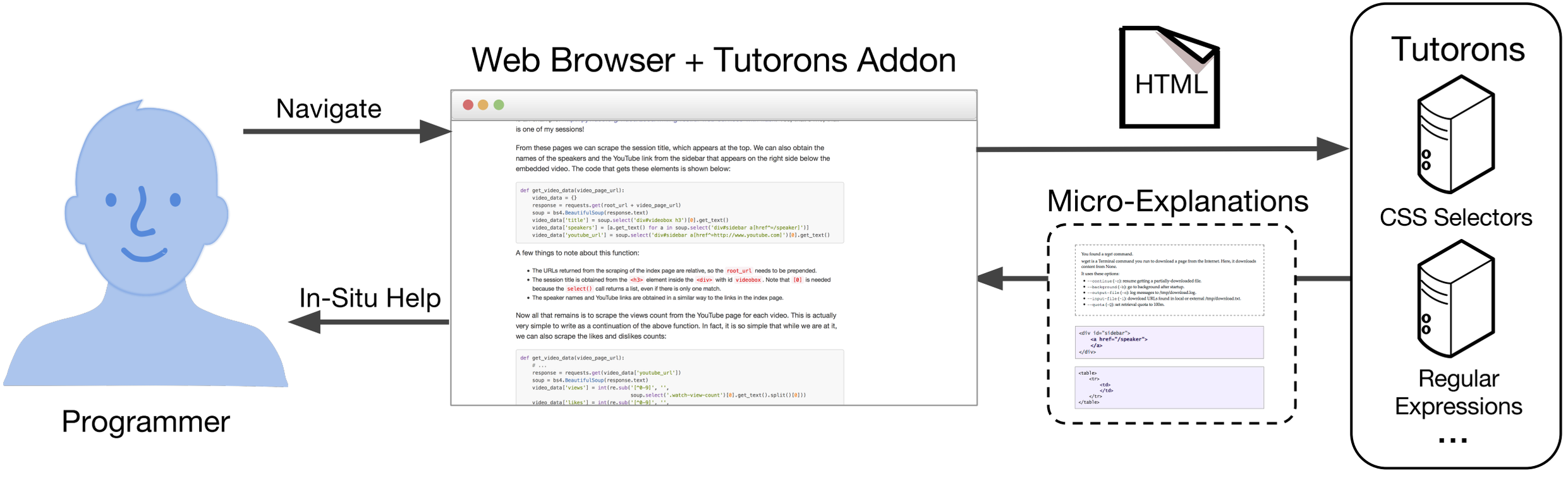
Web Browser + Tutorons Addon
Navigate
In-Situ Help
Programmer
HTML
Micro-Explanations
Tutorons
CSS Selectors
Regular Expressions
...
is one of my sessions!
From these pages we can scrape the session title, which appears at the top. We can also obtain the names of the speakers and the YouTube link from the sidebar that appears on the right side below the embedded video. The code that gets these elements is shown below:
def get_video_data(video_page_url):
video_data = {}
response = requests.get(root_url + video_page_url)
soup = bs4.BeautifulSoup(response.text)
video_data['title'] = soup.select('div#videobox h3')[0].get_text()
video_data['speakers'] = [a.get_text() for a in soup.select('div#sidebar a[href^=/speaker]')]
video_data['youtube_url'] = soup.select('div#sidebar a[href^=http://www.youtube.com]')[0].get_text()
A few things to note about this function:
The URLs returned from the scraping of the index page are relative, so the root_url needs to be prepended.
The session title is obtained from the <h3> element inside the with id videobox. Note that [0] is needed because the select() call returns a list, even if there is only one match.
The speaker names and YouTube links are obtained in a similar way to the links in the index page.
Now all that remains is to scrape the views count from the YouTube page for each video. This is actually very simple to write as a continuation of the above function. In fact, it is so simple that while we are at it, we can also scrape the likes and dislikes counts:
def get_video_data(video_page_url):
# ...
response = requests.get(video_data['youtube_url'])
soup = bs4.BeautifulSoup(response.text)
video_data['views'] = int(re.sub('[^0-9]', '',
soup.select('.watch-view-count')[0].get_text().split()[0]))
video_data['likes'] = int(re.sub('[^0-9]', '',
You found a wget command.
wget is a Terminal command you run to download a page from the Internet. Here, it downloads content from None.
It uses these options:
--continue (-c): resume getting a partially-downloaded file.
--background (-b): go to background after startup.
--output-file (-o): log messages to /tmp/download.log.
--input-file (-i): download URLs found in local or external /tmp/download.txt.
--quota (-Q): set retrieval quota to 100m.

<table>
<tr>
<td>
</td>
</tr>
</table>