Clustering with Sklearn

姓　名：余　平　　学　号：201814848

（一）实验任务

（1）检验 sklearn 中聚类算法在 tweets 数据集上的聚类效果。

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

（二）实验数据集

数据集一共包含 2472 行，代表 2472 个测试样本， cluster 的个数是 89 。

（三）实验过程

(1)根据 sweet 构建字典，建立向量空间模型，每一行用向量表示。词典的大小是 5097，将每一个样本表示成一个 5097 维的向量。

（2）评价标准（NMI） 在 Sklearn 包中有已经定义好的函数，调用这个函数来评估自己的聚类效果。

（三）实验结果

（1）K-Means

```
K-means accuracy: 0.7881140298093793
```

（2）Affinity propagation

```
AffinityPropagation accuracy: 0.7834777200368181
```

（3）Mean-shift

```
MeanShift accuracy: 0.7468492000608157
```

（4）Spectral clustering

```
SpectralClustering accuracy: 0.6782978969064626
```

（5）DBSCAN

```
DBSCAN accuracy: 0.7009526046894612
```

（6）Agglomerative clustering

```
AgglomerativeClustering accuracy: 0.7800394104591923
```

（四）实验总结 通过本次实验，熟悉了 Sklearn clustering 中的各种聚类方法和聚类的过程，不同的聚类方法得到不同的聚类效果。