初步思路

1. 预处理，把回目编号删除，然后根据标点符号等，把每一句话标记出来；
2. 利用开源分词组件“结巴分词”(<https://github.com/fxsjy/jieba>)对红楼梦全书进行分词，统计词频，把出现100次以上的高频词语标记出来；
3. 统计每一章高频词语出现的频次；
4. 前80回、后40回各选15回分别作为训练数据，通过机器学习得出用词特征；
5. 根据机器学习训练出的用词特征，计算每章是否和所属前80回或后40回相同，如果相同，则说明前80回和后40回很可能是两个人写的。换句话说，如果给出前80章的某一章，机器能算出它属于前80回而不是后40回；给出后40回的某一章，机器算出它属于后40回而不是前80回，则显然前80回和后0回遣词造句方面不同，从而得出不是同一个作者。
6. 利用相同算法分析《三国演义》、《水浒传》、《西游记》会得出什么结论呢？

参考文献

[1]黎晨(知乎).用机器学习判定红楼梦后40回是否曹雪芹所写[EB/OL].(2016-6-27)[2018-3-28].https://zhuanlan.zhihu.com/p/21421723

[2]楼宇(知乎). 用 Python 分析《红楼梦》[EB/OL].(2017-11-23)[2018-3-28] .https://zhuanlan.zhihu.com/p/29209681

[3]. 小智(知乎). 通过数据挖掘能分析《红楼梦》各回的真伪吗[EB/OL].(2016-6-24)[2018-3-28] . <https://www.zhihu.com/question/19768898/answer/107358050>

[4]. 永琳(知乎).有哪些现代分析方法可以用于解决“红楼梦续写争议”？有哪些例子？[EB/OL].(2016-6-5)[2018-3-28] . <https://www.zhihu.com/question/47059344/answer/104326214>