full music**数据预处理**：(IPYNB文件里有一些特征数值分布的可视化)  
tempo keypo   pularity year直接minmax标准化  
loudness将小于-40的值改为-40、duration\_ms经过z-score后将16个大于15的值改为15后，对这两个特征进行minmax标准化

分类降维：

先对所有原始数据做了个热力图，然后根据音乐特征、声音特征、音乐描述分三类，分别进行pca降维

音乐特征取前三个主成分：信息比分别占0.41160853, 0.2152998 , 0.19853013，和为0.8254384567075188

声音特征取前三个主成分：信息比分别占0.48617794, 0.24233794, 0.12938711，和为0.8579029955463436

描述只用第一个主成分即可，信息占比为0.84965615

由此就取得了每支歌曲对应的7个新特征，即每支歌曲都对应上了一个7维的向量

最后用余弦相似度衡量歌曲之间的相似度

欧氏距离能够体现个体数值特征的绝对差异，所以更多的用于需要从维度的数值大小中体现差异的分析，如使用用户行为指标分析用户价值的相似度或差异。

余弦距离更多的是从方向上区分差异，而对绝对的数值不敏感，更多的用于使用用户对内容评分来区分兴趣的相似度和差异，同时修正了用户间可能存在的度量标准不统一的问题（因为余弦距离对绝对数值不敏感）。  
<https://blog.csdn.net/huangfei711/article/details/78469614>

这篇博客提到了余弦相似度和欧氏距离的差异，由此可以论证为什么我们衡量音乐的相似度要用音乐向量的余弦相似度而非欧氏距离：**因为同一类风格的音乐可能会因为其风格的激进或柔和导致其对应的“音乐向量”存在数值上的差异，但是这种情况能够被更注重于分辨不同“音乐向量”的方向差异而非数值差异的余弦相似度规避，或者说，每个“音乐向量”的方向，就是这首曲子的风格**