**任务2.新闻的预处理**

新闻的预处理因为需要去掉停留词和进行分词，所以主要运用了jieba进行分词。

对于去停留词，我使用了在网上下载的1892字的中文停用词表、891字的英文停用词表以及哈工大的767字的中文停用词表。

对于分词.词性标注以及去停用词，我用了这样的代码



对于进行词性的筛选，我使用了这样的代码



预处理其实第一步应该是保证数据的质量，所以需要去掉不合要求的数据（太短的空的），清理数据（将其中的网页代码或者js代码去除）。

第二步才应该是分词词性筛查操作。

接下来你可以思考你所生成的词列表，如何转化成可以学习的向量或矩阵（看一下one-hot和其他词嵌入模型如word2vec）