Basic Webscraping

Kari Kusler

Exercises

1. Read the HTML content of the following URL with a variable called webpage: https://money.cnn.com/data/us_markets/ At this point, it will also be useful to open this web page in your browser.
2. Get the session details (status, type, size) of the above mentioned URL.

```
## <session> https://money.cnn.com/data/us_markets/
##   Status: 200
##   Type:   text/html; charset=utf-8
##   Size:   95783
```

3. Extract all of the sector names from the "Stock Sectors" table (bottom left of the web page.)

Stock Sectors	3 Month % Change
Communications	+3.24%
Consumer Durables	-9.28%
Consumer Non-Durables	+2.55%
Commercial Services	-2.74%
Electronic Technology	-4.44%
Energy Minerals	-7.68%
Finance	-4.86%
Health Services	+4.40%
Retail Trade	-3.52%
Technology Services	-6.28%
Transportation	-4.62%
Utilities	+1.80%

4. Extract all of the "3 Month % Change" values from the "Stock Sectors" table.

x
+3.24%
-9.28%
+2.55%
-2.74%
-4.44%
-7.68%
-4.86%
+4.40%
-3.52%
-6.28%
-4.62%
+1.80%

5. Extract the table "What's Moving" (top middle of the web page) into a data-frame.

Gainers & Losers	Price	Change	% Change
MYLMylan NV	36.43	5.06	+16.13%

Gainers & Losers	Price	Change	% Change
SYMCSymantec Corp	22.54	2.52	+12.59%
TRIPTripAdvisor Inc	57.00	3.76	+7.06%
BKNGBooking Holdings I...	1,949.46	78.34	+4.19%
AMDAdvanced Micro Dev...	20.68	0.78	+3.92%
TJXTJX Companies Inc	54.77	-54.78	-50.00%
NKTRNektar Therapeutic...	36.08	-3.92	-9.80%
MARMarriott Internati...	114.55	-6.13	-5.08%
TTWOTake-Two Interacti...	123.68	-3.02	-2.38%
VRSNVerisign Inc	157.15	-3.65	-2.27%

6. Re-construct all of the links from the first column of the "What's Moving" table. Hint: the base URL is "https://money.cnn.com"

links
https://money.cnn.com/quote/quote.html?symb=MYL
https://money.cnn.com/quote/quote.html?symb=SYMC
https://money.cnn.com/quote/quote.html?symb=TRIP
https://money.cnn.com/quote/quote.html?symb=BKNG
https://money.cnn.com/quote/quote.html?symb=AMD
https://money.cnn.com/quote/quote.html?symb=TJX
https://money.cnn.com/quote/quote.html?symb=NKTR
https://money.cnn.com/quote/quote.html?symb=MAR
https://money.cnn.com/quote/quote.html?symb=TTWO
https://money.cnn.com/quote/quote.html?symb=VRSN

7. Extract the titles under the "Latest News" section (bottom middle of the web page.)

titles
Starbucks isn't done growing in America or China
Colorado voters reject limitations on fracking
Fed meets to weigh future rate hikes
Papa John's is still struggling to bring customers back
Applebee's is betting on stress eaters, and it's paying off
Foxconn hiring plans for Wisconsin plant under scrutiny
The myth of Donald Trump, CEO president
The dismantling of GE continues: It is selling yet another business
Lowe's is closing 47 stores in the US and Canada
Skip the fancy restaurant and try a food hall on your next business trip
German manufacturer Schaeffler is closing two UK plants over Brexit
Iran is still exporting oil as sanctions deadline passes

8. To understand the structure of the data in a web page, it is often useful to know what the underlying attributes are of the text you see. Extract the attributes (and their values) of the HTML element that holds the timestamp underneath the "What's Moving" table.

```
##                                                                                  stream
## "time_208526|264170|40451792|226354|44307|268937|213004|196573|273612|282500"
##                                                                              streamjstime
##                                                                            "1541538214000"
##                                                                          streamdateformat
```

```
##                                                                    "g%3Ai%3Asa%20x"
```

9. Extract the values of the blue percentage-bars from the "Trending Tickers" table (bottom right of the web page.) Hint: in this case, the values are stored under the "class" attribute.

pctbar
bars pct100
bars pct60
bars pct60
bars pct60
bars pct50
bars pct50
bars pct50
bars pct40
bars pct40
bars pct40

10. Get the links of all of the "svg" images on the web page.

links
/.element/cnnm-3.0/img/logo/logo_cnn.svg
/.element/cnnm-3.0/img/logo/logo_cnn.svg