实验 2-3 kettle 实现源数据的预处理

建议课时：60 分钟

一、 实验目的

 熟练使用 kettle 工具做数据预处理；

 掌握 kettle 中常用步骤的配置；

 能按清洗目标准确实现数据清洗工作；

二、 实验环境

Kettle7.0

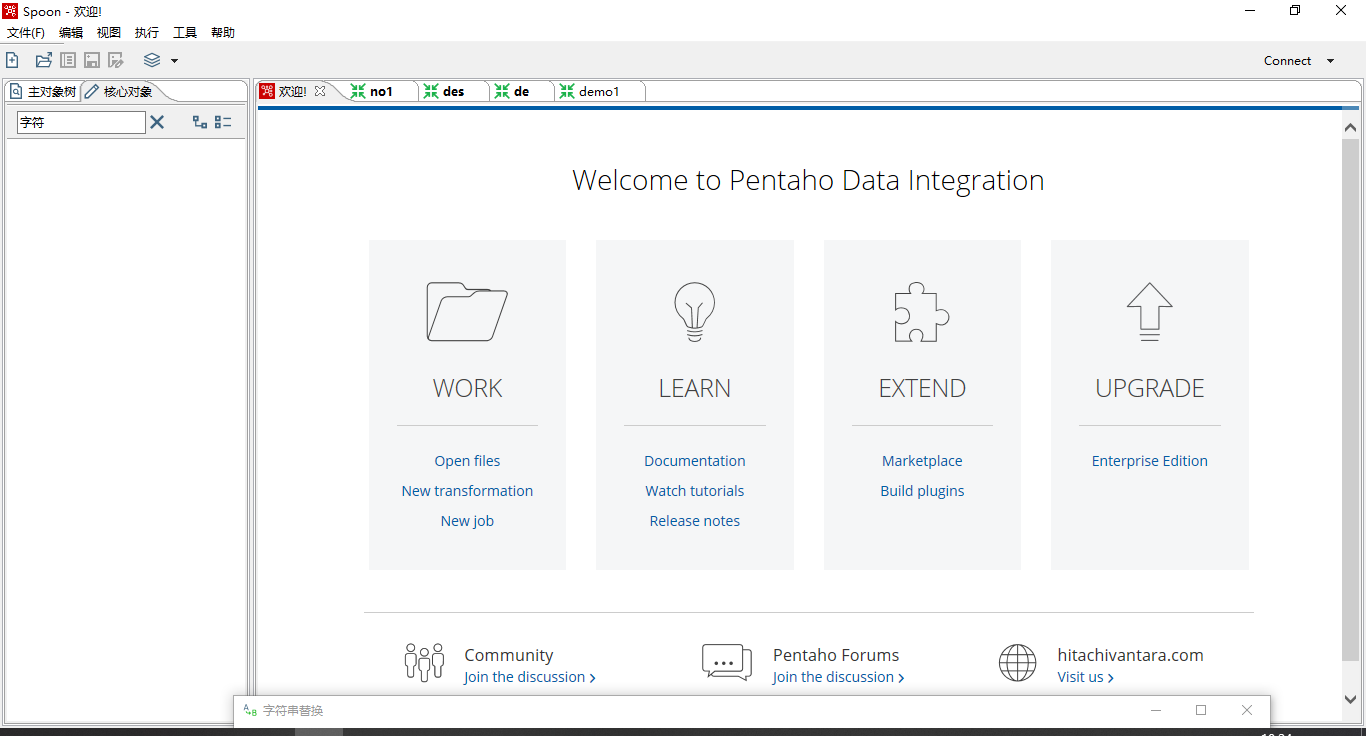
三、 实验步骤

本节实验主要是通过 kettle 实现源数据的预处理。

具体实验步骤如下：

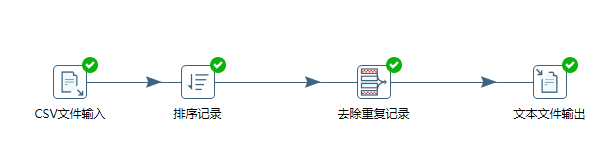
1. 下载安装 kettle7.0 工具

2. 运行 Spoon.bat 文件启动 kettle



3. 新建转换去除手机销售信息表的重复记录

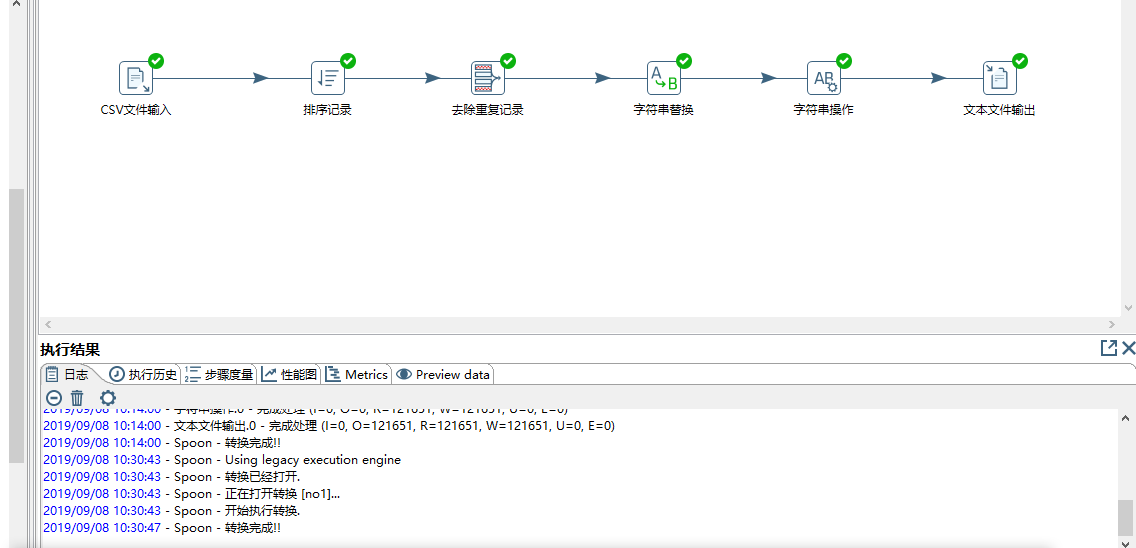
首先导入目标文件，之后进行排序，排序之后才能清洗（去重），之后导出文件：



**(操作流程图)**

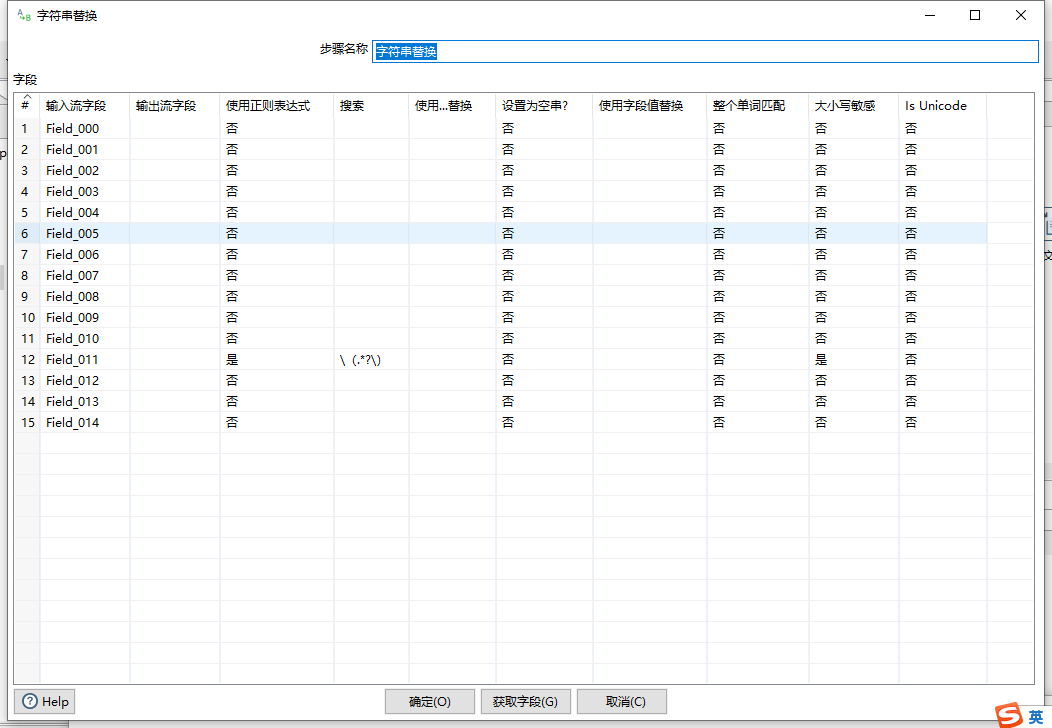
4. 在上述转换的基础上清洗手机销售信息表中型号字段中的数据

要求：去除该字段中的所有空格，方便后续聚合统计，字母统一大小写，去除该字段中的所有特殊字符（各种标点符号）



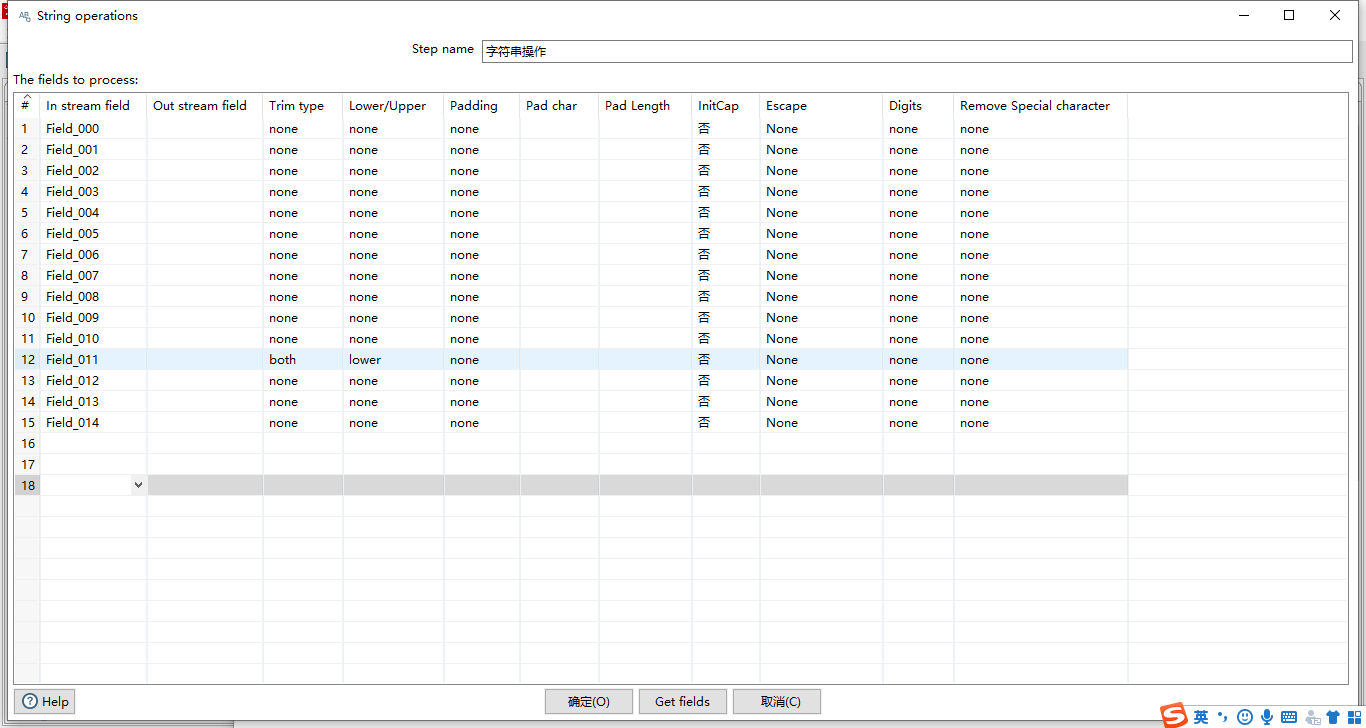
**(操作流程图)**

思路：因为该列字段有的位置存在品牌英文，有的没有，所以直接去掉括号里的英文，使数据保持一致。使用正则表达式搜索目标字段之后进行替换。



**（操作步骤）**

对于英文的品牌直接进行大小写统一（字符转换）：

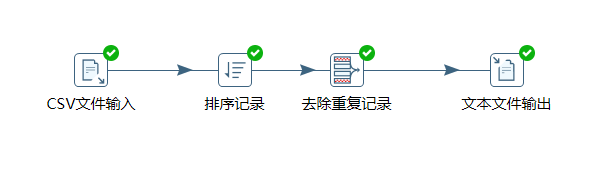


**（操作步骤）**

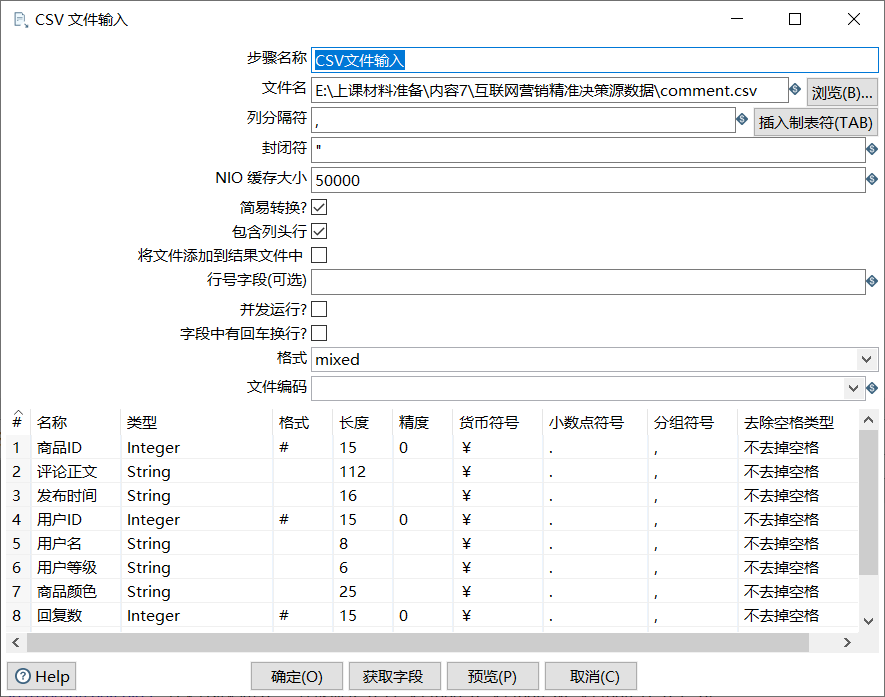
5. 新建转换去除用户评论信息表的重复记录

**步骤如下：**

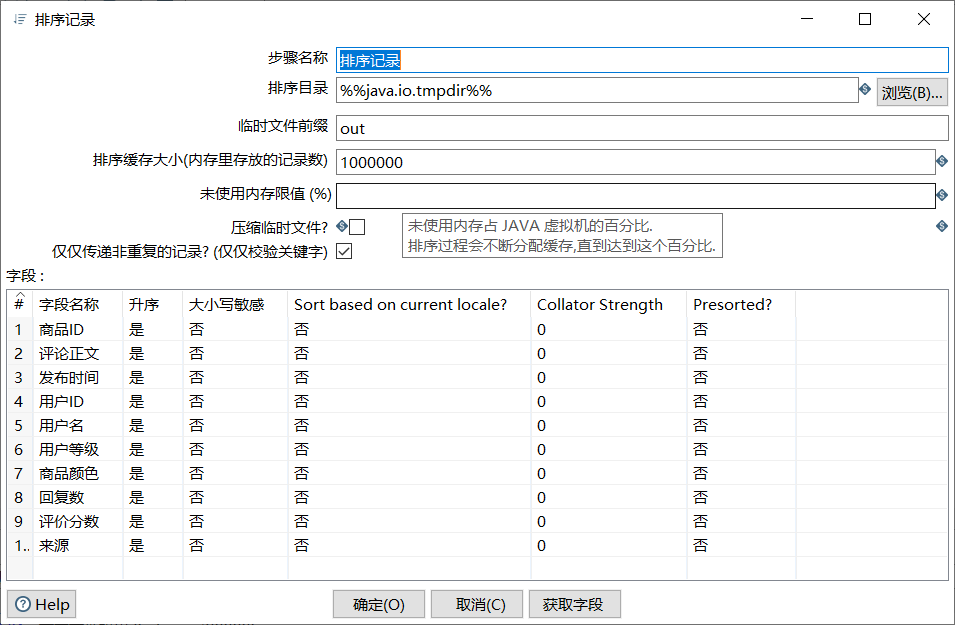
（1）首先先按照如图所示建立好连接顺序



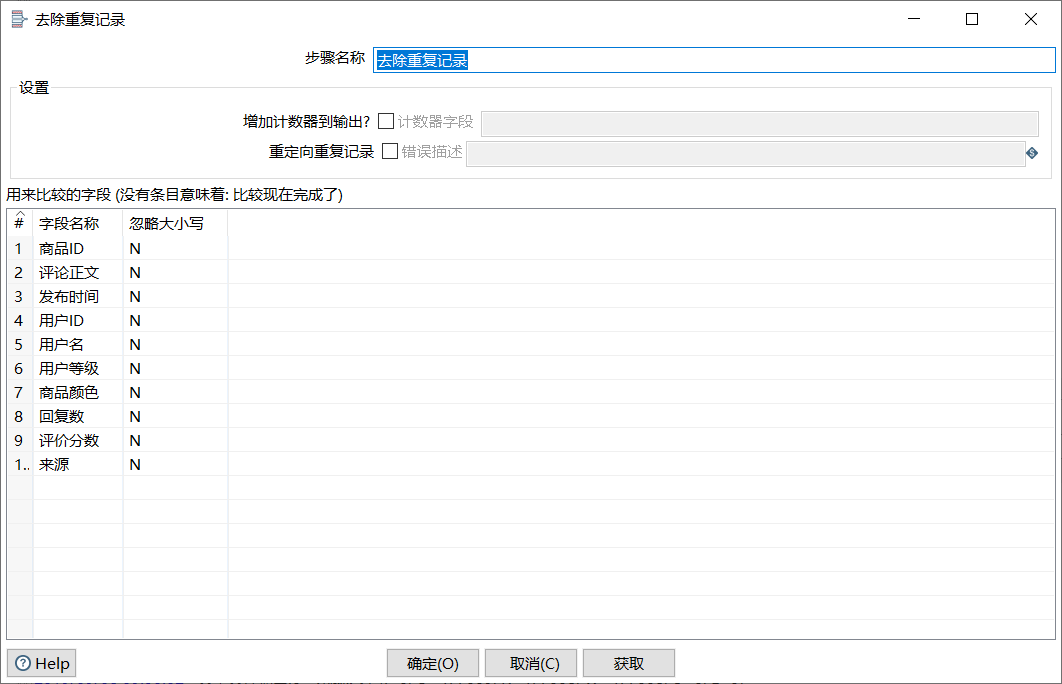
（2）在csv文件输入中选择要进行操作的.csv文件



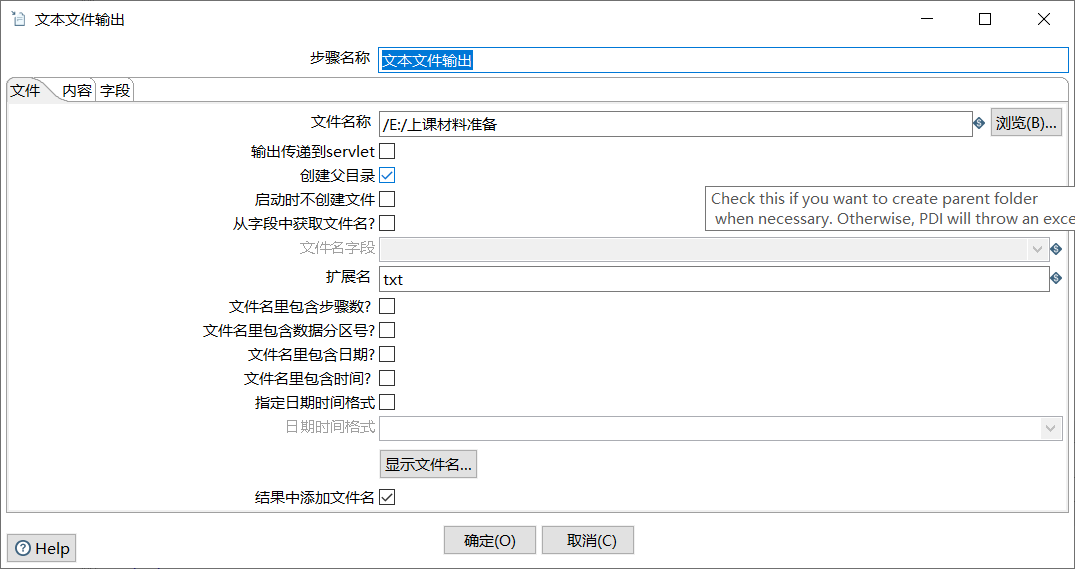
（3）在排序记录中获取字段



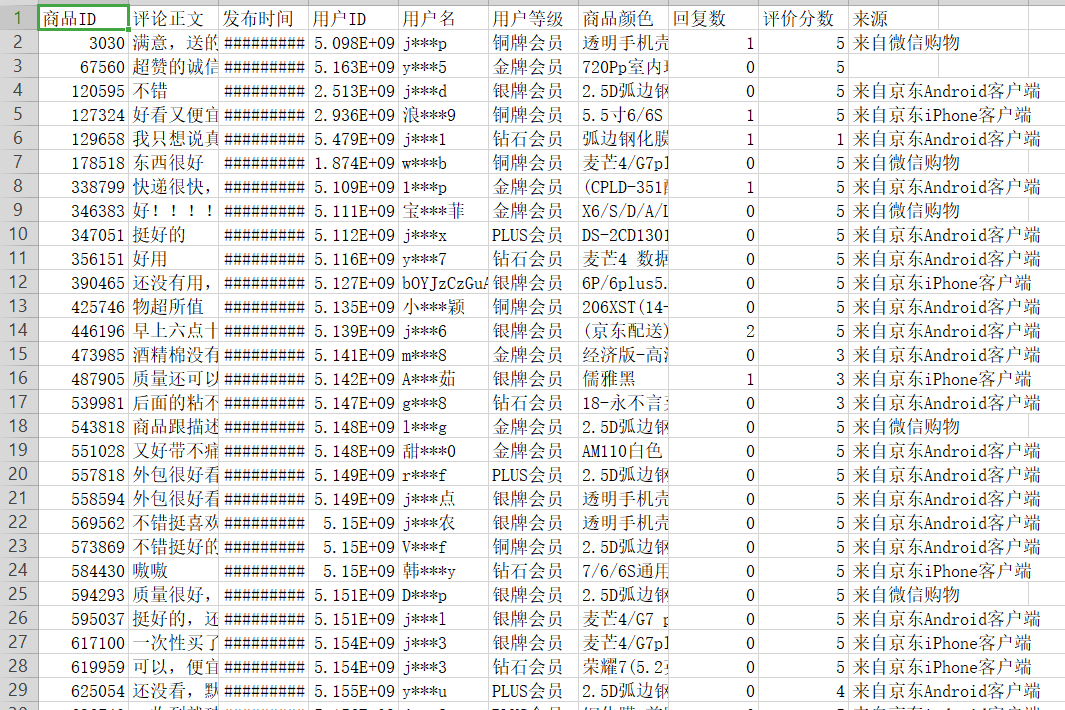
（4）在去除重复记录中获取字段



（5）在文本文件输出中选择输出的路径，进行输出



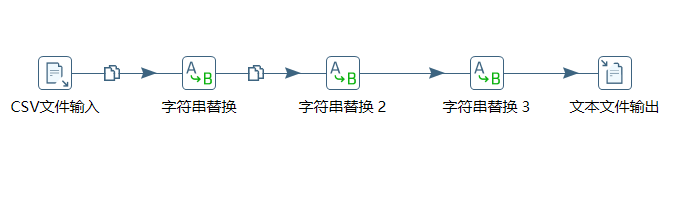
（6）转换成csv文件，生成的最终结果如下：



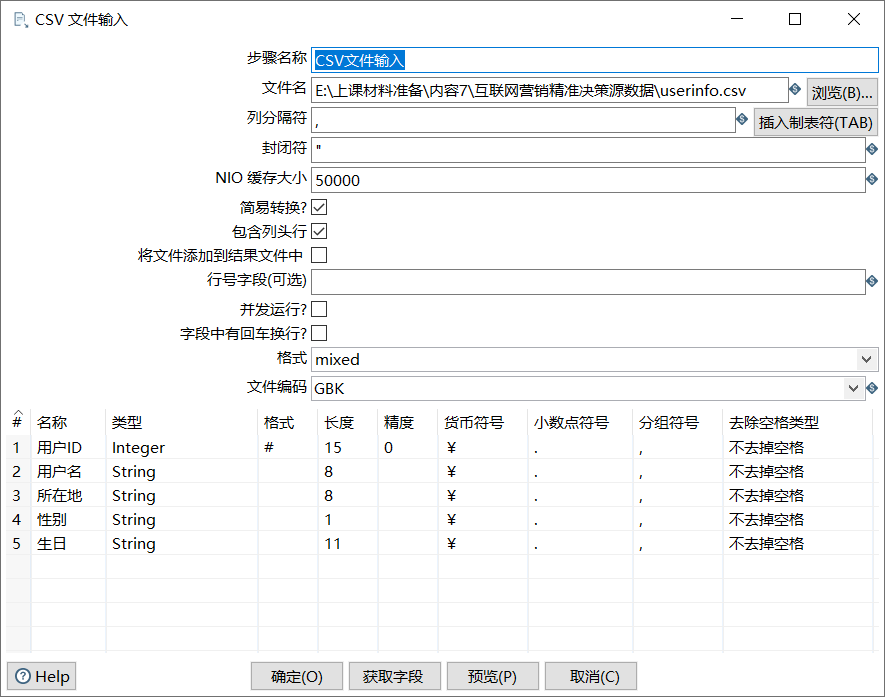
6. 新建转换处理用户信息表中出生日期字段（将 2019 年 5 月 20 日转换为2019-5-20）

**步骤如下：**

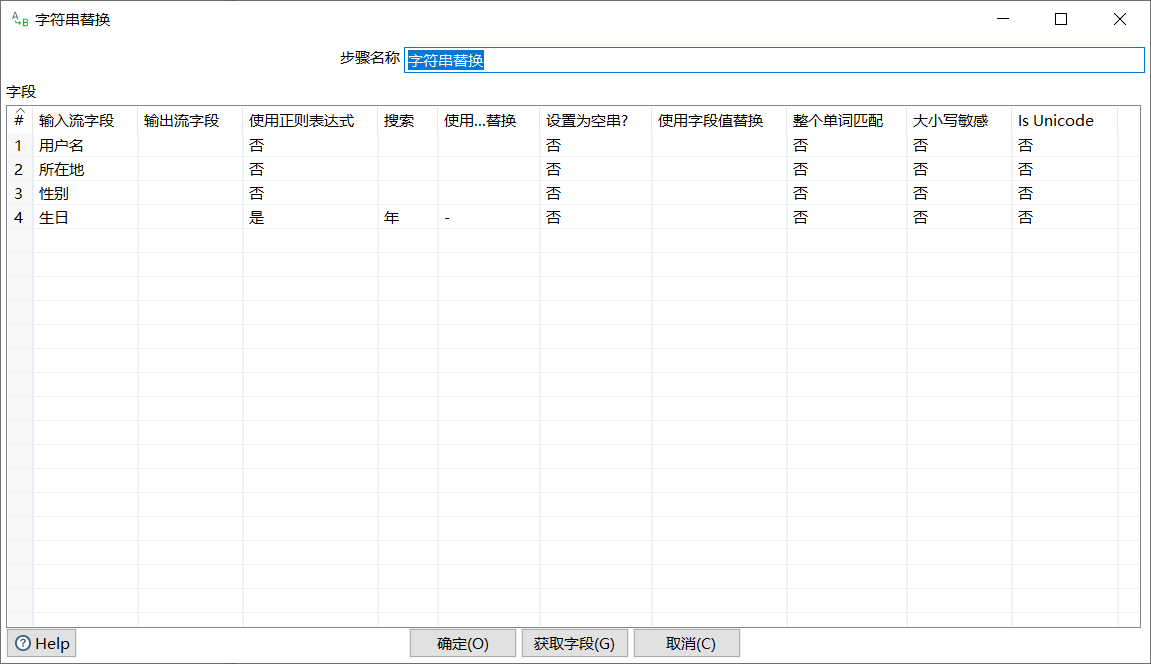
（1）首先先按照如图所示建立好连接顺序

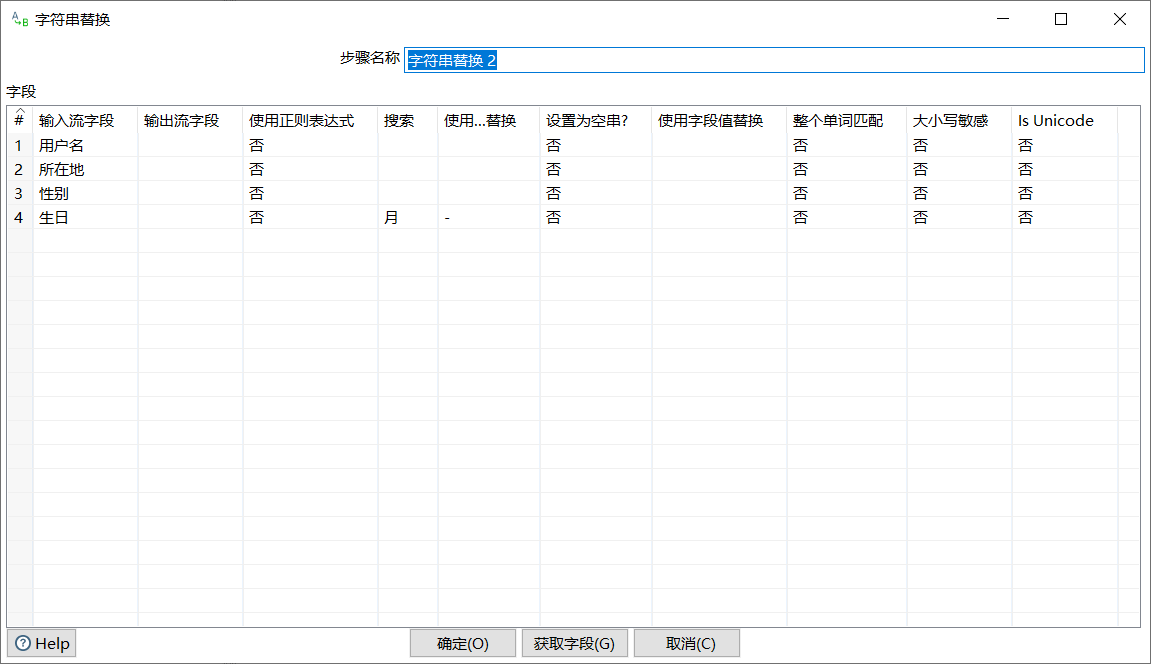


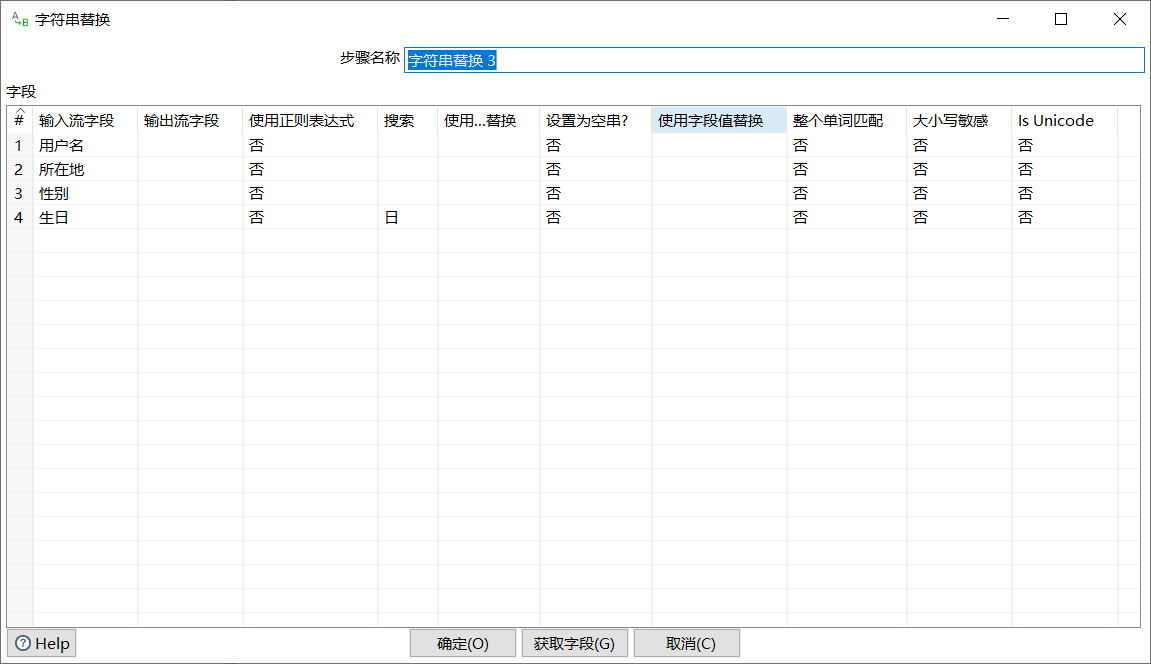
1. 在csv文件输入中选择要进行操作的.csv文件



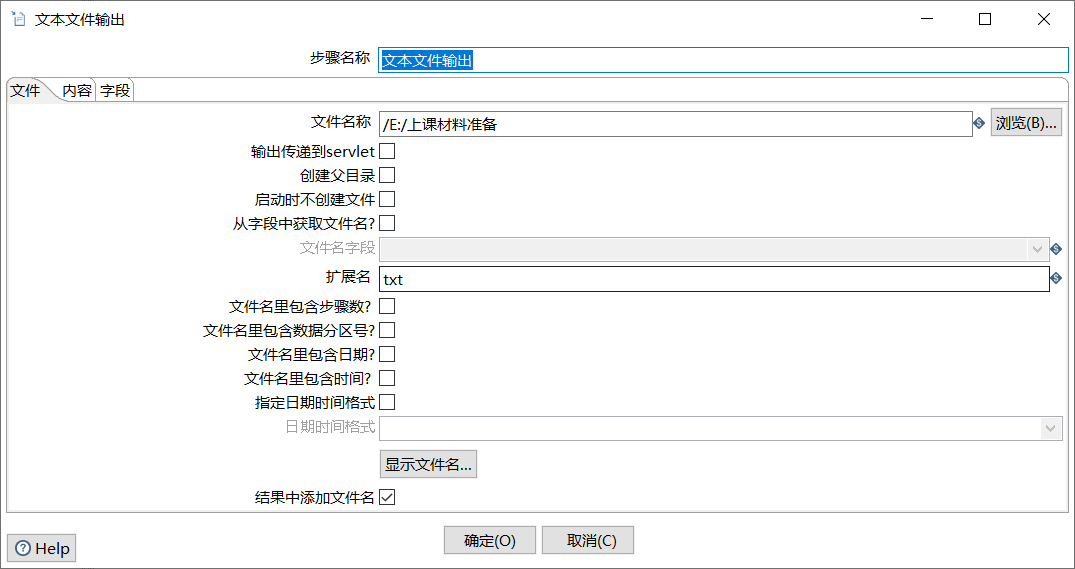
（3）建立3个替换将生日列中的年，月，日分别进行替换，将年，月替换成“-”，将日替换成空



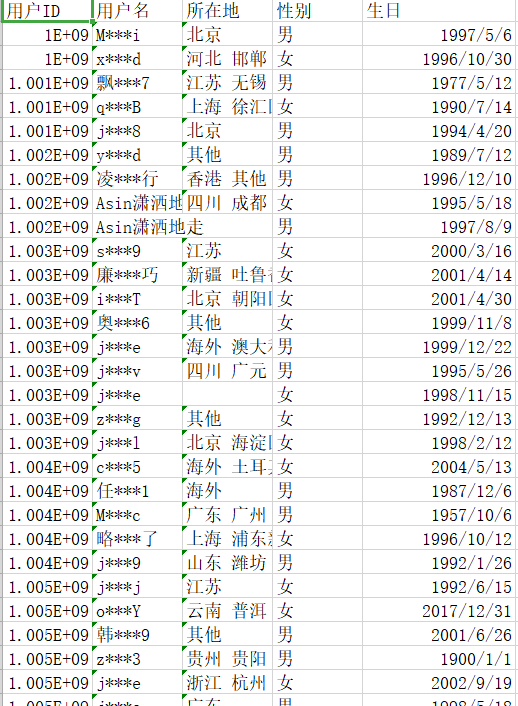




（4）在文本文件输出中选择输出的路径，进行输出

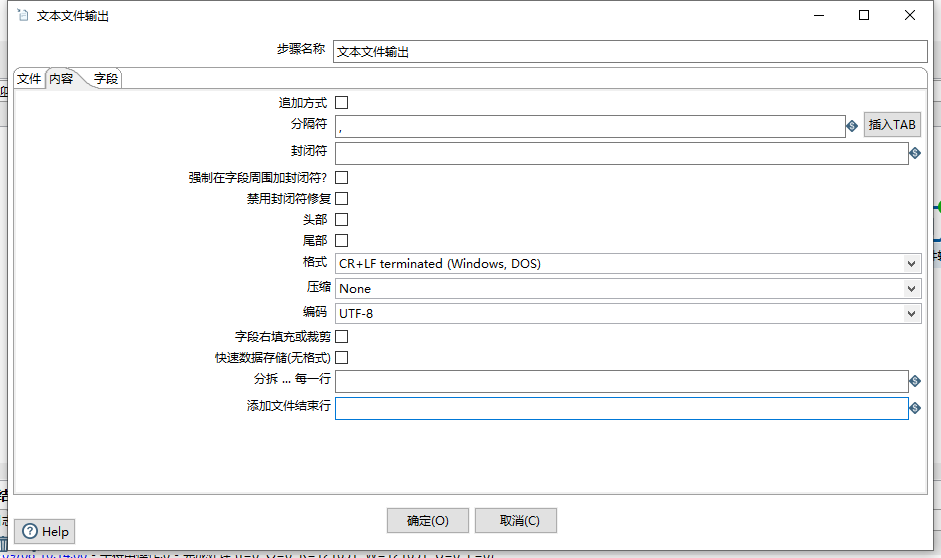


（5）转话成csv文件，输出结果如下：



7. 在每个转换中的文本文件输出步骤中，指定分隔符为英文状态下的逗号，去除输出文件的表头数据，编码格式保持一致，最终的编码格式为 utf-8。

输出：



**（操作步骤）**

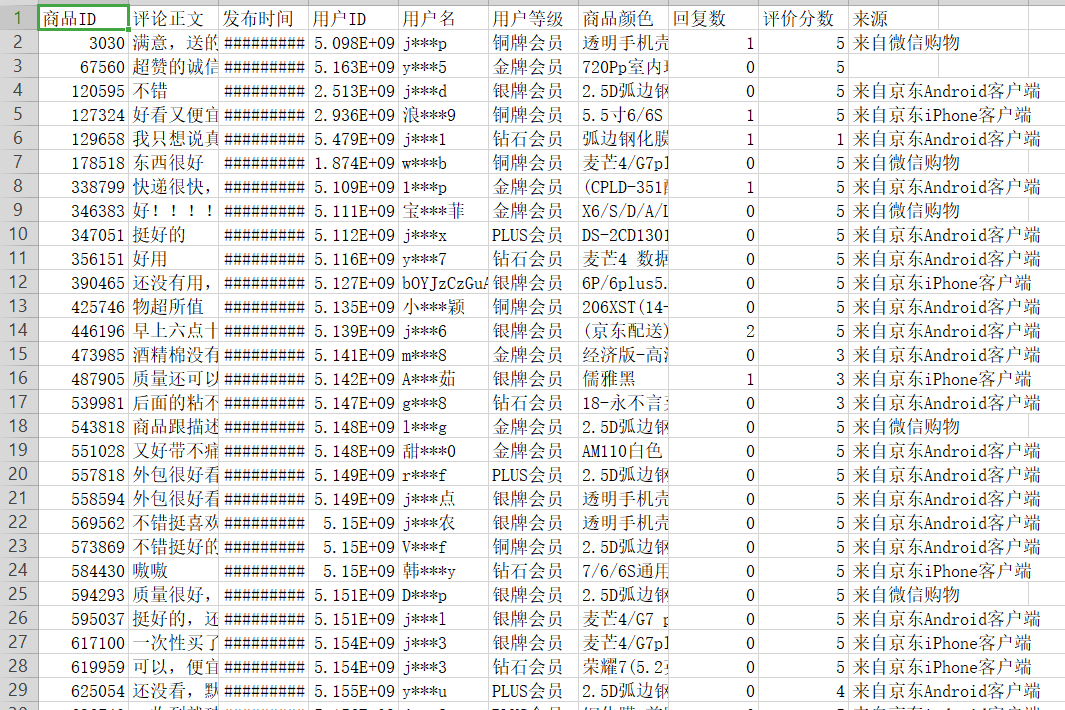
四、 实验成果

本次实验完成后，需要得到以下结果：

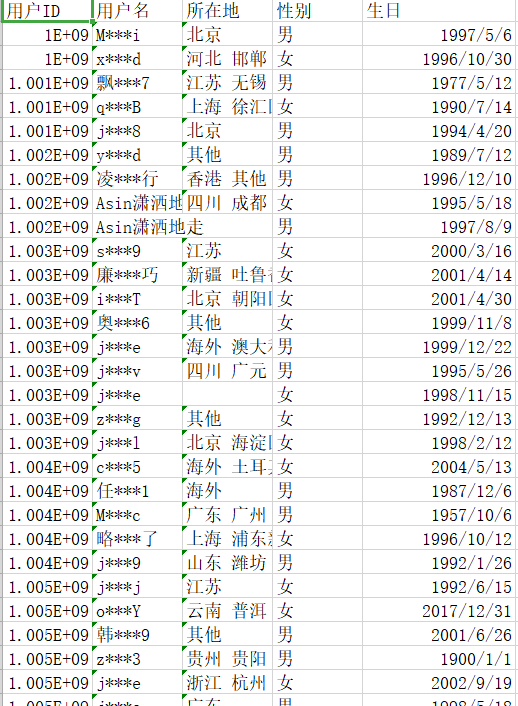
 去除手机销售信息表的重复记录；



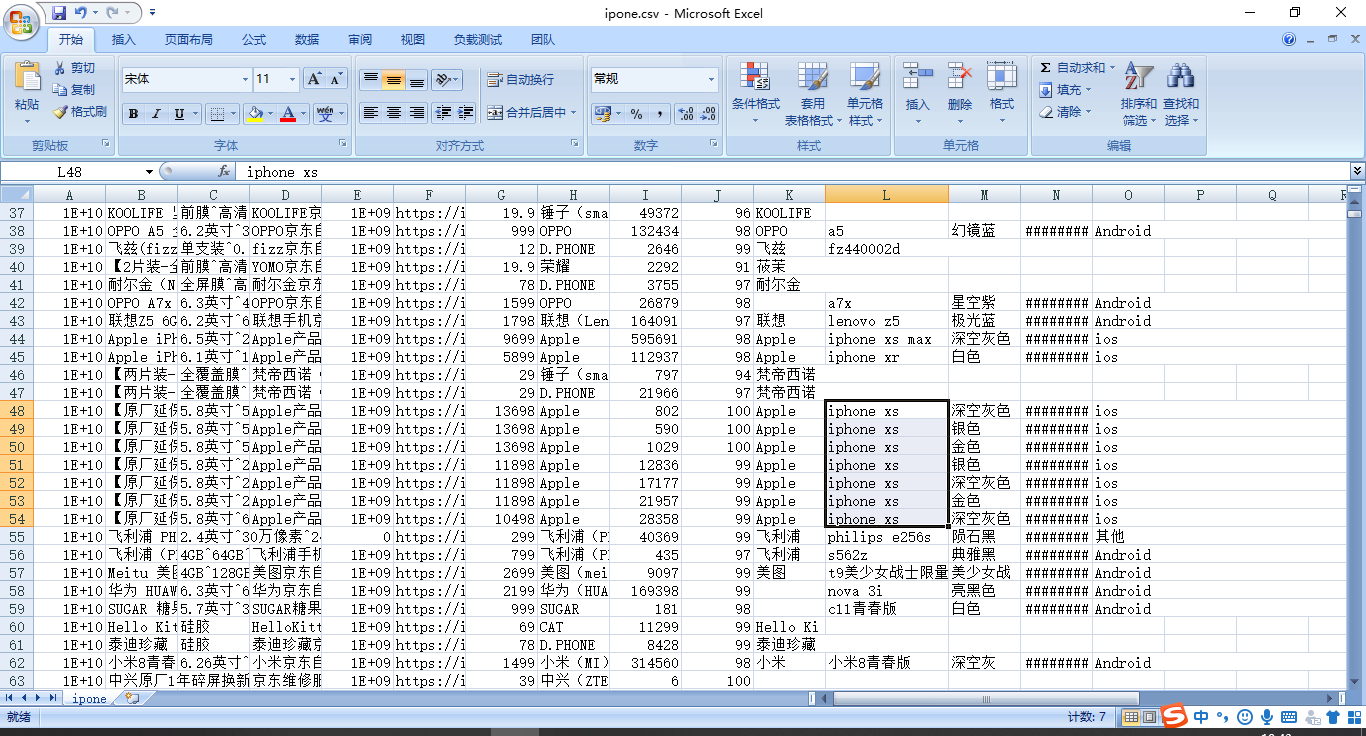
 去除用户评论表的重复记录；



年月份格式一致；



 清洗手机销售信息表中型号字段数据（同一型号数据格式保持一致）；



**（结果截图：无括号，全部小写，格式一致）**