Clustering with sklearn

一、 实验目的

学习 scikit-learn、聚类算法、NMI、PCA 和 Matplotlib，回顾 tf-idf。

二、 实验要求

测试 sklearn 中以下聚类算法在 tweets 数据集上的聚类效果。

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

三、 实验环境

windows 10

python 3.6.5

scikit-learn 0.19.1

Numpy (>=1.6.1)

Scipy (>=0.9)

四、 实验过程

1. 读取训练集文本内容 。读取给定数据集文件夹中每一个文档后，将文本内容写入一个 txt，每一行为一个文档，方便后面词频矩阵的处理。

2. 文本预处理。读取之前存放所有文本内容的 txt，将其内容去空格，去标点，并进行分词等预处理。（以上两步老师已做处理）

3. 特征提取。使用 scikit-learn 工具调用 CountVectorizer()和 TfidfTransformer()函数计算 TF-IDF 值，将文本转为词频矩阵，矩阵元素 a[i][j] 表示 j 词在 i 类文本下的词频。将词频矩阵保存在 tfidf_result.txt(weight.pkl)文档中。

4. 聚类方法。根据聚类方法，调用 sklearn.cluster，设置参数等，并保存该聚类模型（对测试集使用）。

5. 进行分类。使用 clf.fit_predict 方法进行预测。

6. 用 NMI(Normalized Mutual Information)作为评价指标，进行评测。

7. 用 PCA 降维，pca = PCA(n_components=2) #降维两维

   new_weight = pca.fit_transform(weight) # 重新计算成二维形式

8. 用 matplotlib 进行可视化

五、 出现的问题及解决

1. 文件读写问题：

(1)file = open('data/Tweets.txt', 'rb')

TypeError: the JSON object must be str, not 'bytes' 解决办法：r

(2)text_file = open('data/Tweets_text.txt', 'wb', encoding='utf-8')

ValueError: binary mode doesn't take an encoding argument 解决办法：w

file = open('data/Tweets.txt', 'r',encoding='utf-8')

text_file = open('data/Tweets_text.txt', 'w',encoding='utf-8')

r/w 模式，可以指定编码，也可以不指定，windows 下默认是 gbk 编码。

rb/wb 模式直接读取二进制，与编码没有关系，加上就报错。

2．GaussianMixture 模型中

covariances = np.empty((n_components, n_features, n_features))

MemoryError

找了一些解决办法：需要设置 covariance_type 参数

clf = GaussianMixture(n_components=89,covariance_type='diag')

疑问：covariance_type 默认的就是'diag'，不写这一项就报错

六、实验结果

```
KMeans
    聚类数为85的NMI值为:0.7893025683289238 运行时间:60.527125120162964 本地时间：2018-12-08 23:03:55
    聚类数为86的NMI值为:0.7868741451151905 运行时间:60.27579879760742 本地时间：2018-12-08 23:04:55
    聚类数为87的NMI值为:0.7915634585025494 运行时间:62.60656428337097 本地时间：2018-12-08 23:05:58
    聚类数为88的NMI值为:0.7801190036547079 运行时间:62.19315791130066 本地时间：2018-12-08 23:07:00
    聚类数为89的NMI值为:0.7868365025021115 运行时间:63.0054976940155 本地时间：2018-12-08 23:08:03
MiniBatchKMeans
    NMI值为:0.6559083621643357 运行时间:1.7473254203796387 本地时间：2018-12-08 23:11:07
AffinityPropagation
    NMI值为:0.7855844431117215 运行时间:28.947553873062134 本地时间：2018-12-08 23:12:34
MeanShift
    NMI值为:-1.6132928326584306e-06 运行时间:20.132158756256104 本地时间：2018-12-08 23:15:22
SpectralClustering
    NMI值为:0.683594084430078 运行时间:3.936443567276001 本地时间：2018-12-08 23:15:45
HierarchicalClustering
    NMI值为:0.7800394104591925 运行时间:19.091907739639282 本地时间：2018-12-08 23:16:12
DBSCAN
    NMI值为:0.10801213485085728 运行时间:9.540486097335815 本地时间：2018-12-08 23:16:43
Gaussian mixtures
    NMI值为:0.7678904578420558 运行时间:7.678448915481567 本地时间：2018-12-08 23:17:02
Birch
    NMI值为:0.795088793201375 运行时间:13.850955724716187 本地时间：2018-12-08 23:17:19

SpectralClustering
    NMI值为:0.6630059256278761 运行时间:4.043200254440308 本地时间：2018-12-09 16:55:25
Birch
    NMI值为:0.795088793201375 运行时间:13.998553991317749 本地时间：2018-12-09 17:08:30
```