两个目录：

src：正式代码

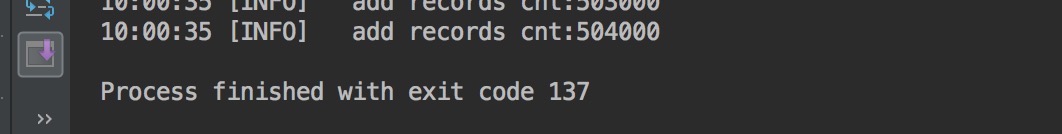
test目录：每个功能有简单测试，测试数据

每个代码至少一个人review

其他：

1. 面向对象
2. 变量名命名规范： 以下划线连接每个字符，不用大写，cal\_tf\_idf ， 尽量不出现任何warn
3. 编码统一为utf-8
4. sklearn是Python的，我可能python多点，如果是GBDT，用xgboost，可能c++
5. 适当打log日志，log是否真正输出要容易控制，python的日志见下：

|  |
| --- |
| 1. **def log\_debug**(s):  **if** log\_level == 1:  **print** time.strftime("%H:%M:%S",time.localtime(time.time())) +\  " [DEBUG] " + s |

也就是有log级别，以及格式是：

打印日志时间 + log级别 + log信息

数据量大的时候，建议每隔1000或者10000条数据打印一次信息，以确定运行进度

6、代码版本维护：0.1开始，以后每次更新增加0.1