计算机应用的基础知识：文本表示综述及其改进 PDF转换可能丢失图片或格式，建议阅读原文 https://www.100test.com/kao_ti2020/136/2021_2022__E8_AE_A1_E7_AE_97_E6_9C_BA_E5_c98_136377.htm 文本表示综述及其改进 主要内容:现阶段文本表示的主要技术已有的工作对我们的启发已有的改进工作的介绍我们的改进(可行性?)计算机如何解决文本分类问题?一个中文文本表现为一个由汉字和标点符号组成的字符串，由字组成词，由词组成短语，进而形成句、段、节、章、篇等结构。 自然语言理解借助统计学这个有力的工具现阶段文本表示的主要技术向量空间模型 特征项的粒度选择预处理去除停用词 特征选择 特征项权重计算特征重构VSM向量空间模型（Vector Space Model）Salton的概念文档（Document）特征项（Term） 特征项的权重（Term Weight） 向量空间模型（VSM） 相似度（Similarity） 特征项的粒度字 简单高效,国家标准GB2312-80 中定义的常用汉字为6763个 . 表示能力比较差，不能独立地完整地表达语义信息。词 词是最小的能够独立运用的语言单位 . 词的个数在10万个以上,面临复杂的分词问题 特征项的粒度(2)短语特征 和词相比频率更低,表现力更强概念特征“ 爸爸 ”=“ 父亲 ”,在自动文摘领域很有帮助N元组特征“ 中国人民银行 ”2元组: 中中国 国人 人民 民银 银行 行 主要用于自动纠错.特征项的粒度(3)重复串特征? 分词程序的统计逼进新的粒度?David Lewis 的结论: 单个word作为特征效果好于phrase和cluster of phrase以及cluster of words.phrase的低频率和高同义性(synonymy)大大的影响其性能 .(抵消了phrase的低歧义性的好处)而cluster of words的效果不佳主要的原因应该还是训练集不够大的缘故 .

预处理去除停用词虚词,助词出现频率高,对于表达意义的贡献却不大.如:“着”、“了”、“过”、“的”、“地”、“得”统计词频时过滤掉这些停用词. 停用词无用吗?红楼梦作者考证 李贤平 1987利用120回中每一回用的47个虚字（之，其，或，亦……，呀，吗，咧，罢……；的，着，是，在，……；可，便，就，但，……，儿等）出现的频率进行聚类.前80回基本聚成一类,后40回聚类情况较零散.得出结论:前80回与后40回之间有交叉。前80回是曹雪芹据《石头记》写成，中间插入《风月宝鉴》，还有一些别的增加成分。后40回是曹雪芹亲友将曹雪芹的草稿整理而成，宝黛故事为一人所写，贾府衰败情景当为另一人所写。 特征选择 目标表达力强频率较高区分度高合理的特征评价函数消除干扰,提高分类准确率 特征空间降维,减少运算量特征选择(2)文档频次 (DF) 根据预先设定的阈值去除那些文档频次特别低和特别高的特征项。 合理的阈值往往难以得到 ! 互信息(MI)出现频率差异很大的特征项的互信息大小不具有可比性 !(即低频特征具有较高的MI)同时,训练集中不同类别相差较大时,低频词也有较大MI.实践证明,互信息方法是效果最差的特征选择方法!特征选择(3) χ 2统计量:用于度量特征项w 和类别C 之间的独立性 对低频特征项的区分效果也不好 !信息增益(IG):该特征项为整个分类所提供的信息量 将长文档和短文档视为等同.频率信息.特征选择性能比较:特征项权重计算布尔权重词频权重 TFIDF权重(为什么?)权重计算(2)TFC权重: 对TFIDF进行归一化LTC权重:降低TF的作用(最常用)(不区分长短文章)熵权重: (效果最好)特征重构LSI：(Latent Semantic Indexing) 一词多义和多词一义的等现象使得文本特征项向量中的不同分量之间互相关

联. LSI 方法 通过矩阵奇异值分解(SVD),将特征项空间中的文本矩阵转换成概念空间中的正交矩阵 ,概念空间中各个特征之间是相互独立的. 进行奇异值分解过程中信息损失过大，因此在自动文本分类问题中往往性能不佳!VSM潜在的问题:长文档(黄萱菁) VSM把文档看成是文档空间的一个向量 ,实践结果表明对长文档来说不适宜.长文档内容比较丰富,必须对文档长度进行正规化出现问题.解决的办法是对长文档进行分割。(如何定义“ 长 ”?) 特征项的独立性假设 如果我们把词看做特征项,那么词之间的独立性即意味着一个词的上下文信息对于这个词的信息没有任何作用!这显然是与人们的常识相违背的. 思路启发 1:特征项顺序关系 (卜东波) 中文文本由特征项的频率及相互的顺序表达 .考虑顺序信息,必然使用有向指针,使得文本变成复杂的图结构 .由于难以定义合理的距离函数描述两个由图结构表示的文本是否相似,因此不得不舍弃顺序信息 . 要尽可能的考虑特征项间的顺序关系. 思路启发 2:上下文信息贡献 (鲁松)引入信息增益方法确定上下文各位置的信息量后，构造上下文位置信息量函数，最终通过多项式积分确定85%信息量的上下文边界，即汉语核心词语最近距离[-8, 9]和英文[-16, 13]位置之间的上下文范围。 要尽可能的考虑特征项的上下文贡献. 思路启发 3:分类器集成分类器集成思想的提出: 不同的分类算法往往适用于不同的特定问题,没有最优的分类算法. 希望把不同的方法综合在一起,尽可能的减小错误的发生.