**Homework1贝叶斯应用经典案例 实验报告**

**实验目的：**

1788年，包含85篇文章的著名的《联邦党人文集》集结出版。《联邦党人文集》出版的时候，Hamilton坚持匿名发表，于是，这些文章到底出自谁人之手，成了一桩公案。在之后的1810年和1818年，Hamilton和Madison分别列出了两份作者名单。在85篇文章中，有73篇文章的作者身份较为明确，其余12篇存在争议。而Hamilton和Madison的写作风格又极其接近。实验要求用贝叶斯方法对这12篇disputed的文章进行分类，分成Hamilton和Madison两大类，来推测它们的作者。

**实验材料：**

《联邦党人文集》的85篇文章，按文章编号顺序存在了The Federalist Papers.txt文件中。原网址是<https://avalon.law.yale.edu/subject_menus/fed.asp>。

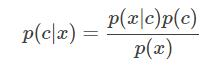
**实验思路：**

1. 把这85篇文章从网上下载下来或者爬取下来，合到一个文本文件中，作为数据集。
2. 写一个用来分类、提取、排序85文章的程序splitter.py，将85篇文章中作者为Hamilton、Madison和disputed的三类文章提取出来，每一类按文章编号顺序做成一个列表，三类合起来做成一个大的列表，存入一个新的文本文件。
3. 写一个用来推测11篇未知作者的主程序classifier.py。该程序先要提取出在Hamilton和Madison两个作者的文章中出现频率高的且相对有意义的词，作为特征词。再根据11篇文章中特征词的出现情况，依据一定的算法，推测出对应的作者。

**实验原理：**

如何依据特征词来推断出作者？首先依次统计Hamilton所有文章中各特征词出现的频率、Madison所有文章中各特征词出现的频率、各特征词在两个作者文章中出现的频率比，均以列表的形式呈现。然后，在11篇disputed的文章中依个检索所有特征词，在每篇文章中，某个词出现了就记为1，未出现就记为0，这样又做成一个列表。

依据贝叶斯公式：



而根据朴素贝叶斯的思想，把每个特征词的出现看成是独立的。假设每个特征相互独立，即每个词相互独立，不相关。则：



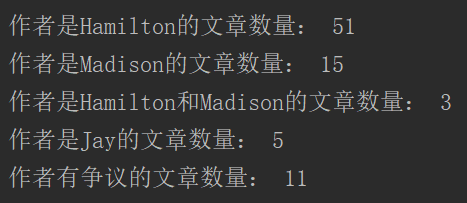
代入到这个案例中，代表作者是Hamilton，代表作者是Madison，而x1到xn则代表各个特征词。易得：

若 >1，则判定为作者是Hamilton；反之，判定为Madison。

其中，和很容易就能得到，重点是中间 那一串。而之前得到了各特征词在两个作者文章中出现的频率比，以及各个特征词是否出现在该篇争议文章中（用0、1来表示），可以利用np.power(a,b) for a,b in zip()这个函数，作用是对于a，b中的每一项，用a的b次方表示，这样若该词出现了，则对原频率比做一次方，保持不变；而若该词未出现，做零次方，得1。之后对所有项的结果做乘积，就是的值。

**实验结果：**

对85篇文章做分类的结果是：



说明：因为不知道那3篇作者是Hamilton和Madison的文章中，两位作者各写了多少、各写了哪些部分，为了避免混淆，把这三篇单独分出来一个类别，并且不加到后续的数据集列表中，而作者是Jay的5篇文章也是这样处理。这样一来有争议的文章只有11篇，而网上说的是有12篇，不知道为何。

对11篇争议文章的作者推断结果是：



程序的运行结果显示了筛选出的特征词列表、总共的特征词个数、两位作者文章中各个特征词的出现次数、频率、频率比（由于列表太长只截取了前几项）、以及每篇争议文章编号及其预测作者。

可以看到，共有11篇争议文章，其中第49、50、51等10篇被推测为作者是Madison，只有第56篇被推测为作者是Hamilton。查阅资料发现，历史学家推测这11篇文章均为Madison所写，那么这个程序的推测准确率为10/11 = 90.91%。