**第一个demo**

爬虫采用PHP编写, 下面以糗事百科为例, 来看一下我们的爬虫长什么样子:

**安装**

**1、通过GitHub下载**

require\_once \_\_DIR\_\_ . '/../autoloader.php';

use phpspider\core\phpspider;

**2、通过composer下载**

composer require owner888/phpspider

require './vendor/autoload.php';

use phpspider\core\phpspider;

**3、加上一段很讨厌的注释，别问我为什么，我就是这么讨厌 ^\_^**

/\* Do NOT delete this comment \*/

/\* 不要删除这段注释 \*/

$configs = array(

'name' => '糗事百科',

'domains' => array(

'qiushibaike.com',

'www.qiushibaike.com'

),

'scan\_urls' => array(

'http://www.qiushibaike.com/'

),

'content\_url\_regexes' => array(

"http://www.qiushibaike.com/article/\d+"

),

'list\_url\_regexes' => array(

"http://www.qiushibaike.com/8hr/page/\d+\?s=\d+"

),

'fields' => array(

array(

// 抽取内容页的文章内容

'name' => "article\_content",

'selector' => "//\*[@id='single-next-link']",

'required' => true

),

array(

// 抽取内容页的文章作者

'name' => "article\_author",

'selector' => "//div[contains(@class,'author')]//h2",

'required' => true

),

),

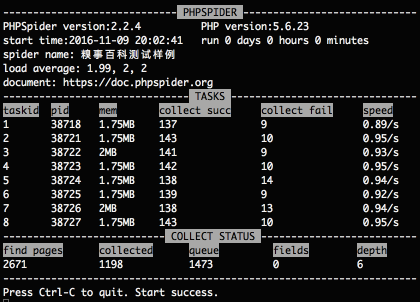
);

$spider = new phpspider($configs);

$spider->start();

爬虫的整体框架就是这样, 首先定义了一个$configs数组, 里面设置了待爬网站的一些信息, 然后通过调用$spider = new phpspider($configs);和$spider->start();来配置并启动爬虫.

**运行界面如下:**



$configs对象如何定义, 后面会作详细介绍.^\_^