数据预处理

预处理完的csv文件保存在 neo4j - community 中import文件夹。

1. \n的去除（person.csv, movie.csv）

因为是sql导出的csv，带有\n，在可视化中需要去除）

2. " - " 变成“ ,” （person.csv）

Hong Kong - China
Hong Kong
Hong Kong, China

3. 清理重复数据（movie.csv, person.csv）

这一步非常重要，因为数据集中的结点是assert成unique的。

wps / excel点击删除重复行清理即可。

4. 创建问题模版的表格（question_model.csv）

id	问题类型		例子
Person2Movie	演员的作品		xx演员演了哪些电影?
Movie2Person	电影的演员		xx电影的演员有哪些?
Movie2Genre	电影的风格		xx电影是什么风格?
MovieItem.Rate	电影的评分		xx电影评分是多少?
Person2Movie,Movie2Genre	演员的电影的类型		xx演员演过哪些类型的电影?
MovieItem.releasedate	电影的上映时间		xx电影的上映时间是?
a: Person2Movie ,b:Person2Movie	两个演员共同参演的电影		A演员和B演员合作过哪些电影?
Person2Movie, len(list)	演员的电影总数		xx演员出演过多少部电影?

5. movie.csv 对rating进行清洗，发现有文字数据，替换为0。

0代表此电影暂无评分。