实验报告

学院：数据科学与计算机学院　　专业：计算机科学与技术　　科目：自然语言处理

学号：17341213　　姓名：郑康泽

总结

我尝试了 jieba、SnowNLP、thulac、pynlpir 这四种中文分词工具，首先这四种分词工具的区别还是很大，但感觉 jieba 分的应该是最好的。以下是我的一些问题：

1) SnowNLP 工具分词很奇怪，会把人的姓和名分开，我觉得正常情况下姓名应该是不用分的；
2) 还有文中有“20 日”这个组合词，个人觉得是不用分的，但是 jieba、SnowNLP、thulac 这三种分词工具都把该组合词分成了“20”和“日”了；
3) “中国人民政治协商会议”应该是一个整体，但有些分词工具还是会把它分开，可能是它们的词库中没有这个词；
4) 基本四种分词工具对于“单个字的形容词+名词”这种组合词都会将它们分开，但我个人觉得可以不分，比如“副主席”、“新时代”等这种词就是一个整体；

总之，中文分词工具还是有一些不完善的地方，但整体还是不错的，按照它们的分法是可以将文章读懂。中文分词工具还是挺强大的。