**Homework 3: Clustering with sklearn**

**姓名**：史金婉 **学号：**201834874

**教师：**尹建华  **项目：**测试sklearn中的算法在tweets数据集上的聚类效果

1. **实验目的**

1.了解sklearn工具中的各种聚类算法；

2.掌握NMI(Normalized Mutual Information)评价指标；

3.掌握sklearn中的七种算法并能利用它对文档进行分类

1. **实验任务**
2. 测试sklearn中以下聚类算法在tweets数据集上的聚类效果：

K-means，AffinityPropagation，MeanShift，AgglomerativeClustering，SpectralClustering，DBSCAN，GaussianMixture；

2.使用NMI(Normalized Mutual Information)作为评价指标

1. **实验数据**

The Tweets dataset is in format of JSON.

1. **实验步骤**
2. K-means算法：

kmeans = KMeans(n\_clusters=89, max\_iter=50, n\_init=10, init='k-means++')

1. AffinityPropagation 算法：

affinity\_propagation = AffinityPropagation(damping=.55, max\_iter=200, convergence\_iter=15, copy=False)

1. MeanShift 算法：

mean\_shift = MeanShift(bandwidth=0.65, bin\_seeding=True)

1. AgglomerativeClustering 算法：

agglomerative\_clustering = AgglomerativeClustering(n\_clusters=89)

1. SpectralClustering 算法：

spectral\_clustering = SpectralClustering(n\_clusters=89, n\_init=10)

1. DBSCAN 算法：

dbscan = DBSCAN(eps=0.5, min\_samples=1, leaf\_size=30)

1. GaussianMixture 算法：

gaussian\_mixture = GaussianMixture(n\_components=89)

1. **实验结果**

K-means accuracy:0.79

AffinityPropagation accuracy: 0.78

MeanShift accuracy: 0.75

SpectralClustering accuracy: 0.68

AgglomerativeClustering accuracy: 0.78

DBSCAN accuracy: 0.70

GaussianMixture accuracy: 0.79