

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：

```

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

```

实验运行结果展示：

```

Run: clustering x clustering x
D:\Program Files\Anaconda3\python.exe D:/repository/Homework3/clustering.py
The nmi of K-Means is : 0.7917095906857613
The nmi of Affinity propagation is : 0.7855884431117214
The nmi of Mean-Shift is : -1.6132928326584306e-06
The nmi of Spectral clustering is : 0.6591101070038539
The nmi of Ward hierarchical clustering is : 0.7800394104591925
The nmi of Agglomerative clustering is : 0.7424366187602991
The nmi of DBSCAN is : 0.10801213485085728
The nmi of Gaussian mixtures is : 0.8060886330660539

Process finished with exit code 0

```

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 值得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。