Pandas

Data Mining

Ester Vidaña Vila

laSalle ENG
Universitat Ramon Llull

Pandas

- Su nombre proviene de el concepto “Panel de Datos” o “Panel Data”.
- Es una librería de software para Python que nos ofrece estructuras de datos (pandas DataFrame) y operaciones para manipular estas estructuras y series temporales.
- Nos permite leer datos de listas, diccionario, de otro DataFrame, tablas de Excel, ficheros CSV, JSON, tablas SQL, etc.
- También se pueden crear DataFrames a partir de diccionarios, listas, tuplas, etc.

Funciones interesantes de pandas

- **df = pd.DataFrame(datos):** crea un pandas dataframe sobre la variable df.
- **df.head():** muestra las cinco primeras filas del dataframe df.
- **df.tail():** muestra las cinco últimas filas del dataframe df.
- **df.index**: muestra las filas que tiene el dataframe df.
- **df.columns**: muestra el nombre de las columnas de df.
- **df.dtypes**: muestra el tipo de variable de cada columna de df.
- **df.shape**: nos dice el número de filas x columnas que tiene el df.

Cómo acceder a determinada fila o columna

- **df[[”col1", ”col2"]]:** devuelve los datos de df las dos **columnas** que se han pasado entre [ ].

- **df.iloc[[5]]:** devuelve los datos de df de la **fila** 5 (se ha especificado dentro de los [ ]).

- **df.iloc[[1,3,5],[0,3]]:** devuelve los datos de df de las filas 1, 3 i 5 i las columas 0 (el índice), 1 y 2.

Cómo eliminar una fila o columna

- **df.drop(5):** eliminamos la fila 5 del dataframe df.

- **df.drop(”Col1”, axis = 1**): eliminamos la columna 1 del dataframe df.

- **df.reset_index():** tras eliminar columnas, podemos resetear el índice del dataframe con esta función. Al hacerlo, crea una nueva columna con el índice antiguo por defecto. Si no queremos mantener este índice, dentro de los paréntesis ponemos el parámetro **drop = True**

Ordenar o contar

- **df.sort_values(“Col1”):** nos permite ordenar las filas del dataframe según el orden ascendete de los valores de “Col1”.
- **df.sort_values(“Col1” , ascending=False):** nos permite ordenar las filas del dataframe según el orden ascendete de los valores de “Col1”.
- **df[”Col1"].unique():** devuelve los valores únicos que hay en toda la columna ”Col1” del dataframe df.
- **df[”Col1"].nunique():** devuelve el número de elementos únicos que hay en toda la columna ”Col1” del dataframe df.
- **df[”Col1"].value_counts():** devuelve el número de elementos que hay para cada valor único de la columna ”Col1” del dataframe df.

Operaciones estadísticas básicas

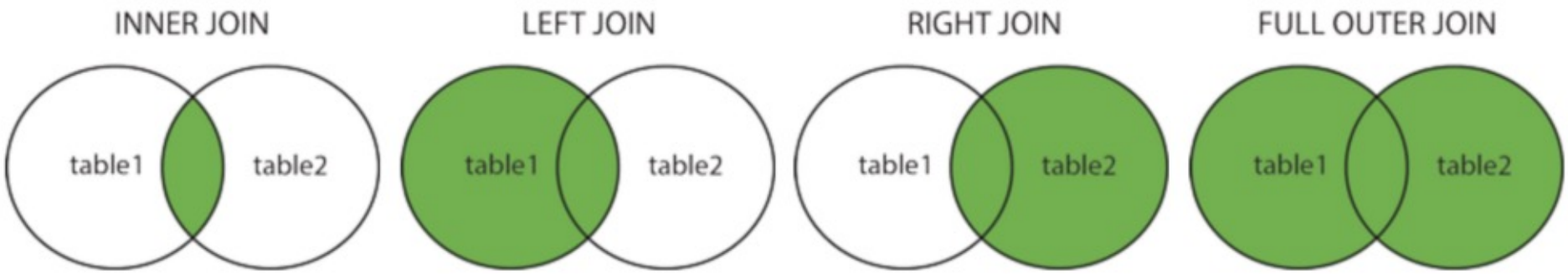
- **df[”Col1"].sum():** devuelve la suma de todas los valores de la columna ”Col1”.
- **df[”Col1"].min():** devuelve el valor mínimo de los valores de la columna ”Col1”.
- **df[”Col1"].max():** devuelve el valor máximo de los valores de la columna ”Col1”.
- **df[”Col1"].mean():** devuelve el valor promedio de los valores de la columna ”Col1”.
- **df.describe():** devuelve una tabla con valores estadísticos básicos (máximo, mínimo, etc.) de todas las columnas numéricas del dataset.

Gestionar valores NaN

- **df.dropna():** devuelve el dataframe df únicamente con aquellas **filas** que no contienen valores NaN.
- **df.dropna(axis = 1):** devuelve el dataframe df únicamente con aquellas **columnas** que no contienen valores NaN.
- **df.dropna(axis=1, thresh=3):** devuelve el dafatrame df únicamente con aquellas columnas que contienen más de 3 valores que no son NaN.
- **df.fillna(0):** reemplaza los valores NaN del dataframe por ceros.
- **df.describe():** devuelve una tabla con valores estadísticos básicos (máximo, mínimo, etc.) de todas las columnas numéricas del dataset.

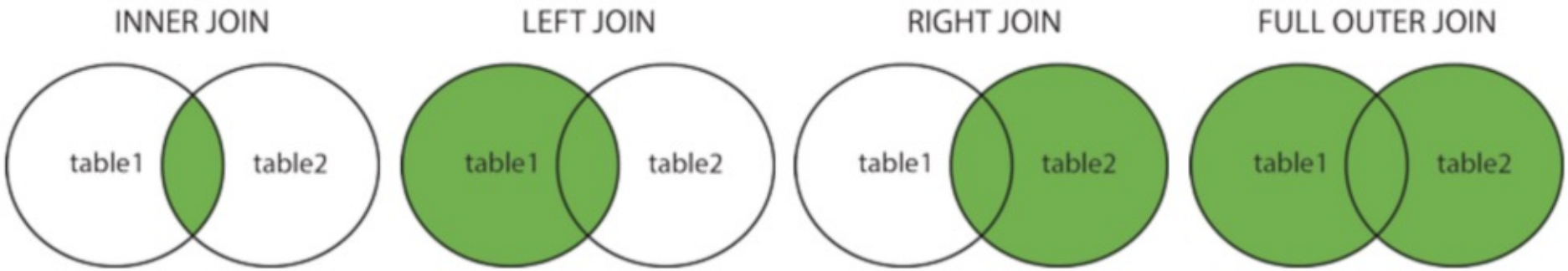
Concatenación de DataFrames

- **Concatenación vertical:**
  - pd.concat([dataframe_a, dataframe_b], ignore_index=True)
  - Ignore_index = True hace que las filas no tengan en cuenta su índice anterior, sino que se añaden unas debajo de otras siguiendo números consecutivos.
  - Ignore_index = False mantiene los índices de los DataFrames antiguos.
- **Concatenación horizontal, teniendo en cuenta un parámetro (on):**
  - pd.merge(dataframe_a, dataframe_b, on=”columna_comun", how="inner")



Fuente: https://en.wikipedia.org/wiki/Join_(SQL)

Tipos de Merge (atributo *how*)



- **inner**: valor por defecto de merge(). Muestra las filas que tienen el mismo valor en ambos DataFrames.
- **left**: muestra las filas del DataFrame de la izquierda más los que tienen el mismo valor en el DataFrame de la izquierda y el de la derecha.
- **right**: muestra las filas del DataFrame de la derecha más los que tienen el mismo valor en el DataFrame de la izquierda y el de la derecha.
- **outer**: muestra TODAS las filas indiferentemente de si éstas están en el DataFrame de la izquierda o el de la derecha.
- Más información: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.merge.html

Lectura y escritura de ficheros

- **Escritura en CSV:**

mi_dataframe.to_csv(”path_escritura/nombre_fichero.csv")

- Tened en cuenta que no se crea el directorio de destino, por tanto, debéis crearlo previamente o usar uno existente.
- Se puede modificar el separador de valores añadiendo el argumento “*sep*” dentro de los paréntesis (e.g. **mi_dataframe.to_csv(”path_escritura/nombre_fichero.csv”, sep = “\t”)**)
- Todos los argumentos modificables se pueden leer en:

https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.to_csv.html

- **Lectura de CSV:**

mi_dataframe = pd.read_csv(”path_lectura/nombre_fichero.csv")

Lectura y escritura de ficheros

- **Escritura en EXCEL:**

mi_dataframe.to_excel("path_escritura/nombre_fichero.xlsx", index=False, sheet_name=”Hoja1")

- Todos los argumentos modificables se pueden leer en:

https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.to_excel.html#pandas.DataFrame.to_excel

- **Lectura desde EXCEL:**

mi_dataframe = pd.read_Excel(“path_lectura/nombre_fichero.xlsx”, sheet_name='Hoja1')

Lectura y escritura de ficheros

- **Lectura desde un HTML:**

dataframe_wikipedia = pd.read_html("https://es.wikipedia.org/wiki/Wikipedia")

  - Devuelve una lista de DataFrames.
- Todos los argumentos modificables se pueden leer en:

https://pandas.pydata.org/docs/reference/api/pandas.read_html.html

- **Escritura en HTML:**

mi_dataframe.to_html(”path_escritura/nombre_fichero.html", index=False)

- La documentación oficial de las funciones generales de Pandas puede encontrarse en: https://pandas.pydata.org/docs/reference/general_functions.html