Motivation:
We attempt to classify different posts on the Lily BBS. Given a new post such as '南京行货和水货手机的主要购买渠道...', we attempt to predict which forum (like '计算机系', '手机天地',...) it may belong to. In this task, given the Lily BBS posts data for 10 different forums, it is required to do some preparations on the original data, namely, sentences tokenization, deleting the stop words and extracting TF-IDF (Term Frequency–Inverse Document Frequency) features.

In assignment 1, we only need to do data processing (sentences tokenization, deleting the stop words and extracting TF-IDF features.), and classification task is not required.

Dataset:
The zip file contains 10 txt files (10 classes, one for each forum). Each txt file contains more than 100 lines, and each line represents a raw post. You may regard each line in each txt file as an instance/example of the corresponding class.

Task Description
1. First, please do the tokenization job to get words list from the Chinese sentences. You may use some existing tools like jieba for python, or IKAnalyzer for java, to help you accomplish the tokenization job.
2. Second, please delete the stop words. We provide a Chinese stop words list for your reference, and you can download it.
3. Third, please extract TF-IDF (Term Frequency–Inverse Document Frequency) features from the raw text data on the basis of step 1 and step 2 above. Each raw post is mapped to a TF-IDF feature vector and all vectors are the same length. You may learn or review TF-IDF here. Please implement extracting TF-IDF features by yourself, and Do NOT invoke other existing codes or tools.
4. Four, please generate 10 new txt files which only contain TF-IDF features for each class, each line in each new txt file represents the TF-IDF feature of a post (instance/example) in the corresponding class. Please name the 10 new txt files as same as the original 10 txt files that we provide.

The format of txt files in the result folder is shown as below: (only to illustrate the format, not the true answer)
D_Computer.txt
Line 1: 1.8262 0.0665 0 2.6031 0 0 2.4644 0.7164 ...
Line 2: 0 2.1743 0.9495 0 0.7056 0 0 2.6424 ...
...
Mobile.txt
...
Each number represents the TF-IDF value for a word in the post and 0 means that the word does not appear in the post. The word dictionary is based on the 10 txt files that we provide.