web_scraping_NYT

December 30, 2019

[1]:
```
# import libraries
import pandas as pd
from bs4 import BeautifulSoup
import requests

# fetches the required webpage from the URL and stores the results in a␣
 ↪response object
r = requests.get(
    'https://www.nytimes.com/interactive/2017/06/23/opinion/trumps-lies.html')
# parses the HTML into a special object to read it and make sense of the␣
 ↪structure
soup = BeautifulSoup(r.text, 'html.parser')
# collecting all information with the sepcific HTML tag
results = soup.find_all('span', attrs={'class': 'short-desc'})

# break down information contents into the required substituents
lie_records = []
for result in results:
    date = result.find('strong').text[:-1] + ' , 2017'
    lie = result.contents[1][1:-2]
    explanation = result.find('a').text[1:-1]
    url = result.find('a')['href']
    lie_records.append((date, lie, explanation, url))

# make the list of records more readable through converting to a dataframe
df = pd.DataFrame(lie_records, columns=['date', 'lie', 'explanation', 'url'])
# convert date into pandas datetime format for the sake of consitency and␣
 ↪calculability
df['date'] = pd.to_datetime(df['date'])
# check dataframe
df.head()
# save dataframe in csv format
df.to_csv('lies.csv', index=False, encoding='utf-8')
```

[2]:
```
%reload_ext version_information
%version_information BeautifulSoup4, requests, pandas
```

[2]:

Software	Version
Python	3.7.4 64bit [MSC v.1915 64 bit (AMD64)]
IPython	7.10.2
OS	Windows 10 10.0.17763 SP0
BeautifulSoup4	4.8.2
requests	2.22.0
pandas	0.25.3
Wed Dec 25 23:24:40 2019 Mountain Standard Time	