AMAZON BestSeller商品抓取文档

一、主要修改点

1.新增spider: amazon\_bestseller.py

(1) bestseller入口：

lpush spider:targets:amz:bestseller <https://www.amazon.com/Best-Sellers/zgbs/>

(2) category解析逻辑：详见第二节

(3) goods解析逻辑: 同amazon\_category

2.新增setting

PROXY\_LUMINATI Luminati代理配置

3.修改middlerwares

ProxyMiddleware: 增加对PROXY\_LUMINATI配置读取与初始化

4.修改pipelines

新增CSVPipeline: 将item写入CSV文件

5.修改items

修改GoodsItem: 新增数据字段，修改数据模型

二、Category解析

1、一级分类列表

分析<https://www.amazon.com/Best-Sellers/zgbs/>得到，过滤出Baby/Electronic分类

2、递归下级分类

parse\_category先处理当前分类商品，然后递归调用parse\_category处理下一级分类，通过level控制分类级别

3、TOP100商品列表

(1). 进入某一级分类，默认展示TOP1~50商品， 直接处理

(2). TOP51~100商品需要翻页后再处理

三、luminati接入文档：

1、登录https://luminati-china.biz/

激活要购买的分区通道zone

2、本地搭建Luminati Proxy Manger

wget -qO- https://luminati-china.biz/static/lpm/luminati-proxy-latest-setup.sh | bash

3、启动Luminati Proxy Manger：

启动：luminati --daemon --www\_whitelist\_ips “0.0.0.0/0” --whitelist\_ips “0.0.0.0/0”

停止: luminati --stop-daemon

4、登录Luminati Proxy Manger 将本机ip添加到对应的zone白名单

http://127.0.0.1:22999/

常规设置--->管理员白名单ip/代理白名单ip--->Add IP

5、https配置：访问amazon需要https

http://127.0.0.1:22999/proxy/24000/general

Connection to Super Proxy: Default改为https

6、scrapy爬虫中使用ip代理

middlewares.py--->ProxyMiddleware--->process\_request

request.meta['proxy'] = "127.0.0.1:24000" #容器部署network\_mode使用host模式

四、本地部署

1、安装pyenv和virtualenv

2、pyenv安装python3.6.6

3、创建amz-env: pyenv virtualenv 3.6.6 amz-env

4、启用amz-env: pyenv activate amz-env

5、安装requirements.txt

6、启动爬虫： scrapy crawl amazon\_bestseller

五、容器部署

1、沿用之前的Dockerfile和docker-compose.yml, 将Dockerfile中spider改为

amazon\_bestseller， docker-compose.yml中指定network\_mode为host模式。

2、启动爬虫： docker-compose up -d