**# Python依赖**

python3.6

```

scrapy==1.5.0 #升级至1.5

pymysql==0.7.9 #升级至0.7.9

redis==2.10.6 #升级至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：是否使用ftp服务器

**# 启动**

启动方式一：一个master，多个slave

```

# master

cd scrapy-urun

scrapy runspider scraper/spiders/master.py

# slave

cd scrapy-urun

scrapy runspider scraper/spiders/slave.py

```

启动方式二：使用engine.py导入外部数据，master主机上也跑slave.py

```

# master主机上运行engine.py

cd scrapy-urun & python engine.py

# 在master主机和其他slave机子上

cd scrapy-urun

scrapy runspider scraper/spiders/slave.py

```

**# 注意事项**

文章和微博的爬虫使用的是不同的pipeline，使用文章爬虫时，在settings文件中启用。因此，每一个爬虫都需要一套代码。

```

ITEM\_PIPELINES = {

'exts.pipelines.ZipfilePipelineWithFTP': 300,

}

```

使用微博爬虫时，启用

```

ITEM\_PIPELINES = {

'exts.pipelines.WeiboPipelineWithFTP': 300,

}

```