
Universidad de Sonora

División de Ciencias Exactas y Naturales

Física Computacional

# Limpieza y preparación de datos usando Emacs

*Autor:*
Carlos Alí Medina Leal
*Profesor:*
Carlos Lizárraga Celaya

30 de Agosto del 2016

# 1. Un breve resumen...

En esta práctica veremos cómo limpiar datos y prepararlos para ser usados en programas que leen datos como excel, libreOffice y especializados para el análisis de datos, tales como Python, R, etc. Para esto, usaremos el editor de texto Emacs, que nos resultará muy útil para estos casos en los que tenemos muchísimos datos y texto basura.

# 2. Introducción

Actualmente existen muchas opciones de editores de texto, pero el editor Gnu Emacs es históricamente uno de los editores más versátiles y potentes y que fue desarrollado por Richard Stallman.

Para traer los datos a limpiar utilizamos una página que proporciona datos de sondeos de la atmósfera, creado por la Universidad de Wyoming, usando el siguiente script:

```
# Para bajar datos (adaptar a un sólo año)
#!/bin/bash
# Despues de editar: chmod 755 script1.sh
# Para ejecutar: ./script1.sh

IFS=":"
LOOPY=2015
LISTM="1:2:3:4:5:6:7:8:9:10:11:12"
LISTD="1:2:3:4:5:6:7:8:9:10:11:12:13:14:15:16:17:18:19:20:21:22:23:
24:25:26:27:28:29:30:31"

# Script para el año 2015, dentro del URL:  YEAR=2015
# Solo el sondeo de las 12Z
H="12"
for i in $LISTM ; do
    for j in $LISTD ; do
/usr/local/bin/wget "http://weather.uwyo.edu/cgi-bin/sounding?region=na
conf&TYPE=TEXT%3ALIST&YEAR=2015&MONTH=$i&FROM=$j$H&TO=$j$H&STNM=76692";
        /bin/sleep 10
    done
done
```

Este script sirve para bajar un año de datos en una estación en particular, se puede modificar para bajar 2, 3, 10 años de datos, y para cambiar la estación de la cual quieres los datos; en mi caso bajé doce meses de datos de la estación número 72293, que es San Diego.

Después de modificar el script, debemos ir a la terminal y cambiar los "derechos" del archivo, cambiando su modo y escribiendo "*chmod* 755 nombre-del-script.sh" y después se ejecuta con ".\"

# 3. Desarrollo y limpieza

Cuando se descargan los datos, se bajan miles de datos que no nos interesa analizar, por lo que debemos filtrar los datos que nos interesen y enviarlos a un nuevo archivo, todo esto se logra con el comando:

```
egrep -v 'PRES|hPa' soundings-archivo.txt | egrep'76692|Show|LIFT|SWEAT|
K|Totals|virtual|CAPV|Lifted|thickness|Precip'>nombre-del-archivo.csv
```

Este comando filtra los datos que etiquetamos y los envía a un archivo ".csv"(comma-separated values), el cual tendremos que limpiar y separar por comas, como el archivo lo requiere.

Es aquí cuando necesitamos de emacs, el cual nos servirá para empezar a limpiar los datos y a separarlos por comas. Lo primero que se hace al abrir el archivo es borrar los primeros datos que no ocupemos, como los datos que nos dicen la estación y la parte de:

```
<LINK REL="StyleSheet" HREF="/resources/select.css" TYPE="text/css">
```

Para esto, se emplea el siguiente proceso:

1. Se presiona "Ctrl + space"para comenzar a seleccionar lo innecesario
2. Al llegar al final de la selección, se presiona "Ctrl + W"para eliminar lo seleccionado, a lo que después se le presiona "Ctrl + Y"para deshacer
3. Se viaja hasta el principio del archivo en la primera línea, con las teclas "Esc y Menor que"
4. Se presiona "Esc y x", para después escribir el comando "query-replace"ó "query-rep"para reemplazar lo que no queramos

5. nos preguntará qué deseamos reemplazar y presionamos “Ctrl + Y”para elegir el texto anteriormente seleccionado
6. Nos preguntará con qué queremos reemplazar lo seleccionado y escribiremos un espacio, y después presionaremos “!”para señalar que sí estamos seguros y lo hará de inmediato

Lo que queda será limpiar el archivo de las palabras y símbolos que no sean datos, repitiendo el proceso pero reemplazando con “, ”para finalmente tener datos separados con comas, listos para leer con cualquier programa analizador de datos. Recordar también el nombre de cada columna para al inicio escribirla y nombrarlas.

# 4. Bibliografía

$N$C State University/ Climate Education for K-12/ Structure of the Atmosphere : $http : //climate.ncsu.edu/edu/k12/.AtmStructure$, Recuperado el 1 de Septiembre del 2016.