Design Decisions

Performance

Async vs Sync

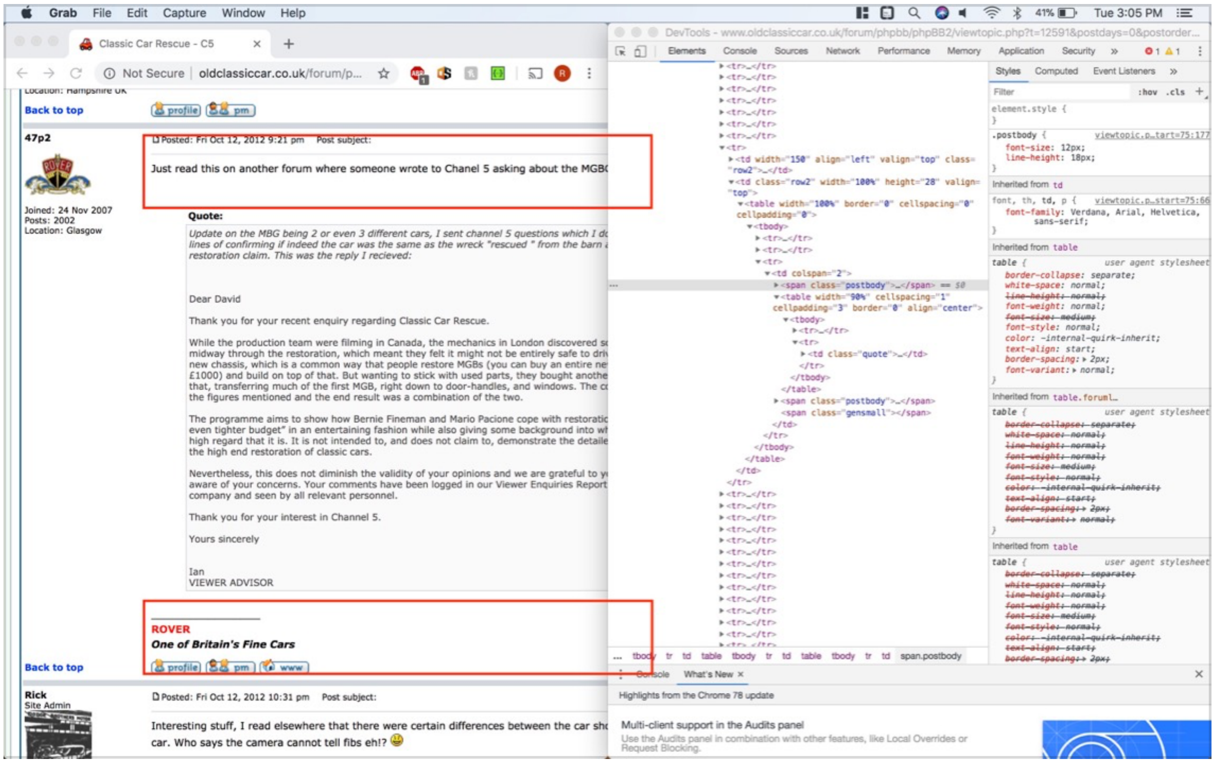
In a synchronous web scraper, each website is visited in serial. This means while the page data is being fetched, the program is waiting, essentially doing nothing. This is lower the program's throughput and therefore performance. This may be fine for this use case, since the thread is only nine pages long. However, when dealing with enourmous websites, excecution time increases dramatically.

In order to improve this, I used a async http request library `grequests` in order to be able to fetch data from multiple web pages in parallel. Which means that there is less detriments to the program's throughput.

Modularity and Scalability

Returning as Lists in Raw Data

Prior to any custom transformations, the data is stored in list. This might not make sense at first since ideally each needed datapoint maps to one element per post. However, this is not the case for the example website since that when quotes are present there are multiple post bodies.


Grab File Edit Capture Window Help
41% Tue 3:05 PM
Classic Car Rescue - C5
Not Secure | oldclassiccar.co.uk/forum/p...
DevTools - www.oldclassiccar.co.uk/forum/phpbb/phpBB2/viewtopic.php?t=12591&postdays=0&postorder...
Elements Console Sources Network Performance Memory Application Security »
1 1
Styles Computed Event Listeners »
Filter :hov .cls +
Back to top
profile pm
Console What's New ×

Location: Hampshire UK

47p2

Joined: 24 Nov 2007
Posts: 2002
Location: Glasgow

Posted: Fri Oct 12, 2012 9:21 pm Post subject:

Just read this on another forum where someone wrote to Chanel 5 asking about the MGB

Quote:

Update on the MBG being 2 or even 3 different cars, I sent channel 5 questions which I d
lines of confirming if indeed the car was the same as the wreck "rescued " from the barn
restoration claim. This was the reply I recieved:

Dear David

Thank you for your recent enquiry regarding Classic Car Rescue.

While the production team were filming in Canada, the mechanics in London discovered s
midway through the restoration, which meant they felt it might not be entirely safe to dri
new chassis, which is a common way that people restore MGBs (you can buy an entire ne
£1000) and build on top of that. But wanting to stick with used parts, they bought anothe
that, transferring much of the first MGB, right down to door-handles, and windows. The c
the figures mentioned and the end result was a combination of the two.

The programme aims to show how Bernie Fineman and Mario Pacione cope with restoratic
even tighter budget" in an entertaining fashion while also giving some background into w
high regard that it is. It is not intended to, and does not claim to, demonstrate the detaile
the high end restoration of classic cars.

Nevertheless, this does not diminish the validity of your opinions and we are grateful to y
aware of your concerns. Your comments have been logged in our Viewer Enquiries Report
company and seen by all relevant personnel.

Thank you for your interest in Channel 5.

Yours sincerely

Ian
VIEWER ADVISOR

ROVER
One of Britain's Fine Cars

profile pm www

Rick
Site Admin

Posted: Fri Oct 12, 2012 10:31 pm Post subject:

Interesting stuff, I read elsewhere that there were certain differences between the car sh
car. Who says the camera cannot tell fibs eh!?

```
▶<tr>…</tr>
▶<tr>…</tr>
▶<tr>…</tr>
▶<tr>…</tr>
▶<tr>…</tr>
▶<tr>…</tr>
▶<tr>…</tr>
▼<tr>
  ▶<td width="150" align="left" valign="top" class=
  "row2">…</td>
  ▼<td class="row2" width="100%" height="28" valign=
  "top">
    ▼<table width="100%" border="0" cellspacing="0"
    cellpadding="0">
      ▼<tbody>
        ▶<tr>…</tr>
        ▶<tr>…</tr>
        ▼<tr>
          ▼<td colspan="2">
...          ▶<span class="postbody">…</span> == $0
            ▼<table width="90%" cellspacing="1"
            cellpadding="3" border="0" align="center">
              ▼<tbody>
                ▶<tr>…</tr>
                ▼<tr>
                  ▶<td class="quote">…</td>
                  </tr>
                </tbody>
              </table>
            ▶<span class="postbody">…</span>
             <span class="gensmall"></span>
           </td>
         </tr>
       </tbody>
     </table>
   </td>
 </tr>
▶<tr>…</tr>
▶<tr>…</tr>
▶<tr>…</tr>
▶<tr>…</tr>
▶<tr>…</tr>
▶<tr>…</tr>
▶<tr>…</tr>
▶<tr>…</tr>
▶<tr>…</tr>
▶<tr>…</tr>
▶<tr>…</tr>
▶<tr>…</tr>
▶<tr>…</tr>
▶<tr>…</tr>
... tbody tr td table tbody tr td table tbody tr td span.postbody
```

```
element.style {
}
.postbody {                     viewtopic.p…tart=75:177
  font-size: 12px;
  line-height: 18px;
}
Inherited from td
font, th, td, p {               viewtopic.p…start=75:66
  font-family: Verdana, Arial, Helvetica,
      sans-serif;
}
Inherited from table
table {                      user agent stylesheet
  border-collapse: separate;
  white-space: normal;
  line-height: normal;
  font-weight: normal;
  font-size: medium;
  font-style: normal;
  color: -internal-quirk-inherit;
  text-align: start;
  border-spacing: ▸ 2px;
  font-variant: ▸ normal;
}
Inherited from table.foruml…
table {                      user agent stylesheet
  border-collapse: separate;
  white-space: normal;
  line-height: normal;
  font-weight: normal;
  font-size: medium;
  font-style: normal;
  color: -internal-quirk-inherit;
  text-align: start;
  border-spacing: ▸ 2px;
  font-variant: ▸ normal;
}
Inherited from table
table {                      user agent stylesheet
  border-collapse: separate;
  white-space: normal;
  line-height: normal;
  font-weight: normal;
  font-size: medium;
  font-style: normal;
  color: -internal-quirk-inherit;
  text-align: start;
  border-spacing: ▸ 2px;
```

Highlights from the Chrome 78 update

Multi-client support in the Audits panel

Use the Audits panel in combination with other features, like Local Overrides or Request Blocking.

Since this pattern is present on this website, it could be present on others. With modularity in mind, it's better to keep the data raw at first so that this scraper can be reused on multiple websites. The engineer can then modify the data as needed using custom transformations in the Scraper class' `data_config` paramter.

Config Schema

```
data_config = {
    'post_id': { # field the data will be saved in result record object
        'path': [
            { 'selector': 'span', 'options': { 'class': 'name' } }
        ],
        'selector': 'a',
        'options': {},
        'attribute': 'name',
        'transform': lambda res: res[0] # optional custom function to
manipulate response
    },
}
```

The scraper is designed to follow the path of the target's ancestors first before getting all of the targeted elements. Each object feeds into a typical BeautifulSoup `find` function. Written this way, not only does the scraper remain DRY, but it allows engineers to use the scraper without much knowledge of BeautifulSoup.

Transform

Raw data is collected initially and can be transformed using custom functions specified by the engineer.

```
[u'Posted: Mon Sep 24, 2012 4:53 pm\xa0\xa0 \xa0Post subject: Classic Car Rescue –
C5']
=> u'Mon Sep 24, 2012 4:53'
```

Rather than hardcode these transformations into the webscraper, which would reduce modularity, the engineer has the ability to pass in custom transformations for their own needs. This means that the scraper can be used on multiple websites.

Seperation of Nav, Body, and Data

The locations of the navigation, the body (where the data is located) and what data the engineer wants to scrape is different for every website. To maintain modularity, I decided to seperate their logic.