第十一章 自然语言处理作业

作业内容：

1. 复习前面章节的内容；
2. 完成第四次实验报告

《红楼梦》前八十回和后四十回到底是不是同一个作者？现在我们看到的版本大多写的是：曹雪芹写了前八十回，高鹗续写了后四十回。事实上，红学上关于《红楼梦》的作者争议一直很大，存在着很多种版本。

“红学史”上“高鹗续作《红楼梦》后四十回”的说法，最早出于清代文学家张问陶。当时，张问陶有《赠高兰墅(鄂)同年》诗。诗题自注云：“传奇《红楼梦》八十回以后俱兰墅所补。”《绘境轩读画记》记载说，曹雪芹“《红楼梦》小说，称古今平话第一。嘉庆时，汉军高进士鄂酷嗜此书，续作四十卷附于后，自号为‘红楼外史’。”

清代著名《红楼梦》评论家张新之指出，《红楼梦》80回前后在思想、结构上、人物性格发展上都具有高度的一致性，他在《红楼梦读法》中写道：“一部《石头记》，计百二十回，沥沥洋洋，可谓繁矣，而实无一句闲文。有谓此书只八十回，其余四十回乃出另手，吾不能知。但观其中结构，如常山蛇，首尾相应，安根伏线，有牵一发浑身动摇之妙，且此句笔气，前后略无差别——重以父兄命，万金赠，使闲人增半回，不能也。何以耳为目，随声附和者之多？”

出自：<http://www.360doc.com/content/18/0112/19/46551251_721410413.shtml>

本次实验纯粹在文本相似度的角度分析前八十回和后四十回的作者是否为同一人。

提示：本次实验因为仅从数据分析角度考量出现任何结果都是可能，有理即可。后续如果有同学对这个项目感兴趣可以多参考一些文献从各种角度去解读这一问题。

提交内容：一份完整的实验报告、py代码的压缩文件。

要求：以Python进行编程，思考我们在以前课程中所所学的方法，寻找最佳解决方案。

提交截止时间、地址和格式：

周一班提交地址：http://xzc.cn/Ol368x3Br5；（5.30 24:00）（前）

周五班提交地址：http://xzc.cn/Pn2F76nmLn；（6.03 24:00）（前）

命名方式：序号+姓名+学号！

若有修改，可以将修改后的内容按照地址重新上传一份，并在文件名尾加上“修改版”/“最终版”等标识。（在作业截止日前）