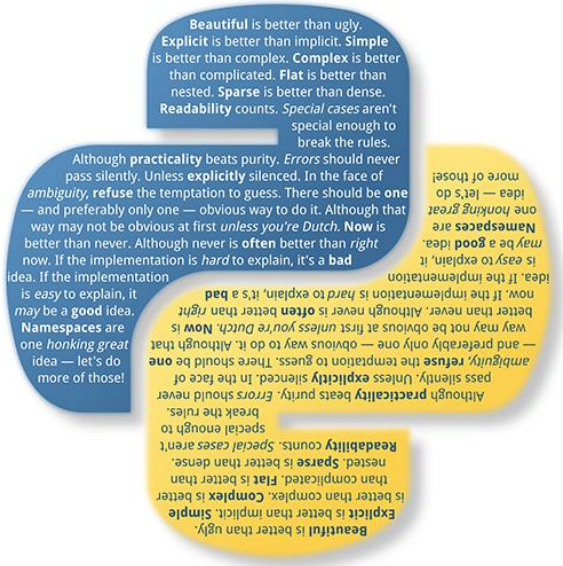
CDA数据分析师
CERTIFIED DATA ANALYST

CDA大数据分析师就业班
之
Python机器学习



python™

贝叶斯分析

新闻分类


首页
百家号
国内
国际
军事
社会
财经
娱乐
体育
互联网
科技
游戏
时尚
女人
汽车
房产
个性推荐
热点要闻
个性推荐
进入推荐版
习近平引领中国经济向高质量发展阶段迈进
厉害了，这份成绩单真亮眼！
习近平主席视察王杰生前所在连在全军引起强烈反响
见证中国经济新高地 领航新征程 暖新闻
从全面开放看中国经济新格局 新时代新气象新作为
浙江省农科院同事眼中的王一成
韩媒：5架中国飞机今日飞入韩国防空识别区
中国空军编队飞越对马海峡赴日本海远洋训练
最高法：推动家事审判与少年审判工作协同发展
两副部从外省调入，华东大省近期密集人事调整
中国2018年起取消钢材等部分产品出口关税 日媒挑拨
十九大后首个中央经济工作会议，意义非同寻常
2017百度热搜：已经秃顶成佛，依旧关注苍生
港媒："港生秦国歌拒肃立"系预谋 媒体假扮亲友入场
共青团安徽省委："大眼睛"苏明娟属兼职 无行政级别
1架上海飞马德里航班起飞不久后返航：乘客身体不适
人民日报再批形式主义:一个乡迎接检查打印费就10多万
9省份职业卫生技术服务机构检查：涉弄虚作假严重
特朗普演说将批中国经济"侵略"？白宫急忙否认
日本机场将普及使用"人脸识别"办理出入境手续
胜局已定！印度执政党将在莫迪老家赢得关键选举
习近平会见来京述职的林郑月娥
理上网来·辉煌十九大
十九大·理论新视野
热搜新闻词 HOT WORDS
这些经济大事习近平要求明年重点干
十九大后，习近平对"四风"问题再出重拳
第2架C919大型客机完成首飞
十部门公布清洁取暖规划
机器人产业迈向中高端
地方环保税税额标准公布
明年养老金将迎十四连涨
WIFI全覆盖将推广到高铁
四价HPV疫苗上市遭"抢"
11月份70个城市房价下降
百家号 BAIJIA
财经 | 文化 | 娱乐 | 体育
加入百家号
腾讯京东终于牵手唯品会，电商格局再生变数
这就是法拉第造车工厂，网友：贾跃亭还好
网上有害信息举报

评论分析


酒店详情
酒店点评(3027)
立即预订
luya****
2013-12-23
总评: 5.0 卫生：5 服务：5 设施：5 位置：5
豪华房
价格便宜 性价比高 交通便捷 靠近市区 服务不错。[详情]
有用(0)
luya****
2013-12-23
总评: 5.0 卫生：5 服务：5 设施：5 位置：5
高级房
价格公道 性价比高 交通便捷 酒店餐厅很好吃 服务也很到位。[详情]
有用(0)
luya****
2013-12-23
总评: 5.0 卫生：5 服务：5 设施：5 位置：5
豪华房
五星级酒店而言 价格便宜 性价比高 交通便捷 服务到位。[详情]
有用(0)
1100****
2013-12-23
总评: 5.0 卫生：5 服务：5 设施：5 位置：5
高级房
价格合理，出行方便[详情]
有用(0)
酒店回复： 2013-12-24
尊敬的顾客您好，感谢您入住上海明悦大酒店并对我们酒店做出的肯定，期待您的下次光临！
300720****
2013-12-23
总评: 3.8 卫生：5 服务：5 设施：3 位置：2
高级单人房
在携程订购的话给的房间都是最小的。别的还行[详情]
有用(0)
来自：手机用户
酒店回复： 2013-12-24
尊敬的顾客您好，感谢您入住上海明悦大酒店并对我们酒店做出的肯定，期待您的下次光临！
109216****
2013-12-23
总评: 5.0 卫生：5 服务：5 设施：5 位置：5
高级单人房
还不错。[详情]
有用(0)

贝叶斯介绍

- 贝叶斯(Thomas Bayes，1701—1761)：英国牧师、业余数学家。为了证明上帝的存在，他研究概率统计学原理。但生前没发表科学论著。
- 《机遇理论问题中一个问题的解》在其逝世2年后发表，开创了贝叶斯分析的崭新统计思维新式。但当时没受到重视（超越时代，数学发展史中有很多类似的情况）。
- 20世纪中叶以后，由于经典统计遭遇困难(扔硬币)，逐渐进入全盛时期被发展为一种关于统计推断的系统理论和方法，称为“贝叶斯方法”，由这种方法得到的统计推断全部结果，称为“贝叶斯统计学”。
- 信奉贝叶斯统计，乃至鼓吹贝叶斯观点是统计推断唯一正确新法的学者，形成数理统计学中的贝叶斯学派（Bayesian school）

数理统计学处理的信息

总体信息：当前总体样本符合某种分布。比如抛硬币，二项分布。学生的某一科的成绩符合正态分布。
样本信息：通过抽样得到的部分样本的某种分布。
抽样信息=总体信息+样本信息
基于抽样信息进行统计推断的理论和方法称为经典统计学。

先验信息：抽样之前，有关推断问题中未知参数的一些信息，通常来自于经验或历史资料。
基于总体信息+样本信息+先验信息进行统计推断的方法和理论，称为贝叶斯统计学。

先验信息存在吗？

戴维·萨尔斯伯格统计学名著《The Lady Tasting Tea——How Statistics Revolutionized Science in the Twentieth Century》（《女士品茶》）

例 1.1.2 英国统计学家 L. J. Savage (1961) 提出一个令人信服的例子说明先验信息有时是很重要的, 且看下面两个统计试验:

(1) 一位常饮牛奶和茶的女士说她能辨别出先倒进杯子里的是茶还是牛奶, 对此做了 10 次试验, 她都说对了.

(2) 一位音乐家说他能够从一页乐谱辨别出是海顿 (Haydn) 还是莫扎特 (Mozart) 的作品, 在 10 次试验中, 他都说对了.

在上面两次试验中, 如果认为试验者是猜对的, 每次成功概率为 0.5, 则 10 次都猜中的概率为 $(0.5)^{10} \approx 0.000\,976\,6$, 这是一个很小的概率, 几乎不可能发生. 故每次猜对的概率为 0.5 的假设被否定. 他们每次说对的概率比 0.5 要大得多, 这不能认为是猜测, 而是经验帮了忙. 可见经验 (先验信息) 在推断中不可忽视, 应当加以利用.

古典学派和贝叶斯学派的争论

古典频率学派认为概率来源于统计，需要通过统计来得到概率。
抛硬币

贝叶斯学派认为有些情况概率是来源于统计，有些情况概率来源于先验知识。
特朗普竞选总统

矛盾点在于是否承认先验概率

贝叶斯定理

贝叶斯定理告诉我们如何交换条件概率中的条件与结果

即如果

已知P(X|H)，要求P(H|X)，那么可以使用下面的计算方法：

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

贝叶斯定理

$P(H|X)$给定观测数据样本X，假设H是成立的概率
比如X是一份具有特定特征的邮件，H是垃圾邮件。它里面包含很多的单词（特征），然后我们判断这封邮件属于垃圾邮件的概率是多少。

$P(H|X)$是后验概率。比如一份特定邮件中，是垃圾邮件的概率。

$P(H)$是H的先验概率。比如总体邮件中垃圾邮件的概率。
$P(X)$是X的先验概率。比如总体邮件中带有特定特征的邮件概率。

可以通过抽样来计算先验概率。抽样的数量越大，得到的结果越接近于真实的概率分布-大数定理。

贝叶斯定理

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

举例：
$P(H)$垃圾邮件的先验概率。
$P(X)$特定特征的先验概率。
$P(X|H)$在垃圾邮件中，包含特定特征(比如“办证”)邮件的概率。
$P(H|X)$包含特定特征(比如“办证”)的邮件属于垃圾邮件的概率。

贝叶斯定理-举例

邮件：总体100，正常70，垃圾30。
“办证”在正常邮件中出现10次，在垃圾邮件中出现25次

假设X为“办证”，H为垃圾邮件
$P(X|H)$=25/30=5/6
$P(H)$=30/100=3/10
$P(X)$=35/100=7/20

$$P(H|X)=\frac{P(X|H)P(H)}{P(X)}=\frac{\frac{5}{6}\times\frac{3}{10}}{\frac{7}{20}}=\frac{5}{7}$$

包含“办证”这个词的邮件属于垃圾邮件的概率为5/7

贝叶斯定理-举例

分词

办证
理财
投资
资讯
北京
学习
天气
……

筛选

办证
理财
投资
资讯

	办证	理财	投资	资讯
总共100	15	15	35	40
正常70	5	5	20	25
垃圾30	10	10	15	15

贝叶斯定理-举例

	办证	理财	投资	资讯
总共100	15	15	35	40
正常70	5	5	20	25
垃圾30	10	10	15	15

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

在有多个特征的情况下会使得统计量巨大。
比如需要计算办证对于垃圾邮件的影响
计算办证+理财对于垃圾邮件的影响
计算办证+理财+投资对于垃圾邮件的影响
计算办证+理财+投资+资讯对于垃圾邮件的影响
……
所有特征需要计算2^n次，n是特征数

朴素贝叶斯(Naive Bayes)

假设：特征X1，X2，X3……之间都是相互独立的

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} = \frac{P(X_1|H)P(X_2|H)\dots P(X_n|H)P(H)}{P(X_1)P(X_2)\dots P(X_n)}$$

多项式模型

“代开发票。增值税发票，正规发票。”
分词后为向量：
（“代开”，“发票”，“增值税”，“发票”，“正规”，“发票”）
重复的词语我们视为其出现多次：

$$P(\text{（“代开”，“发票”，“增值税”，“发票”，“正规”，“发票”）}|S)$$
$$= P(\text{“代开”}|S)P(\text{“发票”}|S)P(\text{“增值税”}|S)P(\text{“发票”}|S)P(\text{“正规”}|S)P(\text{“发票”}|S)$$
$$= P(\text{“代开”}|S)P^3(\text{“发票”}|S)P(\text{“增值税”}|S)P(\text{“正规”}|S)$$ 注意这一项：$P^3(\text{“发票”}|S)$

$$P(\text{“发票”}|S) = \frac{\text{每封垃圾邮件中出现“发票”的次数的总和}}{\text{每封垃圾邮件中所有词出现次数（计算重复次数）的总和}}$$

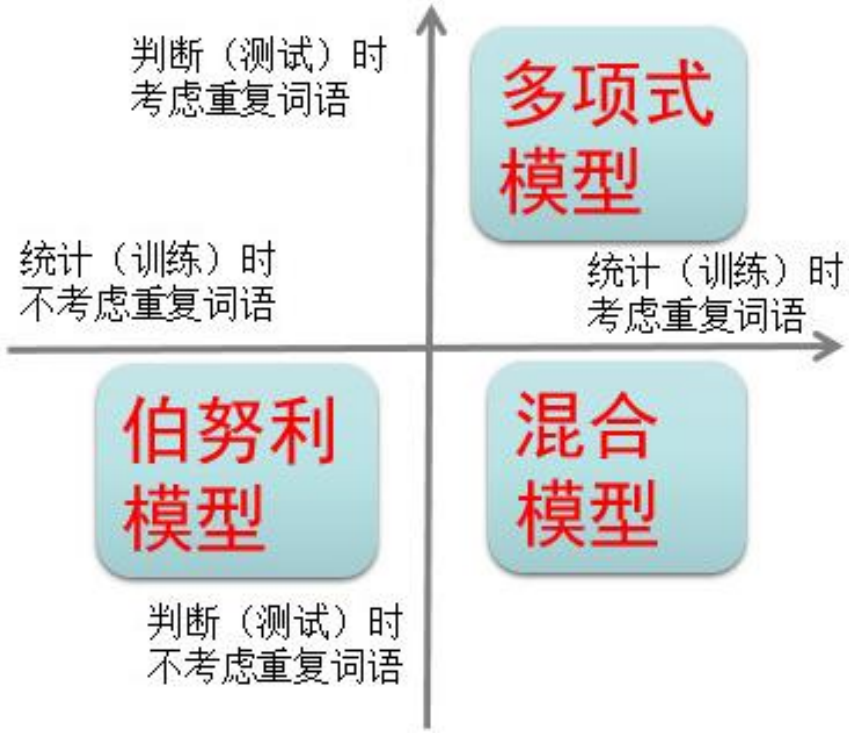
伯努利模型

“代开发票。增值税发票，正规发票。”
分词后为向量：
（“代开”，“发票”，“增值税”，“正规”，“发票”）
重复的词语我们视为其出现一次：

$$P(\text{“发票”}|S) = \frac{\text{出现“发票”的垃圾邮件的封数}}{\text{每封垃圾邮件中所有词出现次数（出现了只计算一次）的总和}}$$

混合模型

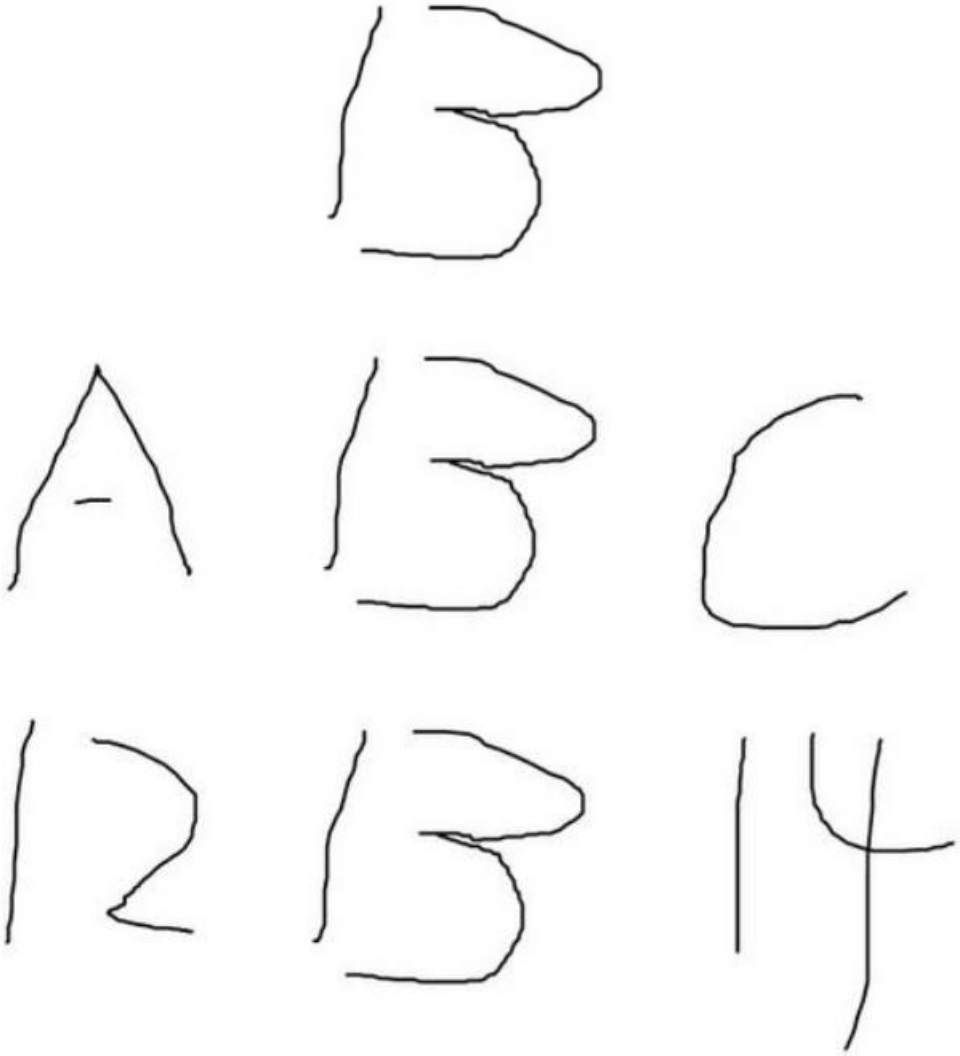
在计算句子概率时，不考虑重复词语出现的次数，但是在统计计算词语的概率P(“词语”|S）时，却考虑重复词语的出现次数，这样的模型可以叫作混合模型。



高斯模型

有些特征可能是连续型变量，比如说人的身高，物体的长度，这些特征可以转换成离散型的值，比如如果身高在160cm以下，特征值为1；在160cm和170cm之间，特征值为2；在170cm之上，特征值为3。也可以这样转换，将身高转换为3个特征，分别是f1、f2、f3，如果身高是160cm以下，这三个特征的值分别是1、0、0，若身高在170cm之上，这三个特征的值分别是0、0、1。不过这些方式都不够细腻，高斯模型可以解决这个问题。

大脑中的贝叶斯



大脑中的贝叶斯

Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe.

研表究明，汉字的序顺并不定一能影阅响读，比如当你看完这句话后，才发这现里的字全是乱的。

Prof. Daniel Kahneman的研究

贝叶斯- iris

一.iris数据集简介

iris数据集的中文名是安德森鸢尾花卉数据集，英文全称是Anderson's Iris data set。iris包含150个样本，对应数据集的每行数据。每行数据包含每个样本的四个特征和样本的类别信息，所以iris数据集是一个150行5列的二维表。

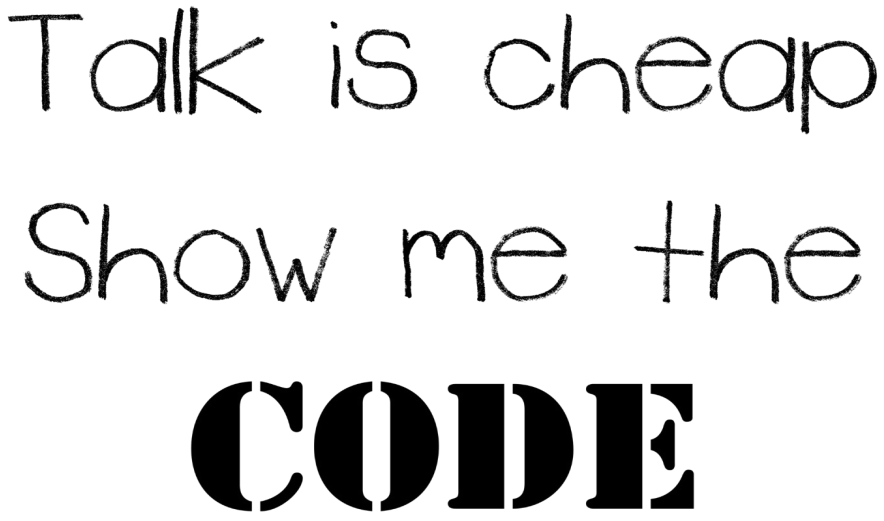
通俗地说，iris数据集是用来给花做分类的数据集，每个样本包含了花萼长度、花萼宽度、花瓣长度、花瓣宽度四个特征（前4列），我们需要建立一个分类器，分类器可以通过样本的四个特征来判断样本属于山鸢尾、变色鸢尾还是维吉尼亚鸢尾（这三个名词都是花的品种）。

iris的每个样本都包含了品种信息，即目标属性（第5列，也叫target或label）。

样本局部截图：

费雪鸢尾花卉数据集

花萼长度	花萼宽度	花瓣长度	花瓣宽度	属种
5.1	3.5	1.4	0.2	*setosa*
4.9	3.0	1.4	0.2	*setosa*
4.7	3.2	1.3	0.2	*setosa*
4.6	3.1	1.5	0.2	*setosa*
5.0	3.6	1.4	0.2	*setosa*
5.4	3.9	1.7	0.4	*setosa*
4.6	3.4	1.4	0.3	*setosa*
5.0	3.4	1.5	0.2	*setosa*



贝叶斯-英文新闻分类

Talk is cheap

Show me the

CODE

词袋模型(Bag of Words)

Bag-of-words model (BoW model) 最早出现在自然语言处理（Natural Language Processing）和信息检索（Information Retrieval）领域。该模型忽略掉文本的语法和语序等要素，将其仅仅看作是若干个词汇的集合，文档中每个单词的出现都是独立的。BoW使用一组无序的单词(words)来表达一段文字或一个文档。

词袋模型(Bag of Words)

基于文本的BoW模型的一个简单例子如下：
首先给出两个简单的文本文档如下：

John likes to watch movies. Mary likes too.

John also likes to watch football games.

基于上述两个文档中出现的单词，构建如下一个词典 (dictionary)：

{"John": 1, "likes": 2,"to": 3, "watch": 4, "movies": 5,"also": 6, "football": 7, "games": 8,"Mary": 9, "too": 10}

上面的词典中包含10个单词, 每个单词有唯一的索引, 那么每个文本我们可以使用一个10维的向量来表示。如下：

[1, 2, 1, 1, 1, 0, 0, 0, 1, 1]

[1, 1,1, 1, 0, 1, 1, 1, 0, 0]

该向量与原来文本中单词出现的顺序没有关系，而是词典中每个单词在文本中出现的频率。

TF-IDF

提取文章关键词：

1.提取词频(Term Frequency，缩写TF)。但是出现最多的词可能是“的，是，在”等对文章分类或搜索没有帮助的停用词(stop words)。
2.假设我们把停用词都过滤掉了，只考虑有意义的词。可能会遇到这样一个问题，“中国”，“蜜蜂”，“养殖”这三个词的TF一样。作为关键词，它们的重要性是一样的吗？
3.显然不是这样。因为"中国"是很常见的词，相对而言，"蜜蜂"和"养殖"不那么常见。如果这三个词在一篇文章的出现次数一样多，有理由认为，"蜜蜂"和"养殖"的重要程度要大于"中国"，也就是说，在关键词排序上面，"蜜蜂"和"养殖"应该排在"中国"的前面。

TF-IDF

所以，我们需要一个重要性调整系数，衡量一个词是不是常见词。如果某个词比较少见，但是它在这篇文章中多次出现，那么它很可能就反映了这篇文章的特性，正是我们所需要的关键词。

用统计学语言表达，就是在词频的基础上，要对每个词分配一个"重要性"权重。最常见的词（"的"、"是"、"在"）给予最小的权重，较常见的词（"中国"）给予较小的权重，较少见的词（"蜜蜂"、"养殖"）给予较大的权重。这个权重叫做"逆文档频率"（Inverse Document Frequency，缩写为IDF），它的大小与一个词的常见程度成反比。

TF-IDF

$$\text{词频(TF)} = \text{某个词在文章中的出现次数}$$

$$\text{词频(TF)} = \frac{\text{某个词在文章中的出现次数}}{\text{文章的总词数}}$$

$$\text{词频(TF)} = \frac{\text{某个词在文章中的出现次数}}{\text{该文出现次数最多的词的出现次数}}$$

TF-IDF

$$\text{逆文档频率(IDF)} = \log\left(\frac{\text{语料库的文档总数}}{\text{包含该词的文档数} + 1}\right)$$

$$\text{TF} - \text{IDF} = \text{词频(TF)} \times \text{逆文档频率（IDF）}$$

TF-IDF

还是以《中国的蜜蜂养殖》为例，假定该文长度为1000个词，"中国"、"蜜蜂"、"养殖"各出现20次，则这三个词的"词频"（TF）都为0.02。然后，搜索Google发现，包含"的"字的网页共有250亿张，假定这就是中文网页总数。包含"中国"的网页共有62.3亿张，包含"蜜蜂"的网页为0.484亿张，包含"养殖"的网页为0.973亿张。则它们的逆文档频率（IDF）和TF-IDF如下：

	包含该词的文档数（亿）	IDF	TF-IDF
中国	62.3	0.603	0.0121
蜜蜂	0.484	2.713	0.0543
养殖	0.973	2.410	0.0482

贝叶斯-练习

Talk is cheap

Show me the

CODE

贝叶斯拼写检查器

Talk is cheap

Show me the

CODE