一、程序的运行调用顺序





creatTrainDictionary()



GetCategoryTermList()

calculateFinalTFIDF()

二、关于计算训练集和测试集（预测集）中的TFIDF的说明。

训练集一般情况下是预先收集并标注好的数据集合，一般用来训练分类器并输出分类模型，而测试集或者预测集一般是未标注好的数据集合，一般是以数据流的形式以单个文档的形式进入分类器（例如利用爬行器的采集的预测集就是一个数据流的形式进入分类器，这也是在分类系统中常见的情况，这种数据集合称之为动态的数据集合），利用分类模型对测试集（预测集）进行预测时需要计算测试集（预测集）的词的TFIDF，这里就遇到一个问题关于测试集（预测集）中IDF的计算。IDF计算时是取训练集中IDF还是 单独计算测试集（预测集）中的IDF？

In Text classification, the creation of TF-IDF for the testing documents is performed using the IDF from the train documents.  
  
For instance, if we only want to classify one document (using the same structure), the TF-IDF for the document should be based on the occurrences of terms in the document and the IDF previously computed based on the training collection. In the same example, if IDF is based on the test document alone all the features will become 0, as all the terms appear in all documents (one) of the test collection.

For this state if your dataset is big enough you could using just training set for IDF. in the test phase if the new term be in train set use the IDF of training and if the term is new use the number of train set documents for calculate IDF. For some purposes you could use smoothing methods for having better results.

You should compute the IDF (inverse document frequency) for every term using the training set. You should then use the same IDF for the documents in your test set. The TF on the other hand depends on the concrete document at hand that you try to classify, so it will be different for different documents in the test and train set.