Форум сообщества Альт Линукс

Автор Тема: wget - возник вопрос (Прочитано 2894 раз)

Speccyfighter

Мастер

Сообщений: 10 259

wget - возник вопрос

« : 24.01.2014 06:55:53 »

Преамбула:

Спойлер

Иногда хочется иметь зеркало сайта, который по-сути сборник вопросов и ответов.
Здесь сразу приходит мысль о wget.
Но вот загвоздка, по команде

Код: [Выделить]

```
wget -rk -np адрес_сайта
```

wget возвращает

Код: [Выделить]

```
403 FORBIDDEN
```

Сразу приходит в голову мысль, что на сервере стоит защита.
Понятно, что никакие чьи-то данные нам не нужны, а вот чтобы рядом локально был эдакий справочник вопросов и ответов, дело полезное.

Собсно здесь вопрос:
как сделать локальное зеркало сайта?

И сам себе ответил:

гугл подсказал ответ тут и тут.

Суть в том, чтобы заставить wget замаскироваться под браузер.
Например посмотреть user agent своего браузера введя в адресной строке Firefox 'about:', и ниже Firefox сообщит:

Код: [Выделить]

```
Идентификатор сборки: Mozilla/5.0 (X11; Linux i686; rv:24.0) Gecko/201
```

Как подсказывают в сторонке, последние два параметра точно лишними не будут:

Цитировать

> -r - скачивать сайт полностью, автоматически переходя по ссылкам;
> -l 2 - глубина скачивания(проходить по ссылкам на 2 уровня вниз);
> -k - после скачивания преобразовать все ссыки в html-файлах относительно локальной файловой системы;
> -np - не выходить за пределы заданной директории;
> –wait 5 - ждать 5 секунд после каждого запроса. Многие сайты вычисляют программы для автоматического скачивания по времени между запросами. Но тут
> есть проблема - сайт может отслеживать запросы, повторяющиеся через определенные промежутки времени.
> Это и сгубило ScrapBook - он делает паузу в 3 секунды между запросами, и этот
> параметр не настраивается(по крайней мере поверхностный поиск результатов не
> дал).
> В wget есть параметр –randomwait, который в случайном порядке устанавливает
> промежуток между запросами, умножая –wait на число от 0.5 до 1.5. Это позволяет
> обойти проверки на запросы, повторяющиеся с определенной периодичностью.

Но здесь сразу провести коррекцию по 'man wget'

В конечном счёте можно нарисовать команду

Код: [Выделить]

```
wget -rk -np -nc -U "Mozilla/5.0 (X11; Linux i686; rv:24.0) Gecko/2011
```

которая и выполнит зеркалирование сайта.

Полезные линки

Спойлер

wget | Николай Беляшов
http://beliashou.by/blog/archives/tag/wget
wget - программа для скачивания по http и ftp | Скачивание “умных” сайтов с помощью wget
http://wget.org.ua/wget114.html
Wget - [1] :: Программы :: Компьютерный форум Ru.Board
http://forum.ru-board.com/topic.cgi?forum=5&topic=10066&start=0#8
Лучший софт для чтения конференции в офф-лайне - [3] :: Программы :: Компьютерный форум Ru.Board
http://forum.ru-board.com/topic.cgi?forum=5&topic=19&start=26
HTTrack Website Copier - [7] :: Программы :: Компьютерный форум Ru.Board
http://forum.ru-board.com/topic.cgi?forum=5&topic=24178&glp#lt

« *Последнее редактирование: 24.01.2014 07:13:56 от Speccyfighter* »

Записан

Страницы: [1]

ПЕЧАТЬ

« предыдущая тема следующая тема »