一.向量空间模型(vector space model)

向量空间模型是指把对文本内容的处理简化为向量空间中的向量运算，并且它以空间上的相似度表达语义的相似度，直观易懂。当文档被表示为文档空间的向量，就可以通过计算向量之间的相似性来度量文档间的相似性。文本处理中最常用的相似性度量方式是余弦距离。

VSM基本概念：

（1） 文档(Document):泛指一般的文本或者文本中的片断(段落、句群或句子),一般指一篇文章,尽管文档可以是多媒体对象,但是以下讨论中我们只认为是文本对象,本文对文本与文档不加以区别"。

（2） 项(Term):文本的内容特征常常用它所含有的基本语言单位(字、词、词组或短语等)来表示,这些基本的语言单位被统称为文本的项,即文本可以用项集(Term List)表示为D(T1,T2,,,,Tn)其中是项,1≤k≤n"

（3） 项的权重(TermWeight):对于含有n个项的文本，项常常被赋予一定的权重表示他们在文本D中的重要程度。

（4） 相似度(Similarity)：两个文本之间的(内容)相关程度(Degree of Relevance)常常用他们之间的相似度来度量,当文本被表示为向量空间模型时,我们可以借助与向量之间的某种距离来表示文本间的相似度"常用向量之间的内积进行计算，或者用夹角的余弦值表示。

         可以看出，对向量空间模型来说，有两个基本问题：即**特征项的选择和项的权重计算**。

**1.特征项选择**

用来表示文档内容的项可以是各种类别，对汉语来说，有字、词、短语，甚至是句子或句群等更高层次的单位。项也可以是相应词或短语的语义概念类。

项的选择必须由处理速度、精度、存储空间等方面的具体要求来决定。特征项选取有几个原则：一是应当选取包含语义信息较多，对文本的表示能力较强的语言单位作为特征项；二是文本在这些特征项上的分布应当有较为明显的统计规律性，这样将适用于信息检索、文档分类等应用系统；三是特征选取过程应该容易实现，其时间和空间复杂度都不太大。实际应用中常常采用字、词或短语作为特征项。

由于词汇是文本最基本的表示项,在文本中的出现频度较高,呈现一定的统计规律,在考虑到处理大规模真实文本所面临的困难,一般选择词汇或短语作为特征项,但是直接选用文本中的词或词组作为文本特征项也会存在以下问题:

(1) 文本中存在一些没有实在意义但使用频率很高的虚词和功能词,如中文中“的”、“把”、“了”等,常常把一些真正有分类作用的实词淹没掉了。解决这个问题的方法是把这些词组织成一个禁用词表,或者进行权重计算时,使它们的权重很低,通过取阀值将它们丢弃。采用禁用词表时,词表的选择很关键,很难全面地包括所有的禁用词,并且语言是不断发展的,禁用词表也是随着训练文本集合的不同而不同,某个词在这里不是禁用词,到另外一类文本中可能就成了禁用词。另一方面考虑到,最能代表一篇文章实际意义的词,往往是那些实词,如形容词、动词、名词,而且同一个词,当处于不同词性时,可能分别属于和不属于禁用词表。例如：“他高兴地走了”(副词“地”应是禁用词),“地很不平”(名词“地”不应作为禁用词)"针对这个现象,提出了只提取形容词、动词和名词作为特征项,并尝试着取代禁用词表方法.

(2) 采用词语作为特征项时还会出现所谓的同义现象,同义现象是指:对于同一个事物不同的人会根据个人的需要、所处的环境、知识水平以及语言习惯有着不同的表达方式,因此所采用的词汇也有很大的不同。所以经常出现两个文本所用的词汇有所不同,但实际上两者是相似的，这就是词的同义现象造成的。例如电脑和计算机是同一个概念,应该属于同一个特征项,目前最常用的解决方案是采用概念词典来解决这个问题。

**分词：**

确定了特征项单位以后,接下来要做的就是把文本分割成特征项的表示。我们知道,词是最小的能够独立活动的有意义的语言成分。然而,汉语是以字为基本的书写单位,文本中词与词之间没有明确的分隔标记,而是连续的汉字串,显而易见,自动识别词边界,将汉字串分为正确的词串的汉语分词问题无疑是实现中文信息处理各项任务的基础与关键。中文词语分析一般包括3个过程:预处理过程的**词语粗切分、切分排歧与未登陆词识别、词性标注**。目前中文词语分析采取的主要步骤是:先采取最大匹配、最短路径、概率统计、全切分等方法,得到一个相对最好的粗分结果,然后进行排歧、未登陆词识别,最后标注词性。在实际系统中,这三个过程可能相互交叉、反复融合,也可能不存在明显的先后次序。可以将现在的分词算法分为3大类:基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法。

**特征值抽取：**

一篇文章在经过了分词处理之后,会产生很多词条。如果一个文档所有词条都被作为其特征,将会使特征项异常庞大,而且这样的特征项会使得每个特征项所含信息非常平滑,有用信息反而不会突出。因此我们需要进行特征项选取,把词条中最能代表某类文本信息的词条挑选出来,作为文本的特征项。实验结果表明简化特征项不但不会使分类结果准确率降低,而且还会使结果更加准确。特征项选择一般使用统计方法,利用各种计算公式,计算词代表的信息含量,确定一个阀值,将低于阀值的词语过滤掉。或者确定一个特征项数目n,保留处于信息含量在前n位的词条。

特征抽取算法是文本自动分类中的一项关键技术和瓶颈技术,如何从原始文本特征集合中选择最能表示文本主题内容的特征子集,是文本特征抽取算法的研究目标。目前,有多种特征抽取算法被用于文本自动分类的研究中,但这些算法都有其优点和缺点,没有公认的最优方法,需要针对具体系统进行对比来确定最优方法。

特征选择可以从两个方面提高系统性能一是分类速度,通过特征选择,可以大大减少特征集合中的特征数,降低文本向量的维数,简化计算,防止过度拟合,提高系统运行速度。二是准确率,通过适当的特征选择,不但不会降低系统准确性,反而会使系统精度提高。

在文本处理中,一些常用特征提取评估函数有文档频数(document frequency)、信息增益(information gain)、期望交叉熵(expected cross entropy)、互信息(mutual information)、统计(CHI)、文本证据权(the weight of evidence for text)等。

（1） 文档频数DF

它是最简单的评估函数,值为训练集合中该单词发生的文本数。DF评估函数的理论假设稀有单词可能不包含有用信息,也可能太少而不足以对分类产生影响,也可能是噪音,因此可以删去。显然它在计算量上比其他评估函数小很多,但是实践运用中它的效果却很好.DF的缺点是稀有单词可能在某一类文本中并不稀有,也可能包含着重要的判断信息,错误的舍弃,可能影响分类器的精度。因此,在实际运用中一般并不直接使用DF。

（2） 信息增益(information Gain)

信息增益表示文档中包含某一特征值时文档类的平均信息量。它定义为某一特征在文档中出现前后的信息熵之差。假定c为文档类变量,C为文档类的集合,d为文档,f为特征(以下各节同此)。对于特征f,其信息增量记为IG(f)，计算公式如下：

       IG(f)=H(C)-H(C|f)

**2.特征项赋权**

为了兼顾查全率和查准率，检索系统在对特征项进行赋权时，应同时包含提高查全率和查准率的赋权因子。特征项赋权因子由频率因子（TF）、文档集因子（DF）和规格化因子三部分组成。

TF-IDF 权重

特征项的权重计算是文本相似度计算中的一个非常重要的环节。一篇文本中的特征项数目众多，要想得到比较准确的对文本内容的数学化表示，我们需要对能显著体现文本内容特征的特征项赋予高权重，而对不能可以体现文本内容特征的特征项赋予低权重。从效率方面来说，特征项权重的计算是文本相似度计算中的主要工作，它的效率也直接影响文本相似度计算的整体效率。

经典的 TF-IDF 权重是向量空间模型中应用最多的一种权重计算方法，它以词语作为文本的特征项，每个特征项的权重由 TF 权值和 IDF 权值两个部分构成。

TF-IDF 是基于统计的权重计算方式，在全局文本集包含的语料特征足够的情况下，这种基于统计学的方法经过实践检验是一种有效的特征项权重衡量方法。其局限性在于它的准确度受全局文本集的影响较大：全局文本集越大，语料越完备，所得的权重也就越准确，但相应地计算效率也会随着全局文本集的增大而降低。