结合sklearn做情感分析（分类算法的应用）

1. 选定类标签，好或坏
2. 选择特征，一般以词为单位。也可以选择双词或三词
3. 特征过多，需要做特征的降维操作。可以通过**统计方法**找到信息量丰富的特征。统计方法包括：词频、文档频率、互信息、信息熵、卡方统计等等。选择特征的数量视情况而定。
4. 把语料文本变成使用特征表示。

在使用分类算法进行分类之前，第一步是要把**所有**原始的语料文本转化为特征表示的形式。如果在NLTK 中，如果选择**所有**词作为特征，其形式是这样的：[ {“手机”: True, “非常”: True, “好用”: True, “！”: True} , **positive**] 注意表达的结构！

如果选择信息量丰富的词作为特征，其形式是这样的：[ {“好用”: True} , **positive** ]

1. 将数据集随机分为训练集7/10、开发测试集1/10、测试集2/10。比例可调整（第7步用）
2. 用不同的分类算法给训练集构建分类器，用开发测试集检验分类器的准确度（选出最佳算法后可以调整特征的数量来测试准确度）
3. 通过NLKT调用sklearn算法

其实方法很容易。只要以下五步。

1. 仅仅使用开发集（Development Set）。

2. 用分类算法训练里面的训练集（Training Set），得出分类器。

3. 用分类器给开发测试集分类（Dev-Test Set），得出分类结果。

4. 对比分类器给出的分类结果和人工标注的正确结果，给出分类器的准确度。

5. 使用另一个分类算法，重复以上三步。

在检验完所有算法的分类准确度之后，就可以选出最好的一个分类算法了。

在选出最好的分类算法之后，就可以测试不同的特征维度对分类准确度的影响了。一般来说，特征太少则不足以反映分类的所有特点，使得分类准确率低；特征太多则会引入噪音，干扰分类，也会降低分类准确度。所以，需要不断的测试特征的数量，这样才可以得到最佳的分类效果。

1. 选择出开发集中最佳的分类算法和特征维度，使用测试集检验得出情感分类的准确度。

在终于得到最佳分类算法和特征维度（数量）之后，就可以动用测试集。

直接用最优的分类算法对测试集进行分类，得出分类结果。对比分类器的分类结果和人工标注的正确结果，给出分类器的最终准确度

def score(classifier):  
   classifier = SklearnClassifier(classifier) #在nltk 中使用scikit-learn 的接口  
   classifier.train(train) #训练分类器

   pred = classifier.batch\_classify(testSet) #对开发测试集的数据进行分类，给出预测的标签

   return accuracy\_score(tag\_test, pred) #对比分类预测结果和人工标注的正确结果，给出分类器准确度