**[PYTHON CIRCLE](https://www.pythoncircle.com/)**

[Practice](https://www.pythoncircle.com/practice/)   [Python Books](https://www.pythoncircle.com/post/646/python-books/)   [Archive](https://www.pythoncircle.com/archive/)   [Tools](https://www.pythoncircle.com/tools/home/)   [Contact](https://www.pythoncircle.com/contact/)   [Subscribe](https://www.pythoncircle.com/subscribe/)

### **Recommended Books:**



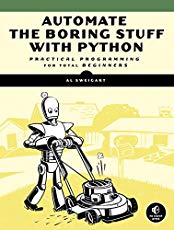
[**Python Tricks: A Buffet of…**](https://aax-us-east.amazon-adsystem.com/x/c/QhRpFxpp9m1igttfcHVmbiwAAAFsHsYhJQEAAAFKAVWiohM/https:/www.amazon.com/Python-Tricks-Buffet-Awesome-Features-ebook/dp/B0785Q7GSY/ref=sm_n_ma_dka_IN_pr_ran_nodl?adId=B0785Q7GSY&creativeASIN=B0785Q7GSY&linkId=b8d6e601e22f9fdef3f32fedf790fe25&tag=dogstore42-20&linkCode=w58&ref-refURL=https%3A%2F%2Fwww.pythoncircle.com%2Fpost%2F678%2Fpython-script-14-scraping-news-headlines-using-python-beautifulsoup%2F&slotNum=0&imprToken=BC4QTcRyKk8-GwkIgsVa2w&adType=smart&adMode=manual&adFormat=card&impressionTimestamp=1563884396684)

By **Dan Bader**

**$4.99**

Rated 5 out of 5

[Buy Now](https://aax-us-east.amazon-adsystem.com/x/c/QhRpFxpp9m1igttfcHVmbiwAAAFsHsYhJQEAAAFKAVWiohM/https:/www.amazon.com/Python-Tricks-Buffet-Awesome-Features-ebook/dp/B0785Q7GSY/ref=sm_n_ma_dka_IN_pr_ran_nodl?adId=B0785Q7GSY&creativeASIN=B0785Q7GSY&linkId=b8d6e601e22f9fdef3f32fedf790fe25&tag=dogstore42-20&linkCode=w58&ref-refURL=https%3A%2F%2Fwww.pythoncircle.com%2Fpost%2F678%2Fpython-script-14-scraping-news-headlines-using-python-beautifulsoup%2F&slotNum=0&imprToken=BC4QTcRyKk8-GwkIgsVa2w&adType=smart&adMode=manual&adFormat=card&impressionTimestamp=1563884396684)



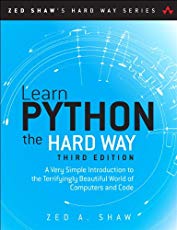
[**Automate the Boring Stuff with…**](https://aax-us-east.amazon-adsystem.com/x/c/QjDRsX1Cfnu9SWxz_glPwyAAAAFsHsYhMwEAAAFKAdOKfR0/https:/www.amazon.com/Automate-Boring-Stuff-Python-Programming/dp/1593275994/ref=sm_n_ma_dka_IN_pr_ran?adId=1593275994&creativeASIN=1593275994&linkId=e20ca59e70652fc2911925afbd6c3952&tag=dogstore42-20&linkCode=w58&ref-refURL=https%3A%2F%2Fwww.pythoncircle.com%2Fpost%2F678%2Fpython-script-14-scraping-news-headlines-using-python-beautifulsoup%2F&slotNum=1&imprToken=djcibfeY4kCVuvyO9LZ1Vg&adType=smart&adMode=manual&adFormat=card&impressionTimestamp=1563884396793)

By **Al Sweigart (Paperback - Apr 14, 2015)**

**$16.83**~~$29.95~~

Rated 4.5 out of 5

[Buy Now](https://aax-us-east.amazon-adsystem.com/x/c/QjDRsX1Cfnu9SWxz_glPwyAAAAFsHsYhMwEAAAFKAdOKfR0/https:/www.amazon.com/Automate-Boring-Stuff-Python-Programming/dp/1593275994/ref=sm_n_ma_dka_IN_pr_ran?adId=1593275994&creativeASIN=1593275994&linkId=e20ca59e70652fc2911925afbd6c3952&tag=dogstore42-20&linkCode=w58&ref-refURL=https%3A%2F%2Fwww.pythoncircle.com%2Fpost%2F678%2Fpython-script-14-scraping-news-headlines-using-python-beautifulsoup%2F&slotNum=1&imprToken=djcibfeY4kCVuvyO9LZ1Vg&adType=smart&adMode=manual&adFormat=card&impressionTimestamp=1563884396793)



[**Learn Python the Hard Way:…**](https://aax-us-east.amazon-adsystem.com/x/c/QsKDRIQVZJ990o4NVkr3gJ4AAAFsHsYhQAEAAAFKAYxPvU0/https:/www.amazon.com/Learn-Python-Hard-Way-Introduction/dp/0321884914/ref=sm_n_ma_dka_IN_pr_ran?adId=0321884914&creativeASIN=0321884914&linkId=21308b6c7c1bd764b018191509fc1def&tag=dogstore42-20&linkCode=w58&ref-refURL=https%3A%2F%2Fwww.pythoncircle.com%2Fpost%2F678%2Fpython-script-14-scraping-news-headlines-using-python-beautifulsoup%2F&slotNum=2&imprToken=vV-heKSdruD-t1jOILMIgA&adType=smart&adMode=manual&adFormat=card&impressionTimestamp=1563884396857)

By **Zed A. Shaw (Paperback - Oct 11, 2013)**

**$31.13**~~$39.99~~

Rated 4 out of 5

[Buy Now](https://aax-us-east.amazon-adsystem.com/x/c/QsKDRIQVZJ990o4NVkr3gJ4AAAFsHsYhQAEAAAFKAYxPvU0/https:/www.amazon.com/Learn-Python-Hard-Way-Introduction/dp/0321884914/ref=sm_n_ma_dka_IN_pr_ran?adId=0321884914&creativeASIN=0321884914&linkId=21308b6c7c1bd764b018191509fc1def&tag=dogstore42-20&linkCode=w58&ref-refURL=https%3A%2F%2Fwww.pythoncircle.com%2Fpost%2F678%2Fpython-script-14-scraping-news-headlines-using-python-beautifulsoup%2F&slotNum=2&imprToken=vV-heKSdruD-t1jOILMIgA&adType=smart&adMode=manual&adFormat=card&impressionTimestamp=1563884396857)



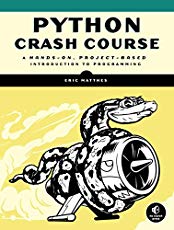
[**Learning Python, 5th Edition**](https://aax-us-east.amazon-adsystem.com/x/c/Qsdp_AEi3-DL2Wth3NjDTm4AAAFsHsYhKAEAAAFKARDymsA/https:/www.amazon.com/Learning-Python-5th-Mark-Lutz/dp/1449355730/ref=sm_n_ma_dka_IN_pr_ran?adId=1449355730&creativeASIN=1449355730&linkId=548fe208dfe3ea9ec71966682bd61293&tag=dogstore42-20&linkCode=w58&ref-refURL=https%3A%2F%2Fwww.pythoncircle.com%2Fpost%2F678%2Fpython-script-14-scraping-news-headlines-using-python-beautifulsoup%2F&slotNum=3&imprToken=eWLU-E5RBNlx3Qw88w20SQ&adType=smart&adMode=manual&adFormat=card&impressionTimestamp=1563884396594)

By **Mark Lutz (Paperback - Jul 6, 2013)**

**$44.58**~~$74.99~~

Rated 4 out of 5

[Buy Now](https://aax-us-east.amazon-adsystem.com/x/c/Qsdp_AEi3-DL2Wth3NjDTm4AAAFsHsYhKAEAAAFKARDymsA/https:/www.amazon.com/Learning-Python-5th-Mark-Lutz/dp/1449355730/ref=sm_n_ma_dka_IN_pr_ran?adId=1449355730&creativeASIN=1449355730&linkId=548fe208dfe3ea9ec71966682bd61293&tag=dogstore42-20&linkCode=w58&ref-refURL=https%3A%2F%2Fwww.pythoncircle.com%2Fpost%2F678%2Fpython-script-14-scraping-news-headlines-using-python-beautifulsoup%2F&slotNum=3&imprToken=eWLU-E5RBNlx3Qw88w20SQ&adType=smart&adMode=manual&adFormat=card&impressionTimestamp=1563884396594)



[**Python Crash Course: A…**](https://aax-us-east.amazon-adsystem.com/x/c/QntRS4B3_Bi1TWDXhGm3SNgAAAFsHsYh6gEAAAFKAfIH7M8/https:/www.amazon.com/Python-Crash-Course-Hands-Project-Based/dp/1593276036/ref=sm_n_ma_dka_IN_pr_ran?adId=1593276036&creativeASIN=1593276036&linkId=cf33a24f1ed8ef0c6a13bb7069d98245&tag=dogstore42-20&linkCode=w58&ref-refURL=https%3A%2F%2Fwww.pythoncircle.com%2Fpost%2F678%2Fpython-script-14-scraping-news-headlines-using-python-beautifulsoup%2F&slotNum=4&imprToken=I84ot3GWhbUkHbWK3RBO8A&adType=smart&adMode=manual&adFormat=card&impressionTimestamp=1563884397017)

By **Eric Matthes (Paperback - Nov 1, 2015)**

**$26.37**~~$39.95~~

Rated 4.5 out of 5

[Buy Now](https://aax-us-east.amazon-adsystem.com/x/c/QntRS4B3_Bi1TWDXhGm3SNgAAAFsHsYh6gEAAAFKAfIH7M8/https:/www.amazon.com/Python-Crash-Course-Hands-Project-Based/dp/1593276036/ref=sm_n_ma_dka_IN_pr_ran?adId=1593276036&creativeASIN=1593276036&linkId=cf33a24f1ed8ef0c6a13bb7069d98245&tag=dogstore42-20&linkCode=w58&ref-refURL=https%3A%2F%2Fwww.pythoncircle.com%2Fpost%2F678%2Fpython-script-14-scraping-news-headlines-using-python-beautifulsoup%2F&slotNum=4&imprToken=I84ot3GWhbUkHbWK3RBO8A&adType=smart&adMode=manual&adFormat=card&impressionTimestamp=1563884397017)

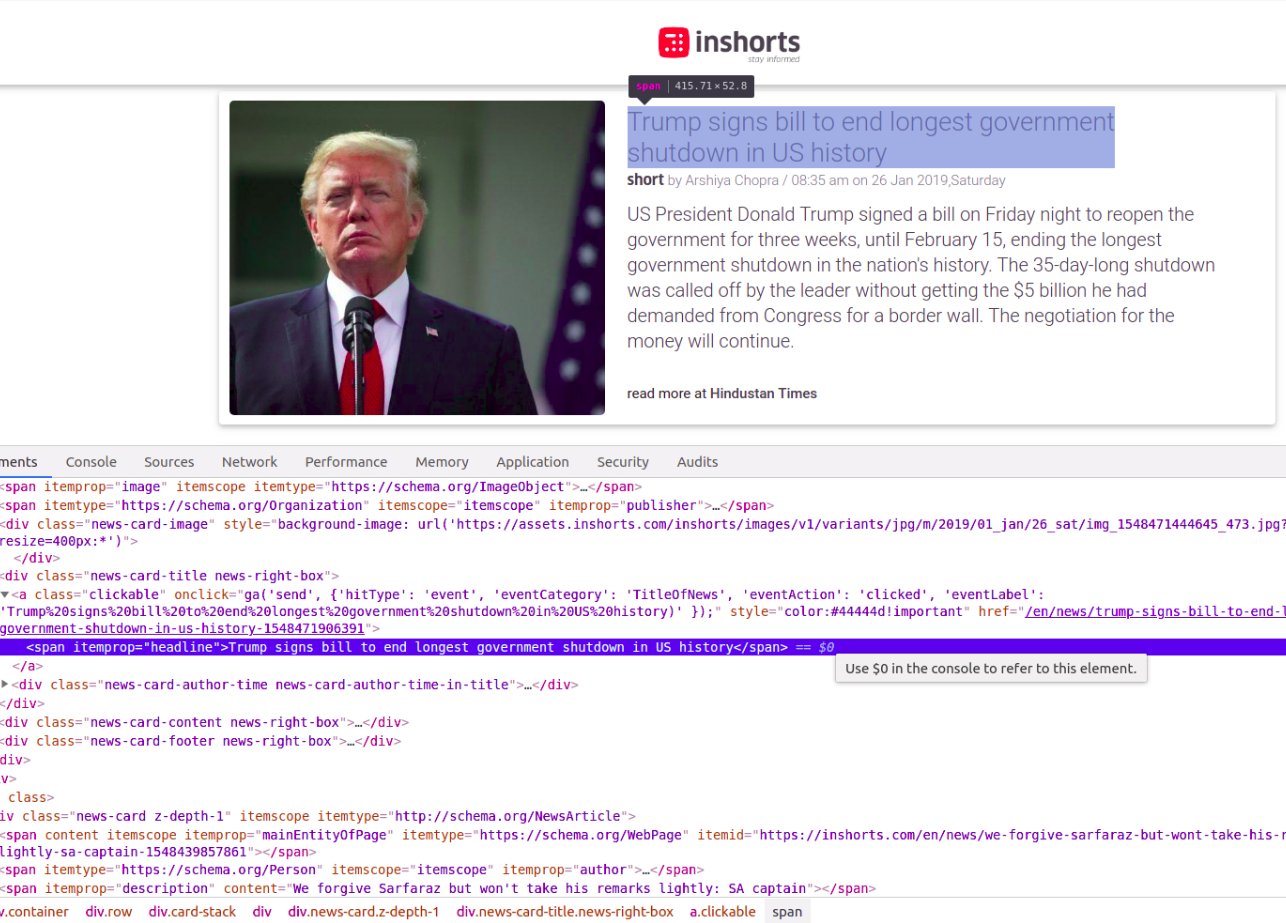
[Python Script 14: Scraping news headlines using python beautifulsoup](https://www.pythoncircle.com/post/678/python-script-14-scraping-news-headlines-using-python-beautifulsoup/)

[beautifulsoup](https://www.pythoncircle.com/tag/beautifulsoup/" \t "_blank)    [0](https://www.pythoncircle.com/post/678/python-script-14-scraping-news-headlines-using-python-beautifulsoup/#comments)    2115



We have already written few articles about web scraping using beautifulsoup and requests in python. This is yet another article where we will scrape news headlines from a news website.

For this article we have chosen the website inshorts.com. Lets start reading the news from their homepage https://inshorts.com/en/read/. To scrape headlines, we need to inspect the headline html element.



As we can see all the headlines are inside a span html tag with attribute name itemprop and its value as headline. In beautifulsoup, we can find all elements with given attribute value using the method find\_all(attrs={"attribute\_name": "attribute\_value"}) .

Before starting, we strongly recommend to [create a virtual environment](https://www.pythoncircle.com/post/404/virtual-environment-in-python-a-pocket-guide/) and install below dependencies in it.

beautifulsoup4==4.6.0

lxml==4.3.0

requests==2.21.0

#### ****Scrape news website's homepage:****

Lets start by getting the response from the homepage url.

url = 'https://inshorts.com/en/read'

response = requests.get(url)

Create a seperate function to print headlines from the response text. This will be helpful later on as well (Remember [DRY](https://en.wikipedia.org/wiki/Don%27t_repeat_yourself) principal).

def print\_headlines(response\_text):

soup = BeautifulSoup(response\_text, 'lxml')

headlines = soup.find\_all(attrs={"itemprop": "headline"})

for headline in headlines:

print(headline.text)

Call print\_headlines function and pass response.text to it.

url = 'https://inshorts.com/en/read'

response = requests.get(url)

print\_headlines(response.text)

Code so far would be

import requests

from bs4 import BeautifulSoup

import json

def print\_headlines(response\_text):

soup = BeautifulSoup(response\_text, 'lxml')

headlines = soup.find\_all(attrs={"itemprop": "headline"})

for headline in headlines:

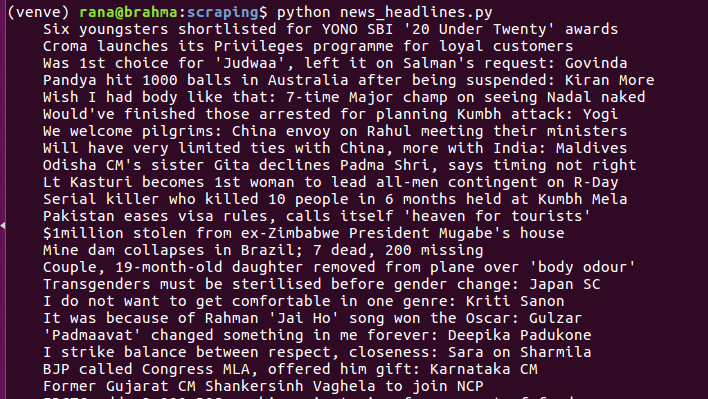
print(headline.text)

url = 'https://inshorts.com/en/read'

response = requests.get(url)

print\_headlines(response.text)

Save this code in a file with name, lets say news\_headlines.py. Activate the [virtual environment](https://www.pythoncircle.com/post/404/virtual-environment-in-python-a-pocket-guide/) and run the script using command python news\_headlines.py. Script will print the headlines shown on first page on terminal.

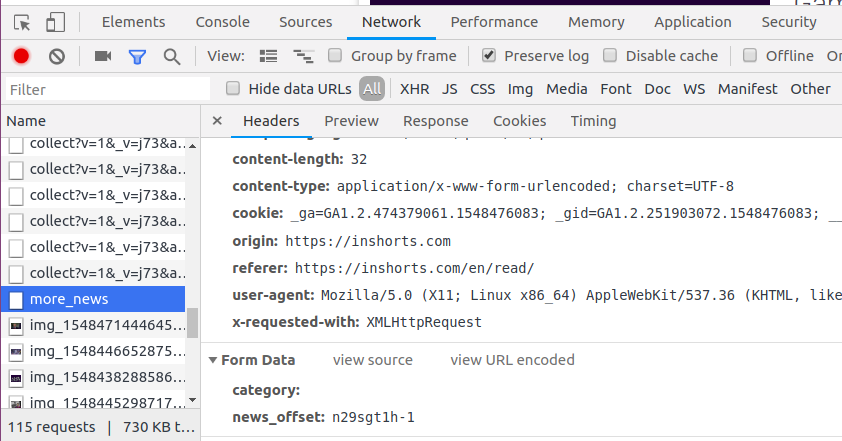


Code written so far will print headlines shown on first page only. What if we want to fetch more headlines than that.

#### ****Fetching more headlines:****

On the news website's homepage, you will see a load more button at the bottom. Open the devtool on chrome by pressing F12 and click on network tab. Here you can see all requests and responses.

When you click the **Load More** button, a request is sent to the server with 2 key values in form data which you can see in screenshot below.



Value of news\_offset variable can be found from the source code of homepage. Open the source code of homepage and search for text min\_news\_id. Use value of this variable in news\_offset.

#### ****Post request with form data:****

URL used to load more news headlines is https://inshorts.com/en/ajax/more\_news. Lets send the post request to this URL with required form data to fetch more headlines. We will send post requests inside a while loop until we keep getting 200 OK status.

url = 'https://inshorts.com/en/ajax/more\_news'

news\_offset = "apwuhnrm-1"

while True:

response = requests.post(url, data={"category": "", "news\_offset": news\_offset})

if response.status\_code != 200:

print(response.status\_code)

break

response\_json = json.loads(response.text)

print\_headlines(response\_json["html"])

news\_offset = response\_json["min\_news\_id"]

Since the response returned is JSON string with two keys, min\_news\_idand html, we will parse the response into json object and get values of these two keys. min\_news\_id will be used to send next post request and html text will be used to get headlines by passing this text to the print\_headlines function we defined earlier.

#### ****Complete Code:****

Complete python code to get news headlines is also available on [Github](https://github.com/anuragrana/Python-Scripts/blob/master/news_headlines.py" \t "_blank).

import requests

from bs4 import BeautifulSoup

import json

def print\_headlines(response\_text):

soup = BeautifulSoup(response\_text, 'lxml')

headlines = soup.find\_all(attrs={"itemprop": "headline"})

for headline in headlines:

print(headline.text)

def get\_headers():

return {

"accept": "\*/\*",

"accept-encoding": "gzip, deflate, br",

"accept-language": "en-IN,en-US;q=0.9,en;q=0.8",

"content-type": "application/x-www-form-urlencoded; charset=UTF-8",

"cookie": "\_ga=GA1.2.474379061.1548476083; \_gid=GA1.2.251903072.1548476083; \_\_gads=ID=17fd29a6d34048fc:T=1548476085:S=ALNI\_MaRiLYBFlMfKNMAtiW0J3b\_o0XGxw",

"origin": "https://inshorts.com",

"referer": "https://inshorts.com/en/read/",

"user-agent": "Mozilla/5.0 (X11; Linux x86\_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/71.0.3578.98 Safari/537.36",

"x-requested-with": "XMLHttpRequest"

}

url = 'https://inshorts.com/en/read'

response = requests.get(url)

print\_headlines(response.text)

# get more news

url = 'https://inshorts.com/en/ajax/more\_news'

news\_offset = "apwuhnrm-1"

while True:

response = requests.post(url, data={"category": "", "news\_offset": news\_offset}, headers=get\_headers())

if response.status\_code != 200:

print(response.status\_code)

break

response\_json = json.loads(response.text)

print\_headlines(response\_json["html"])

news\_offset = response\_json["min\_news\_id"]