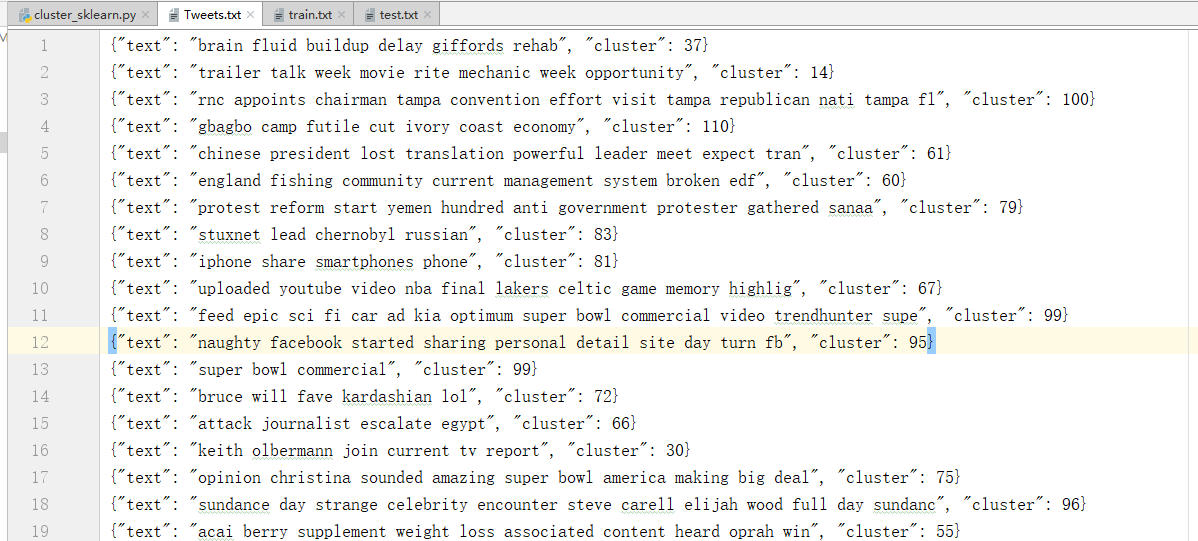
**作业三实验记录**

* **实验要求**

在给定的Tweet数据集上利用sklearn库函数取实现一些聚类算法。

* **实验步骤**

首先，根据给定的Tweet数据来划分training set和test set。具体的数据格式如下：{‘text’: ‘XXX’ , ‘cluster’: ‘XX’}



由于是要实现一些聚类算法给定的数据，因此类的大小是随机的，本次试验中，首先去逐行读取给定文件，按照随机原则，将文件的80%分给training set，剩余的20%给了test set。结果统计如下： training set的大小是2002个， test set的大小是470,，然后聚类算法的类的个数应该是training set中的数据的label，经统计可得是87，并且删除那些test中不在train中类下的数据。

然后对得到的数据构建Vector Space Model，本次实验中，得到的词典大小是4486，最终将train的tf\_vector存入了train.pkl 文件中，将test的tf\_vector存入了test.pkl 文件中。

然后调用sklearn库函数，调用不同的聚类算法进行聚类。

* **实验结果以及说明**

对于聚类算法的衡量指标有很多，常用的有纯度（purity）、熵（entropy）、准确率（accuarcy）、归一化互信息（NMI）、调整兰德指数（ARI）等，本次实验结果采用归一化互信息（NMI）进行评测，它在聚类算法中可以衡量对同一个数据集不同的划分的之间的相似程度，它的值介于0~1之间，当两者的差别很小的时候，这个值趋向于1，反之则趋向于0。

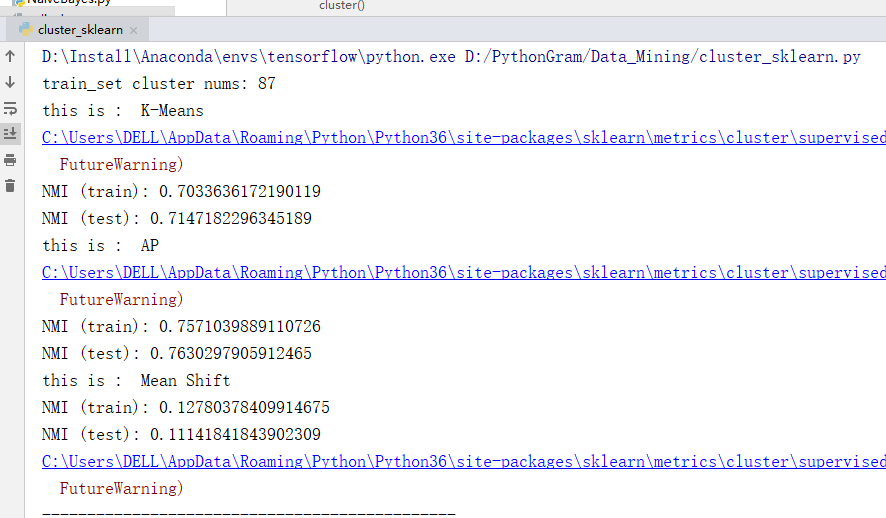
K-means聚类需要预先设置 K 的大小，经统计训练集的聚类标签数量，发现label有 87 个，所以先验 K 的值为 87。

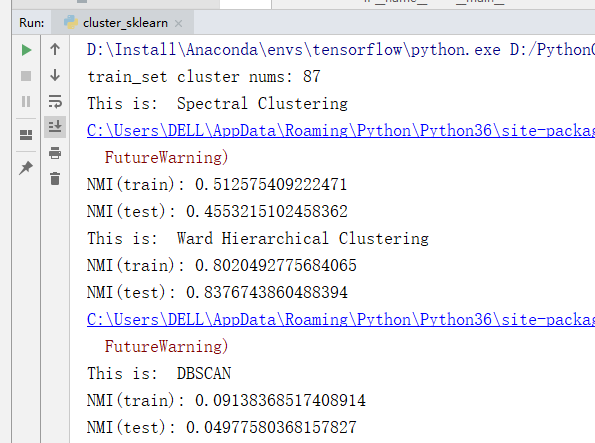
AP 和MS聚类算法不需要设置类簇的数目，直接input进入训练数据即可。

sklearn的一些聚类算法中，如spectral clustering、Agglomerative Clustering 和 DBSCAN，没有 predict() 方法，经查阅资料得知，我们需要调用只能使用 fit\_predict() 方法来进行预测。因为他们的结构在加入新的数据时会发生改变，所以在预测的时候，就相当于再训练，而不能用原来训练好的模型来预测，这样产生的predict和真实值进行比较评测。

对于GM聚类算法，在计算过程中基于概率论的算法，对一些分类较为模糊的点有很好的描述。如果数据集的各特征值均服从高斯分布，可以用高斯混合模型聚类，但是在运行过程中总会报出内存溢出的错误，我认为是类簇个数较多导致的。 经查阅资料，在以后的实践过程中可以采用降维的方法，比如无监督降维（PCA）和有监督降维（LDA），即将原高维空间中的数据点映射到低维度的空间中再进行聚类。

本次实验结果如下：





* **实验总结**

本次实验的过程整体完成了聚类的功能，除了前期的处理类似于作业一和二中的处理手段，在聚类算法实现的过程中，直接调用sklearn库函数大大简化了工作量，但是同时也会发现不同的函数由于算法内部的设计会有不同的结构。因此在调用的时候，我查阅了相关的技术文档，并且和同学讨论，对算法有了进一步的了解。