Introduction to Data Science Project 1

Posted: 11/3/2020; 2020; **Due:** 21/3/2020

This projected

Introduction

You've been hired by a new space weather startup looking to disrupt the space weather reporting business. Your first project is to provide better data about the top 50 solar flares recorded so far than that shown by your competitor SpaceWeatherLive.com. To do this, they've pointed you to this messy HTML page from NASA (available here also) where you can get the extra data your startup is going to post in your new spiffy site. To focus on the main tasks of this project, the needed data from NASA has been scraped and it will be provided as text comma delimited file (NASA.csv).

Of course, you don't have access to the raw data for either of these two tables, so as an enterprising data scientist you will scrape this information directly from each HTML page using all the great tools available to you in Python. By the way, you should read up a bit on Solar Flares, coronal mass ejections, the solar flare alphabet soup, the scary storms of Halloween 2003, and sickening solar flares.

Part 1: Data scraping and preparation

Step 1: Scrape your competitor's data

Use Python to scrape data for the top 50 solar flares shown in SpaceWeatherLive.com. Steps to do this are:

1. `conda install` the following Python packages: `beautifulsoup4, requests, pandas, numpy`; Use Anaconda for Windows.
2. Use `requests` to get (as in, HTTP GET) the URL
3. Extract the text from the page
4. Use `BeautifulSoup` to read and parse the data, either as html or lxml

5. Use *prettify()* to view the content and find the appropriate table
6. Use *find()* to save the aforementioned table as a variable
7. Use *pandas* to read in the HTML file. HINT make-sure the above data is properly typecast.
8. Set reasonable names for the table columns, e.g., rank, x_classification, date, region, start_time, maximum_time, end_time, movie. *Pandas.columns* makes this very simple.

The result should be a data frame, with the first few rows as:

Out[406]:

	rank	x_class	date	region	start_time	max_time	end_time	movie
0	1	X28+	2003/11/04	486	19:29	19:53	20:06	MovieView archive
1	2	X20+	2001/04/02	9393	21:32	21:51	22:03	MovieView archive
2	3	X17.2+	2003/10/28	486	09:51	11:10	11:24	MovieView archive
3	4	X17+	2005/09/07	808	17:17	17:40	18:03	MovieView archive
4	5	X14.4	2001/04/15	9415	13:19	13:50	13:55	MovieView archive
5	6	X10	2003/10/29	486	20:37	20:49	21:01	MovieView archive
6	7	X9.4	1997/11/06	8100	11:49	11:55	12:01	MovieView archive
7	8	X9.3	2017/09/06	2673	11:53	12:02	12:10	MovieView archive

Step 2: Tidy the top 50 solar flare data

Your next step is to make sure this table is usable using pandas:

1. Drop the last column of the table, since we are not going to use it further.
2. Use datetime import to combine the date and each of the three time-columns into three datetime columns. You will see why this is useful later. *iterrows()* should prove useful here.
3. Update the values in the dataframe as you do this. *Set_value* should prove useful.
4. Set regions coded as - as missing (NaN). You can use *dataframe.replace()* here.

The result of this step should be a data frame with the first few rows as:

Out[407]:

	rank	x_class	start_time	max_time	end_time	region
0	1	X28+	2003-11-04 19:29:00	2003-11-04 19:53:00	2003-11-04 20:06:00	486
1	2	X20+	2001-04-02 21:32:00	2001-04-02 21:51:00	2001-04-02 22:03:00	9393
2	3	X17.2+	2003-10-28 09:51:00	2003-10-28 11:10:00	2003-10-28 11:24:00	486
3	4	X17+	2005-09-07 17:17:00	2005-09-07 17:40:00	2005-09-07 18:03:00	808
4	5	X14.4	2001-04-15 13:19:00	2001-04-15 13:50:00	2001-04-15 13:55:00	9415
5	6	X10	2003-10-29 20:37:00	2003-10-29 20:49:00	2003-10-29 21:01:00	486
6	7	X9.4	1997-11-06 11:49:00	1997-11-06 11:55:00	1997-11-06 12:01:00	8100
7	8	X9.3	2017-09-06 11:53:00	2017-09-06 12:02:00	2017-09-06 12:10:00	2673

Step 3: Scrape the NASA data

Next you need to load the data provided in NASA.csv. To get additional data about these solar flares. This table format is described here: http://cdaw.gsfc.nasa.gov/CME_list/radio/waves_type2_description.htm, and here:

Tasks

1. Upload the provided csv file into a Dataframe, and pay attention to writing the reference path to the file NASA.csv
2. Make sure that the file has been uploaded properly by printing the first few rows of the uploaded data.

The result of this step should be like:

Out[421]:

	Start_Date	Start_Time	End_Date	End_Time	Start_Frequency	End_Frequency	Flare_Location	Flare_Region	Importance	CME_Date	CME_Time	CPA
0	4/1/1997	14:00	4/1/1997	14:15	8000.0	4000.0	S25E16	8026	M1.3	4/1/1997	15:18	74.0
1	4/7/1997	14:30	4/7/1997	17:30	11000.0	1000.0	S28E19	8027	C6.8	4/7/1997	14:27	NaN
2	5/12/1997	5:15	5/14/1997	16:00	12000.0	80.0	N21W08	8038	C1.3	5/12/1997	5:30	NaN
3	5/21/1997	20:20	5/21/1997	22:00	5000.0	500.0	N05W12	8040	M1.3	5/21/1997	21:00	263.0
4	9/23/1997	21:53	9/23/1997	22:16	6000.0	2000.0	S29E25	8088	C1.4	9/23/1997	22:02	133.0
5	11/3/1997	5:15	11/3/1997	12:00	14000.0	250.0	S20W13	8100	C8.6	11/3/1997	5:28	240.0
6	11/3/1997	10:30	11/3/1997	11:30	14000.0	5000.0	S16W21	8100	M4.2	11/3/1997	11:11	233.0
7	11/4/1997	6:00	11/5/1997	4:30	14000.0	100.0	S14W33	8100	X2.1	11/4/1997	6:10	NaN

Step 4: Tidy the NASA table

Now, we tidy up the NASA table. Here we will code missing observations properly, recode columns that correspond to more than one piece of information, and treat dates and times appropriately.

1. Recode any missing entries as NaN. Refer to the data description in http://cdaw.gsfc.nasa.gov/CME_list/radio/waves_type2_description.htm to see how missing entries are encoded in each column. Be sure to look carefully at the actual data, as the NASA descriptions might not be completely accurate.
2. The CPA column (cme_angle) contains angles in degrees for most rows, except for halo flares, which are coded as "Halo". Create a new column that indicates if a row corresponds to a halo flare or not, and then replace Halo entries in the cme_angle column as NA.
3. The width column indicates if the given value is a lower bound. Create a new column that indicates if width is given as a lower-bound, and remove any non-numeric part of the width column.
4. Combine date and time columns for start, end and cme so they can be encoded as datetime objects.

The output of this step should be like this:

Start_DateTime	End_DateTime	Start_Frequency	End_Frequency	Flare_Location	Flare_Region	Importance	CME_DateTime	CPA	Width	Speed	Plots	Is_halo	Width_Low
1997-04-01 14:00:00	1997-04-01 14:15:00	8000.0	4000.0	S25E16	8026	M1.3	1997-04-01 15:18:00	74.0	79	312.0	PHTX	False	
1997-04-07 14:30:00	1997-04-07 17:30:00	11000.0	1000.0	S28E19	8027	C6.8	1997-04-07 14:27:00	NaN	360	878.0	PHTX	True	
1997-05-12 05:15:00	1997-05-14 16:00:00	12000.0	80.0	N21W08	8038	C1.3	1997-05-12 05:30:00	NaN	360	464.0	PHTX	True	
1997-05-21 20:20:00	1997-05-21 22:00:00	5000.0	500.0	N05W12	8040	M1.3	1997-05-21 21:00:00	263.0	165	296.0	PHTX	False	
1997-09-23 21:53:00	1997-09-23 22:16:00	6000.0	2000.0	S29E25	8088	C1.4	1997-09-23 22:02:00	133.0	155	712.0	PHTX	False	
1997-11-03 05:15:00	1997-11-03 12:00:00	14000.0	250.0	S20W13	8100	C8.6	1997-11-03 05:28:00	240.0	109	227.0	PHTX	False	
1997-11-03 10:30:00	1997-11-03 11:30:00	14000.0	5000.0	S16W21	8100	M4.2	1997-11-03 11:11:00	233.0	122	352.0	PHTX	False	

Part 2: Analysis

Now that you have data from both sources, let's start the analysis part.

Question 1: Replication

Can you replicate the top 50 solar flare table in SpaceWeatherLive.com exactly using the data obtained from NASA? That is, if you get the top 50 solar flares from the NASA table based on their classification (e.g., X28 is the highest), do you get data for the same solar flare events?

Include code used to get the top 50 solar flares from the NASA table (*be careful when ordering by classification*). Write a sentence or two discussing how well you can replicate the SpaceWeatherLive data from the NASA data.

Question 2: Integration

Write a function that finds the best matching row in the NASA data for each of the top 50 solar flares in the SpaceWeatherLive data. Here, you have to decide for yourself how you determine what is the best matching entry in the NASA data for each of the top 50 solar flares.

In your submission, include an explanation of how you are defining the best matching rows across the two datasets in addition to the code used to find the best matches. Finally, use your function to add a new column to the NASA dataset indicating its rank according to SpaceWeatherLive, if it appears in that dataset.

Question 3: Analysis

Prepare one plot that shows the top 50 solar flares in context with all data available in the NASA dataset. Here are some possibilities (you can do something else)

1. Plot attributes in the NASA dataset (e.g., starting or ending frequencies, flare height or width) over time. Use graphical elements (e.g., text or points) to indicate flares in the top 50 classification.

2. Do flares in the top 50 tend to have Halo CMEs? You can make a barplot that compares the number (or proportion) of Halo CMEs in the top 50 flares vs. the dataset as a whole.

3. Do strong flares cluster in time? Plot the number of flares per month over time, add a graphical element to indicate (e.g., text or points) to indicate the number of strong flares (in the top 50) to see if they cluster.

Submission

Prepare an Jupyter Notebook file that includes for each step in Part 1: (a) code to carry out the step discussed, (b) output showing the output of your code, similar to the examples above, and (c) a short description of how your code works. For questions 1 and 2 of Part 2, follow the instructions there. For Question 3 of part 2 provide: (a) a short description (2-3 sentences) of what the intent of your plot is (think in terms of our discussion on how we show variation, co-variation in terms of central trend, spread, skew etc.), (b) code to produce your plot, (c) a short text description of your plot, and (d) a sentence or two of interpretation of your plot (again think of variation, co-variation, etc.).

You are also required to prepare a PowerPoint presentation (Max. 5 slides) to present how did you tackle this project to complete it properly and the main findings of each part. Note that the presentation and discussion are compensated with 5 pts