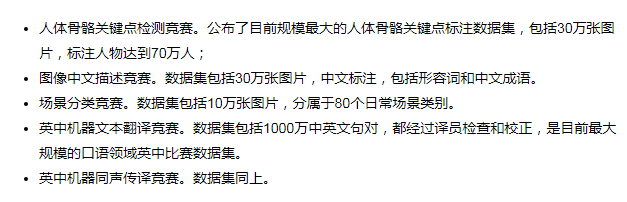
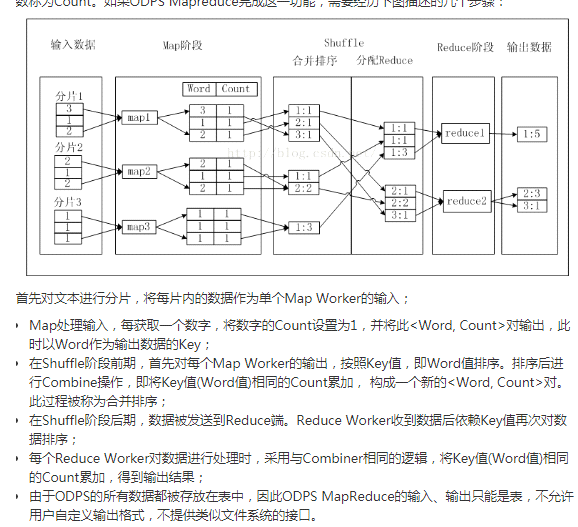
TIANCHI\_4362@aliyun-tianchi

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”**

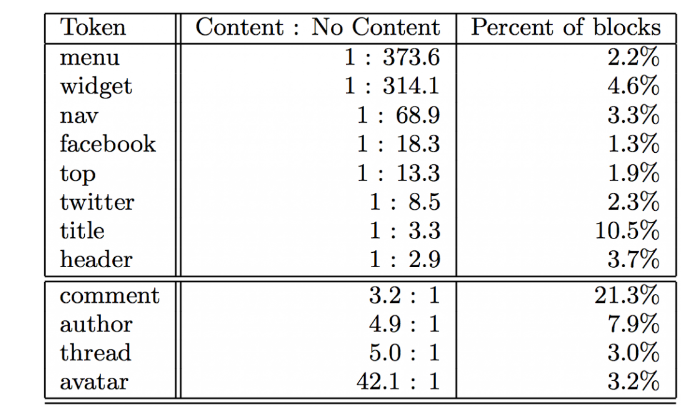
<https://yq.aliyun.com/articles/68227>问题解答

沙箱

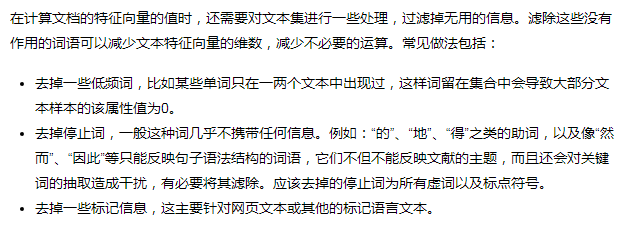
使用自己的库产生故障

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

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

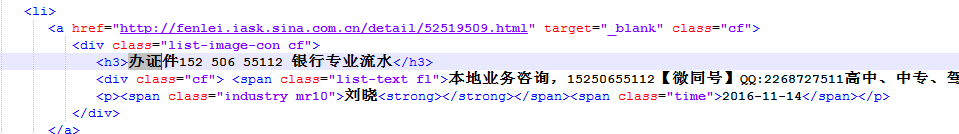
Fake card 59219 0.0538

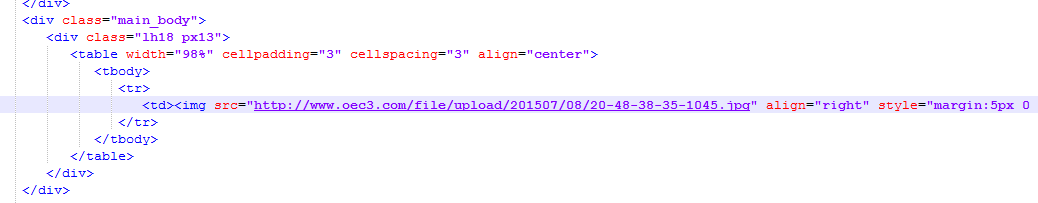
Sexy 23833 0.021

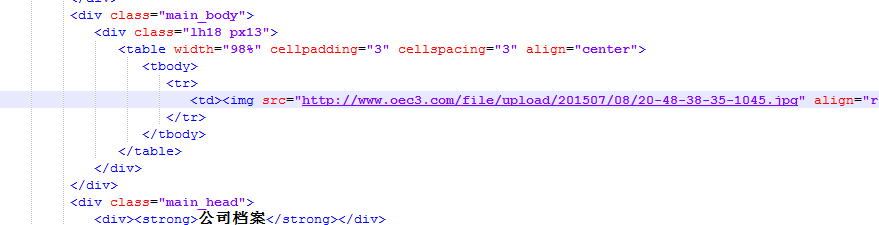
Normal 986623 0.89

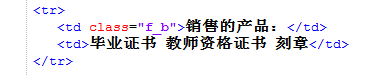
Total : 1099999

Fake card:





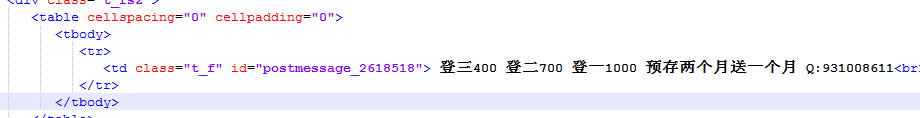


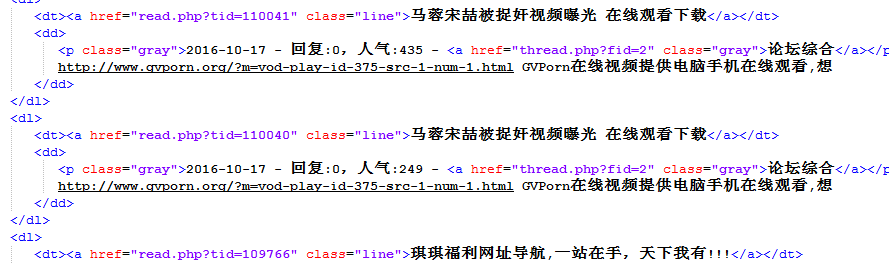


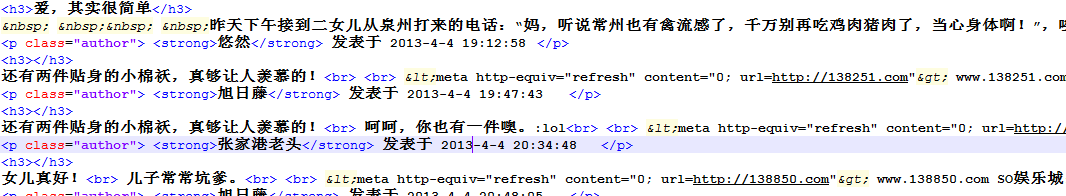
办假证，替考 word2vec

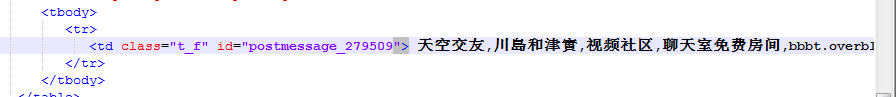
赌和投注

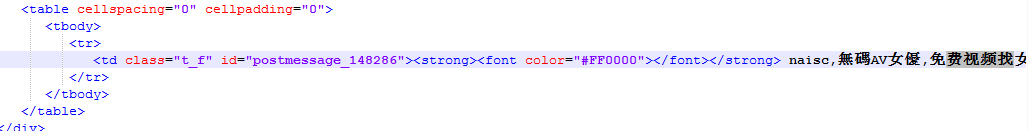






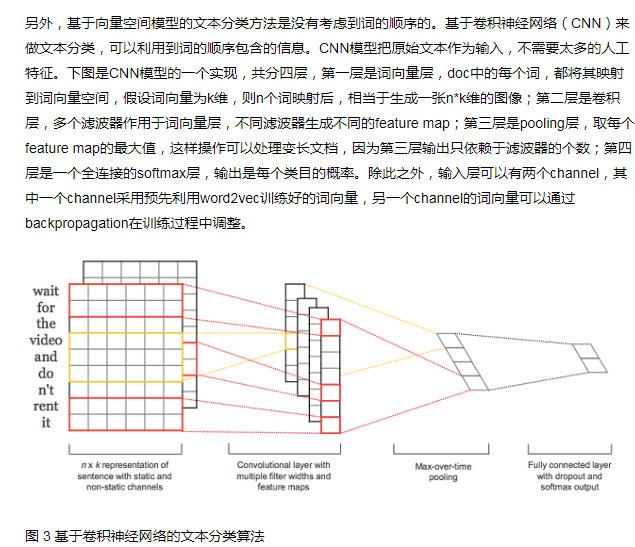


Sexy



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

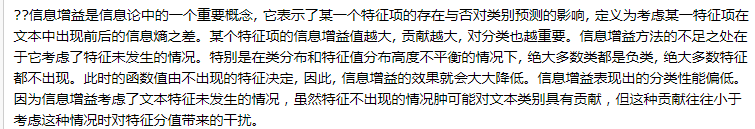
一起词频统计，剔除低频词，特征表示，归一化，svm分类

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

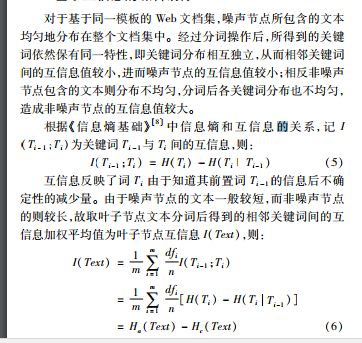


word2vec 聚类，pca降维，LDA,TF-idf

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

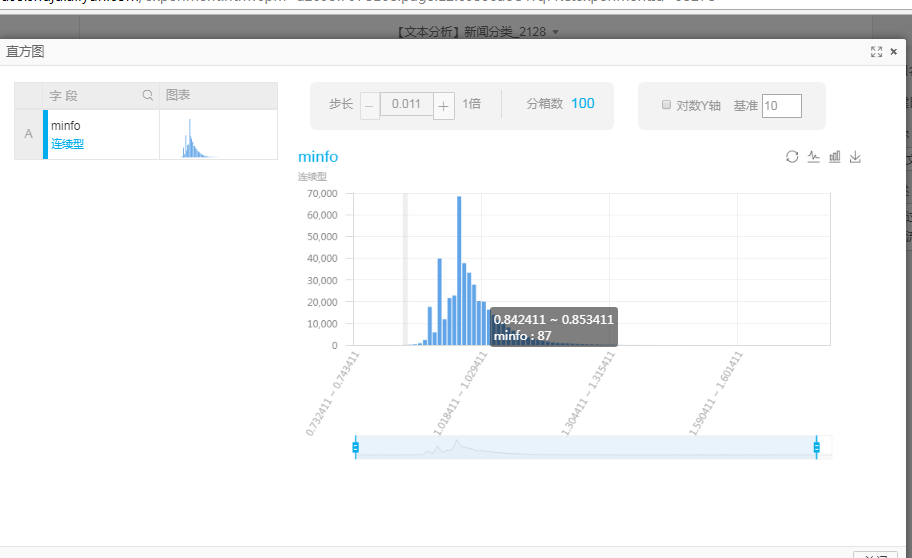
1.word2vec聚类







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

