**DC\_Crawler数据库设计说明书**

|  |  |  |
| --- | --- | --- |
| 版本 | 作者 | 修改日期 |
| 0.1 | 李全锋 |  |
|  |  |  |

# 1.说明

本系统采用MongoDB数据库存储。

# 2.数据库设计

## 2.1集合种类说明

项目数据存储在名为DC\_Crawler的数据库下，该数据库有以下集合：

|  |  |
| --- | --- |
| 集合名 | 说明 |
| product | 单个商品信息 |
| imgs | 图片信息 |
| category | 商品类目分类信息 |
|  |  |
|  |  |

## 2.2集合详细说明

下列各表中红色字段为必须字段，”\_id“字段为MongoDB默认生成,不作说明。

product集合详细字段

|  |  |  |  |
| --- | --- | --- | --- |
| 字段名 | 类型 | 说明 | 格式备注 |
| title | String | 商品名称 | “探路者情侣款冬季冲锋衣 男女三合一冲锋衣TAWB92604/TAWB91603” |
| price | Float | 价格 | 299.0 |
| shop | String | 所在商店 | “探路者官方旗舰店” |
| originSite | Dict | 从哪个网站上抓取来的 | {“domain”:”www.taobao.com”,  “title”:”淘宝”  } |
| category | List | 商品的类别，从大到小依次排序，每个元素为一个字典，存储了某一类别的\_id和name | [{“\_id”:xxx,”name”:“男装” },  {“\_id”:xxx,”name”:“户外运动”},  {“\_id”:xxx,”name”:“冲锋衣”},  …] |
| keyWords | Set | 该集合存储了所有描述该商品的关键字 | set([“图书”,”考研”,”英语”,”正品”,”包邮”,…]) |
| time | Dict | 时间信息,包括商品首次抓取时间，最后更新时间；  存储格式为时间戳(计算方式为time.time()\*1000) | {“create” : 1418728287685,  “update” : 1418728387623  } |
| isAvailable | Boolean | 是否有货 | True |
| evaluate | Int | 总评价数 | 288 |
| img | Dict | 商品图片信息 | {“$ref”:”imgs”,  “$id”:ObjectId  } |

imgs集合详细字段

|  |  |  |  |
| --- | --- | --- | --- |
| 字段名 | 类型 | 说明 | 格式备注 |
| url | String | 图片的url链接，（可能是本机url，也可能是云服务器上的，暂时为本地服务器） |  |
| type | Int | 图片的类型，不同值表示的意思：  0：代表product图片  1：其他类型图片 | 尺吋默认220x220 |
|  |  |  |  |

category集合详细字段

|  |  |  |  |
| --- | --- | --- | --- |
| 字段名 | 类型 | 说明 | 格式备注 |
| name | String | 该类目的名称 | “户外运动” |
| super | Dict | 上级目录信息,包括上级目录名称和ObjectId，如果没有上级该字段为None | {“\_id”:ObjectId,  “name”:”男装”  } |
| lower | List | 下级目录列表，包含所有下级目录的id和name字段,如果没有下级该字段为None | [{“\_id”:xxx,”name”:xxx},  {“\_id”:xxx,”name”:xxx},  {“\_id”:xxx,”name”:xxx},  …] |
| level | Int | 该类目的级别，最高级为0,次高级为1,依次递推。 |  |