

#由于内存问题，对比时选用四类数据

['alt.atheism', 'talk.religion.misc','comp.graphics', 'sci.space']

#控制变量，使用相同的特征提取器【vect=TfidfVectorizer(stop\_words='english')】

#使用两种分类器【mnb=MultinomialNB(alpha=0.1)，gnb=GaussianNB()】

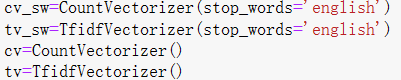
#预测且评价【测试集，占20%】



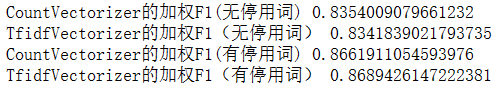
#实验1结论：性能上，MultinomialNB表现更优，且接收稀疏矩阵传入，大大减少了时空复杂度，因此该选用MultinomialNB。对于后续实验，选定分类器为MultinomialNB，此时由于其接收稀疏矩阵的特性，可以选取所有的数据集，而不用固定4类



#控制变量，使用相同的分类器【mnb=MultinomialNB()】

#使用两种特征提取器

#预测且评价【新闻数据全部类别，测试集，占20%】



#结论：【回答问题二】一方面，在同有停用词或同无停用词的条件下，CountVectorizer与TfidfVectorizer的性能差异不大。【回答问题 三】另一方面，无论CountVectorizer或TfidfVectorizer都被停用词的引入显著提高了预测效果





#结论：有平滑小小的提升