实验报告

学号：201814849　　姓名：张延超　　班级：2018 级学硕

1、 实验任务

（1） 测试 sklearn 中以下聚类算法在 Tweet 数据集上的聚类效果
（2） 使用 NMI（Normalized Mutual Information）作为评价指标

2、 实验步骤

（1）K-Means

class sklearn.cluster.**KMeans**(***n_clusters***=*8*, ***init***=*'k-means++'*, ***n_init***=*10*, ***max_iter***=*300*, *tol=0.0001*, *precompute_distances='auto'*, *verbose=0*, ***random_state***=*None*, *copy_x=True*, *n_jobs=None*, *algorithm='auto'*)

（2）Affinity Propagation

class sklearn.cluster.**AffinityPropagation**(***damping***=*0.5*, ***max_iter***=*200*, *convergence_iter=15*, *copy=True*, *preference=None*, *affinity='euclidean'*, *verbose=False*)

（3）Mean-Shift

sklearn.cluster.**mean_shift**(***X***, ***bandwidth***=*None*, *seeds=None*, *bin_seeding=False*, *min_bin_freq=1*, *cluster_all=True*, *max_iter=300*, *n_jobs=None*)

（4）Spectral Clustering

class sklearn.cluster.**SpectralClustering**(***n_clusters***=*8*, *eigen_solver=None*, ***random_state***=*None*, ***n_init***=*10*, *gamma=1.0*, *affinity='rbf'*, *n_neighbors=10*, *eigen_tol=0.0*, *assign_labels='kmeans'*, *degree=3*, *coef0=1*, *kernel_params=None*, *n_jobs=None*)

（5）Agglomerative Clustering

class sklearn.cluster.**AgglomerativeClustering**(***n_clusters***=*2*, *affinity='euclidean'*, *memory=None*, ***connectivity***=*None*, *compute_full_tree='auto'*, ***linkage***=*'ward'*, *pooling_func='deprecated'*)

linkage 的参数选项有 4 种：ward、complete、average、single。在本次实验中，使用了 3 种参数，分别为 ward、average、complete。

connectivity=kneighbors_graph(X,n_neighbors=200,include_self=False)
connectivity = 0.5 * (connectivity + connectivity.T)
创建邻接矩阵，并且保证邻接矩阵是对称的。

（6）DBSCAN

class sklearn.cluster.**DBSCAN**(***eps***=*0.5*, ***min_samples***=*5*, *metric='euclidean'*, *metric_params=None*, *algorithm='auto'*, *leaf_size=30*, *p=None*, *n_jobs=None*)

（7）Gaussian Mixtures

*class sklearn.mixture.GaussianMixture(**n_components**=1, **covariance_type**='full', tol=0.001, reg_covar=1e-06, max_iter=100, n_init=1, init_params='kmeans', weights_init=None, means_init=None, precisions_init=None, random_state=None, warm_start=False, verbose=0, verbose_interval=10*)

（8）评价指标 NMI

sklearn.metrics.**normalized_mutual_info_score(*labels_true*, *labels_pred*,** *average_method='warn'*)

用于评价聚类算法的性能

3、 实验结果

聚类算法	NMI
K-Means	**81.66%**
Affinity Propagation	**76.44%**
Mean-Shift	**78.66%**
Spectral Clustering	**72.25%**
Agglomerative Clustering(Ward)	**80.74%**
Agglomerative Clustering(Average)	**75.58%**
Agglomerative Clustering(Complete)	**56.86%**
DBSCAN	**79.38%**
Gaussian Mixtures	**79.84%**