Elementos Esenciales de Programación para Científicos en Formación

Taller Nº 2

Fecha de entrega: Martes 9 de junio de 2020, 12:30hs

1. Enunciado

Se administra un sistema de registro de variables hidrometeorológicas, entre los que está la medición de temperatura. Así, nos brindan los datos recolectados en diferentes momentos para distintos sensores distribuidos en una región. Estas mediciones están almacenadas en un archivo de texto en formato CSV.

Se pide **diseñar** e **implementar** un programa que cargue un archivo de estas características y compute promedios a lo largo de una ventana deslizante de tamaño fijo.

Por ejemplo, podría invocarse de la siguiente forma desde una consola linux:

```
python3 taller2.py entrada.csv salida.csv 20 prom
```

Mientras que desde ipython (ventana inferior derecha de Spyder) se ejecutaría así:

```
%run taller2.py entrada.csv salida.csv 20 prom
```

Se recibirán como parámetros por línea de comandos el nombre del archivo de entrada, el nombre del archivo de salida y el tamaño de la ventana, en ese orden. Opcionalmente, se indicará el método elegido para tratar los valores faltantes (marcados como `NA`), que son comunes para el tipo de sensores utilizados: default (`def`), promedio (`prom`), mediana (`med`) y distribución (`dist`).

La llamada por línea de comandos de más arriba indica que se debe cargar el archivo de entrada `entrada.csv`, escribir los promedios en `salida.csv`, con un tamaño de la ventana deslizante de 20, y usando el promedio (`prom`) para completar los valores faltantes.

2. Formato de entrada

Cada línea del archivo de entrada tiene los siguientes campos:

```
timestamp,temp_1,temp_2,temp_3,temp_4,...,temp_n
```

El primer campo de cada fila es un *timestamp*, que se encuentra ordenado crecientemente, y que nos indica la fecha y la hora en la que se realizó la medición. Este campo se encuentra en el formato ISO-8601. Por ejemplo, el *timestamp* que representa las 11:09hs del 23 de febrero de 2020 es `2020-02-23T11:09:00`.

Los campos `temp_i` son los valores de temperatura de cada sensor para esa medición, cuya cantidad no varía entre las distintas filas de un archivo, aunque sí puede hacerlo entre distintos archivos. Ocasionalmente algún sensor puede fallar al momento de realizar la medición, en cuyo caso el archivo de entrada tendrá el texto `NA` en lugar del valor de la temperatura.

3. Cálculo del promedio

El cálculo del promedio se realiza según el procedimiento esquematizado en la Figura 1. Allí, se indica con `m` la cantidad de sensores (cantidad de columnas sin contar el timestamp), `w` la cantidad de mediciones (cantidad de filas) y `n` el tamaño de la ventana deslizante. Los promedios se calculan para cada sensor independientemente, tomando `n` mediciones contiguas dentro de la ventana.

Como se indicó anteriormente, se podrán usar cuatro métodos para tratar con los valores `NA`:

- **Default:** Si alguno de los valores de la ventana fuera `NA`, el promedio debe dar `NA`
- **Promedio:** Los `NA` de una ventana se completarán con el promedio de ese sensor para esa misma ventana. Si todos fueran `NA`, el valor será `NA`
- **Mediana:** *(opcional)* Los `NA` de una ventana se completarán con la mediana de ese sensor para esa misma ventana. Si todos fueran `NA`, el valor será `NA`
- **Distribución:** *(opcional)* Los `NA` se completarán con nuevos valores generados según una distribución estadística estimada a partir del resto de los valores de la ventana. ¿Qué cambiaría si se usara la columna completa de valores para estimar la distribución? ¿Sería correcto?

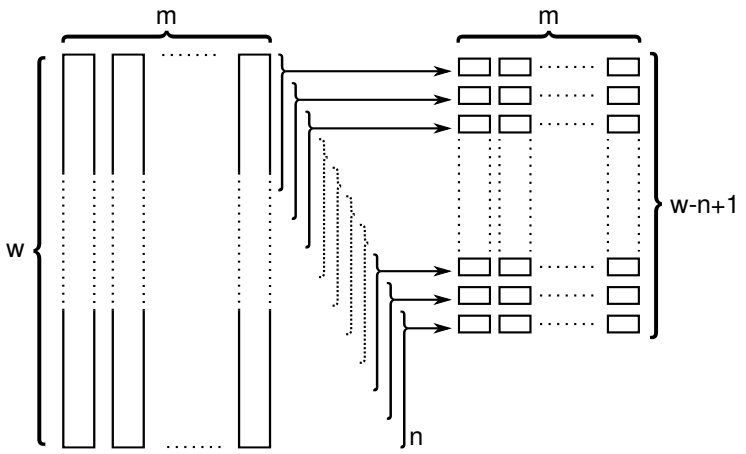

Figura 1: Cálculo del promedio por ventana deslizante

En cualquier caso, el valor del promedio por cada ventana deberá aparecer en la salida como un número con dos cifras decimales significativas. La cantidad de filas del archivo resultante (es decir, la cantidad de ventanas de tamaño `n`) estará determinada por el valor `w-n+1`.

4. Formato de salida

Cada uno de los promedios debe ser almacenado en el archivo de salida en el orden en el que son calculados, y respetando el siguiente formato:

```
lapso,promedio_1,promedio_2,promedio_3,...,promedio_n
```

Aquí, `lapso` indica la cantidad de segundos entre la primera medición y la última para la ventana, ya que se supone que los valores están ordenados temporalmente.

Se puede utilizar `datetime.datetime.strptime(timestamp, '%Y-%m-%dT%H:%M:%S')` para interpretar los *timestamps* según ISO-8601. Luego, si `t1` y `t2` son variables que contienen los timestamps de la primera y última fila, el lapso en segundos se obtiene como `(t2 - t1).total_seconds()`.

El archivo de salida debe tener tantas líneas como ventanas deslizantes de `n` elementos haya en el archivo de entrada, o ninguna en caso de no contar con suficientes mediciones en el archivo de entrada. Por ejemplo, si la entrada fuera un archivo con el siguiente contenido:

```
2020-05-03T02:17:24,NA,47.60,39.31,2.69,31.56
2020-05-03T02:17:26,19.13,49.44,54.12,24.34,31.85
2020-05-03T02:19:12,46.12,76.07,31.86,14.37,17.18
2020-05-03T02:20:40,15.67,55.12,61.51,64.71,40.26
```

La salida correspondiente para una ventana deslizante de tamaño 3, y usando el método default para tratar los valores faltantes, sería:

```
108.0,NA,57.70,41.76,13.80,26.86
194.0,26.97,60.21,49.16,34.47,29.76
```

Condiciones de entrega:

- Generar un único archivo Python con la solución, con los comentarios correspondientes.
- Se evaluará la correctitud del código producido, su claridad y legibilidad; y el uso de la herramienta `git`.
- Asegurarse de que todos/as los/as docentes del curso tengan permisos de lectura (*Developer*) en el repositorio de Gitlab (ver usuarios en el Campus). Se descargará la última versión de los archivos directamente de ahí, luego de ser informados de que el taller se encuentra listo.
- Enviar el aviso por correo electrónico a la lista de los docentes de la materia: **elemprog-doc@dc.uba.ar**.
- El mail deberá tener el siguiente *subject*:
  "[Taller 2]: <apellido> <DNI>". Por ejemplo, "[Taller 2]: Howard 36114601".
- **Importante**: Sólo se admite la entrega por medio de `Gitlab`. Se buscarán el/los archivos en la carpeta correspondiente (en este caso, dentro de `Taller2`).
  No se aceptarán las entregas de código por mail o sólo en forma impresa.