## 数据的获取

打开网页

点击翻页

循环翻页

点击翻页

提取数据

循环

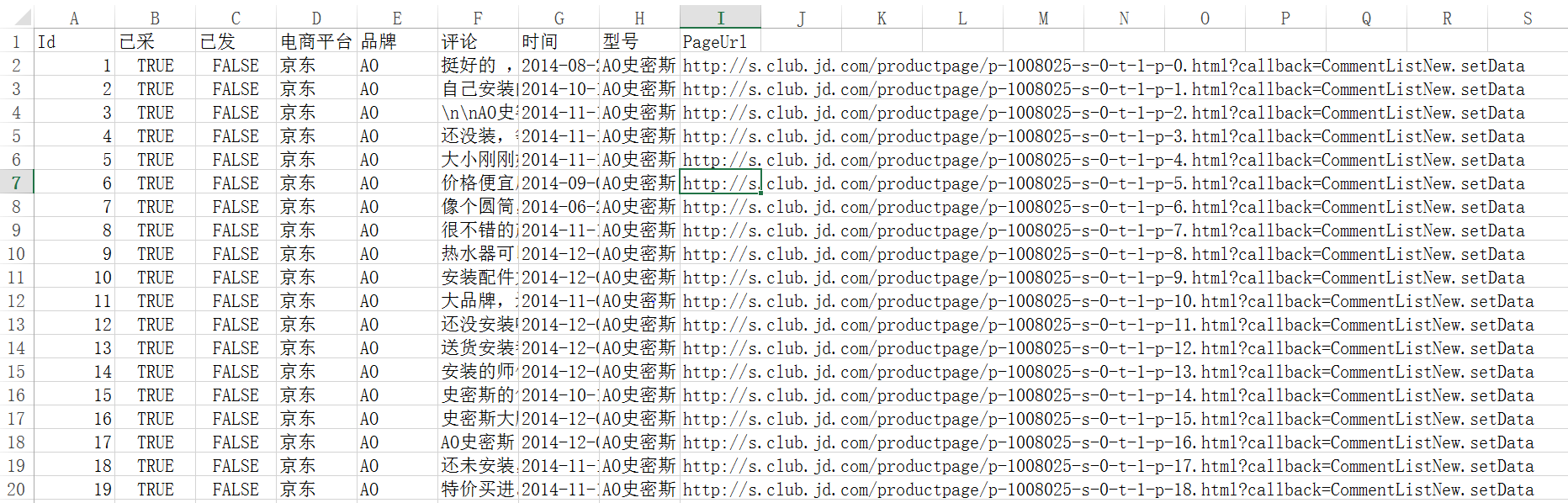
点击元素

循环

循环翻页

爬取数据流程设计

使用八爪鱼爬虫软件从京东上爬取数据，共获取20多万条有用的评论，并将数据保存在EXCEL中，获取的数据如下所示：



## 数据预处理

获取的数据中对我们最有用的是其中的评论，但在对评论进行分析之前先要对其进行预处理。

预处理主要分为三个步骤：

1. 文本去重

文本去重就是要去除文本评论中重复的部分，很多去重方法都存在着缺陷，这里采取最简单的去重方法，对那些完全重复的预料直接采用最简单的比较删除法，也就是两两对比，完全相同就取出的方法

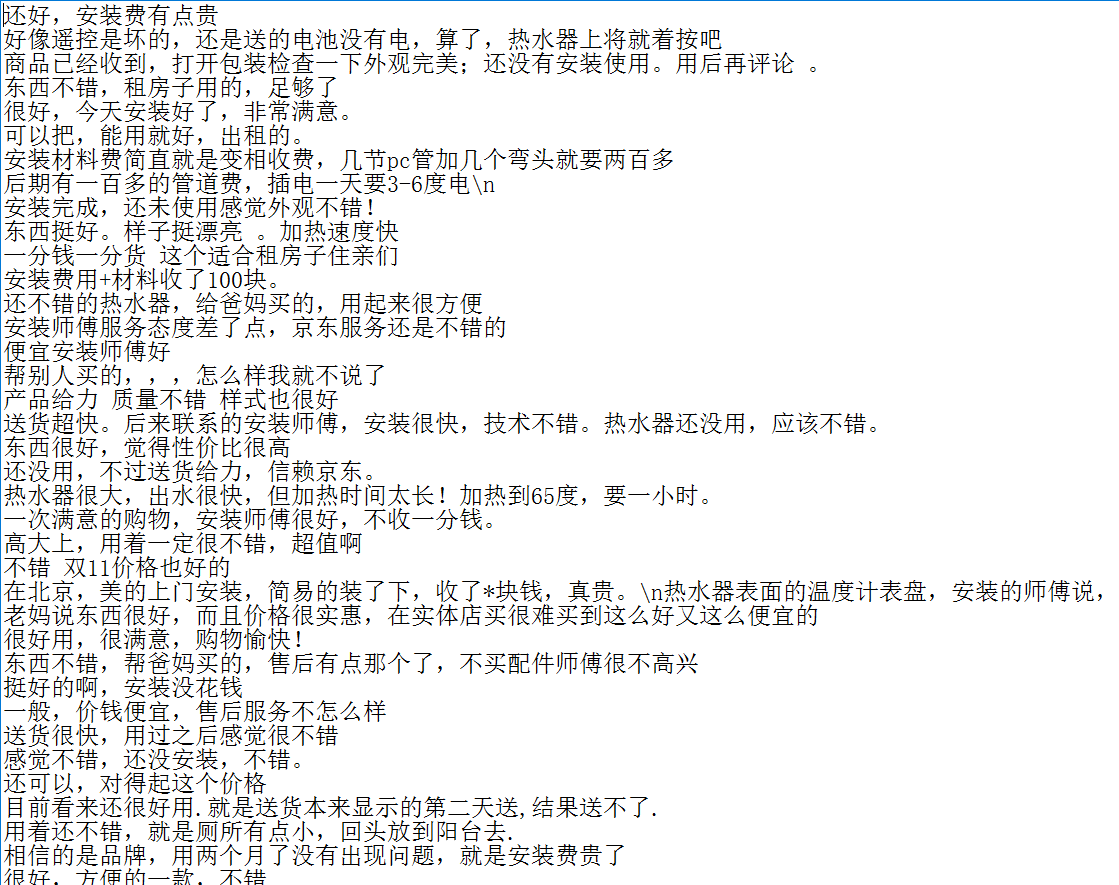
1. 机械压缩去词

有很多评论存在连续重复的语料，比如：好呀好呀好呀好呀，像这样的评论就要采取机械去词的方法，将评论变的更为有意义。此处采取对开头和结尾的连续重复进行压缩去词的处理。

1. 短句删除

经过机械压缩去词处理后，进行最后的预处理步骤：短句删除。因为很多字数过少的评论本身就没有什么意义，比如：很不错、质量差等评论，为此我们采用7个国际字符为下限，对评论进行短句删除。

处理后的评论如下所示：



## 进一步工作

前期的数据准备和预处理工作已经基本完成，后面我们会对之前获得的这些数据建模和诊断，通过情感倾向分析、语义网络分析、LDA主题分析几个步骤对评论进行分析，以期能完成前期的目标。