修订日志：

|  |  |  |
| --- | --- | --- |
| 修改日期 | 修改人 | 备注 |
| 2019/11/17 | 刘晨旭 | 添加部分内容 |
| 11.17 | 王文祥 | 数据格式化处理 |
|  |  |  |
|  |  |  |
|  |  |  |

**数据格式化处理**

数据被爬取到本地后，应对数据进行格式化处理和保存，以便于其他模块保存和使用。

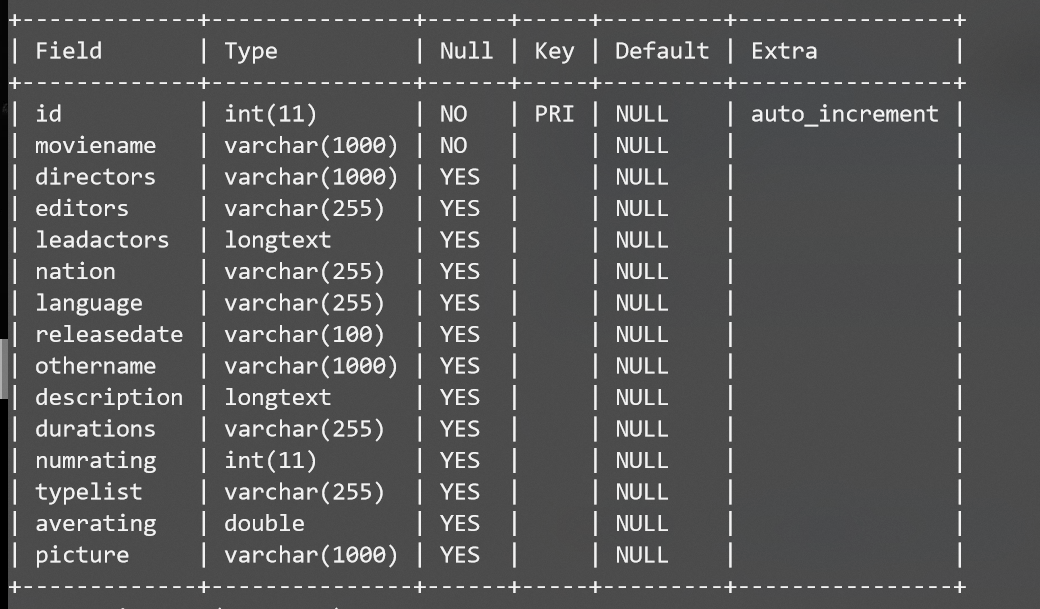
在开发过程中，每一条电影信息数据在流通的过程中都被按照片名、导演、编剧、主演、类型、制片国家/地区、语言、上映日期、片长、又名、简介、豆瓣id、平均评论得分、评论总人数、封面url保存，每一条评论数据在流通的过程中都被按照点赞数、用户链接、用户名、是否看过、推荐等级、打分、评论日期、评论内容、豆瓣id、电影名称的属性保存，并被相继保存为csv、json格式，以便于其他模块读取。

在数据的保存和入库过程中，我们遇到了一些问题。例如，有些外国演员的人名中包含有单引号（如O'Hara），有些电影简介和评论中包含有双引号，这就使得我们在入库语句的内容中不能直接使用单引号或双引号进行入库。针对这种情况，我们在入库语句中加入转义字符，这样入库语句在执行时就不会将其识别为含有特殊意义的元字符，而是认为它是一般字符，便可以顺利完成入库。

**原始数据到MySQL**

前一阶段爬虫爬到的数据是使用csv或json存储的文件，我们需要将其存储到MySQL数据库中。

以movie信息为例，由于数据爬取是在Windows操作系统中完成的，得到了一个csv文件，编码为‘gbk’编码，很不利与后续处理，需要将其转化为utf-8编码格式。尝试了多种方式之后，我们最后选择了一种简单的方式，通过Windows记事本的另存为utf-8的功能完成这个工作。观察该csv文件，我们建立了如下图所示的数据表



然后使用Python代码逐一读取csv文件，将每一行翻译为一条SQL插入语句，存入insert\_movie.sql文件中。在MySQL shell中，执行source insert\_movie.sql，将数据全部插入。

**MySQL到Hbase**

由于我们的系统部分功能需要使用Hbase数据库，需要将数据从MySQl导入到Hbase数据库中。我们使用Sqoop工具完成这一任务。

Sqoop是一款开源的工具，主要用于在Hadoop(Hive)与传统的数据库(mysql、postgresql...)间进行数据的传递，可以将一个关系型数据库（例如 ： MySQL ,Oracle ,Postgres等）中的数据导进到Hadoop的HDFS中，也可以将HDFS的数据导进到关系型数据库中。

以movie表为例，在hbase shell中先创建一个movie表，然后运行命令。

sqoop import --connect jdbc:mysql://47.92.101.31:3306/liuchenxu\_test?zeroDateTimeBehavior=CONVERT\_TO\_NULL --username 'liuchenxu' --password 'Cloud302@' --table 'movie' --hbase-table 'movie' --hbase-row-key 'id' --column-family 'region'

这样就将MySQL中movie数据表导入到了Hbase中，Hbase中movie的行键为movie的id。其他数据表类似。