文本语料

去标点，纯数字等非文本信息

去停用词，加字典

Jieba分词

tfidf

Word2vec

doc2vec

文本向量

贝叶斯

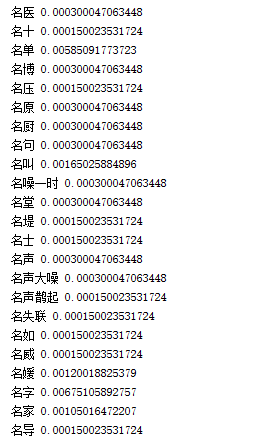
随机森林

scikit-learn工具包

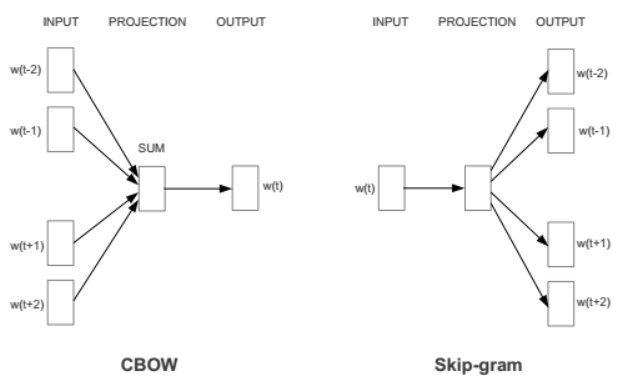
训练

**TF-IDF**模型是:在一个文档中出现次数很多的词相比出现次数少的词在词向量表示中应该得到更高的权值。而IDF归一化起到了减弱在所有文档中总是出现的词的作用。最后的结果就是,稀有的或者重要的词被给予了更高的权值,而更加常用的单词(被认为比较不重要)则在考虑权重的时候有较小的影响。

用来提取关键词。****

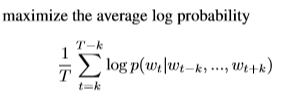


**Word2Vec**模型是把每一个单词表示成一个向量。这些模型一般是基于某种文本中与单词共现相关的统计量来构造。其利用深度学习的思想，可以通过训练，把对文本内容的处理简化为 K 维向量空间中的向量运算，而向量空间上的相似度可以用来表示文本语义上的相似度。在获得词向量后，对词向量进行平均处理，最终获取到句子向量。

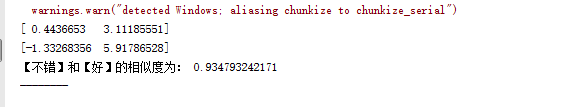


Skip-Gram Model是给定目标词预测上下文的概率值。

Continuous Bag of Words模型(CBOW)是给定上下文预测目标词的概率分布。

在训练前，每一个词都会首先初始化为一个N维的向量，训练过程中，会对输入的向量进行反馈更新，在进行大量语料训练之后，便可得到每一个词相应的训练向量。

1. 求两个词的相似度=numpy.dot(两个词的词向量)
2. 获得词向量，维度可自己设置

**Doc2Vec** 的目的是获得文档的一个固定长度的向量表达。

**朴素贝叶斯**

是标签，目标是找到的最大值，朴素贝叶斯会把值最大的类标签认为是X的类别。

**随机森林** 多棵决策树

**决策树** 计算每个特征的信息量，按照特征的分类能力大小进行排序，从最具有分类能力的特征进行分裂，希望每个子树的叶子结点基本属于同一类。

将生成的多棵树组成随机森林，用随机森林对新的数据进行分类，分类结果按树分类器投票多少而定。

**SVM**