使用Firefox进行抓取

以下列出了使用Firefox进行抓取的提示和建议，以及一系列有用的Firefox附加组件，以简化抓取过程。

检查实时浏览器DOM的注意事项

由于Firefox附加组件在实时浏览器DOM上运行，因此在检查页面源时您实际看到的不是原始HTML，而是在应用某些浏览器清理并执行Javascript代码后修改后的HTML。特别是Firefox `<tbody>` 以向表格添加元素而闻名。另一方面，Scrapy不会修改原始页面HTML，因此如果 `<tbody>` 在XPath表达式中使用，则无法提取任何数据。

因此，在使用Firefox和XPath时，请注意以下事项：

- 在检查DOM时，禁用Firefox Javascript以查找要在Scrapy中使用的XPath
- 从不使用完整的XPath路径，使用基于属性相对和巧妙的人（如 `id` ， `class` ， `width` 等），或任何识别特征等。 `contains(@href, 'image')`
- `<tbody>` 除非您真正知道自己在做什么，否则切勿在XPath表达式中包含元素

用于抓取的有用的Firefox附件

萤火虫

Firebug是Web开发人员中广为人知的工具，它对于抓取也非常有用。特别是，当您需要构造用于提取数据的XPath时，它的Inspect Element功能非常方便，因为它允许您在将鼠标移动到每个页面元素上时查看每个页面元素的HTML代码。

有关如何使用Firebug和Scrapy的详细指南，请参阅使用Firebug进行抓取。

XPather

XPather允许您直接在页面上测试XPath表达式。

XPath的检查

XPath Checker是另一个用于测试页面上XPath的Firefox附加组件。

篡改数据

Tamper Data是一个Firefox附加组件，允许您查看和修改Firefox发送的HTTP请求标头。Firebug还允许查看HTTP标头，但不允许修改它们。

Firecookie

Firecookie使查看和管理cookie变得更加容易。您可以使用此扩展程序创建新Cookie，删除现有Cookie，查看当前网站的Cookie列表，管理Cookie权限等等。