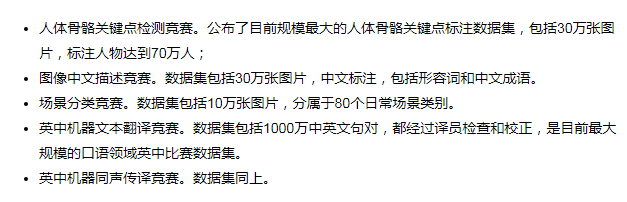
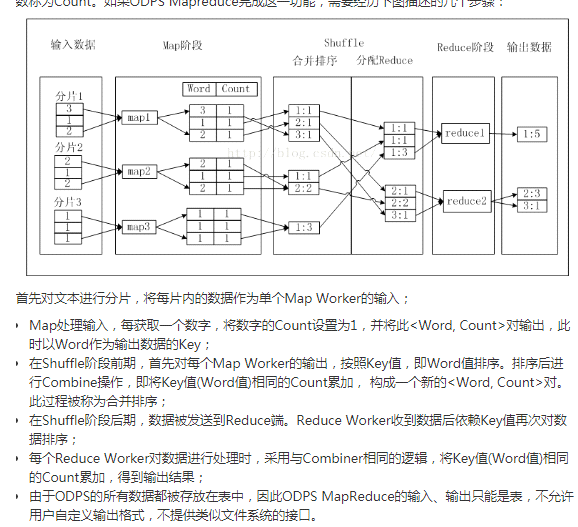
TIANCHI\_4362@aliyun-tianchi

4)8iWEhG



textrank



create table if not exists wcout1 (key string,cnt bigint);

打开主账户控制台，查看accesskey与value

Acecess **”**LTAIEDmUartK5Oak**”** **”MrKgfna9qs0bLQUJlOMm8COIiMbWK7”**

<https://yq.aliyun.com/articles/68227>问题解答

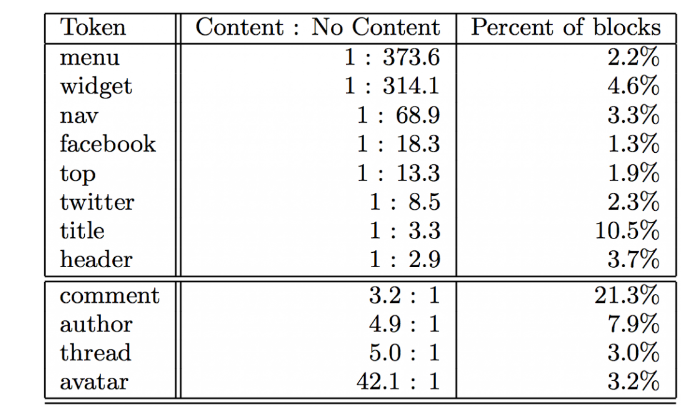
沙箱

使用自己的库产生故障

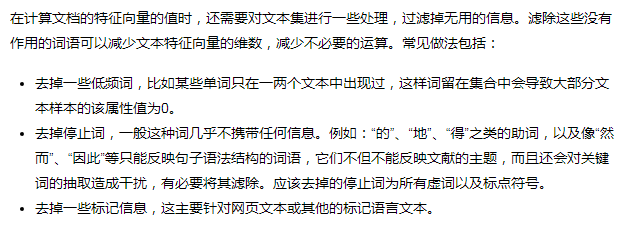
a).adl\_tianchi\_content\_risk\_training\_part1 b).adl\_tianchi\_content\_risk\_training\_part2 c).adl\_tianchi\_content\_risk\_training\_part3 d).adl\_tianchi\_content\_risk\_training\_part4 e).adl\_tianchi\_content\_risk\_training\_part5 f).adl\_tianchi\_content\_risk\_training\_part6 g).adl\_tianchi\_content\_risk\_training\_part7 h).adl\_tianchi\_content\_risk\_training\_part8 i).adl\_tianchi\_content\_risk\_training\_part9

adl\_tianchi\_content\_risk\_training\_part1 adl\_tianchi\_content\_risk\_training\_part2 adl\_tianchi\_content\_risk\_training\_part3 adl\_tianchi\_content\_risk\_training\_part4 adl\_tianchi\_content\_risk\_training\_part5 adl\_tianchi\_content\_risk\_training\_part6 adl\_tianchi\_content\_risk\_training\_part7 adl\_tianchi\_content\_risk\_training\_part8 adl\_tianchi\_content\_risk\_training\_part9

We also extract some additional structured information at this point: the title tag, the meta description tag, and H1/H2 tags.



<http://www.infoq.com/cn/articles/machine-learning-automatic-classification-of-text-data>



以文档频率为例，在特征选择过程中由于某些关键的词语低于了人为设定的阈值，所以会被直接忽视掉，而很多情况这部分词汇能包含较多的信息，对于分类的重要性比较大。怎么能够进一步理解这部分的信息，是急需要解决的问题。一个想法是找到这些使用频率比较低的词语相似的高频词，譬如在讨论“月亮”的古诗词中，包含了很多低频的同义词，如“玉兔”，“婵娟”等，如果我们能把这些低频的词语合并到一个维度，无疑是能够增强分类系统对文档的理解深度的。词向量这一概念能够有效地表示词语之间的相似性，适用于这种方法。

先介绍一下词向量的定义。一种最简单的词向量是one-hot representation，就是用一个很长的向量来表示一个词，向量的长度是词典D的大小N，向量的分量只有一个为1，其他全为0，1的位置对应该词在词典中的索引。这种词向量表示有一些缺点：容易受维数灾难的困扰。另一种词向量是Distributed Representation，它最早是Hinton于1986年提出来的，可以克服one-hot representation的上述缺点。其基本想法是：通过训练将某种语言中的每个词映射成一个固定长度的短向量。所有这些向量构成一个词向量空间，每个向量是该空间中的一个点，在这个空间上引入距离，就可以根据词之间的距离来判断它们之间的（词法、语义上的）相似性了。如何获取Distributed Representation的词向量呢？有很多不同的模型可以用来估计词向量，包括有名的LSA、LDA和神经网络算法。Word2Vec就是使用度比较广的一个神经网络算法实现的词向量计算工具。

s

kmeans聚类

召回率 r = a / (a + c) \* 100%

正确率 p = a / (a + b) \* 100%

F-测度值 F = (2 \* p \* r) /(p + r)

由于在分类结果中，对应每个类别都会有一个召回率和正确率，因此，可以根据每个类别的分类结果评价分类器的整体性能，通常方法有两种：微平均和宏平均。微平均是根据正确率和召回率计算公式直接计算出总得正确率和召回率值。宏平均是指首先计算出每个类别的正确率和召回率，然后对正确率和召回率分别取平均得到总的正确率和召回率。不难看出，宏平均平等对待每一个类别，所以它的值主要受到稀有类别的影响，而微平均平等考虑文档集中的每一个文档，所以它的值受到常见类别的影响比较大。

word2vec向量

一起词频统计，剔除低频词，特征表示，归一化，svm分类

词频需要测试集与训练集一起

Gambling 30324 0.027

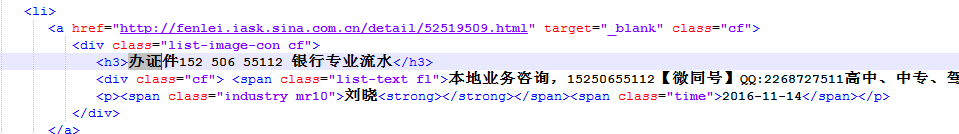
Fake card 59219 0.0538

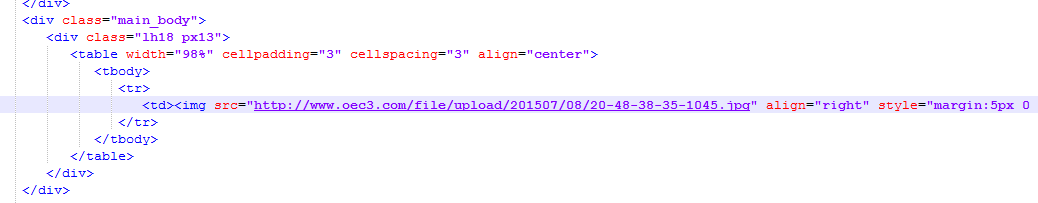
Sexy 23833 0.021

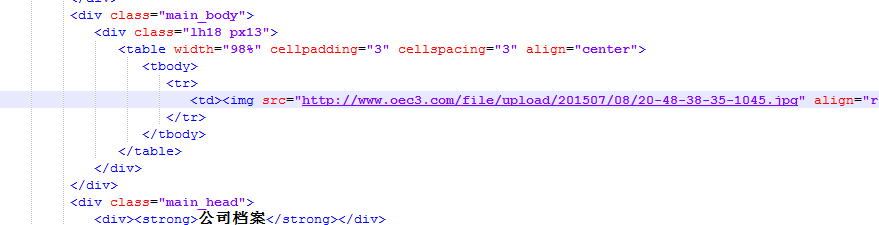
Normal 986623 0.89

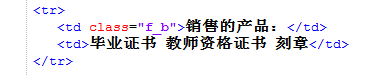
Total : 1099999

Fake card:





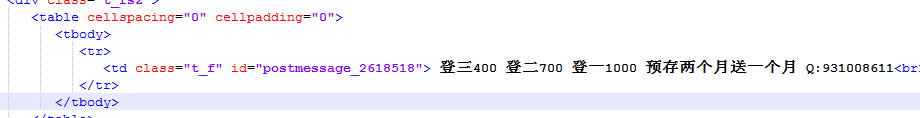


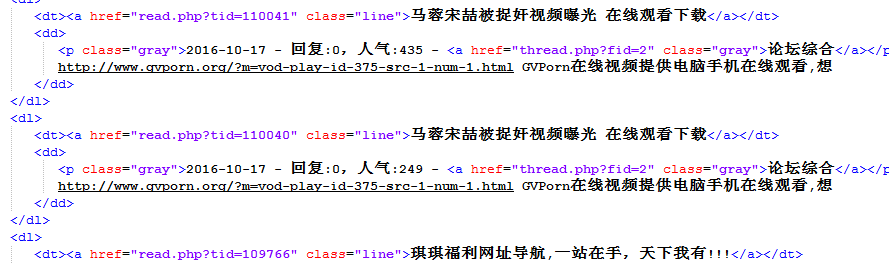


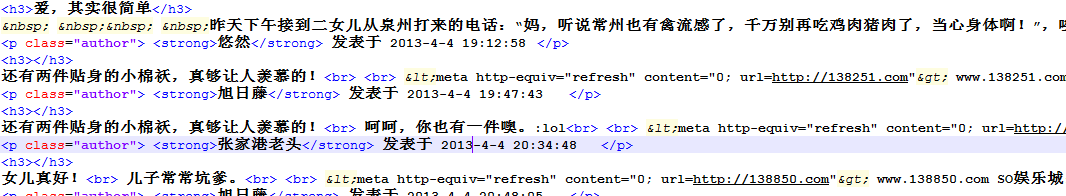
办假证，替考 word2vec

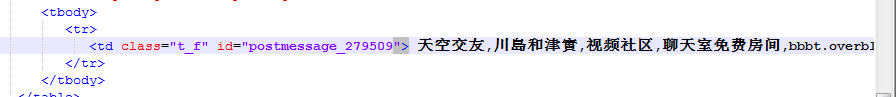
赌和投注

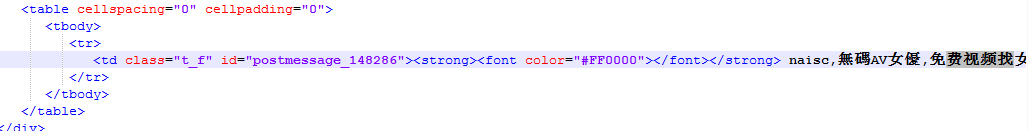






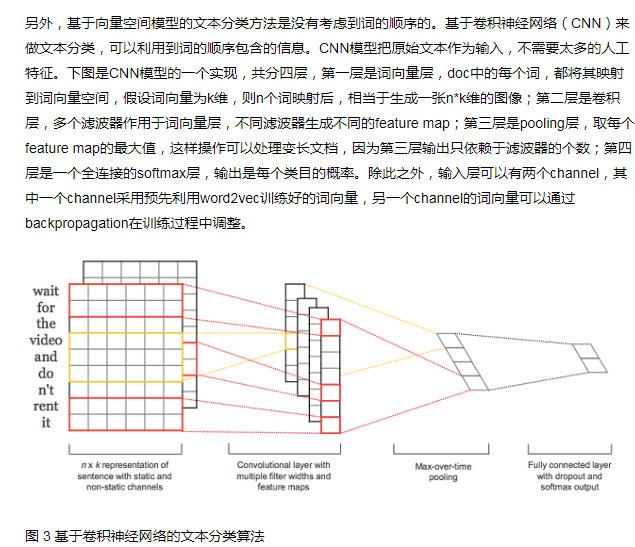


Sexy



如何计算两个文档的相似度二

文本经过分词程序分词后 ,首先去除停用词 ,合并数字和人名等词汇 ;然后统计词频



kmeans聚类

召回率 r = a / (a + c) \* 100%

正确率 p = a / (a + b) \* 100%

F-测度值 F = (2 \* p \* r) /(p + r)

由于在分类结果中，对应每个类别都会有一个召回率和正确率，因此，可以根据每个类别的分类结果评价分类器的整体性能，通常方法有两种：微平均和宏平均。微平均是根据正确率和召回率计算公式直接计算出总得正确率和召回率值。宏平均是指首先计算出每个类别的正确率和召回率，然后对正确率和召回率分别取平均得到总的正确率和召回率。不难看出，宏平均平等对待每一个类别，所以它的值主要受到稀有类别的影响，而微平均平等考虑文档集中的每一个文档，所以它的值受到常见类别的影响比较大。

word2vec向量

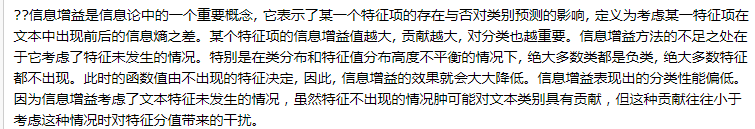
一起词频统计，剔除低频词，特征表示，归一化，svm分类

词频需要测试集与训练集一起

过滤掉一些词

基于向量空间模型的文本自动分类系统的研究与实现

基于归一化词频贝叶斯模型的文本分类方法

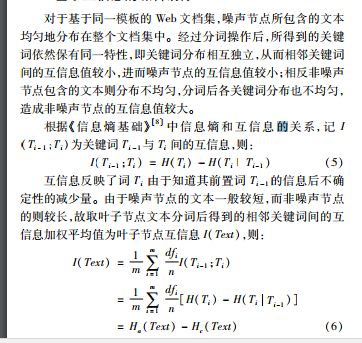


word2vec 聚类，pca降维，LDA,TF-idf

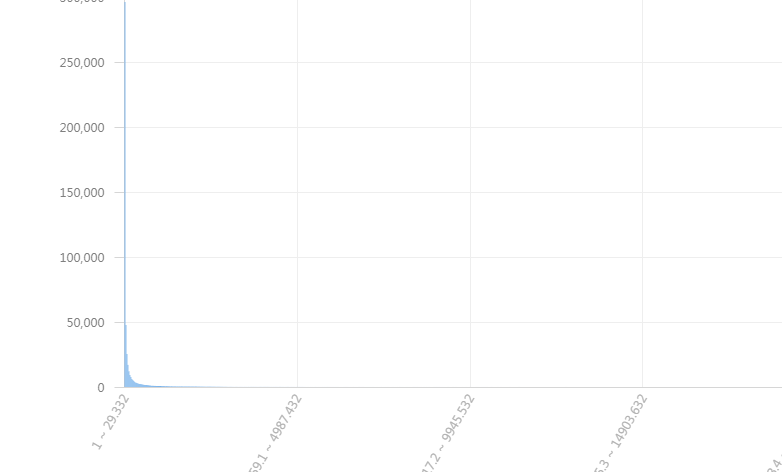
互信息中各词频统计，过滤掉小于500的词频，去除在多个类别同时出现的词，word2vec表示，svm分类

风格多代表这个文字越重要

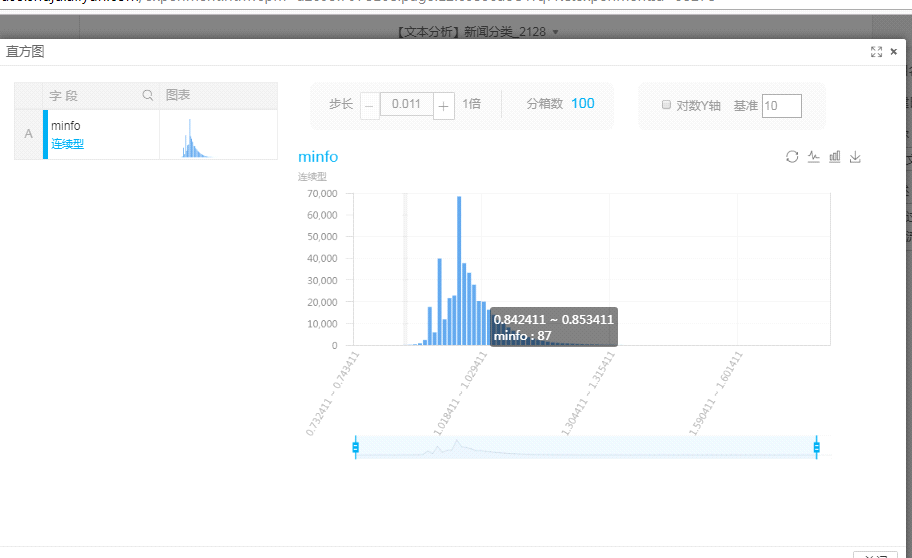
1.word2vec聚类





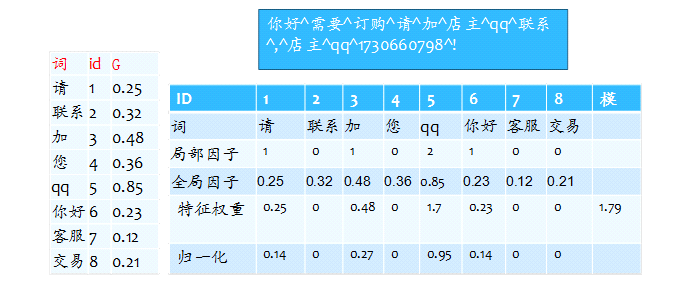


去除normal计算互信息

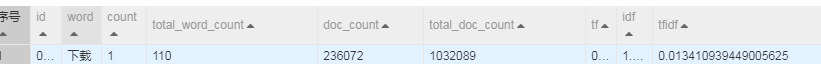


互信息对样本不均匀和低词频信息偏爱

互信息过滤阈值



Tf-idf



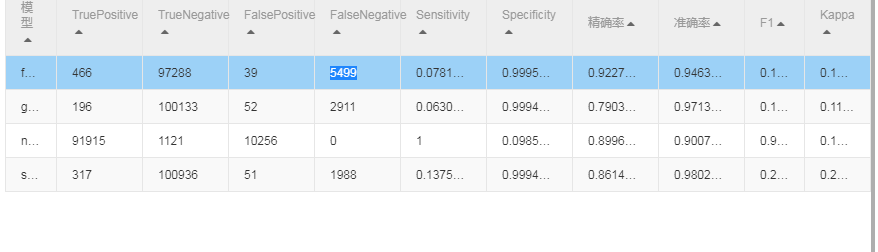
fake card

0.46-0.61 3906

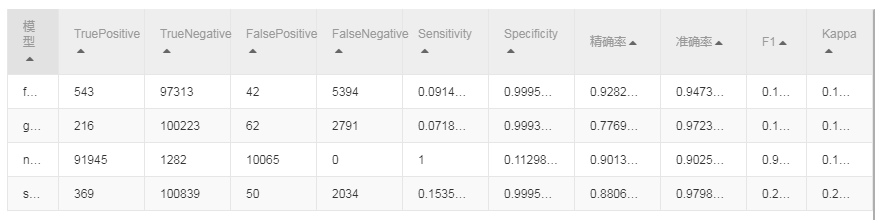
gmabling 0.8-1.2 4000左右

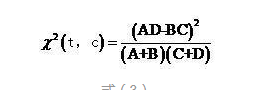
sexy 0.8-1.2 3000左右

全文分析，互信息提取特征 tf-idf



转为词频

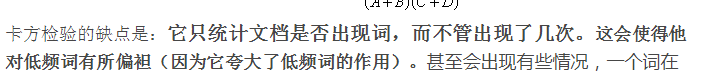


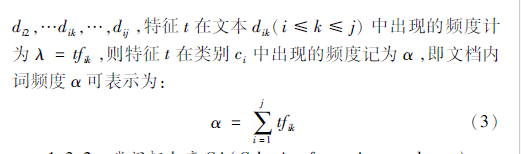


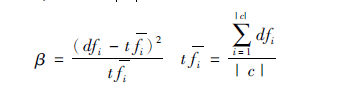
互信息简化公式



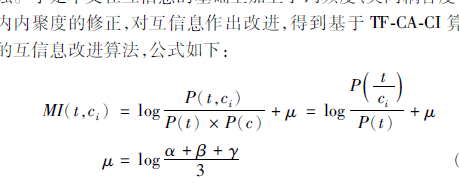


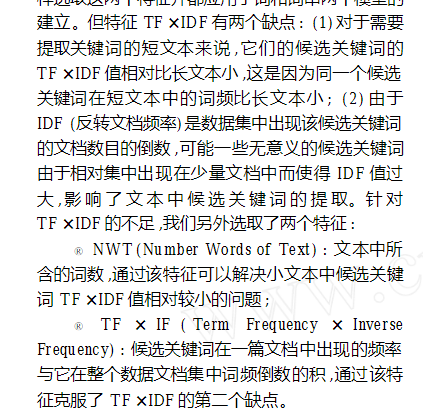












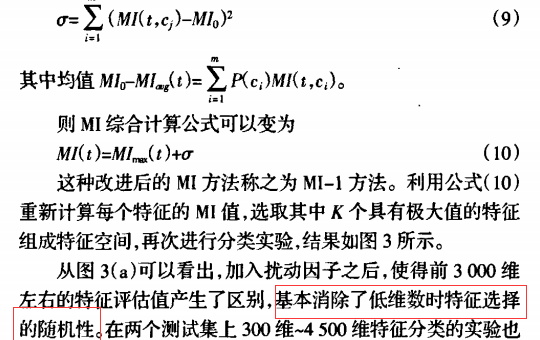
特征抽取的基本思想是利用映射的方法把原始特征项集映射到较低维的空

间中，映射后的特征叫二次特征，是原始特征的某种组合。特征

选择是利用某种评价函数独立地对每个原始特征项进行评估，

然后将它们按评估值地高低排序，从中选取若干个评估值最高

的特征项



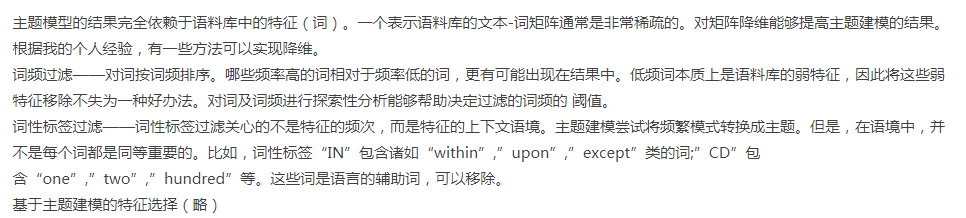
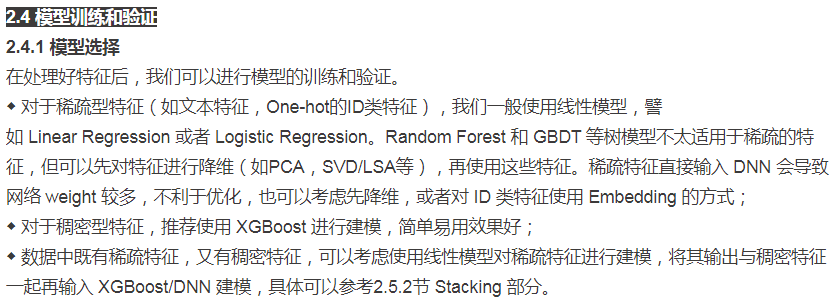
1测试集的tf-idf计算

2.计算互信息类与词的关联，测试集词汇未知

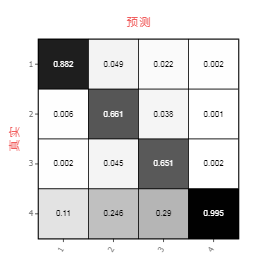
70000normal 4000检测

Lda:

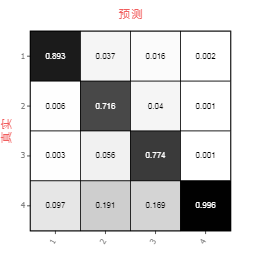
以一定的概率选取上述某个主题，再以一定的概率选取那个主题下的某个单词，



逻辑回归lda 6000

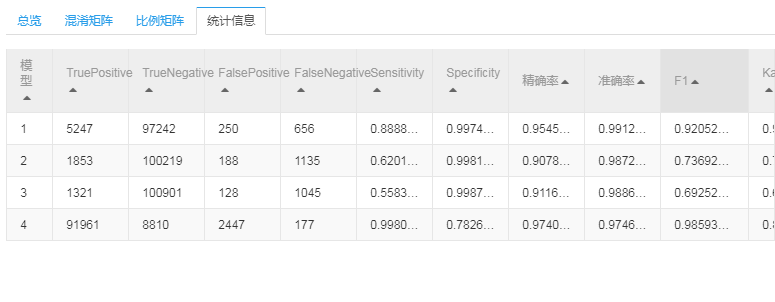


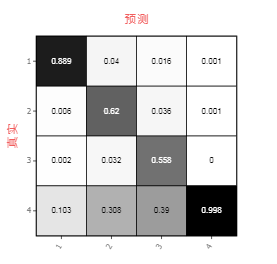
逻辑回归，tf-idf





Lda随机森林





1/全文去除tag 去除risk多个同id

2.获取全文，计算词频，分开训练与测试

3.训练数据》500词频过滤，

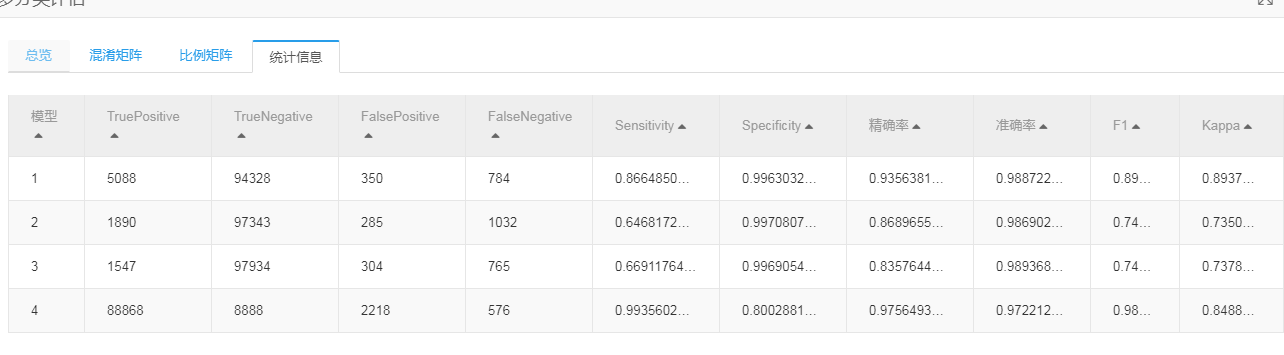
4，去除normal 计算互信息

5，取前1000

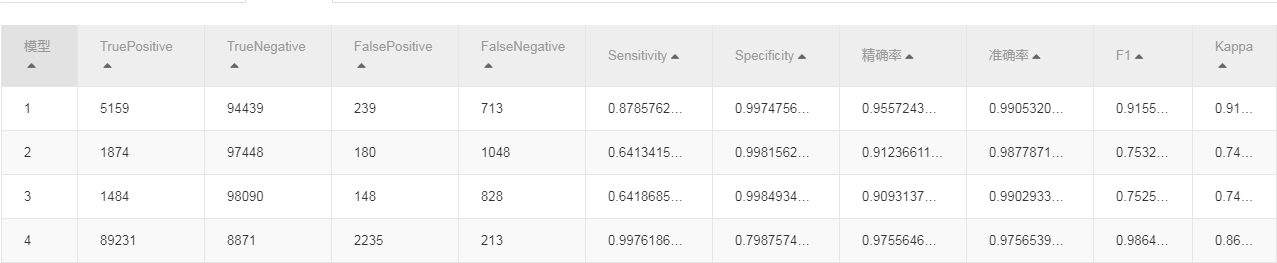
6. lda建模

7逻辑回归

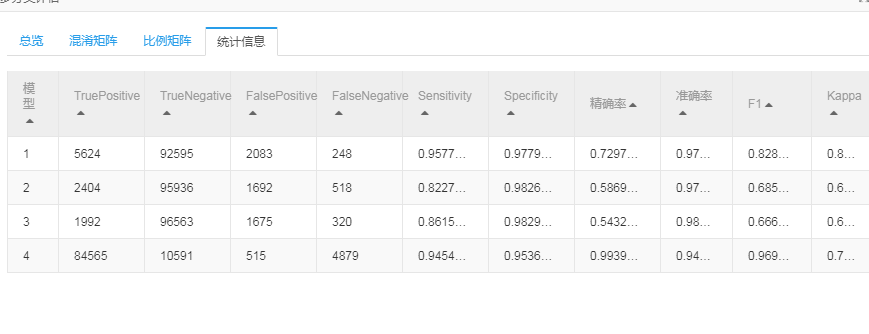
逻辑回归lda 3000



随机森林 lda 3000

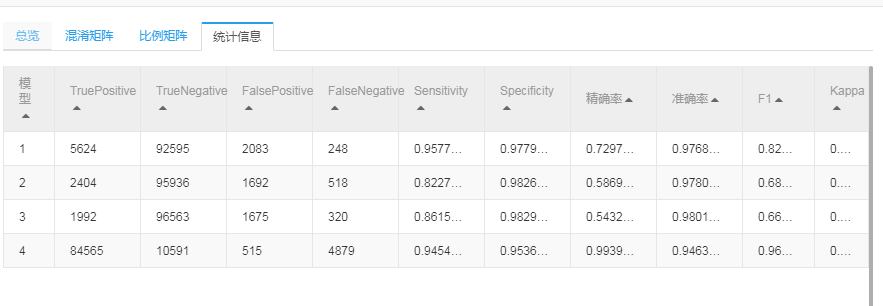


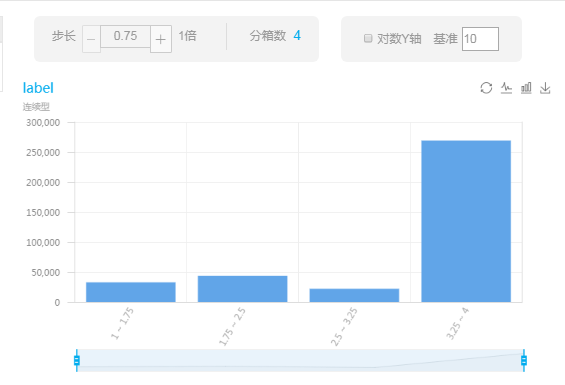
随机森林lda采样





逻辑回归





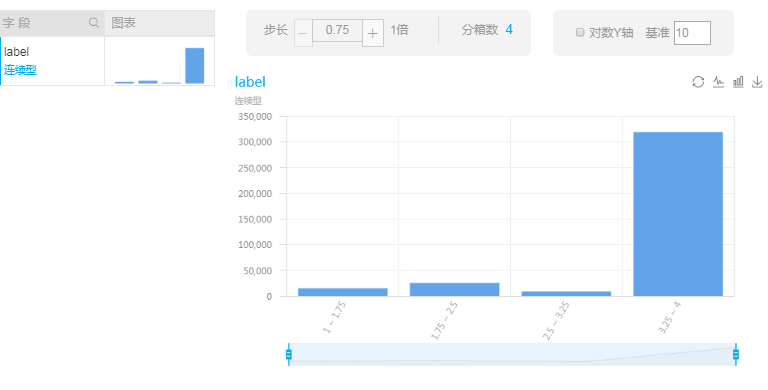
混合



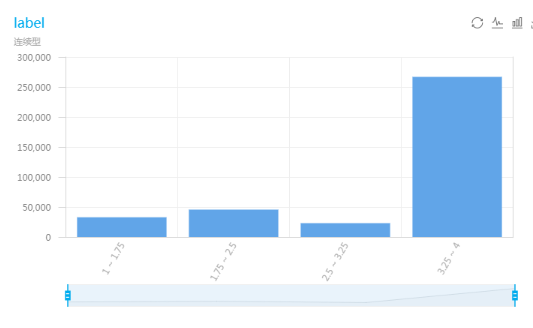
Lranswer 1000,lda,未采样

Lr answer1最后关键词1000，lda,采样

Lr2000 关键词2000，lda



C采样后



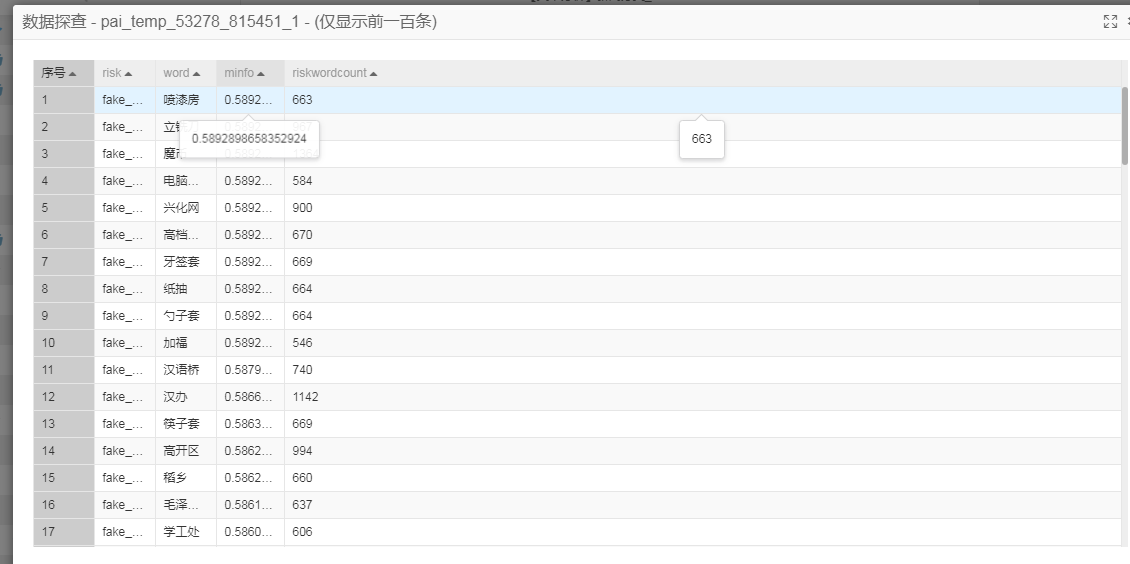
hxbks2ks0643

xxtxxthmhxb0643

TIANCHI\_4142

6C)tRZo2

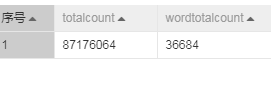
Keywordssn2



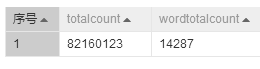


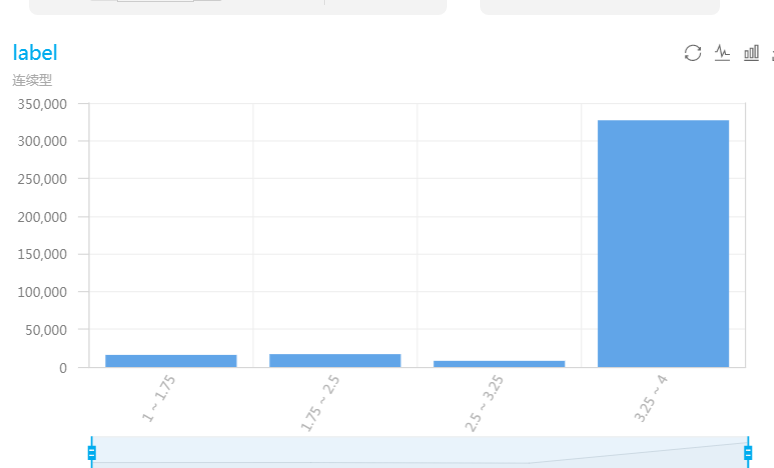


现在



原来





黄第一次提交

16038，17053，8175，327114

过滤100与过滤500，过滤100会有很多噪声数据

互信息加词频会带入噪声数据

Keywordsn2，过滤500，取2000计算互信息，不加词频

Keywordsn1，过滤500，取1000计算互信息，不加词频

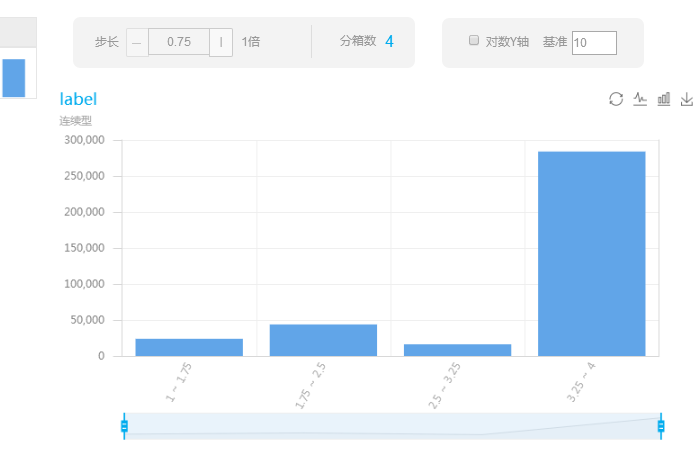
Keywordsf 过滤500，取1000计算互信息，加词频？

W\_keyword过滤500，互信息词频取2000

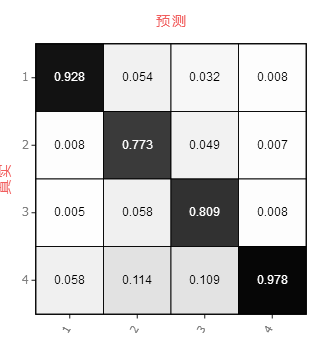
W\_keywords100，过滤100，互信息词频1000

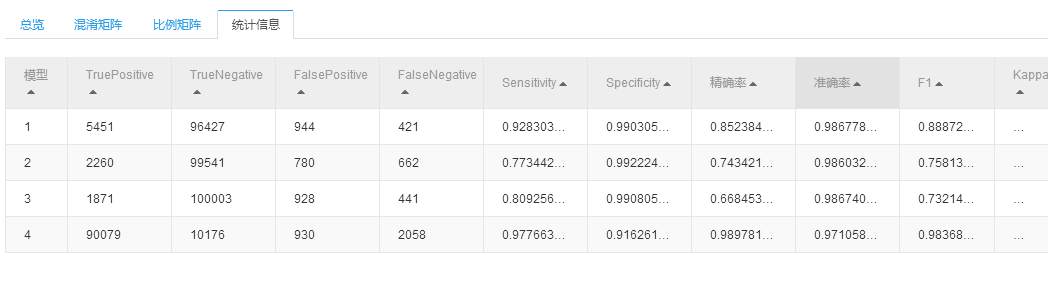
W\_keywords2000,过滤500，互信息2000

采样：l2正则

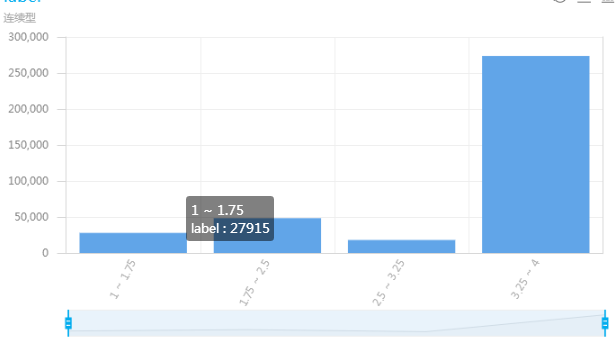


23940 43833 16256 284017

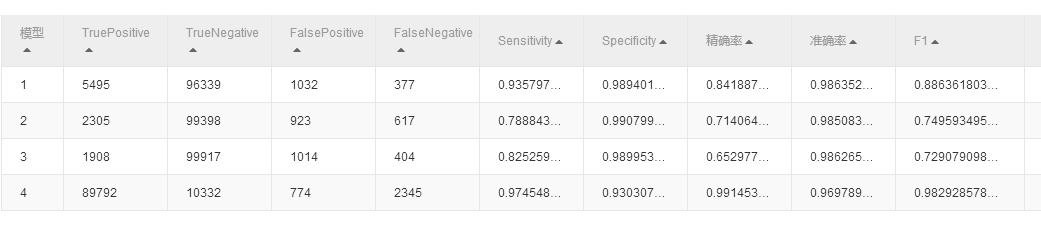




原来



27915 48272 18080 273678



L1正则

25078，48175，17928，276865

最好成绩 0.45

11326 20143 9077 总数328046

次佳成绩 0.42

15037 25365 8834

1500成绩 无重复 0.4149

14451 27384 8650

多模型融合0.4558

Pca后：

17041，27532，8629，314844

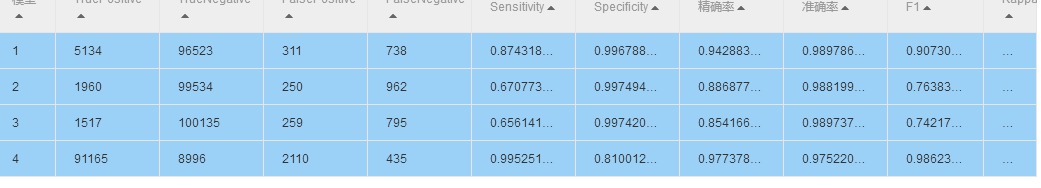
1500 有重复（无重复一样）

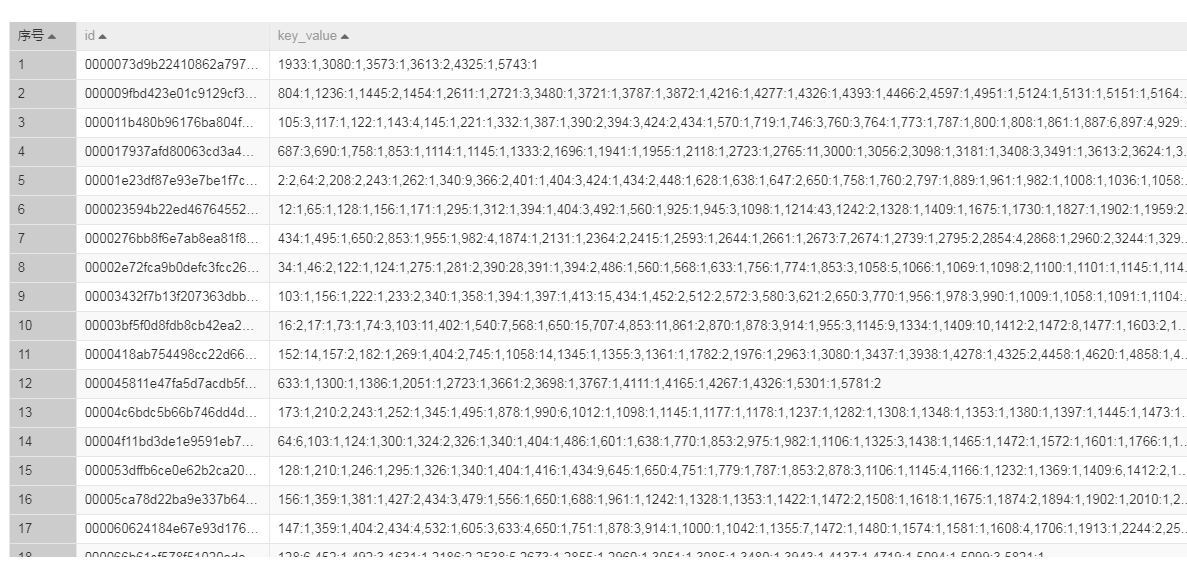
20787 21530 7734

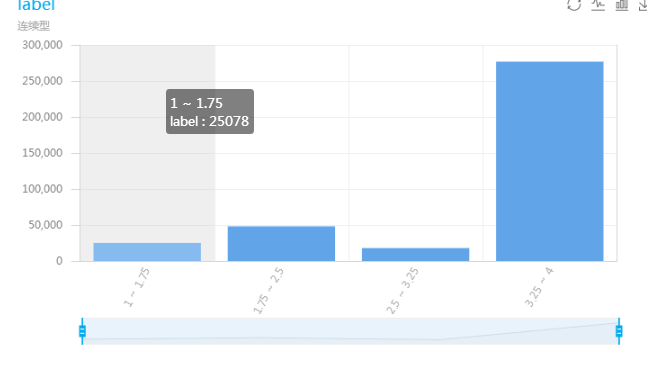
1500增加lda迭代次数 成绩0.3529

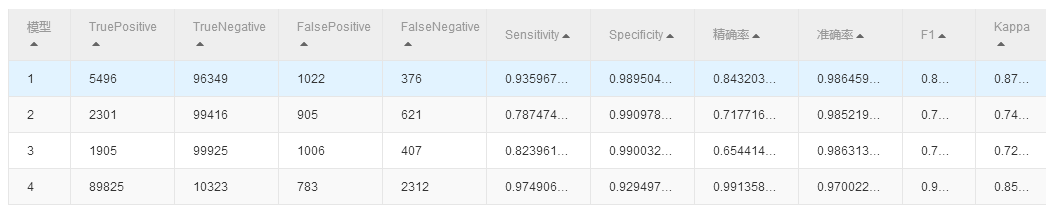
先二分类再多分类

BHA)ABNU5@%49[BN]4(BAH5









去除干扰样本

采样

Lda在wordcount增加时，成绩上升？

逻辑回归以优化总体的精度为目标，不同类别的误分类情况产生的误差是相同的，考虑一个$500:1$的数据集，即使把所有样本都预测为多数类其精度也能达到$500/501$之高，很显然这并不是一个很好的学习效果，因此传统的学习算法在不平衡数据集中具有较大的局限性。

随机过采样则正好相反，即通过多次有放回随机采样从少数类$S\_min$中抽取数据集$E$，采样的数量要大于原有少数类的数量，最终的训练集为$S\_maj+E$。

吉布斯采样：

无论metropolis-hasting算法还是gibbs算法，都需要一个burn in过程，只有在达到平衡状态时候得到的样本才能是平衡状态时候的目标分布的样本，因此，在burn in过程中产生的样本都需要被舍弃。如何判断一个过程是否达到了平衡状态还没有一个成熟的方法来解决，目前常见的方法是看是否状态已经平稳（例如画一个图，如果在较长的过程中，变化已经不大，说明很有可能已经平衡）当然这个方法并不能肯定一个状态是否平衡，你可以举出反例，但是却是实际中没有办法的办法。

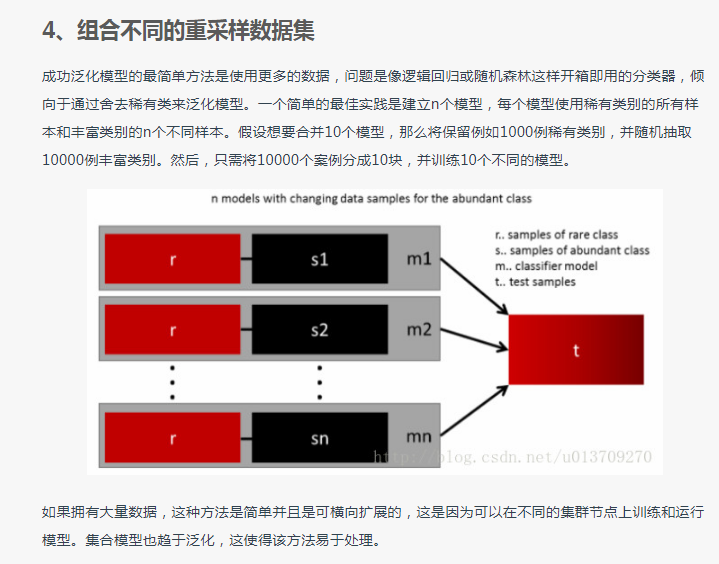
**burn-in阶段可能不需要**

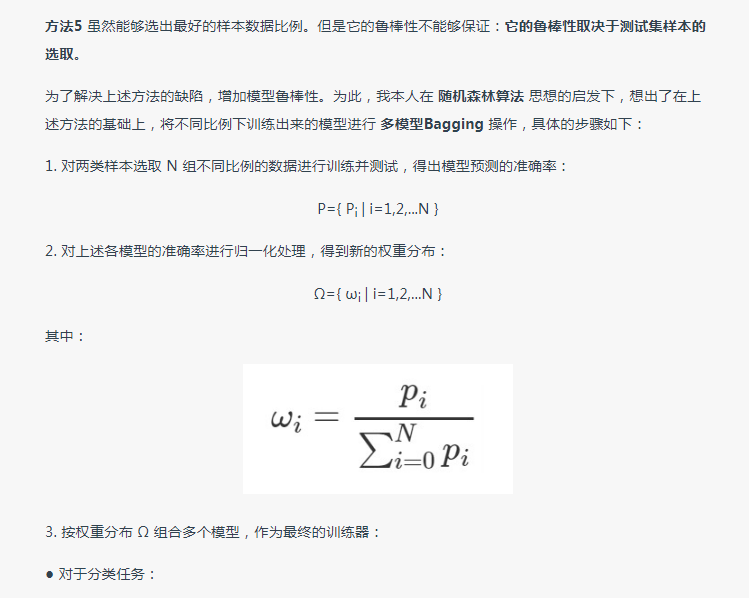
当然可以一条链跑到黑，但是一条链跑到黑只能是写学术 paper 的做法，在工程上可能还是要考虑很实际的速度和效率的问题，做 LDA 的时候我们就得考虑每秒钟能处理多少个请求,这时候不得不设置 burn-in。

So one rule I often follow is to start the next run where the last run ended.

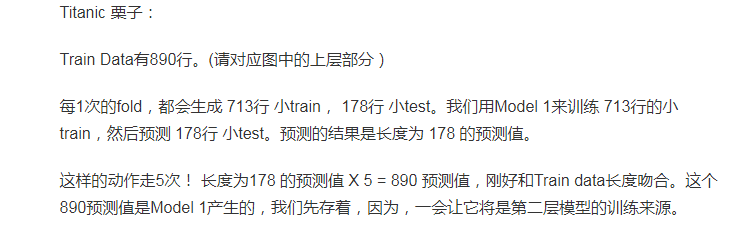
Another possible rule is to start at a point, like the mode, known to have reasonably high probability.

<https://www.leiphone.com/news/201706/dTRE5ow9qBVLkZSY.html>





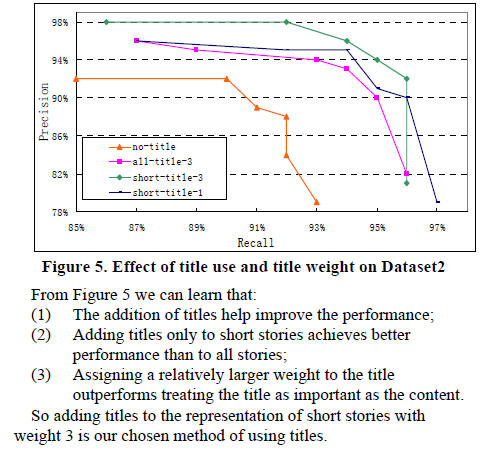
对于Bagging算法来说，由于我们会并行地训练很多不同的分类器的目的就是降低这个方差(variance) ,因为采用了相互独立的基分类器多了以后，h的值自然就会靠近.所以对于每个基分类器来说，目标就是如何降低这个偏差（bias),所以我们会采用深度很深甚至不剪枝的决策树。  
  对于Boosting来说，每一步我们都会在上一轮的基础上更加拟合原数据，所以可以保证偏差（bias）,所以对于每个基分类器来说，问题就在于如何选择variance更小的分类器，即更简单的分类器，所以我们选择了深度很浅的决策树。

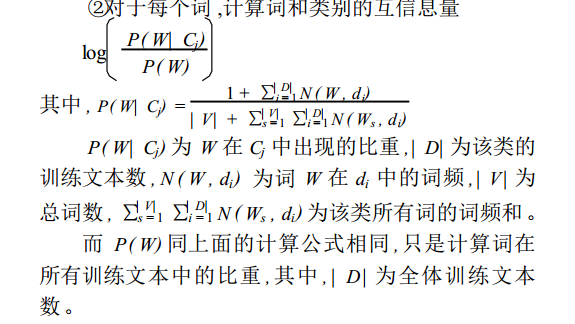


easy ensemble easy ensemble每次从多数类中抽样出和少数类数目差不多的样本，然后和少数类样本组合作为训练集。在这个训练集上学习一个adaboost分类器。 最后预测的时候，是使用之前学习到的所有adaboost中的弱分类器（就是每颗决策树）的预测结果向量（每个树给的结果组成一个向量）和对应的权重向量做内积，然后减去阈值，根据差的符号确定样本的类别。easy ensemble 算法被认为是非监督学习算法，因此它每次都独立利用可放回随机抽样机制来提取多数类样本

BalanceCascade 每次从多样本提取部分，与小样本训练，将正确的剔除，错误的加入剩余样本，继续做，直至稳定

adl\_tianchi\_content\_risk\_training\_part1 adl\_tianchi\_content\_risk\_training\_part2 adl\_tianchi\_content\_risk\_training\_part3 adl\_tianchi\_content\_risk\_training\_part4 adl\_tianchi\_content\_risk\_training\_part5 adl\_tianchi\_content\_risk\_training\_part6 adl\_tianchi\_content\_risk\_training\_part7 adl\_tianchi\_content\_risk\_training\_part8 adl\_tianchi\_content\_risk\_training\_part9





Ngram

Bi-Gram:　　P(T)=p(w1|begin)\*p(w2|w1)\*p(w3|w2)\*\*\*p(wn|wn-1)

Tri-Gram:　　P(T)=p(w1|begin1,begin2)\*p(w2|w1,begin1)\*p(w3|w2w1)\*\*p(wn|wn-1,wn-2)

* 注意上面概率的计算方法：P(w1|begin)=以w1为开头的所有句子/句子总数；p(w2|w1)=w1,w2同时出现的次数/w1出现的次数。以此类推。

一个词在一个网页中频繁出现，就定为属性Automatic Web-Page Classification

by Using Machine Learning Methods

**模型融合使用熟悉**

|  |
| --- |
| Good-quality document summarization can accurately represent the major topic of |
| a Web page. Shen et al. [2004] proposed an approach to classifying Web pages through |
| summarization |

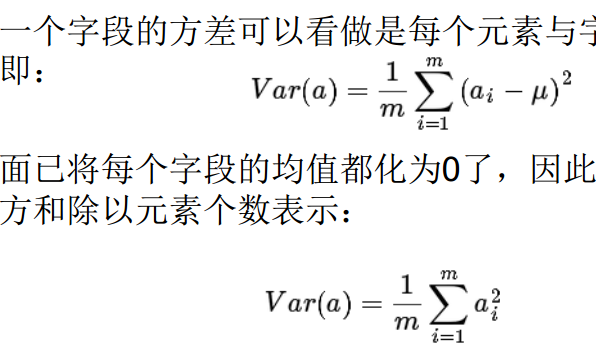
un et al. [2002] showed  
that SVM classifiers using the text on the target page, page title (as separate features),  
and anchor text from parent pages can improve classification compared with a pure text  
classifier

|  |
| --- |
| . Yan et al. [2005] proposed a novel feature selection |
| approach which is more efficient and effective than information gain and *χ* 2 test on |
| large-scale datasets. |

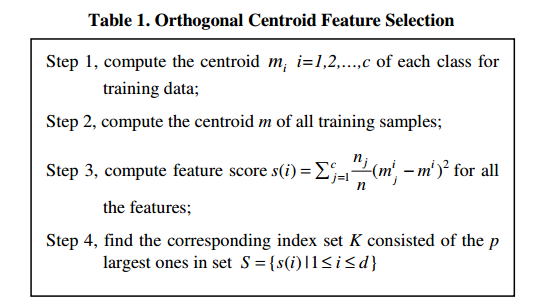
|  |
| --- |
| Search engine spam [Gyongyi and Garcia-Molina 2005a] is a significant concern in ¨ |
| Web information retrieva |

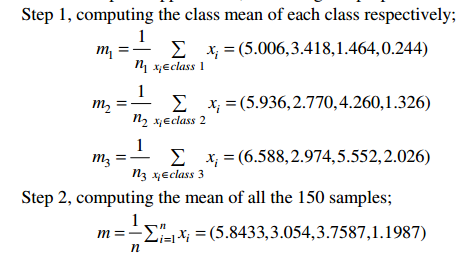
**pca需要数据需要均值为0**

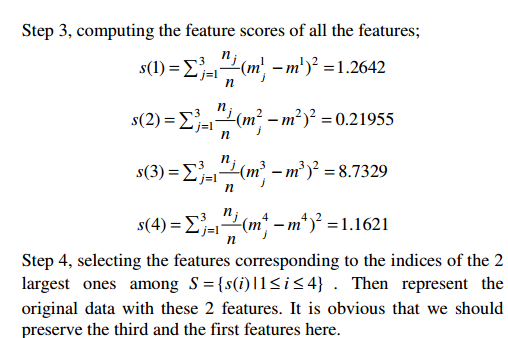
**pca 投影到空间方差最大化，需要归一化方差**

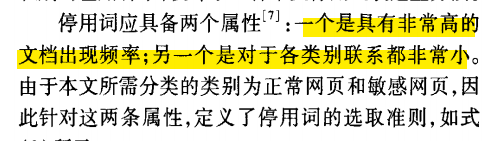


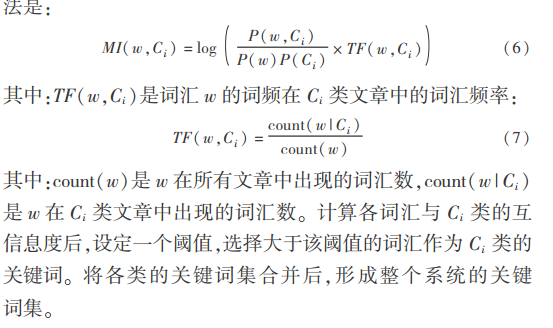
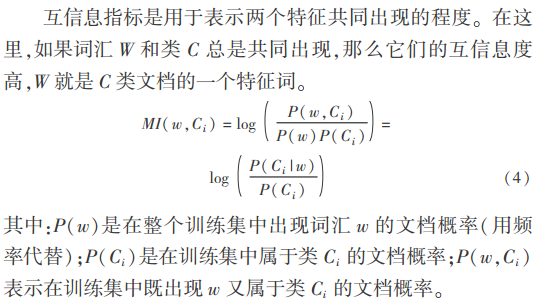
**Svd并不需要方阵，分解为三个矩阵，有各自的意思，pca需要方阵，分解为三个矩阵，特征向量**

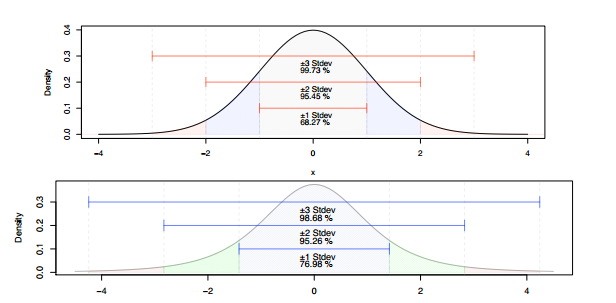












**Unsupervised Detection of Anomalous Text**

**Chebyshev’s inequality theorem (sometime called the Bienaym´e-Chebyshev theorem), which states that, for any finite distribution, a minimum of 1 − k 1 2 of the data will lie within k standard**

**deviations of the mean [Grimmett, 2001; Papoulis, 2002]. So, for any distribution, at least 93.75% of the observations will lie within 4 standard deviations from the mean and 96% lie within 5 standard deviations.**

Mahalanobis distance

马氏距离多元变量，x^2分布

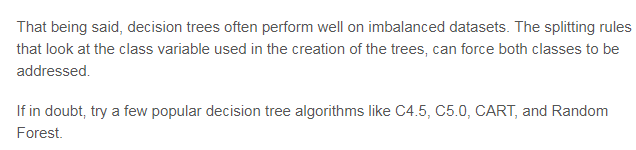
将文章分为不同段，给每一段转化为

**平衡样本**

**不同的分类算法**

**添加号码**

**采样**

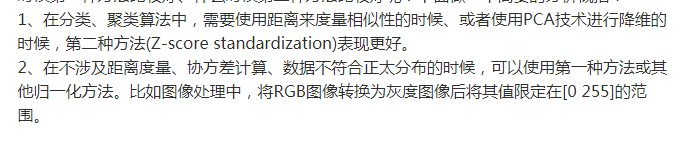


**Kappa，预测与真实结果是否一致**

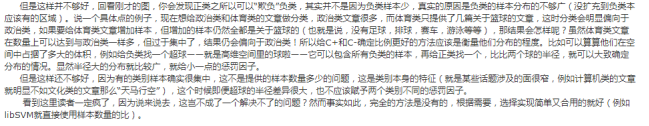
**Lda正规化了没**

**词频特征，lda**

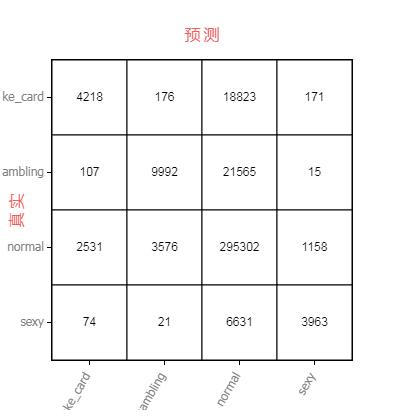
 Divide the more abundant class into L distinct clusters. Then train L predictors, where each predictor is trained on only one of the distinct clusters, but on all of the data from the rare class. to be clear, the data from the rare class is used in the training of all L predictors. Finally, use model averaging for the L learned predictors as your final predictor.



**各种算法**

****

**张**

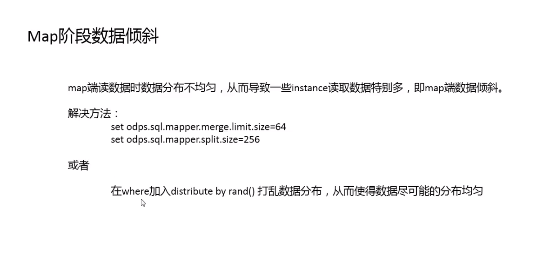


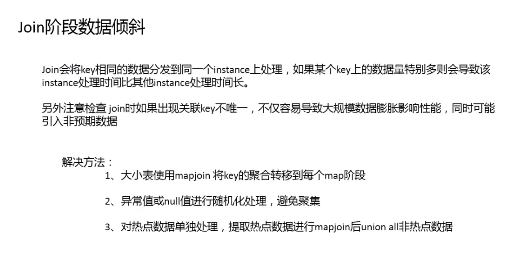
**我**

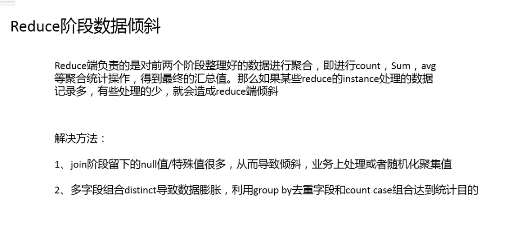
****

**样本不平衡问题不明显**

**预测为normal的不变，每个训练集预测为normal的**





1 prediction\_result lr 1000 [1, 1.75):12214;[1.75, 2.5):22370;[2.5, 3.25):13263;[3.25, 4]:432078

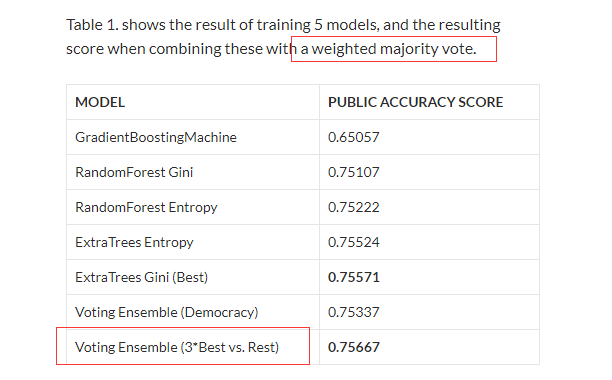
原 label [1, 1.75):12211;[1.75, 2.5):22366;[2.5, 3.25):13257;[3.25, 4]:432091

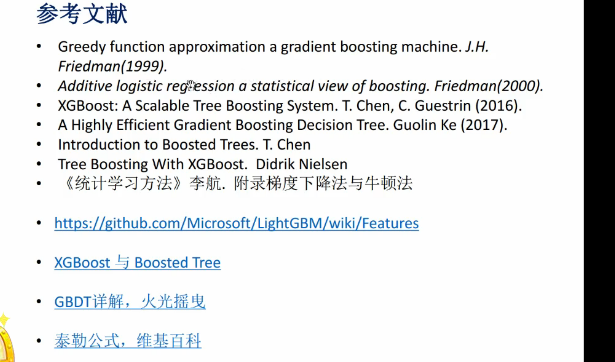
随机森林

1 label [1, 1.75):5829;[1.75, 2.5):12617;[2.5, 3.25):8796;[3.25, 4]:452686

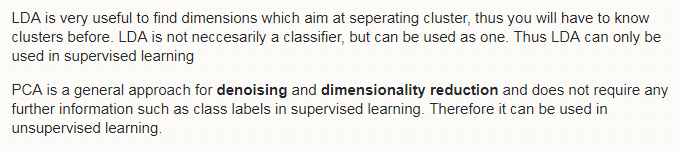
Xgboost:6708 13505 5187

逻辑回归1 label [1, 1.75):7469;[1.75, 2.5):10982;[2.5, 3.25):10642;[3.25, 4]:450835

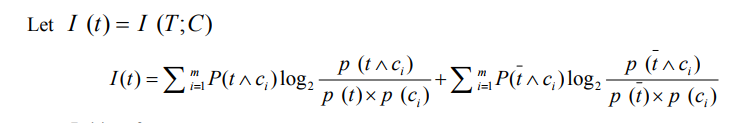




**Kmeans**



所有类一起求这个词的互信息，限制阈值



我所用的是端互信息，

