实验报告

1. 实验内容

测试 sklearn 中的聚类算法在 tweet 数据集上的效果，并使用 NMI 作为评价指标。

2. 实验过程

1）NMI 评价指标：Normalized Mutual Information

$$U(X,Y) = 2R = 2\frac{I(X;Y)}{H(X)+H(Y)}$$

$$I(X,Y) = \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \qquad H(X) = \sum_{i=1}^{n} p(x_i)\, I(x_i) = \sum_{i=1}^{n} p(x_i)\log_b\frac{1}{p(x_i)} = -\sum_{i=1}^{n} p(x_i)\log_b p(x_i),$$

2）在处理文档时，使用了 TfidfVectorizer 来统计 tf-idf。

3）需测试的聚类算法总结：

在 scikit-learn 中的 clustering algorithms （聚类算法）的比较

Method name（方法名称）	Parameters（参数）	Scalability（可扩展性）	Usecase（使用场景）	Geometry (metric used)（几何图形（公制使用））
K-Means（K-均值）	聚类形成的簇的个数	非常大的 n_samples，中等的 n_clusters 使用 MiniBatch code	通用, 均匀的簇大小, 平面几何, 不是太多的簇	点之间的距离
Affinity propagation	damping（阻尼）, sample preference（样本偏好）	n_samples 不可扩展	许多簇，不均匀的簇大小，非平面几何	图形距离（例如，最近邻图）
Mean-shift	带宽	不可扩展的 n_samples	许多簇，不均匀的簇大小，非平面几何	点之间的距离
Spectral clustering	簇的个数	中等的 n_samples，小的 n_clusters	几个簇，均匀的簇大小，非平面几何	图形距离（例如最近邻图）
Ward hierarchical clustering	簇的个数	大的 n_samples 和 n_clusters	很多的簇，可能连接限制	点之间的距离
Agglomerative clustering	簇的个数, 链接类型, 距离	大的 n_samples 和 n_clusters	很多簇，可能连接限制，非欧几里得距离	任意成对距离
DBSCAN	neighborhood 的大小	非常大的 n_samples，中等的 n_clusters	非平面几何，不均匀的簇大小	最近点之间的距离
Gaussian mixtures	很多	不可扩展	平面几何，适用于密度估计	Mahalanobis 与中心的距离
Birch	分支因子, 阈值, 可选全局簇.	大的 n_clusters 和 n_samples	大数据集，异常值去除，数据简化	点之间的欧式距离

3. 实验结果

The NMI of **KMeans** is:0.8260

The NMI of **AffinityPropagation** is: 0.7644

The NMI of **Mean-shift** is:0.7644

The NMI of **Spectral Clustering** is:0.7200

The NMI of **Ward hierarchical clustering** is:0.7991

The NMI of **Agglomerative clustering** is:0.7460

The NMI of **DBSCAN** NMI is: 0.7913

The NMI of **Gaussian Mixture** NMI is:0.7101