**京东商品评论数据采集程序设计方案**

# 一、采集方案设计思路

本次数据采集的目的是获取京东网站上关于颐莲喷雾用户评论数据并对其进行可视化分析。采集京东商品平台上的颐莲喷雾产品评论数据可以分为以下步骤：

（1）分析网站结构：首先，需要分析京东商品评论页面的结构，确定评论数据位于AJAX响应中，由JavaScript动态渲染，这可能需要使用浏览器开发者工具来检查AJAX请求和响应。

（2）模拟请求：使用Python的网络请求库（例如Requests）来模拟发送请求到评论页面的AJAX端点，获取评论数据的JSON响应。

（3）数据提取：从JSON响应中提取评论数据，包括评论文本、用户信息、评分等等。

（4）翻页处理：评论数据一般分页显示，因此需要处理翻页以获取更多评论数据。

（5）存储数据：将采集到的数据存储在数据库中，以便后续分析和可视化。

（6）可视化：创建一个可视化界面或报告，以直观展示数据分析的结果。

# 二、程序运行流程图

程序运行的主要流程包括：程序启动并初始化必要的设置和变量；使用网络请求库（例如Requests）发送初始请求到京东商品评论页面的AJAX端点；接收来自AJAX端点的评论数据信息JSON响应； 解析JSON响应，提取评论数据，包括评论文本和地理位置信息；对采集的数据进行进一步的清洗与处理；最终采集到的评论数据被存储在数据库中，以便后续分析和可视化。

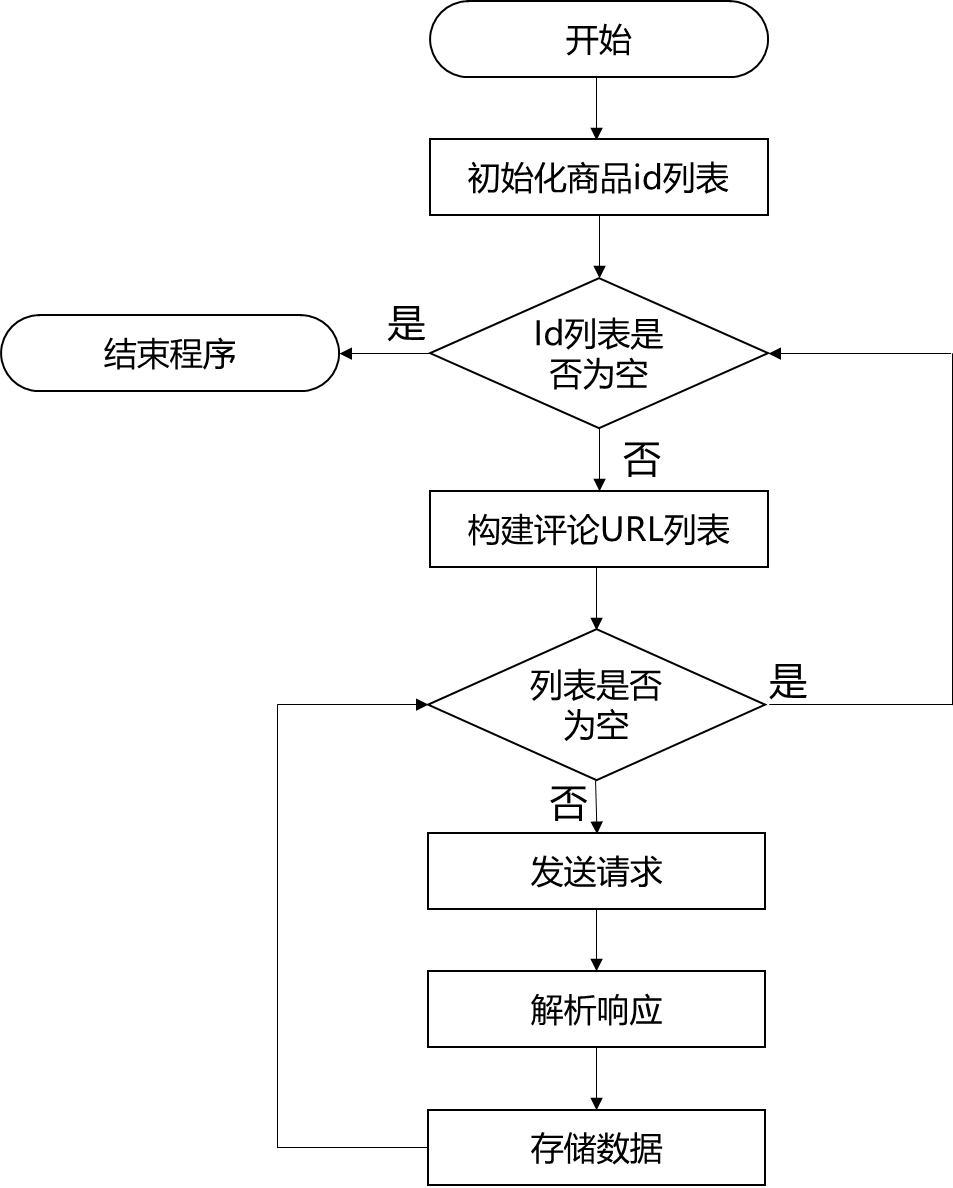


图 1 程序运行流程图

# 三、代码功能描述

## 1.函数入口设计

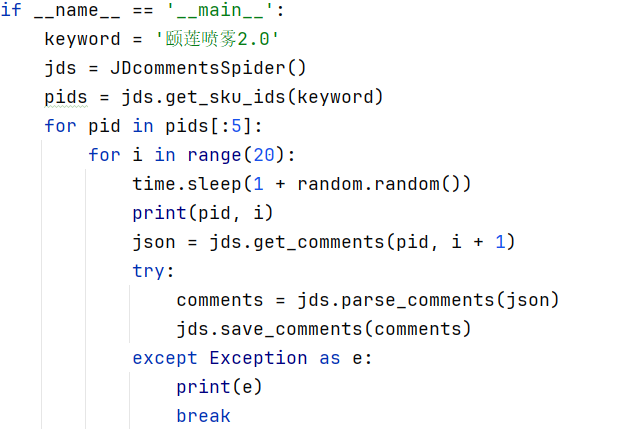


图 2 主函数入口说明

首先确定关键词，这里以“颐莲喷雾2.0”为例，调用get\_sku\_ids()方法访问商品列表页面，获取页面内所有sku-id用于评论爬取，由于每个商品下评论都超过作业要求300条数据，因此选择排名前五的商品，每个商品爬取前200条评论数据（考虑到去重空数据等因素），随后按页遍历获取商品评论数据，对数据进行解析后存储在MySQL数据库中。

为了提高程序的稳定性，防止服务器误将我们的代码当作爬虫程序，每次循环结束都添加了1-2秒的等待。

## 2.获取商品id信息

图形用户界面, 文本, 网站

描述已自动生成

图 3 获取商品id列表

通过对关键词的搜索可直接进入商品列表页面，使用parsel库提供的xml解析工具xpath语法提取页面内所有产品id，具体的xpath语法如上图所示，以带有id的div标签作为锚点，进一步获取其子代li标签的data-sku属性值。

## 3.获取评论数据

获取评论响应的请求时一个get请求，其中涉及关键参数时间戳、产品id和一个uuid，而参数uuid第3、4、5部分明显也是一个时间戳，其值逐渐增大，第一部分与第二部分不够清晰，于是全局搜索分析。

图形用户界面, 应用程序

描述已自动生成

图 4 请求get参数截图

最终关键参数uuid定位在cookie中，名称为\_\_jda与\_\_jdc的cookie键中包含了所需参数，在浏览器控制台中输入hook代码，如图 6所示，重新刷新页面定位到关键参数\_\_jda位置。

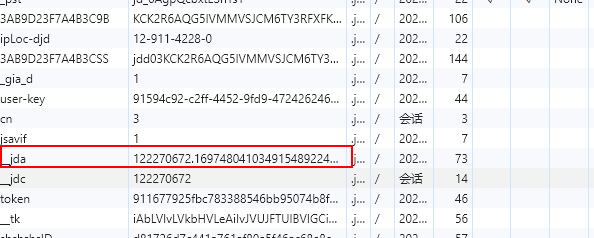


图 5 关键参数定位图1

文本

描述已自动生成

图 6 cookie-hook代码

分析下图 7可知，参数值由js对象生成，其中前两个参数分别对应其属性f与属性h。

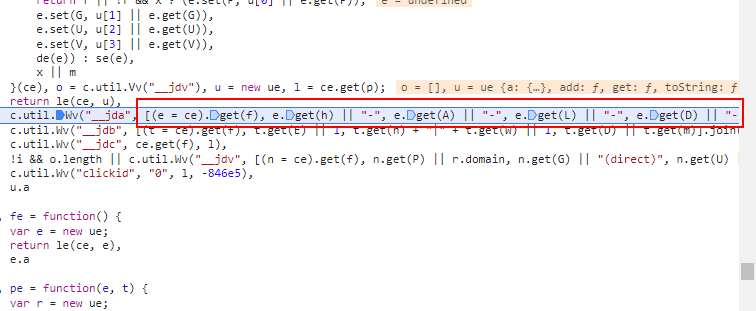


图 7 关键参数定位图2

向上寻找该对象发现属性h对应函数k()，属性f对应函数Q()。

文本

描述已自动生成

图 8 关键参数定位图3

在控制台直接输入可以看到，函数K()由时间戳和一个随机数组成，函数Q()等价于函数re(r.domain)，但是r.domain是定值jd.com，因此Q()函数将恒返回定值122270672，至此所有参数都可知。

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

图 9 关键参数定位图4

如下图 10所示构建了请求参数并发送请求。

图形用户界面, 应用程序, Word

描述已自动生成

图 10 发送请求方法图

## 4.解析响应

图片包含 文本

描述已自动生成

图 11 数据解析方法

从获取的json数据中提取我们需要的数据，先使用响应对象的json()方法将json对象转化为字典，在利用列表推导式创建评论数据列表，其中将时间字符串转化为了时间戳并使用正则表达式清洗了文本中的特殊字符，最终返回评论列表。

# 四、网页数据存储方案

最终将数据存储在MySQL数据库中，存储方法如下：

图形用户界面, 文本

描述已自动生成

图 12 数据存储方法

如上图 12 所示，在数据存储时对数据进行了筛选，去除了空的文本和地点数据，并且设置id为主键自增生成，评论文本内容不重复不为空。

# 五、可视化方法

读取用户评论文本，使用第三方库jieba对中文文本进行分词，并使用stylecloud对制作词云图对关键词进行可视化分析如下图 13 所示：

文本

描述已自动生成

图 13 评论文本可视化词云图

读取用户地理位置分布图，统计各省份用户评论量，使用python第三方库geopandas绘制地理分布可视化图。

地图

描述已自动生成

图 14 用户地理分布可视化图