Homework 3: Clustering with Sklearn

姓名：孙畅畅
学号：201814832

（一）实验任务

（1）测试 sklearn 中以下聚类算法在 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

（1）tweets 数据集
（2）数据集特点：
A：无类别
B：有已知的聚类结果，其中 cluster 的个数是 89
C: 数据集一共包含 2472 行，代表 2472 个测试样本

（三）实验过程

（1）根据已有文本构建字典，建立向量空间模型，将已有文本中改的每一行样本用向量表示。其中构建的词典的大小是 5097，所以实验时，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这个函数，所以在得到聚类后，就可以用这个标准来评估自己的聚类效果与已知的聚类之间的差距。但在实验的过程中，查阅了网上的资料，发现他有多种计算公式，自己也尝试着编码实现了这个方法。最终结果发现，自己实现的方法与已有的库函数的结果相差不大。

（三）实验结果

使用 Sklearn 中国不同的库函数，使用 NMI 指标进行评，得到以下结果：

（1）K-Means

```
Normalized Mutual Information of KMeans: 0.781
```

（2）Affinity propagation

```
Normalized Mutual Information of AffinityPropagation: 0.695
```

（3）Mean-shift

```
Normalized Mutual Information of MeanShift: -0.727
```

（4）Spectral clustering

```
Normalized Mutual Information of SpectralClustering: 0.783
```

（5）DBSCAN

```
Normalized Mutual Information of DBSCAN: 0.108
```

（6）Agglomerative clustering

```
Normalized Mutual Information of AgglomerativeClustering: 0.695
```

（四）实验总结

通过本次实验，熟悉了 Sklearn clustering 中的各种方法，知道了聚类的过程，比较了不同方法的效果的好坏。实验之后，收获很大。