分布式爬虫系统scrapy-exts交接文档

李科健

20180607

1. 说明

爬虫系统基于Python开源爬虫框架scrapy和scrapy-redis，目前支持**微博**和**文章**的两种类型的数据爬取

Scrapy工作流介绍：<http://hao.jobbole.com/python-scrapy/>

Scrapy-redis介绍：<https://www.cnblogs.com/kylinlin/p/5198233.html>

1. 工作流



1. 部署

**# Python依赖**

python3.6

scrapy==1.5.0

pymysql==0.7.9

redis==2.10.6

**# 其他环境依赖**

ftp

redis

mysql

**# 配置**

scrapyer/settings中，有以下重要配置

MYSQL\_\*：mysql服务器的地址，端口，用户名和密码

REDIS\_\*：redis服务器地址，端口

CONCURRENT\_REQUESTS：爬虫的并发数

ITEM\_PIPELINES：使用哪种Pipeline来处理爬回的数据，文章与微博的Pipeline是不同的，使用文章爬虫时，在settings文件中启用。

```

ITEM\_PIPELINES = {

'exts.pipelines.ZipfilePipelineWithFTP': 300,

}

```

使用微博爬虫时，启用

```

ITEM\_PIPELINES = {

'exts.pipelines.WeiboPipelineWithFTP': 300,

}

```

**# 启动**

启动方式一（推荐）：使用engine.py导入外部数据

```

# master主机上运行engine.py

cd scrapy-exts & python engine.py

# 在其他slave机子上

cd scrapy-exts

文章爬虫

scrapy runspider scraper/spiders/slave.py

微博爬虫

scrapy runspider scraper/spiders/weibo\_slave.py

```

启动方式二：一个master，多个slave。这种启动方式，要求每种类型的爬虫有一个master实例，而启动方式一则不需要。

```

# master

cd scrapy-exts

文章爬虫

scrapy runspider scraper/spiders/master.py

微博爬虫

scrapy runspider scraper/spiders/weibo\_master.py

# slave

cd scrapy-exts

文章爬虫

scrapy runspider scraper/spiders/slave.py

微博爬虫

scrapy runspider scraper/spiders/weibo\_slave.py

1. 其他说明

文章的处理逻辑，在scrapy-exts/scraper/spiders/slave.py及scrapy-exts/exts/pipelines.py中的ZipfilePipelineWithFTP

微博的处理逻辑，在scrapy-exts/scraper/spiders/weibo\_slave.py及scrapy-exts/exts/pipelines.py中的WeiboPipelineWithFTP中