



USE PYTHON TO PARSE HTML FOR

LIZZIE SIEGLE



@LIZZIEPIKA



- DEVELOPER
EVANGELIST AT TWILIO
- LIKES [PEOPLE,
PYTHON, AND
PIKACHU]

BEAUTIFUL SOUP 101



@LIZZIEPIK
A

What is Beautiful Soup?

- Web-scraping library
- Extract desired data from page
- Clean content



How will We Use it?


- Find every link on a webpage

Popular Quotes

Search

[Popular](#) [Recent](#) [New](#) [Friends](#) [My Authors](#)

Quotes popular among Goodreads members




"Don't cry because it's over, smile because it happened."
— Dr. Seuss

tags: attributed-no-source, cry, crying, experience, happiness, joy, life, misattributed-dr-seuss, optimism, sadness, smile, smiling

186439 likes

Like




"I'm selfish, impatient and a little insecure. I make mistakes, I am out of control and at times hard to handle. But if you can't handle me at my worst, then you sure as hell don't deserve me at my best."
— Marilyn Monroe

tags: attributed-no-source, best, life, love, mistakes, out-of-control, truth, worst

144136 likes

Like



"Be yourself; everyone else is already taken."
— Oscar Wilde

tags: attributed-no-source, be-yourself, honesty, inspirational, misattributed-oscar-wilde

140215 likes

Like

Some of our Tools

- *Requests* to access HTML
- *bs4* to parse HTML
- *String* to clean text
- *Random* to pick random number



BS4 CODE



SEND HTTP REQUESTS

```
from bs4 import BeautifulSoup
import requests, string, random

def scrape_and_clean(url):
    req = requests.get(url)
    soup = BeautifulSoup(req.content, "html.parser")
```



@LIZZIEPIK
A

Inspec

```
▼<div class="quotes">
  ▼<div class="quote">
    ▼<div class="quoteDetails">
      ▶<a class="leftAlignedImage" href="/author/show/61105.Dr_Seuss">_</a>
      ▼<div class="quoteText">
        "
          "Don't cry because it's over, smile because it happened."
        "
      <br>
      " _
      "
      <a class="authorOrTitle" href="/author/show/61105.Dr_Seuss">Dr. Seuss
      </a>
    </div>
```



Find each `div`

```
#get quotes from page  
div_quotes = soup("div", attrs={"class": "quoteText"})
```



Find each `a` tag

```
quotes = []
for q in div_quotes:
    author = ""
    # If no author, then skip
    try:
        author = q.find("a").get_text() + "\n"
    except:
        continue
```



Multi-line -> Single String

```
quote = ""
# turn multiline quotes/poems into a single string
for i in range(len(q.contents)):
    #find returns
    line = q.contents[i].encode("ascii", errors="ignore").decode("utf-8")
    print("line ", line)
    if (line[0] == "<"): # is tag, ignore characters that aren't part of quote
        break
    else:
        quote += line
```



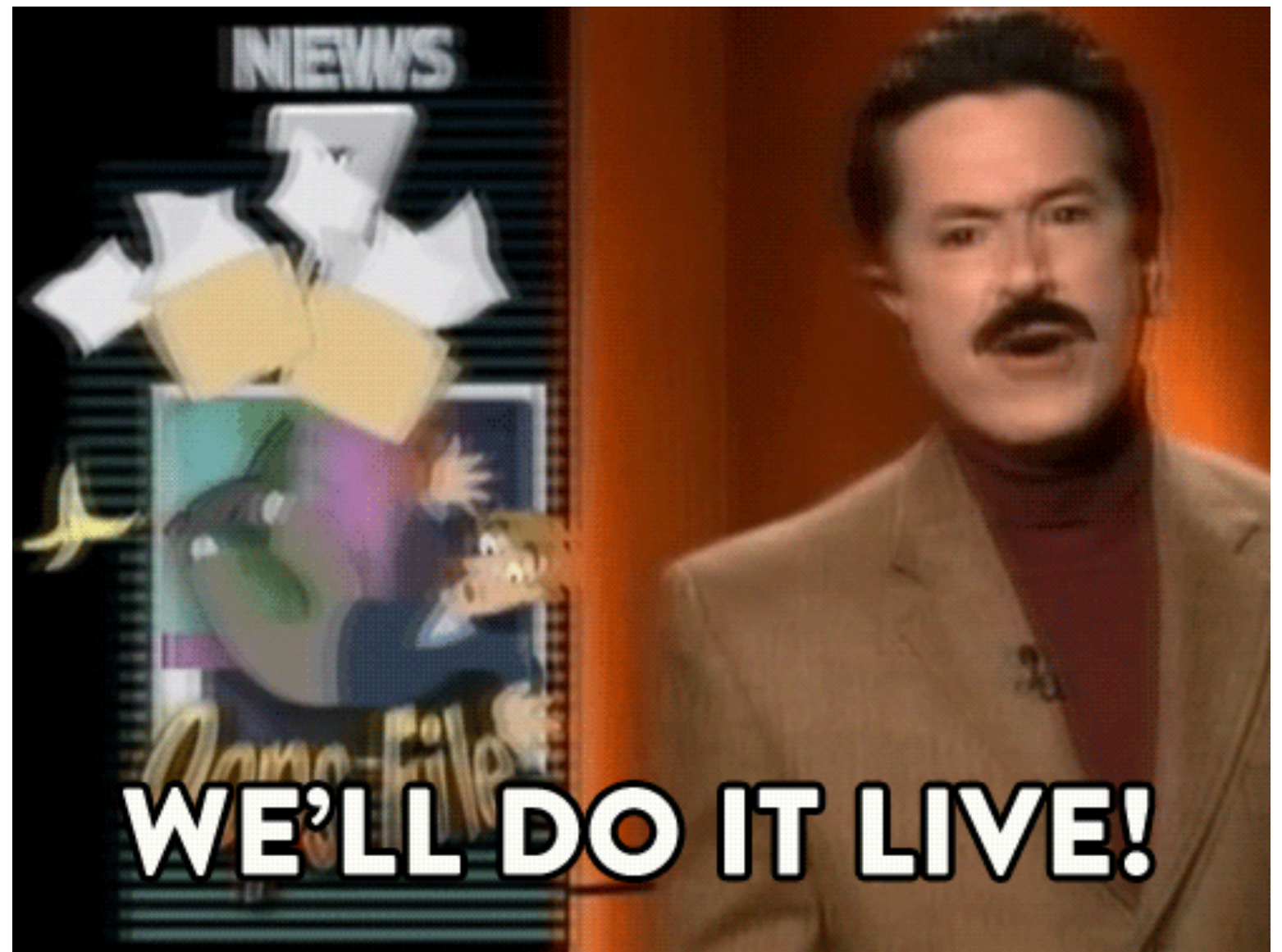
Clean, Format Quotes

```
quote = q.contents[0].encode("ascii", errors="ignore").decode("utf-8")  
quote = "\"" + quote.strip() + "\" "  
quotes += quote + '\n\n' + '-' + author + "#"
```

```
quotes_to_return = filter(lambda x: x in string.printable, quotes) #clean  
return quotes_to_return
```



LIVE
CODE



@LIZZIEPIK
A

Text +1(813)906-5107

“harry potter”

“crazy rich asians”

“jane austen”

etc.



@LIZZIEPIK
A

THANK YOU!

- [HTTPS://WWW.CRUMMY.COM/SOFTWARE/BEAUTIFULSOUP/](https://www.crummy.com/software/beautifulsoup/)
- [HTTPS://WWW.TWILIO.COM/BLOG/PARSE-HTML-FOR-BOOK-QUOTES-PYTHON-BEAUTIFUL-SOUP-WHATSAPP](https://www.twilio.com/blog/parse-html-for-book-quotes-python-beautiful-soup-whatsapp)

LSIEGLE@TWILIO.COM



@LIZZIEPIK
A.