中文文档处理基本过程

中文文档处理的基本流程是这样的：

1）输入每一篇文档；

2）然后“分词”（用软件包）实现；

3）取名词（以后还会增加取动词）

4）在分词结果上去停用词，利用一个停用词表，编写代码，利用这个stop\_words表把不需要的停用词（如啊，了，的等）去掉。

5）建词典（就是把所有文档经过前面4步处理过的单词汇总）6）用词典将每一篇文档利用生成的词典表示成向量的形式。

注意：

1. 虽然有现成的分词工具，能够分词并提取词性，但是，不能批处理的连续处理语料库中的所有文档，所以要编写程序实现全语料库的批处理处理。
2. 不要使用网站上的集成工具（如图像化界面），因为不能实现自动批处理处理多个文档，同时，后面不能改进处理去停用词等功能，所以要自己编写整个框架程序，然后引用网站的分词和词性提取包。

能够下载中文分词包的网站：

1）斯坦福大学自然语言处理的网页，能够下载能够处理英文、中文等语义的分词工具。http://nlp.stanford.edu/software/tagger.shtml

2）中科院的能够处理分词的工具。  
Ictclas

3）pathon中的jieba包。