智能数据挖掘作业5

19200300004 黄铭瑞

# 实验目的

掌握使用层次聚类和k均值聚类的方法，并对20newsgroups数据集进行聚类。

# 实验原理

## 数据集

20newsgroups数据集是用于文本分类、文本挖据和信息检索研究的国际标准数据集之一。数据集收集了大约18000个新闻组文档，均匀分为20个不同主题的新闻组集合。

## 层次聚类

层次的聚类方法（Hierarchical Clustering），字面理解，即层次化的聚类，最终结果是树状结构，换句话说，层次聚类通过计算不同类别数据点间的相似度，来创建一棵有层次的嵌套聚类树。

层次聚类又分为凝聚方法和分裂方法。

### 凝聚方法

凝聚方法是一种自下而上的方法，先将所有样本的每个点都看成一个簇，然后找出距离最小的两个簇进行合并，不断重复到预期簇或者其他终止条件。代表的有Agnes算法：

1. 初始化，每个样本当作一个簇。
2. 计算任意两个簇的距离，找到这两个簇，合并为一个簇。
3. 重复步骤2，直到两个簇的距离超过阈值，或者簇的个数达到规定值，终止。

### 分裂方法

分裂方法是一种自上而下的方法，先将所有样本当作一整个簇，然后找出簇中距离最远的两个簇进行分裂，不断重复到预期簇或者其他终止条件。代表的有Diana算法：

1. 初始化，所有样本归为一类。
2. 同个簇中，计算任意两个样本的距离，找到距离最远的两个样本，作为两个簇的中心。
3. 计算原来簇中其余样本距离这两个中心点的距离，把它归为距离近的中心点所属的那一簇。
4. 重复步骤2、3，直到两簇距离达不到分开的阈值，或者簇数量达到规定之，终止。

## K-Means

K-Means算法的思想很简单，对于给定的样本集，按照样本之间的距离大小，将样本集划分为k个簇。让簇内的点尽量紧密的连在一起，而让簇间的距离尽量的大。总的说，让数据与对应聚类中心的误差平方和最小，即：

其中是簇的均值向量, 为平方误差和。

算法步骤：

1. 初始化聚类中心 。
2. 根据最小距离法则，把划分到为代表的类中。
3. 计算，重新计算 。
4. 重复2、3，直到不再变化，或者不再减小。终止。

## 距离度量

### 类间距离

1. 单连锁：将两个簇中相距最近的两个点的距离作为两个簇的距离。
2. 全连锁：将两个簇中相距最远的两个点的距离作为两个簇的距离。
3. 平均连锁：将两个簇两两点距离的平均值作为这两个簇的距离。
4. Ward’s method：总是使得并类导致的类内误差平方和增量最小，但只能用于欧氏距离。

# 实验过程

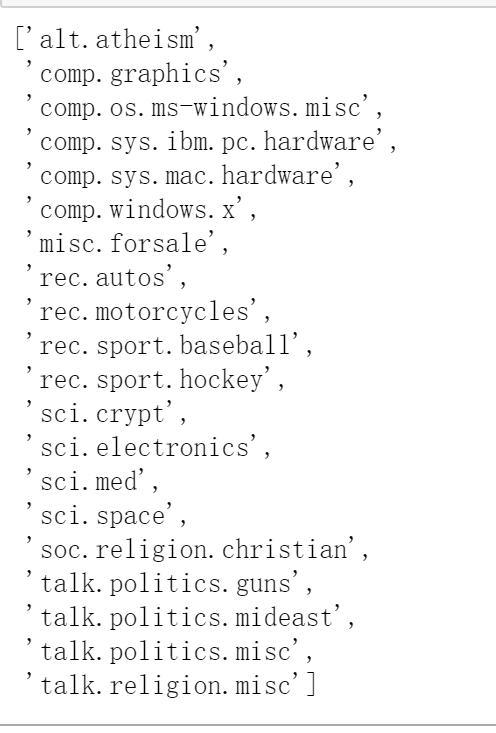
大致过程：导入20newsgroups数据集，转为DataFrame格式，使用nltk里的停止词对数据进行清理，把清理后的文本数据转为TF-IDF向量，分别进行层次聚类和K均值聚类。

## 数据集

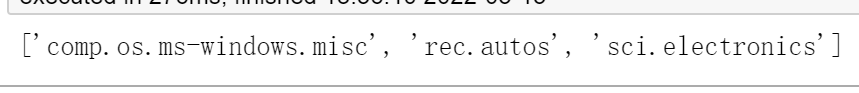
### 导入数据集

使用sklearn.datasets里的fetch\_20newsgroups进行数据集导入，但限于网络原因，使用的是本地下载好的数据集。

原本的数据集分有20个主题。



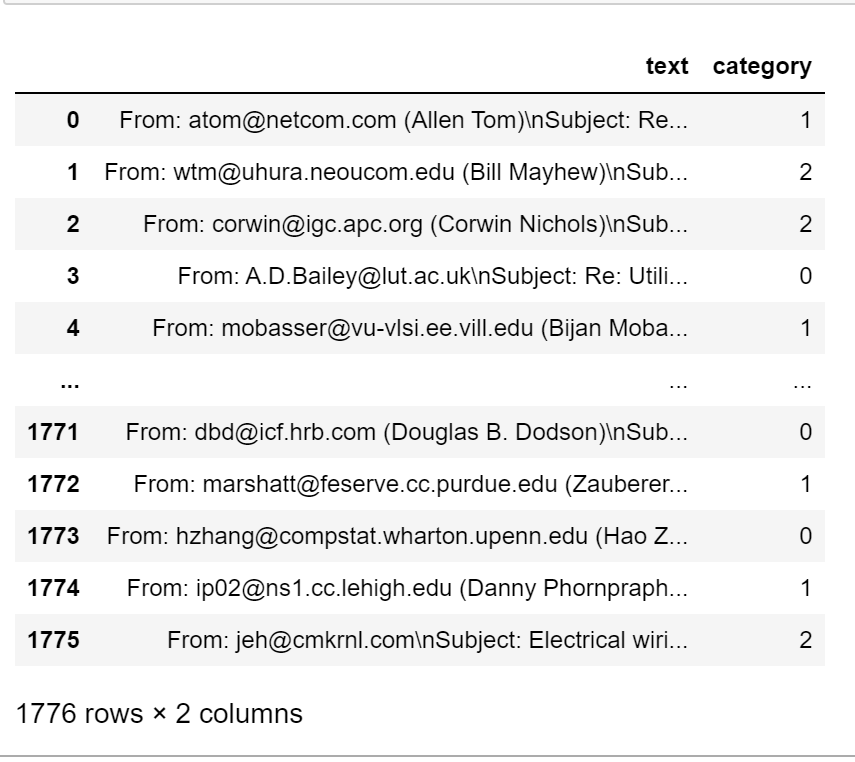
选择其中的三个主题进行实验。



### 转为DataFrame

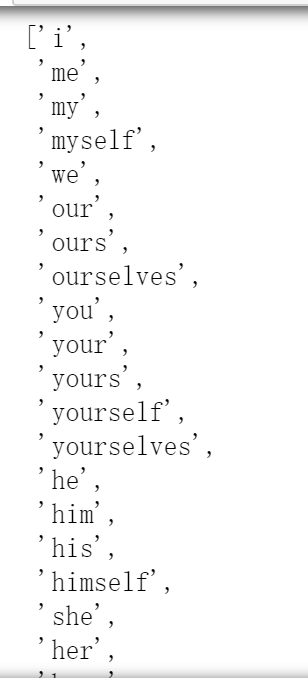
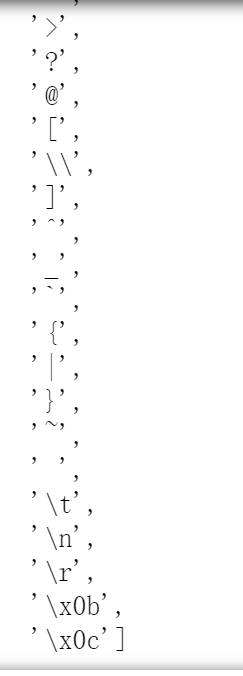
提取出其中的‘data’部分和‘target’部分，并重新命名为‘text’和

‘category’，转为DataFrame。0、1、2分别为‘comp.os.ms-windows.misc’, ‘rec.autos’,‘sci.electronics’的类标签。



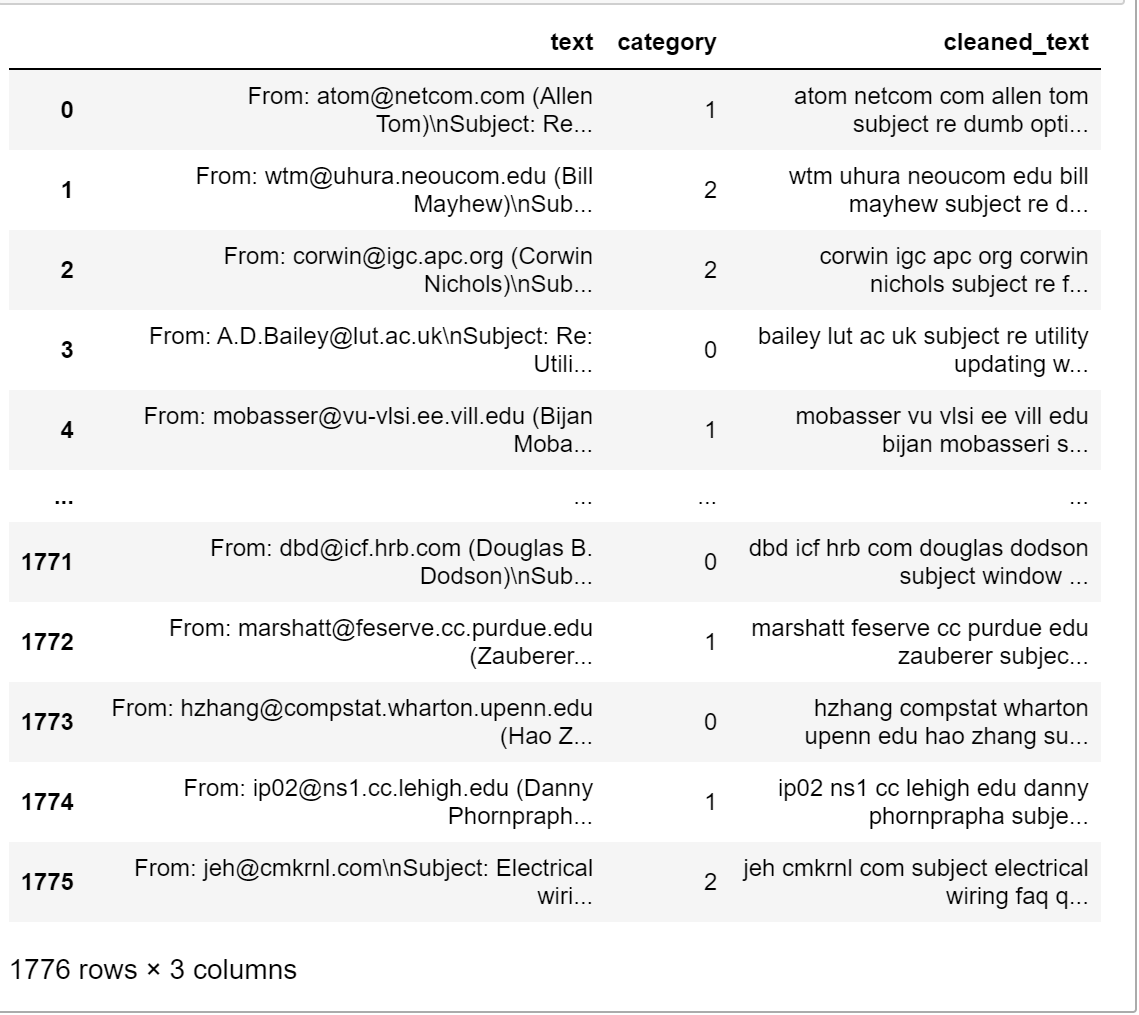
### 停止词设置

使用stopwords.words('english')获取英文停止词，list(string.printable)获取可打印字符，作为数据的筛选器。

，… ，

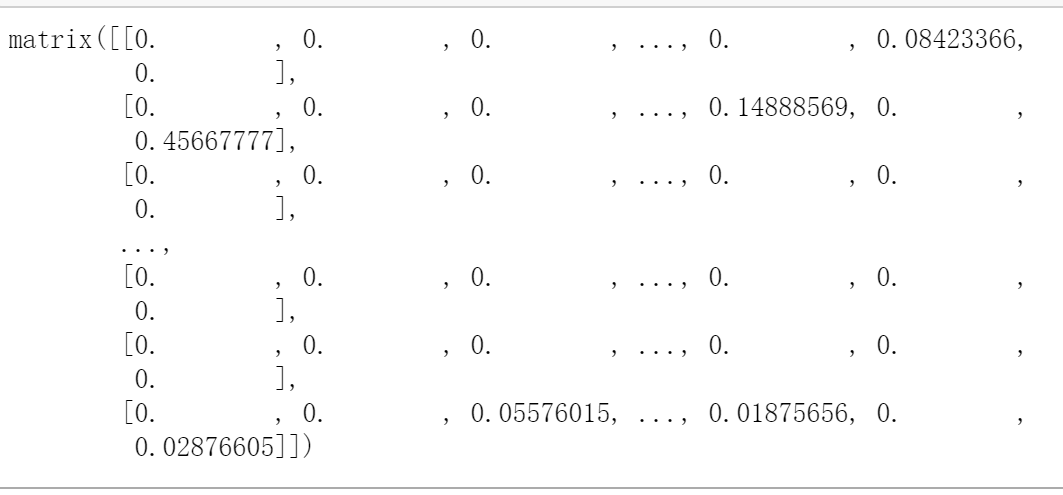
### 数据清洗

把不属于停止词的字符串提取出来，添加到新的列，列标签为‘cleaned\_text ’。

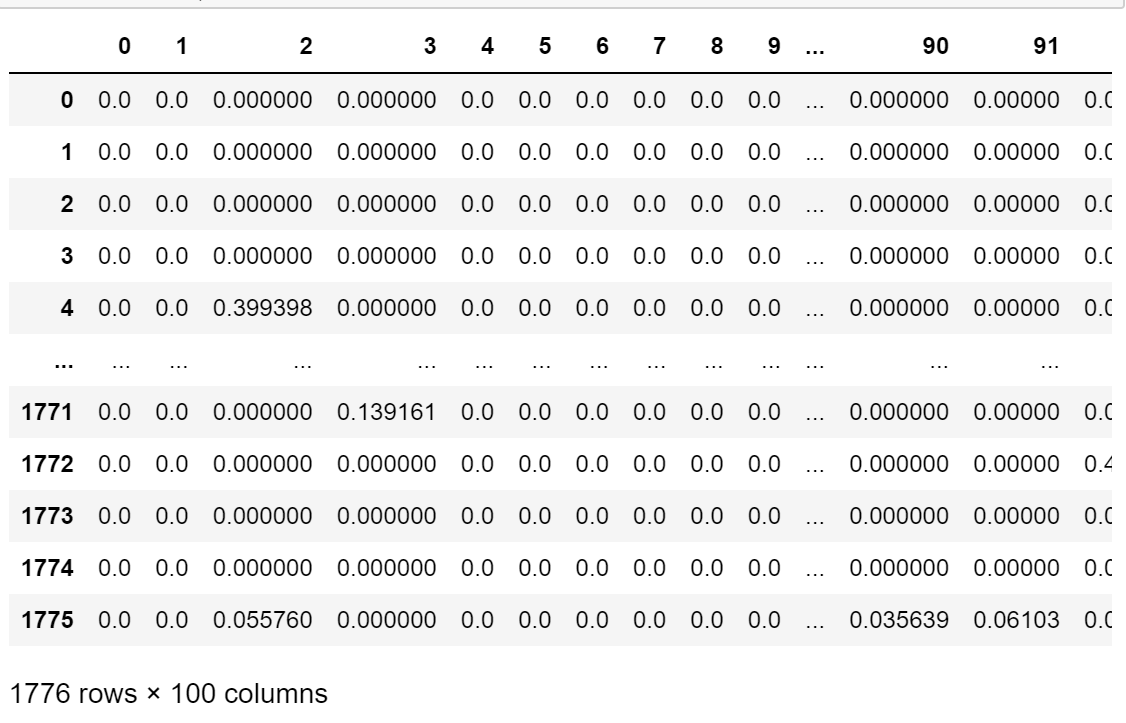


### 转为TF-IDF向量

对清理后的文本数据提取其中的100个TF-IDF特征，转为TF-IDF向量。



把得到的向量转为DataFrame。

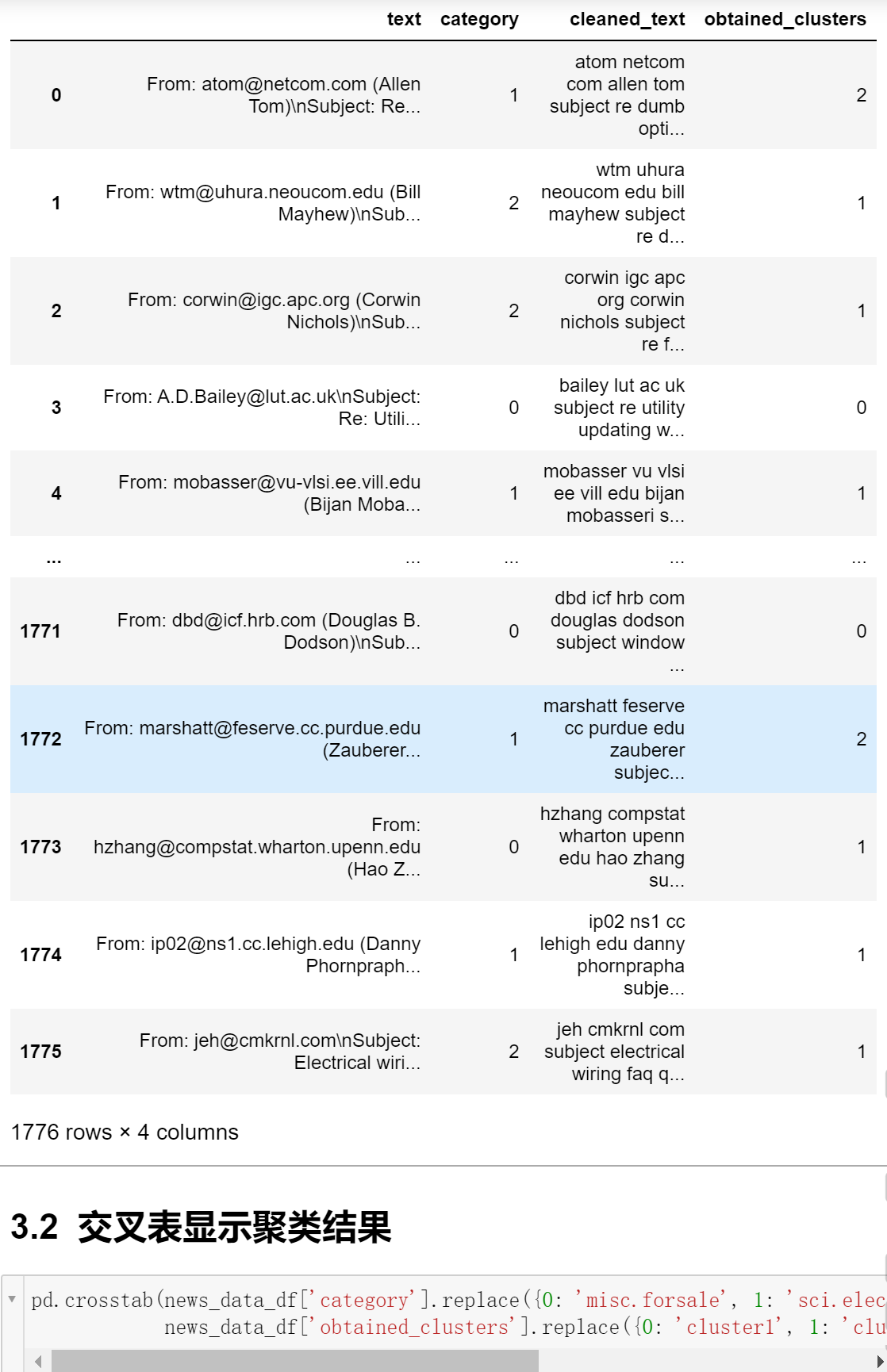


## 层次聚类

调用scipy.cluster.hierarchy里的linkage()，计算欧氏距离，用ward方法使方差和最小化。使用dendrogram()生成聚类树。绘制聚类树图案。

## K-Means

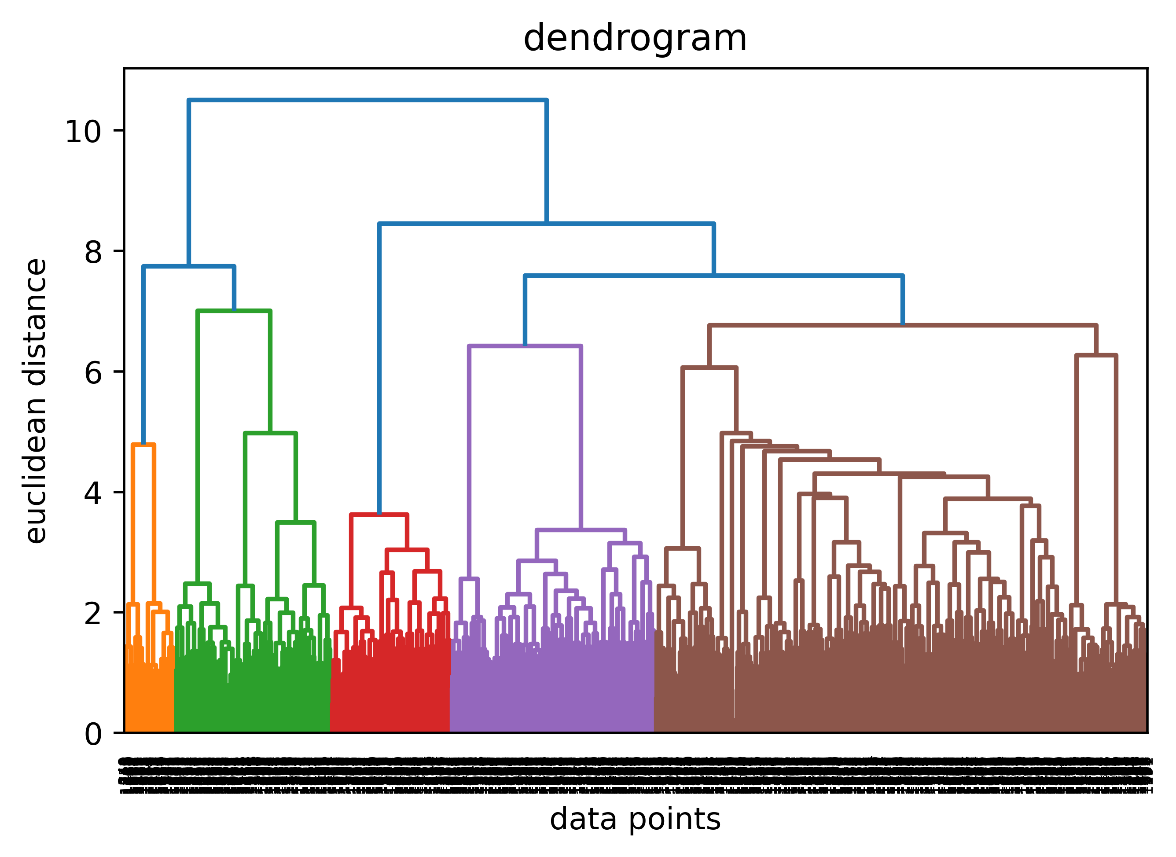
调用sklearn.cluster里的KMeans()，对向量数据进行计算，获取个样本所属的聚类中心‘obtained\_clusters’。



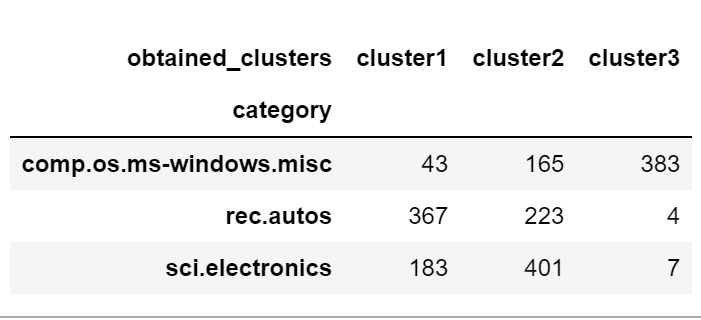
使用交叉表显示出聚类结果。

# 实验结果 与分析

## 层次聚类

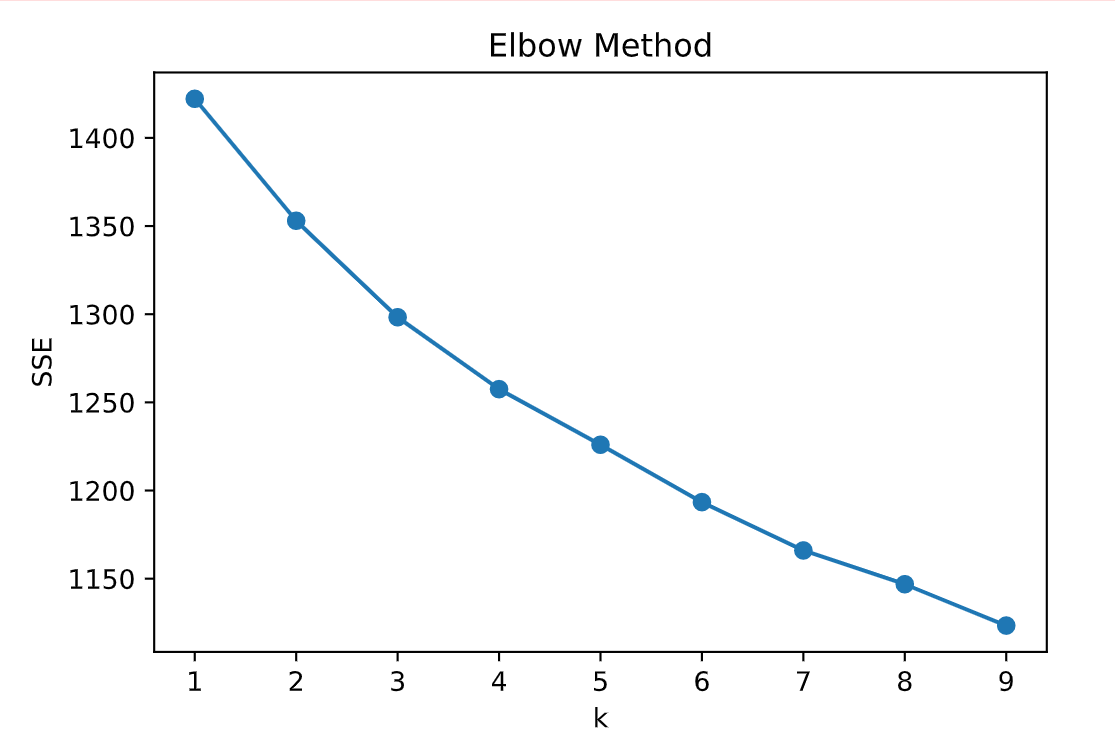


## K-Means



可以看出，第一个聚类簇里，大部分是属于‘rec.autos’这一主题的。第二个聚类簇里，大部分是属于‘sci.electronics’这个主题的。第三个聚类簇里，大部分是属于‘comp.os.ms-windows.misc’这个主题的。

在寻找最佳聚类数的时候，使用手肘法可以更快的定位到最适的k值。随着聚类数k的增大，样本划分会更加精细，每个簇的聚合程度会逐渐提高，那么误差平方和SSE自然会逐渐变小。当k小于真实聚类数时，由于k的增大会大幅增加每个簇的聚合程度，故SSE的下降幅度会很大，当k大于真实聚类簇数时，增加k所得到的聚合程度回报会迅速变小，所以SSE的下降幅度会骤减。可以通过绘制k和SSE关系图，找出那个明显的拐折点，对应的k作为最佳聚类数。



可以看出，在k=3时，有拐点出现。选取3为最佳聚类数。

# 附录

<20newsgroups.html>