1. 对于正常文本的分类,我们采用的是基于SVM的分类模型。SVM属于有监督分类，所以必须需要已有分类好的数据用于训练模型，在这里已分类好的训练数据，我们采用的是搜狗公司提供的文本分类语料库，网址是<http://www.sogou.com/labs/dl/c.html>。
2. 文本分类的训练，首要任务是实现文本的特征向量表示，即把文本用向量表示,在这里采用的是向量空间模型即VSM。为了对文本进行特征向量表示，我们需要对于文本进行分词的预处理,在这里我们采用的Python的结巴分词模块对于SogouC文件夹下的未分词的文本分词后保存到SogouCCut文件夹下。有了分词后的文件，我们可以对这些分词文件进行处理通过卡方检验的方式进行特征选取,不进行特征选取的话，空间向量维度太高，分类效果不好。特征选取的代码在FeatureSelection.py文件中，执行这个文件后我们会对每个类别选取1000个特征词进行表示，搜狗语料库共有10个类别，即选取10000个特征，由于重复共选取了9508个特征,在SVMFeature.txt文件中可以查看。现在已经有了特征,还缺少特征的权重,在这里我们采用TF\*IDF的计算方式计算权重，TF某个词w在文档d中出现次数count(w, d)和文档d中总词数size(d)的比值，IDF为即文档总数n与词w所出现文件数docs(w, D)比值的对数。对于特征权重的计算可以在FeatureWeight.py 和TestFeatureWeight.py文件中查看。训练文本特征向量表示存储在train.svm文件中,测试文本的特征向量表示存储在test.svm文件中。之所以按照这样的格式存储,是为了比较方便的调用LIBSVM([LIBSVM](https://www.csie.ntu.edu.tw/~cjlin/libsvm/))进行分类。