重心放在中文文本预处理的部分，看不同的文本预处理对情感分析结果的影响。

我搜集了一下相关的资料，中文文本预处理的部分包括一些与英文相同的部分，也包括一些不同的部分。主要是以下几点：

相同的包括：1、去除超链接2、去除重复的字3、去除停用词4、去除数字5、标点符号

不同的包括：1、中英文混合文本（目前还没找到简单有效的方法处理,因为其中还夹杂着拼音之类的复杂情况）2、将组合字符以相应的词表示例如((๑•̀ㅂ•́)و✧ 代表棒等等 我找到了搜狗的这种库，可以用简单的字符串匹配来搞)3、替换网络词http://wangci.net/word.html

为了做上面的分析，自己找的数据集想满足几乎是不可能的，只能手动往里面加东西了，估计操作量也不小，我找到一个规模为2000的数据集，那估计得修改2000的数据量。还要涉及上面的内容，工作量目前看来比较大。