，$P(Y)$是先验分布，$P(X|Y)$是似然函数，$P(X)$称为证据，$P(Y|X)$称为后验分布。因此，贝叶斯定理是一种把“先验知识”和“从数据获取的证据”相结合的统计原理。

朴素贝叶斯算法应用了贝叶斯定理，是贝叶斯定理的一个简单应用，尤其是在文本文类应用中已经成为一种准则，算法以贝叶斯定理为基础并衍生，是基于贝叶斯理论的有监督的学习方法，它有一个假设条件，即一个实例在给定类标记的条件下，它的所有属性值相互独立。

因此我们可以得到朴素贝叶斯的算法结构图如下：

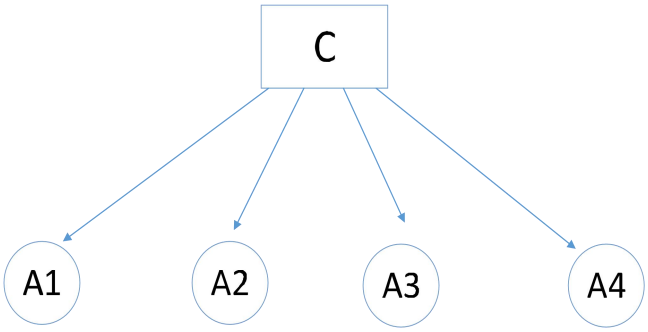


图 1　朴素贝叶斯算法结构图

朴素贝叶斯算法处理分类问题时，给定一个训练集 $D=\{X_1,X_2,...,X_m\}$，其中包含 m 个训练实例，每个训练实例 $X_j=(a_{1j},a_{2j},...,a_{nj},c_j)\in D,(j=1,2,...,m)$ 是一个 $n+1$ 维的向量，$a_{ij},(i=1,2,...,n;j=1,2,...,m)$ 是它的属性 $A_i(i=1,2,...,n)$ 的取值，$y_j\in C$ 为它的类标记，C 为类别，即训练集中所有类标记的集合。

模型求解：

朴素贝叶斯的算法流程如下：

（1）设 $X=\{a_1,a_2,...,a_n\}$ 为一个待分类的未知属性，每个 a 为 X 的一个维度。

（2）有类别集合 $C=\{y_1,y_2,...,y_m\}$。

（3）计算 $P(y_1 \mid x), P(y_2 \mid x), ..., P(y_m \mid x)$

①收集一个标注好的待分类项集合，即训练数据集。
②统计频数并得到在各类别下各维度的条件概率估计。即：

$$P(a_1 \mid y_1), P(a_2 \mid y_1), ..., P(a_n \mid y_1); P(a_1 \mid y_2), P(a_2 \mid y_2), ..., P(a_n \mid y_2); ...;$$

$$P(a_1 \mid y_n), P(a_2 \mid y_n), ..., P(a_n \mid y_m) \tag{1.3}$$

③在给定测试实例 $X = \{a_1, a_2, ..., a_n\}$ 的条件下，计算 $y_k \in C$ 的后验概率，则贝叶斯定理可推导为：

$$P(y_k \mid X) = \frac{P(X \mid y_k)P(y_k)}{P(X)} \tag{1.4}$$

根据算法定理，可将（1.4）改写成：

$$P(y_k \mid X) \propto P(X \mid y_k)P(y_k) \tag{1.5}$$

由于 $P(X)$ 始终为常数，因此只需将分子 $P(X \mid y_k)P(y_k)$ 最大化即可。且朴素贝叶斯的条件假设为各项特征相互独立，因此有：

$$P(X \mid y_k)P(y_k) = P(a_1 \mid y_k)P(a_2 \mid y_k)...P(a_n \mid y_k)P(y_k) = P(y_k)\prod_{i=1}^{n} P(a_i \mid y_k) \tag{1.6}$$

即有：$$P(X \mid y_k) = P(a_1 \mid y_k)P(a_2 \mid y_k)...P(a_n \mid y_k) = \prod_{i=1}^{n} P(a_i \mid y_k) \tag{1.7}$$

（4）根据（1.5）和（1.7）就可得到朴素贝叶斯模型算法针对一个测试实验 x 构建的分类器的公式为：

$$C(X) = \arg\max_{y_k \in C} P(y_k)\prod_{i=1}^{n} P(a_i \mid y_k) \tag{1.8}$$

$C(X)$ 的计算结果表示测试实例 X 最大可能为什么类型的标记。（1.8）式中的 a_i 是测试实例 X 的第 i 个属性的属性值，概率 $P(y_k)$ 和概率 $P(a_i \mid y_k)$ 可以通过计算训练集中不同类标记和属性值组合的出现频率来计算，具体计算方式如下：

$$P(y) = \frac{\sum_{i=1}^{m} \delta(c_i, c)}{m} \tag{1.9}$$

$$P(a_j \mid y)=\frac{\sum_{i=1}^{m}\delta(a_{ij},a_j)\delta(c_i,c)}{\sum_{i=1}^{m}\delta(c_i,c)} \tag{1.10}$$

式子（1.9）、（1.10）中的m表示训练集中训练实例的个数，c_i为第i个训练实例的类标记，a_{ij}为第i个训练实例的第j个属性值，$\delta(c_i,c)$为一个二值函数，当$c_i=c$时为1，否则为0。

附件2中有9211个待分类的数据，其中有11个数据不能准确分类，剩下的9200个数据可以准确分类。在这9200个数据中有9180个数据分类正确；11个不能准确分类的数据中有1个数据被分类。那么：

$$Accuracy=\frac{10+9180}{9211}=99.77\%$$

$$P=\frac{10}{10+11}=47.62\%$$

$$R=\frac{9180}{10000}=91.8\%$$

$$F_1=\frac{2\times(47.62\%\times91.8\%)}{1\times47.62\%+91.8\%}=62.71\%$$

二、对问题二的求解

问题二：热点问题挖掘

模型准备：

该问引进*Tempotron*算法，*Tempotron*是脉冲学习网络的一个经典有监督学习算法，是二分类的学习方法，能对不同的脉冲序列信息进行分类，且需要调整的参数很少。*Tempotron*算法能根据神经元的分类信息，不断调整连接的权值，最终按照神经元是否点火，分类出正确的结果，即神经元发射脉冲点火，则当前模式属于正类；神经元不发射脉冲点火，则当前模式属于负类。

模型建立：

在标准*Tempotron*算法中，神经元模型采用*LIF*模型，在*LIF*模型中，膜电压由传入的脉冲加权和影响，而且当神经元膜电压达到阈值完成点火之后，膜电压会迅速下降到静止电压，膜电压的数学公式是由突触前所有脉冲的加权和组成的，具体公式如下：

$$V(t)=\sum_i \omega_i \sum_{t_i} K(t-t_i)+V_{rest} \tag{2.1}$$

$V(t)$表示在时刻t神经元的膜电压，ω_i表示各个前突触的影响权重，t_i表示第i个输入神经元发射脉冲的时刻，$K(t-t_i)$表示第i个神经元在t_i时刻对突触后膜电位的影响，V_{rest}表示重置膜电压，K函数的数学表达式为：

$$K(t-t_i)=V_0(\exp[-(t-t_i)/\tau]-\exp[-(t-t_i)/\tau_s]) \tag{2.2}$$

τ和τ_s均为时间常数，V_0的计算公式为：

$$V_0=\frac{\tau\tau_s}{\tau-\tau_s}\ln\frac{\tau}{\tau_s} \tag{2.3}$$

K函数随时间推进产生的可模拟神经元电位变化的曲线如图所示：

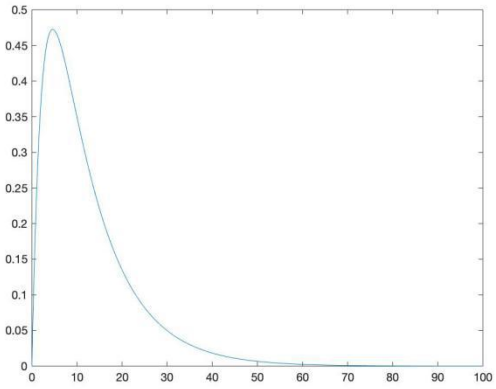


图 1　函数 K 随时间变化图像

在*Tempotron*学习算法中，每个脉冲序列都会被分为正负两类，正类为L^+，负类为L^-，*Tempotron*算法会根据输入的脉冲序列和对应的标签是否匹配进行权值调整，让脉冲序列能在适应的时候触发突触后神经元激活点火。当训练过程中出现了错误点火或错误不点火的情况，*Tempotron*会根据相应的规则调整网络，误差函数为：

$$err=\begin{cases}V_{thr}-V_{t_{\max}} & \text{分类为正类}\\ V_{t_{\max}}-V_{thr} & \text{分类为负类}\end{cases} \tag{2.4}$$

*Tempotron*算法的核心思想是通过不断调整连接的权值ω_i，最终获得期望的脉冲输出。脉冲更新的公式如下：

$$\Delta\omega_i = \lambda \sum_{t_i < t_{\max}} K\left(t_{\max} - t_i\right) \tag{2.5}$$

模型求解：

1.将正负两个模式下的神经元点火时间分别代入*LIF*神经元模型。

2.根据公式$V(t)=\sum_i \omega_i \sum_{t_i<t-dt_i} K\left(t-dt_i-t_i\right)+V_{rest}$（2.6）计算在时间周期[0，$T$]内截至到每个时间步长的膜电压。

3.如果在L^+模式下最大膜电压$V\left(t^+_{\max}\right)$小于阈值，则说明存在L^+类错误，未发放脉冲。

（1）寻找距离$t^+_{\max}$时刻最近的已发放过脉冲的兴奋性突触，同时在当前时刻没有发放脉冲，并且对该突触在之前的训练中没有被延迟过。

（2）对该突触施加延迟dt_i并重新计算膜电压。

（3）若施加延迟dt_i后仍没有脉冲发放，则代入如下公式求权重更新量，增加权重。

$$\Delta\omega_i \begin{cases} -\lambda \sum_{t_i < t_{\max} - dt_i} K\left(t_{\max} - dt_i - t_i\right), \text{若} L^+ \text{类错误} \\ \lambda \sum_{t_i < t_{\max} - dt_i} K\left(t_{\max} - dt_i - t_i\right), \text{若} L^- \text{类错误} \end{cases} \tag{2.7}$$

（4）若脉冲发放，则权重无更新。

4. 如果在L^-模式下最大膜电压$V\left(t^-_{\max}\right)$大于阈值，则说明存在L^-类错误，发放了脉冲。

（1）寻找距离$t^+_{\max}$时刻最近的已发放过脉冲的兴奋性突触，同时在当前时刻没有发放脉冲，并且对该突触在之前的训练中没有被延迟过。

（2）对该突触施加延迟dt_i并重新计算膜电压。

（3）若施加延迟dt_i后仍没有脉冲发放，则代入如下公式求权重更新量，增加权重。

$$\Delta\omega_i \begin{cases} -\lambda \sum_{t_i < t_{\max} - dt_i} K\left(t_{\max} - dt_i - t_i\right), \text{若} L^+ \text{类错误} \\ \lambda \sum_{t_i < t_{\max} - dt_i} K\left(t_{\max} - dt_i - t_i\right), \text{若} L^- \text{类错误} \end{cases} \tag{2.7}$$

（4）若脉冲发放，则权重无更新。

5.重复第二步至第四步，直到两个模式分类正确为止。

三、对问题三的求解

模型准备：

语言模型,给出一句话的前 k 个词，希望它可以预测第 $k+1$ 个词是什么，即给出一个第 $k+1$ 个词可能出现的概率的分布 $p\left(x_{k+1}|x_1,x_2,...x_k\right)$。常用 *PPL* 衡量语言模型收敛情况,下面我们从公式角度来理解一下该指标的意义。

模型建立：

Perplexity 定义：*PPL* 是用在自然语言处理领域 *NLP* 中，衡量语言模型好坏的指标。它主要是根据每个词来估计一句话出现的概率，并用句子长度作 normalize,公式为

$$\begin{aligned} PP(S) &= P\left(w_1 w_2 ... w_N\right)^{-\frac{1}{N}} \\ &= \sqrt[N]{\frac{1}{p\left(w_1 w_2 ... w_N\right)}} \\ &= \sqrt[N]{\prod_{i=1}^{N}\frac{1}{p\left(w_1 w_2 ... w_N\right)}} \end{aligned}$$

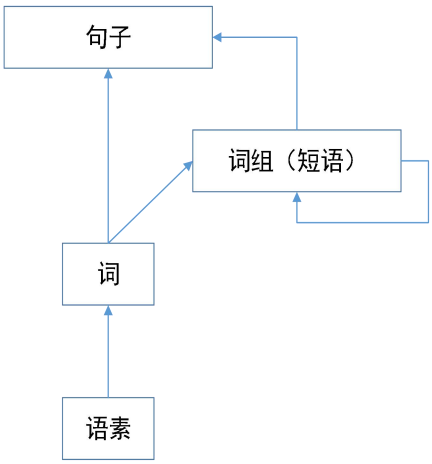
S 代表句子， N 是句子长度， $p\left(w_i\right)$ 是第 i 个词的概率。第一个词就是 $p\left(w_1|w_0\right)$，而 w_0 表示句子的起始，是个占位符。这个式子可以理解为，*PPL* 越小，$p\left(w_i\right)$ 则越大，一句我们期望句子出现的概率就越高。

Perplexity 也可以认为是平均分支系数，即预测下一个词可以有多少种选择。例如，说模型的 *PPL* 下降到 50，可以理解为，在模型生成一句话时下一个词有 50 个合理选择,可选词数越少，我们大致认为模型越准确。这样解释了，为什么 *PPL* 越小，模型越好。

模型求解：

1.词法分析

词是最小的能够独立运用的语言单位，因此，词法分析是其他一切自然语言处理问题的基础，会对后续问题产生深刻的影响。



词法分析的任务是：将输入的句子字串转换成词序列并标记出各词的词性。这里所说的“字”并不仅限于汉字，也可以指标点符号、外文字母、注音符号和阿拉伯数字等任何可能出现在文本中的文字符号，所有这些字符都是构成词的基本单元。从形式上看，词是稳定的字的组合，不同的语言词法分析具体做法是不同的。

中文分词词法分析包括两个主要任务：自动分词：将输入的汉字串切成词串；词性标注：确定每个词的词性并加以标注。

2.自动分词面临的问题

自动分词面临着三个问题：歧义问题、未登录词问题、分词标准问题。

（1）歧义

这里的歧义指的是切分歧义：对同一个待切分字符串存在多个分词结果。分为交集型歧义、组合型歧义和混合歧义。

交集型歧义：字串 abc 既可以切分成 a/bc，也可以切分成 ab/c。其中，a、bc、ab、c 是词。

针对交集型歧义，提出链长这一概念：交集型切分歧义所拥有的交集串的个数称为链长。

混合歧义：以上两种情况通过嵌套、交叉组合等而产生的歧义。

（2）未登录词

词典中没有收录过的人名、地名、机构名、专业术语、译名、新术语等。该问题在文本中的出现频度远远高于歧义问题。未登录词问题是分词错误的主要来源。

（3）分词标准

“缺乏统一的分词规范和标准”这种问题也反映在分词语料库上，不同语料库的数据无法直接拿过来混合训练。

3.自动分词技术方法

（1）最大匹配法

基本思想：先建立一个最长词条字数为 L 的词典，然后按正向（逆向）取句子前 L 个字查词典，如查不到，则去掉最后一个字继续查，一直到找着一个词为止。

（2）最少分词法（最短路径法）

基本思想：假设待切分字串为：$S=c_1c_2...c_n$，其中c_i为单个字，串长为$n(n\geq 1)$。建立一个结点数为$n+1$的切分有向无环图G，若$w=c_ic_{i+1}...c_j(0<i<j\leq n)$是一个词，则在结点$v_{i-1},v_j$之间建立有向边。从产生的所有路径中，选择路径最短的(词数最少的)作为最终分词结果。

（3）基于词的分词方法

基于词的生成模型主要考虑词汇之间以及词汇内部字与字之间的依存关系，大部分基于词的分词方法采用的都是生成式模型。该种分词方法的基本思想很简单：

$$W_{Sep}^{*}=\arg\max W_{sep}\,p\left(W_{Sep}\middle|c_1c_2...c_n\right)$$

即，找到概率最大的切分。

（4）n 元语法模型方法

利用 n 元语法计算整个句子的概率切分。

我们利用二元模型进行计算，即：$p(s)=p(w_1)p(w_2|w_1)p(w_3|w_2)..p(w_n|w_{n-1})$

4. 词性标注

在任何一种语言中，词性兼类问题都普遍存在，汉语中尤为明显。造成词性兼类问题的原因主要有以下几点：

（1）汉语缺乏词形态变化，无法通过词形变化判别词类；

（2）汉语中，常用词兼类现象严重；

（3）没有统一的汉语词类划分标准，有些语料划分很粗糙。

在这样的背景下，词性标注问题往往被转化为序列标注问题来解决。

5. 词法分析评价指标

词法分析中，通常用正确率、召回率、F 值来评价系统的性能。正确率：测试结果中正确结果的个数占系统所有输出结果的比例，即：$P=\frac{n}{N}\times 100\%$

召回率：测试结果中正确结果的个数占标准答案总数的比例，即：$R=\frac{n}{M}\times 100\%$

F 值：正确率和召回率的综合值，即：$F-measure=\frac{(\beta^2+1)\times P\times R}{\beta^2\times P+R}\times 100\%$

通常情况下，取$\beta=1$,称为 F1 值，即：$F1=\frac{2\times P\times R}{\beta^2\times P+R}\times 100\%$

§6 模型的评价与推广

一、对问题一模型的评价

朴素贝叶斯算法依旧依赖于一种常用的错误假设，即随机变量同样重要并且

相互独立；在处理大量数据时，朴素贝叶斯算法能够很好地处理噪声数据和缺失数据，但当数据具有较多数值特征时，数据处理结果并不理想。

朴素贝叶斯算法较为简单、快速、有效，朴素贝叶斯算法有一个极其重要的优点，那就是需要用来训练的例子相对较少，并可以处理好大量的例子。朴素贝叶斯算法关于数据有一对简单的假设，其假设数据集的所有特征都具有相同的重要性和独立性，虽然在实际应用中，这些假设鲜有成立。然而在大多数情况下，当违背这些假设的时候，朴素贝叶斯依然可以很好地应用，甚至在极端事件中，特征之间具有很强的依赖性时，朴素贝叶斯也可以用。

二、对问题二模型的评价

优点：*Tempotron* 算法是二分类的学习算法，能对不同的脉冲序列信息进行分类，*Tempotron* 算法关系简单且可行性很强，运用特别广泛，且算法需要调整的参数很少。

缺点：*Tempotron* 算法对于噪声的处理能力不强，但实际解决的问题存在噪声，需要克服数据中噪声的影响。

三、对问题三模型的评价

最大匹配算法其优点是实现简单，算法运行速度快，缺点是严重依赖词典，无法很好的处理分词歧义和未登录词。

§7 模型的改进

一、对问题一的模型改进

一般情况下，上述的计算方法可以得出一个良好的概率估计值，但是，当属性值的频率无限接近于 0 时，继续使用上述方法对概率的估计是有偏差且过低的，零频率属性值的出现则会直接导致计算出的概率值为 0，级公式（1.8）的计算结果为 0，显然，这样的计算结果是错误的。

下面对公式（1.9）、（1.10）用拉普拉斯平滑估计进行改进，得到改进后的计算公式如下：

$$P(y)=\frac{\sum_{i=1}^{m}\delta(c_i,c)+1}{m+m_c} \tag{1.11}$$

$$P(a_j\mid c)=\frac{\sum_{i=1}^{m}\delta(a_{ij},a_j)\delta(c_i,c)+1}{\sum_{i=1}^{m}\delta(c_i,c)+m_j} \tag{1.12}$$

公式（1.11）、（1.12）中 m_c 为训练集中类标记的个数，m_j 为训练集中第 j 个属性的取值个数。

二、对问题二的模型改进

在有干扰的条件下，*Tempotron* 算法的结果并不理想。改进方式如下：

1.当前模式是正类时，为了降低噪声的干扰，此时可将阈值设定为比标准的阈值更大。如果噪声干扰了最大膜电压，但最大膜电压下降后也并不会降到阈值之下，因此还是会触发点火脉冲。

$$V_{high} = V_{thr} + a \qquad (2.8)$$

V_{thr} 为标准阈值，因为在训练阶段用的是更高的阈值，所以得到的权重会更倾向于将正类的最大膜电压变换到更大，因此即使是在噪声干扰的情况下，正类的模式还是会触发。

2.当前模式是负类时，为了降低噪声的干扰，此时可将阈值设定为比标准的阈值更小。如果噪声干扰了最大膜电压，但最大膜电压上升后也并不会升到阈值之上，因此还是不会触发点火脉冲。

$$V_{high} = V_{thr} - b \qquad (2.9)$$

V_{thr} 为标准阈值，因为在训练阶段用的是更低的阈值，所以得到的权重会更倾向于将负类的最大膜电压变换到更小，因此即使是在噪声干扰的情况下，负类的模式还是不会触发。

三、对问题三的模型改进

改进最短路径法（统计粗分模型），在规则法中，我们介绍了最短路径法，每条边的权重都为 1。改进最短路径法对最短路径法进行了改进，将边的权重改为该词出现频率 $p(w_i)$，那么最终选取切分结果的标准则变为：

$$\max p(W) = \prod_{i=1}^{m} p(w_i) \qquad （1） \quad 其中，\ W = w_1 w_2 ... w_m$$

但是很明显，如果这样的话就没办法用 Dijkstra 最短路径算法了。
为了解决这个问题，将(1)式改写为如下形式：

$$p^*(W) = -\ln[p(W)] = \sum_{i=1}^{m} \{-\ln[p(w_i)]\} \qquad （2）$$

这样，就可以将求 $p(W)$ 的最大值问题转换为求 $p^*(W)$ 的最小值问题，我们所做的，只是将边的权重 $p(w_i)$ 改写为 $-\ln[p(w_i)]$，这样，就可以继续用 Dijkstra 算法进行求解了。

参考文献

[1] 李舟军,范宇,吴贤杰.面向自然语言处理的预训练技术研究综述.计算机科学 2020-03-24
[2] 刘媛.美国自然语言处理技术专利情报分析及启示——基于1999—2018年专利数据.
[3] 葛运东,陈洪梅，姚建民.自然语言处理的技术和产业应用现状与趋势分析.产业与科技论坛.2019-09-01.
[4] 李光明,潘以峰,周宗萍.基于自然语言处理技术的学生管理论坛文本挖掘与分析.智库时代.2019-07-08.
[5] 陈梁.运用自然语言处理技术从电子化病例中提取临床有用信息.重庆医科大学.2019-05-01.
[6]侯佳腾,常薇,林冠峰.基于自然语言理解技术的智能客服机器人的设计与实现[J].电子技术与软件工程,2019(23):238-240.
[7]陆正扬.基于计算机自然语言处理的机器翻译技术应用与简介[J].科技传播,2019,11(22):140-141.
[8]尚瑛杰,董丽亚,何虎.基于脉冲神经网络的迁移学习算法与软件框架[J].计算机工程,2020,46(03):53-59.
[9]李钰. 基于脉冲神经网络的人体摔倒动作识别研究[D].哈尔滨工业大学,2019.
[10]Antonio Benítez-Burraco,Ljiljana Progovac. A four-stage model for language evolution under the effects of human self-domestication[J]. Language and Communication,2020,73.
[11]阮海博. 基于动态视觉传感器的连续脉冲识别算法研究[D].浙江大学,2019.
[12]董蒙. 脉冲神经网络的算法与实现研究[D].哈尔滨理工大学,2018.
[13]余传明,王曼怡,林虹君,朱星宇,黄婷婷,安璐.基于深度学习的词汇表示模型对比研究[J/OL]. 数据分析与知识发现:1-19[2020-05-07].http://kns.cnki.net/kcms/detail/10.1478.G2.20200423.1336.012.html.
[14]Andrew Myles-Wright, Claire Nee. Holding the Child (and Practitioner) in Mind? Youth Justice Practitioners' Experiences Supervising Young People Displaying Sexually Harmful Behavior. 2020, 35(9-10):2055-2081.
[15]Yu-Hsiang Su, Ching-Ping Chao, Ling-Chien Hung, et al. A Natural Language Processing Approach to Automated Highlighting of New Information in Clinical Notes. 2020, 10(8)
[15]张新尉. 微博短文本分类算法的研究与实现[D].北京邮电大学,2016.
[16]孟朋. 自然语言信息隐藏与检测研究[D].中国科学技术大学,2012.
[17]楚彦凌. 基于数据聚类的语言模型研究[D].北京邮电大学,2010.
[18]刘丹,方卫国,周泓.基于贝叶斯网络的二元语法中文分词模型[J].计算机工程,2010,36(01):12-14.
[19]马珍珍,朱建军,张世涛,王嚞华,刘小弟.面向犹豫模糊语言信息的大型群体分类集结模型[J].控制与决策,2019,34(01):167-179.

[20]张亚泉.论自然语言信息处理及模糊自动分类理论[J].图书情报知识,1989(04):36.