操作文档

# 配置环境

基于 Ubuntu 16.04

## 安装 Scrapy 和 PyMongo

sudo apt-get install python-pip python-dev build-essential

sudo apt-get install libxml2-dev libxslt1-dev

sudo pip install Scrapy==1.0.4

sudo pip install pymongo

## 安装 Scrapy-Splash

pip install scrapy-splash

Splash 使用 Splash HTTP API，需要一个 Splash 实例，需要安装 docker

## 安装 docker

移除老版本 docker

sudo apt-get remove docker docker-engine docker.io

配置 apt 允许使用 HTTPS 下载

sudo apt-get update

sudo apt-get install apt-transport-https ca-certificates curl software-properties-common

添加 docker 官方 GPG 密钥

curl -fsSL https://download.docker.com/linux/ubuntu/gpg | sudo apt-key add –

配置使用 docker 稳定 reposiotry

sudo add-apt-repository \

"deb [arch=amd64] https://download.docker.com/linux/ubuntu \

$(lsb\_release -cs) \

stable"

安装 docker 社区版

sudo apt-get update

sudo apt-get install docker-ce

安装完成 docker 之后，下载 splash 镜像

sudo docker pull scrapinghub/splash

## 安装 MongoDB

添加 MongoDB 官方 GPG 密钥

sudo apt-key adv --keyserver hkp://keyserver.ubuntu.com:80 --recv 0C49F3730359A14518585931BC711F9BA15703C6

添加 MongoDB 仓库

echo "deb [ arch=amd64,arm64 ] http://repo.mongodb.org/apt/ubuntu xenial/mongodb-org/3.4 multiverse" | sudo tee /etc/apt/sources.list.d/mongodb-org-3.4.list

安装 MongoDB

sudo apt-get update

sudo apt-get install mongodb-org

启动 MongoDB

sudo systemctl start mongod

# 运行

1. 启动 Splash：

sudo docker run -p 8050:8050 -p 5023:5023 scrapinghub/splash

1. 运行 scrapy/crawler/crawler 下 jd.sh 或者 tb.sh 开始爬虫
2. 爬取的数据存放在 mall.good\_info 集合中，进入 mongo shell 查看已爬取的数据：



