实验报告 3——聚类

实验要求：

测试 sklearn 中以下聚类算法（Kmeans、Affinity propagation、Mean-shift、Spectral clustering、Ward hierarchical clustering、Agglomerative clustering、DBSCAN、Gaussian mixtures）在 tweets 数据集上的聚类效果

评价指标为 NMI（Normalized Mutual Information）

实验步骤为：

1、处理原始数据：按行读取 Tweets.txt 并转换为字典置于列表中，对该列表进行处理，将 text 标签内容置于 test_data 列表，将 cluster 标签内容置于 labels_true 列表中，使用 TfidfVectorizer 将 test_data 列表中字符串向量化置于 X

2、利用 sklearn 中自带的 Kmeans、Affinity propagation、Mean-shift、Spectral clustering、Ward hierarchical clustering、Agglomerative clustering、DBSCAN、Gaussian mixtures 算法函数对 X 聚类，得到 labels

3、使用 sklearn 自带的计算 NMI 的函数，计算 labels_true 和 labels 的 NMI，进行对比

实验结果如下：

聚类算法	NMI
Kmeans	0.6493775338318443
Affinity propagation	0.4839682646264274
Mean-shift	0.48568072835637743
Spectral clustering	0.539574609793204
Ward hierarchical clustering	0.6644202079888861
Agglomerative clustering	0.6644202079888861
DBSCAN	0.4248045834640204
Gaussian mixtures	0.7220048834646094