**实验报告**

**学号：201814849 姓名：张延超 班级：2018级学硕**

1. **实验任务**
2. 测试sklearn中以下聚类算法在Tweet数据集上的聚类效果
3. 使用NMI（Normalized Mutual Information）作为评价指标
4. **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

1. **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%** |