TF-IDF原理。

TF-IDF(Term Frequency-Inverse Document Frequency, 词频-逆文件频率).

是一种用于资讯检索与资讯探勘的常用加权技术。TF-IDF是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。

总结：就是一个word在一篇文章中出现的频率越高，在语料库中的频率越低，越能代表这篇文章，通俗的理解就是当某些词只在特定文章中出现那么他就能有很大的权重可以代表这篇文章。

文本矩阵化，使用词袋模型，以TF-IDF特征值为权重

'''

fit\_transform()就是将fit()和transform()结合了一下

fit\_transform()就是先将数据进行拟合，然后将其标准化

transform()作用是通过找中心和缩放等实现标准化

可以说是：

transform()是一定可以替换成fit\_transform()，但是fit\_transform()不能替换为transform()

'''

from sklearn.feature\_extraction.text import CountVectorizer

vectorizer=CountVectorizer() #构建一个计算词频的对象

text=["I come to China to travel",

"This is a car polupar in China",

"I love tea and Apple ",

"The work is to write some papers in science"]

print(vectorizer.fit\_transform(text)) #vectorizer.fit\_transform(corpus)将文本输入，得到词频矩阵

# # 各个特征代表的词

print(vectorizer.get\_feature\_names())

# # 每个文本的词向量特征

print(vectorizer.fit\_transform(text).toarray())

互信息的原理。

在概率论和信息论中，两个随机变量的互信息（Mutual Information，简称MI）或转移信息（transinformation）是变量间相互依赖性的量度。不同于相关系数，互信息并不局限于实值随机变量，它更加一般且决定着联合分布 p(X,Y) 和分解的边缘分布的乘积 p(X)p(Y) 的相似程度。互信息(Mutual Information)是度量两个事件集合之间的相关性(mutual dependence)。互信息是点间互信息（PMI）的期望值。互信息最常用的单位是bit。