| 1. chat.txt | | |
| --- | --- | --- |
| **数据字段** | **字段含义** | **备注** |
| 会话编号 | 会话编号 |  |
| 顾客ID（已脱敏） | 顾客ID |  |
| 是否顾客发送 | 0: 顾客发送； 1： 客服发送 |  |
| 是否转出 | 是否转给其他客服接待； 0： 该会话未转给其他客服接待； 1： 该会话转接给其他客服接待 |  |
| 是否重复咨询 | 当天内本通会话是否重复咨询； 0： 当天内第一次咨询; 1: 当天内非第一次咨询 |  |
| SKU | 商品编号 |  |

第一步：选择前三个会话进行实验

问题1：读取文件时编码乱码

解决1：vscode终端编码的问题：向settings.json添加

"terminal.integrated.shellArgs.windows": ["-NoExit", "/c", "chcp 65001"]

1. 特殊字符转换

"#E-s[数字x]": "(微笑)",

"#E-j[数字x]": "(愤怒)",

"&nbsp;": " ",

"[数字x]%": "(比例)",

"[金额x]%": "(比例)",

"%": " ",

"#": " ",

"&": " ",

1. 分割会话

问题2：分割不准确，一句话在上个会话中

解决2：添加分隔符的位置不正确

问题3：编码问题：

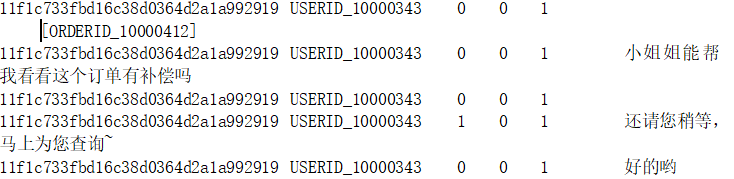
UnicodeEncodeError: 'gbk' codec can't encode character '\u30fb' in position 72: illegal multibyte sequence

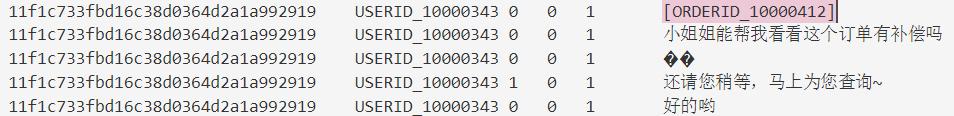
UnicodeEncodeError: 'gbk' codec can't encode character '\ue312' in position 56: illegal multibyte sequence

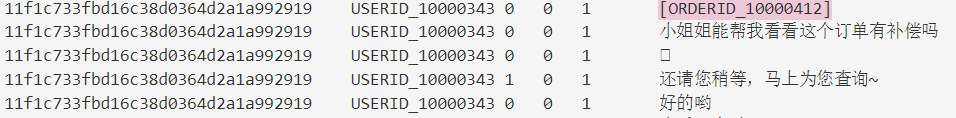
UnicodeEncodeError: 'gbk' codec can't encode character '\ufeff' in position 0: illegal multibyte sequence

换了几种编码方式转为中文编码：

（估计是表情之类的，需要做转码异常处理）







第二步：选择整个chat.txt文件进行实践

1. chat.txt做中文编码的异常处理

直接删掉！（一段会话，一个字符）

问题1：大文件读写，内存溢出，写入失败！

解决1：emmmm，不知道修改了哪里就好了

1. 分割会话、特殊字符转换，得到chat\_zh.txt



1. session，得到chat\_rd.txt

问题3：直接去重是否有影响？

解决3：先分为QAQAQAQ等的形式，再去重。



1. 不去重，直接提取关键字