Amazon数据基本介绍

我们提供的数据服务主要是围绕Amazon电商数据展开的。这里简要介绍一下Amazon电商数据:Amazon.com是全球最大的互联网零售商之一。最初Amazon从图书开始销售,之后零售涉及各类别商品。目前Amazon涉及图书、音像、电子产品等。在全球多个国家和地区有本地化 的Amazon.com。

在我们的数据服务中,主要针对获得Amazon商品数据进行统计分析等工作。围绕Amazon的商品信息展开一系列的工作。因此,一个固定的商品数据模型是必要的。在我们的数据分析任务中,根据需要的分析任务来获取特定的商品信息。目前针对单个商品, 我们需要的信息包括商品描述信息、商品价格、商品评论和卖家。

商品描述信息:针对该商品的详细信息,包括Amazon对商品的特征描述和介绍,商品名还有商品的ASIN码,ASIN码是Amazon用来唯一标识商品的凭证,从数据访问角度来看,给定一 个ASIN即可获得对应的商品信息。

商品价格:针对该商品的价格信息,价格信息包括在某一时间点上的该商品的所有价格信息,有Amazon官方价格,也有第三方卖家的价格。

商品评论:针对该商品的评论信息,每条评论包括评论者姓名、评论时间、评分、评论标题、评论内容等。评论会随着时间不断更新的。当然,对于一些冷门商品可能会没有评论。

卖家:记录在某一时间点上的该商品的所有卖家。除了Amazon官方自营外,可能还有第三方卖家 在销售商品,因此,记录商品对应的所有卖家,信息包括卖家名、卖家链接等。

ASIN:全称AmazonStandardIdentificationNumber,是一个由十个字符(或数字)组成的唯一标识号码,用于亚马逊上的产品标识。

提供的数据

目前我们提供了爬取的大部分数据，这些数据主要包括两类：
1)Amazon的商品数据，Commodity。
2)Amazon的商品的分类数据。 Category。

对于Commodity，访问的url为:http://112.124.1.3:8004/api/commodity，注：通过程序访问的默认返回格式是JSON，如果需要返回XML，可以加入一些参数，后续版本加入。这段请求将会返回最基本的Commodity中所有Category信息。

数据格式

提供的数据具有一定的结构，如下图所示：

```
{
        "ASIN": "",
        "productInfo": [{
                "productDetail": {
                        "UPC": "",
                        "bestSellerRank": ["#1 ['cate1','cate2',...]",""],
                        },
                "name": "",
                "img": "",
                "productDescription": "",
                "feature":["",""],
                "brand": {"link":"", "name":""},
                "timestamp": ""
                }],
        "seller": [
                {"timestamp": Date(),
                "seller":[{"name": "", "link": "", "img": ""}, {}]},
                {}...],
        "category": [
                ["root", "parent", "child"],
                []...]
        "offer": [{"info":[
                                {"seller":{"link":"","name":"","img":""},
                                "price":0,      #may be str '$1.0'
                                "timestamp": Date()},{}...],
                                "timestamp": Date()
                        },
                        {}...]
        "review": [{
                "star": 1,
                "helpRate": "3|15",
                "publishTime": "",
                "summary": "",
                "content": "",
                "consumer": "",
                "profileUrl": ""
                },{}...],
        "stats_info": {
                "keywords": [["word1", 3], #word and word frequency
                        ["word2", 2],...]
                "review_count": 1, #number of reviews
                "star_info": {1: 1,
                        2: 0,
                        3: 0,
                        4: 0,
                        5: 0}  #distribution of reviews' star
                "avg_info": 1 #average star of this commodity
        }
}
```

API说明

目前，我们提供了如下的几个API访问接口：

1）**Url**： http://112.124.1.3:8004/api/commodity/

说明：获取目前数据中的所有的分类，返回分类名列表 。

返回格式： [{'name': 'A>B>C'},{'name': 'D>E>F'}...] 其中 A，B，C等代表分类名，按照一级分类，二级分类排序。

2）**Url**： http://112.124.1.3:8004/api/commodity?category_name="'Shoes>Men>Work $Safety(&page=2&field=['ASIN'])

说明：返回指定分类的商品信息，注意必要参数是category_name，并且这里如果分类中有&符号的话，需要替换成$符号，以防和Http请求参数混淆。可选参数是 page, field，考虑到网络因素，这里一次请求返回20条数据，默认是按照商品评论条数降序返回第一页，所以，如果需要更多的后续数据的话，需要指定返回的页数。field代表了需要返回的属性.例如某次请求需要商品的价格信息。那么可以添加field参数 field=['offer']，这样访问速度会快点。

3）**Url**：http://112.124.1.3:8004/api/commodity/count/?category_name=Clothing%20$ %20Accessories%3EWomen%3EActive

说明：返回指定分类中所包含商品的数量，必要参数是 category_name， 并且这里如果分类中有&符号的话，需要替换成$，以防和Http请求参数混淆。

返回格式：{'category': 'A>B>C', 'count': 576} ，错误情况下为{'error': 'not valid'}。

4）**Url**： http://112.124.1.3:8004/api/commodity/B00AV5K7TI (?field=['offer'])

说明：获取某一个指定ASIN的商品信息。其中 B00AV5K7TI 代表了商品的ASIN码。ASIN是Amazon商品的唯一标示符，一个ASIN确定一个Amazon的商品。

5）**Url**：http://112.124.1.3:8004/api/commodity/field

说明：获取当前商品信息中可以使用的field值，即可以使用的商品属性值。

6）**Url**：http://112.124.1.3:8004/api/commodity/custom?query={}&ret={}(&page=2)

说明：自定义查询，这个需要了解MongoDB的查询语法之后才能熟悉，具体可以参考MongoDB的文档。query: 查询条件 ret: 需要返回的数据 可选：page: 返回指定页的20条数据。

数据访问

1.使用命令行访问API的url即可。

```
XuMM-12:~ macbook$  curl -i http://112.124.1.3:8004/api/commodity/
HTTP/1.1 200 OK
Server: nginx/1.4.1
Date: Fri, 21 Mar 2014 02:06:36 GMT
Content-Type: text/html; charset=utf-8
Content-Length: 6790
Connection: keep-alive
```

产生结果与程序访问相同。

2.程序访问

具体的程序访问可以参见 https://github.com/skymoney/Amazon_REST/tree/master/Demo，Demo中提供了Python和Java的两种接口。（实际上，支持JSON格式和Http请求的语言均可以通过该接口得到数据。）