**根据上次的反馈，地点信息并未进行进一步的结构化，如果有需要，老师可以再通知我。如果上述信息结构还有存在问题或可以优化的地方，老师也可以通知我修改。**

具体输出样例可以参照SampleOut.txt，里面包含了前十个成功抽取的文书结构化信息。当然也可以查看out.txt，里面包含全部成功抽取的文书的结构化信息。

可以参照着输出样例查看下述说明。

每一个作案信息的一级属性名包括：Time（出现一次）、Criminal（出现一次）、Location（出现一次）、Item（出现至少一次）、Value\_Supplement（出现小于等于一次）

Time的二级属性名，包括Complete、Year、Imprecise\_month、Month、Season、Imprecise\_day、Day、Imprecise\_hour、Hour、Minute、Second。其中Complete、Year、Month出现一次，其他的出现小于等于一次

Criminal没有二级属性，其取值由若干个被告人、或非本案处理的同伙作案人的姓名组成，姓名与姓名之间用制表符分隔。每个姓名后可以包含括号进行补充说明，一般是诸如（在逃）、（另案处理）等信息。

Location没有二级属性，是一个完整的地点

Item可能出现多次，每个Item里面包含二级属性Modifier、Name、Value、Color、Brand、Type、Amount。其中Name出现一次，其他属性出现小于等于一次。

Value\_Supplement没有二级属性，是对物品价值的补充说明。（有些文书，对于物品价值的描述没有紧跟在物品后面，只能以这样的形式抽取出来进行补充说明）

**信息格式和所有属性的具体含义如下面的格式描述：**

大致采用了正则表达式的方法描述，

例如(左右)? 表示这里可以出现“左右”，也可以不出现；年[末底终初中]表示可以出现年末，也可以出现年底、年终、年初、年中等。

\t是制表符，下面格式说明中的空格实际上是不存在的，都用\t代替。

如果属性名后面带问号，表示该属性可出现可不出现。

**具体格式如下：**

1. \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* //标志着下一个案件信息的开始
2. D9E739A9-BF89-4A58-9212-45018AAACE8D //案件编号
3. ################ //标志着本案件中下一个作案经过信息的开始
4. Time //时间信息
5. \t Complete: \t XXXXXXXX     //完整的时间
6. \t Year: \t XXXX年      //年份
7. \t Imprecise\_month?: \t 年[末底终初中]      //不精确的关于月份的描述
8. \t Month: \t XX月 (左右)?       //月份
9. \t Season?: \t [春夏秋冬]季?      //季节
10. \t Imprecise\_day?: (\t [上中下]旬)? (\t 月[末底终初中])? (\t 的一天)?     //不精确的关于日的描述
11. \t Day: \t XX日 (\t (左右|前后)(的一天)?)?      //日
12. \t Imprecise\_hour?: \t (凌晨|早晨|早上|晚上|傍晚|上午|中午|下午|深夜|半夜|夜间|夜晚|夜里|夜|早|中|晚)     //不精确的关于时刻的描述
13. \t Hour?: \t XX时(\t (左右|前后|许))?     //时
14. \t Minute?: \t XX分(\t (左右|前后|许))?     //分
15. \t Second?: \t XX秒(\t (左右|前后|许))?     //秒
16. Criminal //罪犯、被告人信息
17. \t 被告人1 \t 被告人2 ……
18. Location
19. \t XXXXXXXXXXXXX
20. Item //物品信息，Item可能出现多次，表示偷盗物品不止一件
21. \t Modifier?: \t XXX的 \t XXX的 \t XXX的 ……//一些定语、修饰语
22. \t Name: \t 物品名 //物品的名称
23. \t Value?: \t XXX元 //物品的价值
24. \t Color?: \t XX色
25. \t Brand?: \t XXXXX牌
26. \t Type?: \t XXXXX(型号?|款)
27. \t Amount?: \t XX //数量，数字+量词
28. Item //物品信息
29. ……
30. Value\_Supplement?
31. \t XXX元 \t XXX元 \t XXX元 …… //关于价值信息的补充
32. ################ //标志着本案件中下一个作案经过信息的开始
33. ……

**补充说明：**

Item本身可以是现金，这个时候一般只包含Name属性，属性取值即为 [数字]+元，数字可以是阿拉伯数字，也可以是中文数字，如果是阿拉伯数字，可能还会出现以逗号分隔三个数字的情况。Value\_Supplement中的钱也是一样的格式。

Amount中的数字也可以是阿拉伯数字和中文数字，量词包括 桶、套、付、对、组、部、只、个、包、辆、袋、件、块、枚、捆、瓶、箱、斤、条、K、k、根、米、盒、张、吨、副、升、位、匹、头、颗、棵、片、株、粒、朵、份、把、顶、架、支、幅、道、面、发、门、台等，还可能是公斤、公升等。

原文中如果出现当年、同年、当月等信息，以尽量替换为了最近出现过的时间，但不保证正确。

**存在的问题：**

不可避免的，采用正则表达式进行信息抽取，很难涵盖所有的情况。大部分文书的信息抽取结果良好，但总是有部分文书存在或多或少的问题。例如，如果定语和物品较为紧密，没有“的”、“牌”等指示词进行分隔，就难以抽取出来。