Data Science – Aufgabe 1 - Scraping

Du wurdest von einem neuen Weltraumwetter-Startup eingestellt, das das Geschäft mit Weltraumwetterberichten revolutionieren möchte. Dein erstes Projekt besteht darin, bessere Daten über die bisher aufgezeichneten Top 50 Sonneneruptionen zu liefern als die, die von deinem Konkurrenten SpaceWeatherLive.com gezeigt werden. Zu diesem Zweck haben sie dich auf diese unübersichtliche HTML-Seite von der NASA (auch hier verfügbar) hingewiesen, von der du die zusätzlichen Daten beziehen kannst, die dein Startup auf eurer neuen schicken Website veröffentlichen wird.

Natürlich hast du keinen Zugang zu den Rohdaten dieser beiden Tabellen, also wirst du als unternehmungslustiger Datenwissenschaftler diese Informationen direkt von jeder HTML-Seite scrapen, unter Verwendung aller großartigen Tools, die dir in Python zur Verfügung stehen. Übrigens solltest du dich ein wenig über Sonneneruptionen, koronale Massenauswürfe, das Alphabet-Suppen-Menü der Sonneneruptionen.

Teil 1: Daten-Scraping und -Vorbereitung

Schritt 1: Daten deines Konkurrenten scrapen

Verwende Python, um Daten für die Top 50 Sonneneruptionen, die auf SpaceWeatherLive.com gezeigt werden, zu scrapen. Schritte dazu sind:

1. pip install oder conda install die folgenden Python-Pakete: beautifulsoup4, requests, pandas, numpy.
2. Verwende requests, um (wie in HTTP GET) die URL zu erhalten

   ! Sollte das Scrapen mit requests so nicht funktionieren suchen Sie im Web nach möglichen Lösungen. (andere bib? Müssen Sie den Server vllt. irgendwie täuschen?)
3. Extrahiere den Text von der Seite
4. Verwende BeautifulSoup, um die Daten zu lesen und zu parsen, entweder als html oder lxml
5. Verwende prettify(), um den Inhalt anzusehen und die entsprechende Tabelle zu finden
6. Verwende find(), um die besagte Tabelle als Variable zu speichern
7. Verwende pandas, um die HTML-Datei einzulesen. TIPP stelle sicher, dass die oben genannten Daten richtig typisiert sind.
8. Setze vernünftige Namen für die Tabellenspalten, z.B. Rang, X-Klassifikation, Datum, Region, Startzeit, Höchstzeit, Endzeit, Film. Pandas.columns macht dies sehr einfach.

Das Ergebnis sollte ein Datenrahmen (DataFrame) sein, mit den ersten Zeilen als:

```
Dimension: 50 × 8
rank x_class date region start_time max_time end_time movie
1  1 X28.0 2003/11/04 0486  19:29  19:53  20:06 MovieView archive
2  2 X20 2001/04/02 9393  21:32  21:51  22:03 MovieView archive
3  3 X17.2 2003/10/28 0486  09:51  11:10  11:24 MovieView archive
4  4 X17.0 2005/09/07 0808  17:17  17:40  18:03 MovieView archive
5  5 X14.4 2001/04/15 9415  13:19  13:50  13:55 MovieView archive
6  6 X10.0 2003/10/29 0486  20:37  20:49  21:01 MovieView archive
7  7  X9.4 1997/11/06  -  11:49  11:55  12:01 MovieView archive
8  8  X9.0 2006/12/05 0930  10:18  10:35  10:45 MovieView archive
9  9  X8.3 2003/11/02 0486  17:03  17:25  17:39 MovieView archive
10  10  X7.1 2005/01/20 0720  06:36  07:01  07:26 MovieView archive
 ... with 40 more rows
```

Schritt 2: Bereinigen der Top 50 Solarflare-Daten

Dein nächster Schritt besteht darin, sicherzustellen, dass diese Tabelle mit pandas verwendbar ist:

1. Entferne die letzte Spalte der Tabelle, da wir sie im Folgenden nicht verwenden werden.
2. Verwende den Import von datetime, um das Datum und jede der drei Zeitangaben in drei datetime-Spalten zu kombinieren. Du wirst später sehen, warum dies nützlich ist. iterrows() könnte sich hier als nützlich erweisen.
3. Aktualisiere die Werte im DataFrame, während du dies tust. set_value könnte sich als nützlich erweisen.
4. Setze Regionen, die als - kodiert sind, auf fehlend (NaN). Du kannst hier dataframe.replace() verwenden.

Das Ergebnis dieses Schritts sollte ein DataFrame mit den ersten Zeilen sein:

```
A dataframe: 50 × 6

rank x_class  start_datetime  max_datetime  end_datetime region

1  1 X28.0 2003-11-04 19:29:00 2003-11-04 19:53:00 2003-11-04 20:06:00 0486

2  2 X20 2001-04-02 21:32:00 2001-04-02 21:51:00 2001-04-02 22:03:00 9393

3  3 X17.2 2003-10-28 09:51:00 2003-10-28 11:10:00 2003-10-28 11:24:00 0486

4  4 X17.0 2005-09-07 17:17:00 2005-09-07 17:40:00 2005-09-07 18:03:00 0808

5  5 X14.4 2001-04-15 13:19:00 2001-04-15 13:50:00 2001-04-15 13:55:00 9415

6  6 X10.0 2003-10-29 20:37:00 2003-10-29 20:49:00 2003-10-29 21:01:00 0486

7  7  X9.4 1997-11-06 11:49:00 1997-11-06 11:55:00 1997-11-06 12:01:00 <NA>

8  8  X9.0 2006-12-05 10:18:00 2006-12-05 10:35:00 2006-12-05 10:45:00 0930

9  9  X8.3 2003-11-02 17:03:00 2003-11-02 17:25:00 2003-11-02 17:39:00 0486

10  10  X7.1 2005-01-20 06:36:00 2005-01-20 07:01:00 2005-01-20 07:26:00 0720

... with 40 more rows
```

Schritt 3: NASA-Daten scrapen

Als Nächstes müssen Sie die Daten auf http://cdaw.gsfc.nasa.gov/CME_list/radio/waves_type2.html (auch verfügbar unter http://www.hcbravo.org/IntroDataSci/misc/waves_type2.html) scrapen, um zusätzliche Daten über diese Sonneneruptionen zu erhalten. Das Format dieser Tabelle wird hier beschrieben: http://cdaw.gsfc.nasa.gov/CME_list/radio/waves_type2_description.htm:

```
The Wind/WAVES type II burst catalog: A brief description

URL:
[http://cdaw.gsfc.nasa.gov/CME_list/radio/waves_type2.html](http://cdaw.gsfc.nasa.gov/CME_list/radio/waves_type2.html).

This is a catalog of type II bursts observed by the Radio and Plasma Wave (WAVES) experiment on board the Wind spacecraft and the associated coronal mass ejections (CMEs) observed by the Solar and Heliospheric Observatory (SOHO) mission. The type II burst catalog is derived from the Wind/WAVES catalog available at
[http://ssed.gsfc.nasa.gov/waves/data_products.html](http://ssed.gsfc.nasa.gov/waves/data_products.html) by adding a few missing events.

The CMEs in this catalog are called radio-loud CMEs because of their ability to produce type II radio bursts. The CME sources are also listed, as derived from the Solar Geophysical Data listing or from inner coronal images such as Yohkoh/SXT and SOHO/EIT. Some solar sources have also been obtained from Solarsoft Latest Events Archive after October 1, 2002:
[http://www.lmsal.com/solarsoft/latest_events_archive.html](http://www.lmsal.com/solarsoft/latest_events_archive.html)

Explanation of catalog entries:

Column 1: Starting date of the type II burst (yyyy/mm/dd format)

Column 2: Starting time (UT) of the type II burst (hh:mm format)

Column 3: Ending date of the type II burst (mm/dd format; year in Column 1 applies)

Column 4: Ending time of the Type II burst (hh:mm format)

Column 5: Starting frequency of type II burst (kHz) [1]

Column 6: Ending frequency of type II burst (kHz) [1]

Column 7: Solar source location (Loc) of the associated eruption in heliographic coordinates [2]

Column 8: NOAA active region number (NOAA) [3]

Column 9: Soft X-ray flare importance (Imp) [4]

Column 10: Date of the associated CME (mm/dd format, Year in Column 1 applies) [5]

Column 11: Time of the associated CME (hh:mm format)

Column 12: Central position angle (CPA, degrees) for non-halo CMEs [6]

Column 13: CME width in the sky plane (degrees) [7]

Column 14: CME speed in the sky plane (km/s)

Column 15: Link to the daily proton, height-time, X-ray (PHTX) plots [8]

Notes

[1] ???? indicate that the starting and ending frequencies are not determined.

[2] Heliographic coordinates. S25E16 means the latitude is 25 deg south and 16 deg east (source located in the southeast quadrant of the Sun. N denotes northern latitudes and W denotes western longitudes. Entries like SW90 indicate that the source information is not complete, but we can say that the eruption occurs on the west limb but at southern latitudes; if such entries have a subscript b (e.g.,
```

NE90b) it means that the source is behind the particular limb. This information is usually gathered from SOHO/EIT difference images, which show dimming above the limb in question. Completely backside events with no information on the source location are marked as “back”.

[3] If the active region number is not available or if the source region is not an active region, the entry is “––”. Filament regions are denoted by “FILA” or “DSF” for disappearing solar filament.

[4] Soft X-ray flare size (peak flux in the 1-8 A channel) from GOES. “–-” means the soft X-ray flux is not available.

[5] Lack of SOHO observations are noted as “LASCO DATA GAP”. Other reasons are also noted if there is no CME parameters measured.

[6] The central position angle (CPA) is meaningful only for non-halo CMEs. For halo CMEs, the entry is “Halo”. For halo CMEs, the height-time measurements are made at a position angle where the halo appears to move the fastest. This is known as the measurement position angle (MPA) and can be found in the main catalog (http://cdaw.gsfc.nasa.gov/CME_List).

[7] Width = 360 means the CME is a fill halo (see [6]). For some entries, there is a prefix “>”, which means the reported width is a lower limit.

[8] ‘PHTX’ (proton, height-time, X-ray) link to three-day overview plots of solar energetic particle events (protons in the >10, >50 and >100 MeV GOES channels).

Links:

The CMEs and the type II bursts can be viewed together using the c2rdif_waves.html movies linked to the starting frequency (Column 5). The c3rdif_waves.html movies are linked to the ending frequencies (Column 6). The CMEs and the GOES flare light curves for a given type II burst can be viewed from the Javascript movies linked to the CME date (Column 10). The height-time plots (linear and quadratic) of the CMEs are linked to the CME speed (Column 14).

PHTX plots are linked to Column 15.

If you have questions, contact: Nat Gopalswamy
(gopals@ssedmail.gsfc.nasa.gov)

This work is supported by NASA’s Virtual Observatories Program

Aufgaben zu Schritt 3

1. Verwenden Sie BeautifulSoup-Funktionen (z.B. find, findAll) und String-Funktionen (z.B. split und integrierte Slicing-Fähigkeiten), um jede Zeile der Daten als langen String zu erhalten. Erstellen Sie an diesem Punkt ein DataFrame, damit es einfacher ist, melt oder wide_to_long für die nächsten Schritte zu verwenden.
2. Verwenden Sie string::split und Listenverständnisse oder Ähnliches, um jede Zeile des Textes in eine Datenzeile zu trennen. Wählen Sie geeignete Namen für die Spalten.

Das Ergebnis dieses Schrittes sollte ähnlich sein wie:

```
Dimension: 482 × 14

start_date start_time end_date end_time start_frequency end_frequency flare_location flare_region

* <chr>  <chr>  <chr>  <chr> <chr> <chr>  <chr>  <chr>

1  1997/04/01  14:00  04/01  14:15  8000  4000 S25E16 8026

2  1997/04/07  14:30  04/07  17:30 11000  1000 S28E19 8027

3  1997/05/12  05:15  05/14  16:00 12000  80 N21W08 8038

4  1997/05/21  20:20  05/21  22:00  5000 500 N05W12 8040

5  1997/09/23  21:53  09/23  22:16  6000  2000 S29E25 8088

6  1997/11/03  05:15  11/03  12:00 14000 250 S20W13 8100

7  1997/11/03  10:30  11/03  11:30 14000  5000 S16W21 8100

8  1997/11/04  06:00  11/05  04:30 14000 100 S14W33 8100

9  1997/11/06  12:20  11/07  08:30 14000 100 S18W63 8100

10 1997/11/27  13:30  11/27  14:00 14000  7000 N17E63 8113

... with 472 more rows, and 6 more variables: flare_classification <chr>, cme_date <chr>, cme_time
<chr>, cme_angle <chr>, cme_width <chr>, cme_speed <chr>
```

Schritt 4: Die NASA-Tabelle bereinigen

Nun bereinigen wir die NASA-Tabelle. Hierwerden wir fehlende Beobachtungen ordnungsgemäß kodieren, Spalten, die mehr als einer Information entsprechen, neu kodieren und Daten und Zeiten angemessen behandeln.

1. Kodieren Sie fehlende Einträge als NaN. Beziehen Sie sich auf die Datenbeschreibung unter http://cdaw.gsfc.nasa.gov/CME_list/radio/waves_type2_description.htm (und oben), um zu sehen, wie fehlende Einträge in jeder Spalte kodiert werden. Achten Sie sorgfältig auf die tatsächlichen Daten, da die NASA-Beschreibungen möglicherweise nicht vollständig genau sind.
2. Die CPA-Spalte (cme_angle) enthält Winkel in Grad für die meisten Zeilen, außer für Halo-Eruptionen, die als Halo kodiert sind. Erstellen Sie eine neue Spalte, die angibt, ob eine Zeile einer Halo-Eruption entspricht oder nicht, und ersetzen Sie dann Halo-Einträge in der cme_angle-Spalte durch NA.
3. Die Breite-Spalte gibt an, ob der angegebene Wert eine Untergrenze ist. Erstellen Sie eine neue Spalte, die angibt, ob die Breite als Untergrenze angegeben wird, und entfernen Sie jeglichen nicht-numerischen Teil der Breite-Spalte.
4. Kombinieren Sie Datums- und Zeitangaben für Start, Ende und CME, damit sie als Datetime-Objekte kodiert werden können.

Das Ergebnis dieses Schrittes sollte ähnlich sein wie dieses:

```
start_datetime  end_datetime start_frequency end_frequency flare_location flare_region importance
cme_datetime  cpa width speed  plot is_halo width_lower_bound

0 1997-04-01 14:00:00 1997-04-01 14:15:00  8000  4000 S25E16 8026 M1.3 1997-04-01 15:18:00 74  79 312
PHTX False False

1 1997-04-07 14:30:00 1997-04-07 17:30:00 11000  1000 S28E19 8027 C6.8 1997-04-07 14:27:00  NaN 360 878
PHTX  True False

2 1997-05-12 05:15:00 1997-05-14 16:00:00 12000  80 N21W08 8038 C1.3 1997-05-12 05:30:00  NaN 360 464
PHTX  True False

3 1997-05-21 20:20:00 1997-05-21 22:00:00  5000 500 N05W12 8040 M1.3 1997-05-21 21:00:00  263 165 296
PHTX False False

4 1997-09-23 21:53:00 1997-09-23 22:16:00  6000  2000 S29E25 8088 C1.4 1997-09-23 22:02:00  133 155 712
PHTX False False

5 1997-11-03 05:15:00 1997-11-03 12:00:00 14000 250 S20W13 8100 C8.6 1997-11-03 05:28:00  240 109 227
PHTX False False
```

Teil 2: Analyse

Jetzt, da Sie Daten von beiden Websites haben, beginnen wir mit einigen Analysen.

Frage 1: Replikation

Können Sie die Tabelle der Top-50-Sonnenereignisse auf SpaceWeatherLive.com genau mit den von der NASA erhaltenen Daten replizieren? Das heißt, wenn Sie die Top-50-Sonnenereignisse aus der NASA-Tabelle basierend auf ihrer Klassifikation erhalten (z.B. ist X28 die höchste), erhalten Sie Daten für dieselben Sonnenereignis-Ereignisse?

Fügen Sie den Code ein, den Sie verwendet haben, um die Top-50-Sonnenereignisse aus der NASA-Tabelle zu erhalten (achten Sie auf die Reihenfolge nach Klassifikation). Schreiben Sie ein oder zwei Sätze darüber, wie gut Sie die Daten von SpaceWeatherLive mit den NASA-Daten replizieren können.

Frage 2: Integration

Schreiben Sie eine Funktion, die die am besten passende Zeile in den NASA-Daten für jedes der Top-50-Sonnenereignisse in den Daten von SpaceWeatherLive findet. Hier müssen Sie selbst entscheiden, wie Sie festlegen, was der am besten passende Eintrag in den NASA-Daten für jedes der Top-50-Sonnenereignisse ist.

In Ihrer Abgabe schließen Sie eine Erklärung ein, wie Sie die am besten passenden Zeilen zwischen den beiden Datensätzen definieren, zusätzlich zu dem Code, den Sie verwendet haben, um die besten Übereinstimmungen zu finden. Verwenden Sie anschließend Ihre Funktion, um eine neue Spalte zum NASA-Datensatz hinzuzufügen, die seinen Rang gemäß SpaceWeatherLive angibt, falls er in diesem Datensatz erscheint.

Frage 3: Analyse

Bereiten Sie eine Grafik vor, die die Top-50-Sonnenereignisse im Kontext mit allen verfügbaren Daten im NASA-Datensatz zeigt. Hier sind einige Möglichkeiten (Sie können etwas anderes tun):

1. Attribute im NASA-Datensatz über die Zeit darstellen (z.B. Anfangs- oder Endfrequenzen, Flare-Höhe oder -Breite). Verwenden Sie grafische Elemente (z.B. Text oder Punkte), um Flares in der Top-50-Klassifikation anzuzeigen.
2. Neigen Flares in den Top 50 dazu, Halo-CMEs zu haben? Sie können ein Balkendiagramm erstellen, das die Anzahl (oder den Anteil) der Halo-CMEs in den Top-50-Flares im Vergleich zum gesamten Datensatz vergleicht.
3. Cluster starke Flares zeitlich? Zeichnen Sie die Anzahl der Flares pro Monat über die Zeit und fügen Sie ein grafisches Element hinzu (z.B. Text oder Punkte), um die Anzahl starker Flares (in den Top 50) anzuzeigen und zu sehen, ob sie sich gruppieren.

Abgabe

Bereiten Sie eine Jupyter-Notebook-Datei vor, die für jeden Schritt in Teil 1 enthält:

(a) Code, um den besprochenen Schritt durchzuführen,

(b) Ausgabe, die die Ausgabe Ihres Codes zeigt, ähnlich wie in den obigen Beispielen, und

(c) eine kurze Prosabeschreibung, wie Ihr Code funktioniert.

Für die Fragen 1 und 2 des Teils 2 folgen Sie den dortigen Anweisungen. Für Frage 3 des Teils 2 liefern Sie:

(a) eine kurze Beschreibung (2 Sätze) dessen, was das Ziel Ihrer Grafik ist (denken Sie in Bezug auf unsere Diskussion darüber, wie wir Variation, Kovariation in Bezug auf zentrale Tendenz, Streuung, Schiefe usw. zeigen),

(b) Code, um Ihre Grafik zu erstellen,

(c) eine kurze Textbeschreibung Ihrer Grafik und

(d) ein oder zwei Sätze der Interpretation Ihrer Grafik (denken Sie wieder an Variation, Kovariation usw.).

Reichen Sie das resultierende .ipynb in Felix ein und präsentieren Sie es.

Gruppenarbeit

Sie sind ermutigt, in Zweiergruppen zu arbeiten.