作业二 实验报告

1. 实验目的

对20 Newsgroups dataset文档集使用Naïve Bayes方法进行分类。

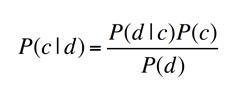
1. 实验环境

Windows 10 + Python3.6.5

1. 实验步骤

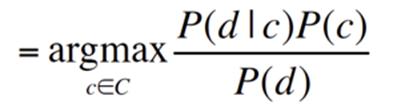
预处理方法同作业一。然后划分训练集与测试集，每个类别取80%的文档作为训练集，20%为测试集，并放入相应的文件夹。

对于一个文档 d 和一个分类 c

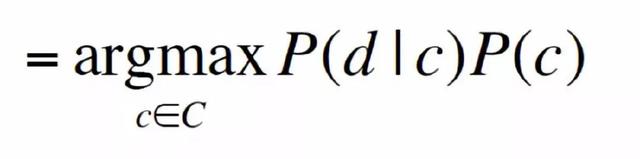


MAP 是最大化后验概率，或者说：最有可能的类别。

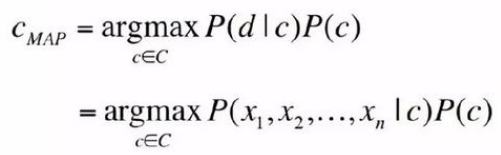
根据贝叶斯规则，转化为



去掉共同的分母，转化为



文档 d 表示为 特征 x1 .. xn

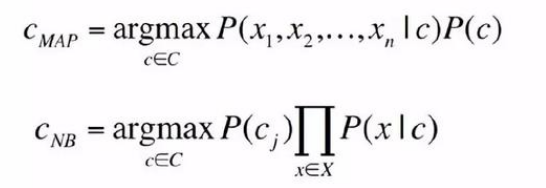


多项朴素贝叶斯独立假设

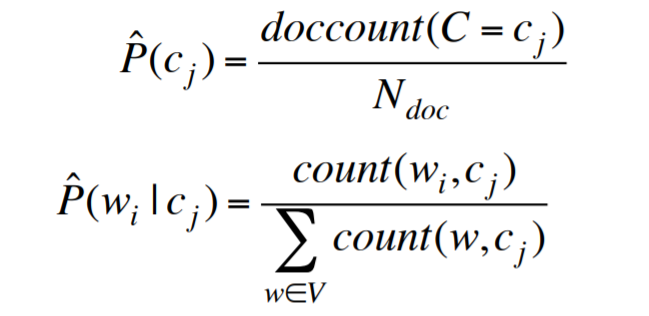
词袋模型假设：假设位置并不重要

条件假设：假设特征概率 P(xi,cj)是独立的，在类别 c 给出的情况下。

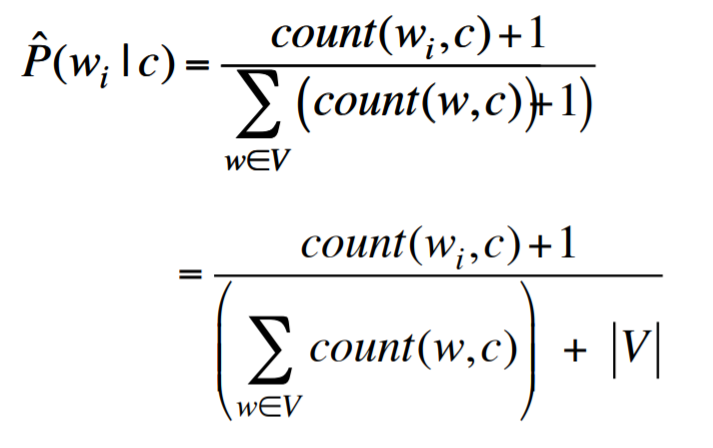
由此可以推出以下等式：



其中P(cj)为第j类文档的先验概率，P(x|c)为第c类文档中单词x出现的概率，以上值均可由频率统计得到：



为避免某一项P(x|c)为0（测试集中包含训练集没有的单词）影响分类结果，为P(x|c)采用平滑的操作：



|V|为训练集词典的大小（不计重复）

1. 实验结果

