"El saber de mis hijos hará mi grandeza"

Universidad de Sonora
Departamento de Física
Licenciatura en Física
Física Computacional-1
2016-2

Limpieza y preparación de datos usando Emacs

Danira Rios Quijada
Profesor: Carlos Lizárraga Celaya

27 de agosto de 2016

Resumen

La estructura del siguiente reporte consiste en explicar brevemente con que datos estamos trabajanndo y como utilizar distintos recursos tanto de bash como de emacs para bajarlos, archivarlos y filtrarlos, teniendo como objetivo que en un futuro empezaremos a trabajar con ellos.

1. Introducción

Cuando reuqerimos analizar con una gran cantidad de datos, siempre es difícil y tedioso trabajar con ellos, sin embargo existen distintas herramientas computacionales para ayudarnos un poco con este trabajo.
Recordemos que siempre que queremos analizar como se comportan o que patrones forman ciertos datos, primero tenemos que saber con que datos necesitamos trabajar, una vez con esto en claro, el siguiente paso es como obtener solo los datos relevantes, recalquemos que en el análisis de datos, una gran parte del trabajo es limpiar la información, pues bien, la actividad de esta semana estuvo enfocada a esta tarea y la aplicación de las herramientas adecuadas.

2. ¿En qué consisten los datos?

Los datos con los que trabajamos en esta actividad fueron obtenidos de la plataforma de ciencias atmosféricas de la Universidad de Wyoming[1], son datos obtenidos de sondeos atmosféricos en varias ubicaciones de Norte América, para la actividad decidí obtener los datos de sondeo de Tucson, Arizona, a lo largo del 2015.
Hay dos tipos de datos en los archivos descargados, primeramente los que se dan como función de la altura, estos estan localizados en una tabla, en las distintas columnas, además de la altura por supuesto, aparecen datos como la presión, la temperatura o la dirección del viento, entre otros.
Otro tipo de datos que aparecen en los archivos son los datos derivables de los anteriores, que son por ejemplo el índice K, el cuál ayuda a determinar si puede haber lloviznas o tormentas, o el índice sweat (severe weather threat index), un índice atribuido a Miller (1972), usado para analizar el potencial de tormentas severas.[2]

3. ¿Cómo obtener los datos?

Decidí bajar datos de un año completo, para obtenerlos modifique el script de 10 años que se nos brindó, teniendo como resultado el siguiente script:[3]

```
#!/bin/bash

# Despues de editar: chmod 755 script1.sh
# Para ejecutar: ./script1.sh

IFS=":"
# Months by number of days
LISTM31="01:03:05:07:08:10:12"
#LISTM31="01:03:05:07"
LISTM30="04:06:09:11"
#LISTM30="04:06"
LISTM28="02"

# Months 31 days
for i in $LISTM31 ; do
wget  "http://weather.uwyo.edu/cgi-bin/sounding?
    region=naconf&TYPE=TEXT%3ALIST&YEAR=2015&
    MONTH=$i&FROM=0112&TO=3112&STNM=72440"
        /bin/sleep 5
done
# Months 30 days
for i in $LISTM30 ; do
wget  "http://weather.uwyo.edu/cgi-bin/sounding?
    region=naconf&TYPE=TEXT%3ALIST&YEAR=2015&
    MONTH=$i&FROM=0112&TO=3012&STNM=72440"
        /bin/sleep 5
done
# Feb. 28 days
for i in $LISTM28 ; do
wget "http://weather.uwyo.edu/cgi-bin/sounding?
    region=naconf&TYPE=TEXT%3ALIST&YEAR=2015&
    MONTH=$i&FROM=0112&TO=2812&STNM=72440"
        /bin/sleep 5
done
done
```

Para correr el script, tuve que cambiar el tipo de archivo, utilizando el comando chmod, cambiandolo de rw- a rwx, para hacerlo ejecutable. Una vez que obtuve los datos de cada mes del año 2015, proseguí a limpiar los datos.

4. Limpieza de datos

Para la limpieza de datos decidí trabajar solo con el mes de enero, debido a que tuve algunas dudas con respecto a como tenían que ser los archivos resultantes.
En la actividad se nos pedía producir dos archivos, uno con los datos en función de la altura,(osea los que se ubican en la tabla), y el otro archivo con los datos derivables de los datos de la tabla,(eliminar los datos de la tabla).
Para el primer archivo utilizé el editor emacs, seleccionando los datos que quería eliminar con ctrl+barra espaciadora, tumbandolos con ctrl+w, después con esc+x, escribí el comando query-replace, puse el contenido tumbado con ctrl+y, para dar la instrucción de que queria remplazar todo eso con un espacio. Siempre trabajé con una copia del archivo original, por cualquier inconveniente, una vez que obtuve el archivo lo guardé como: heightfun.csv.
Para el segundo archivo, modifiqué un script que se nos brindó, para solo quedarnos con la información derivable que queriamos utilizar. El script quedó de la siguiente manera:[3]

```
# Script para filtrar renglones de un archivo que
contengan las cadenas de caracteres dados

#!/bin/bash

egrep -v 'PRES|hPa'  jan2015.txt | egrep
'72440|Show|LIFT|SWEAT|K|Totals|virtual|
CAPV|Lifted|thickness|Precip' > jantuc15.csv
```

Como se muestra en el script, se trabajo desde el archivo original jan2015.txt y el archivo resultante fue jantuc15.csv, el cual después modifiqué a vander.csv.

5. Bibliografía

Referencias

[1] University of Wyoming; Department of Atmospheric Science *Datos, sondeo atmosférico*, (2015), recuperado (2016, 25 de agosto). Desde: `http://weather.uwyo.edu/upperair/sounding.html`

[2] American Meteorological Society, *Meteorology Glossary*, (2012, 25 de abril). Recuperado (2016,27 de agosto), Desde: `http://glossary.ametsoc.org/wiki/Stability_index`

[3] Carlos Lizárraga Celaya, *Actividad dos, curso de computacional 1*, (2016, 25 de agosto).Recuperado (2016, 25 de agosto), Desde: `http://computacional1.pbworks.com/w/page/110289883/Actividad20220282016-229`