Homework3: Clustering with sklearn

姓名：刘敏　　学号：201814820

- **实验要求：**

  测试 sklearn 中 8 种聚类算法在 tweets 数据集上的聚类效果；

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

- **实验过程：**

  聚类算法实现：

  1. 首先读取 json 文件，将文档信息存入 texts 中，将文档类别标签存入 labels 中
  2. 使用 sklearn.feature_extraction.text 中的 CountVectorizer()方法将文档中的单词转换为词频矩阵，再通过 TfidfTransformer()方法计算每个单词的 tfidf 权重，最后将计算得到的 tfidf 权重值数组化，即元素 weight[i][j]表示 j 词在 i 类文本中的 tf-idf 权重
  3. 对所要求实现的 8 个聚类算法，从 sklearn 中导入对应的的聚类算法 KMeans, AffinityPropagation, MeanShift, SpectralClustering, AgglomerativeClustering, DBSCAN，GaussianMixture
  4. 每个聚类算法实现都是先调用 fit()方法进行适配，再预测训练数据标签，最后使用 NMI 指标评价聚类效果

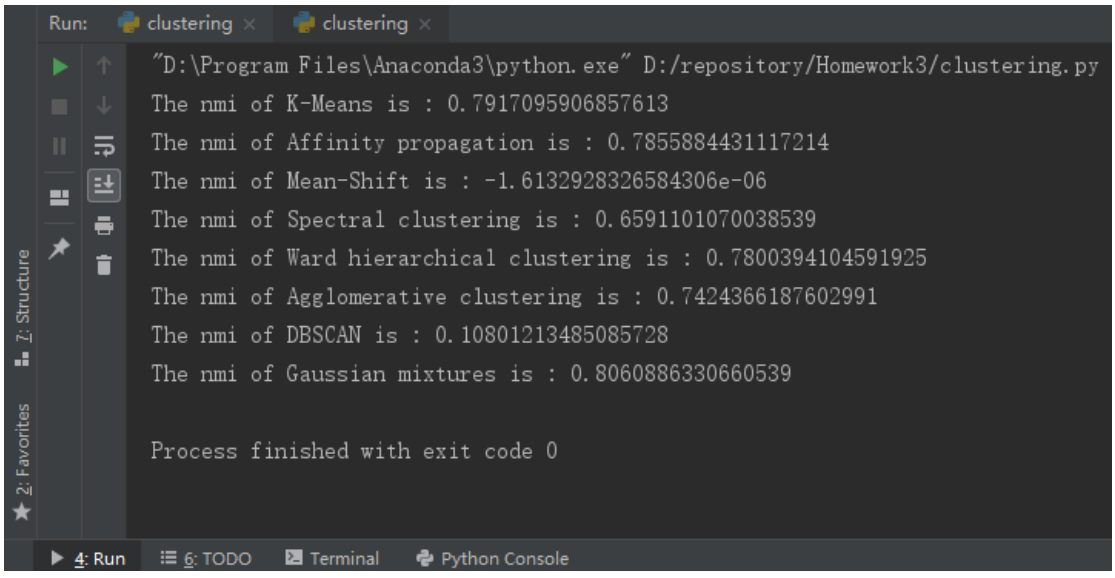
- **实验结果：**

  Tweets.txt 中数据格式如下：

clustering.py × Tweets.txt ×


```
{"text": "brain fluid buildup delay giffords rehab", "cluster": 37}
{"text": "trailer talk week movie rite mechanic week opportunity", "cluster": 14}
{"text": "rnc appoints chairman tampa convention effort visit tampa republican nati tampa fl", "cluster": 100}
{"text": "gbagbo camp futile cut ivory coast economy", "cluster": 110}
{"text": "chinese president lost translation powerful leader meet expect tran", "cluster": 61}
{"text": "england fishing community current management system broken edf", "cluster": 60}
{"text": "protest reform start yemen hundred anti government protester gathered sanaa", "cluster": 79}
{"text": "stuxnet lead chernobyl russian", "cluster": 83}
{"text": "iphone share smartphones phone", "cluster": 81}
{"text": "uploaded youtube video nba final lakers celtic game memory highlig", "cluster": 67}
{"text": "feed epic sci fi car ad kia optimum super bowl commercial video trendhunter supe", "cluster": 99}
{"text": "naughty facebook started sharing personal detail site day turn fb", "cluster": 95}
{"text": "super bowl commercial", "cluster": 99}
{"text": "bruce will fave kardashian lol", "cluster": 72}
{"text": "attack journalist escalate egypt", "cluster": 66}
{"text": "keith olbermann join current tv report", "cluster": 30}
{"text": "opinion christina sounded amazing super bowl america making big deal", "cluster": 75}
{"text": "sundance day strange celebrity encounter steve carell elijah wood full day sundanc", "cluster": 96}
{"text": "acai berry supplement weight loss associated content heard oprah win", "cluster": 55}
{"text": "finished watching black swan twisted mental strain take embrace role natalie portman incredible", "cluster": 101}
{"text": "hot air breaking rahm ballot tcot lt state totally corrupt", "cluster": 21}
{"text": "charlie sheen abuse", "cluster": 68}
{"text": "music featured motorola xoom super bowl commercial", "cluster": 99}
{"text": "rite movie review fan movie deal battle good evil", "cluster": 14}
{"text": "potential replacement steve job", "cluster": 106}
{"text": "yemen president signal won stay freedomwar egypt jan syria feb", "cluster": 79}
{"text": "debt settlement lawyer debt settlement easier lawyer day debt settlement beco", "cluster": 74}
{"text": "aguilera repeat national anthem ap ap christina aguilera flubbed belted", "cluster": 75}
{"text": "skipped romanian crew journalist threatened attacked detained reporting egypt", "cluster": 66}
{"text": "loss weight acai berry capsule benefit", "cluster": 55}
```


实验运行结果展示：



总结：通过几节课的学习，已经对以上几种聚类算法有了大致的了解，加上本次实验可以直接使用 sklearn 中已有的聚类算法实现，所以实验难度降低不少，首先要做的就是对数据的预处理，由于之前做 KNN 的时候已经做过构建 VSM 的相关实

验，所以这次就使用 sklearn 中用于文本特征提取的现有方法直接计算得到每个单词的 tfidf 权重，所以本次工作量不大。但是最终结果有点瑕疵，就是不知道什么原因，在对聚类算法 Mean-shift 计算 NMI 值的时候出现为负的情况，这是我始料未及的，然而还没有找到出错点，会继续更改。