UT2: Práctica 2: Introducción y procesamiento de datos con Python

Objetivos de la actividad

A través de esta actividad vas a conseguir familiarizarte con Python y en concreto con las librerías de Numpy y Pandas.

Descripción de la actividad

La práctica se divide en tres partes o ejercicios. El primero tiene como objetivo que os familiarices con el lenguaje de programación Python, el segundo con el manejo de funciones de cadena para el procesado y limpieza de datos y por último el uso de Pandas y Numpy para el tratamiento de grandes cantidades de datos de forma más avanzada. A continuación, describimos los tres ejercicios.

Primera parte: Introducción a la programación con Python

El objetivo de esta actividad es que adquieras soltura en el uso de Python a través del uso de la herramienta con ejercicios sencillos:

1. Crea una función en Python que calcule el máximo común divisor de dos números.
2. Crea una función en Python que calcule el mínimo común múltiplo de dos números.
3. Basándote en datos aleatorios (los que quieras) crea un gráfico de tarta (pie chart) con Mapplotlib.
4. Basándose en datos aleatorios, crea un Scatter Plot con Mapplotlib. Explica para qué pueden ser útiles estos gráficos.
5. Explica cómo combinar 4 gráficos en una única figura (subplot).

Segunda parte: Procesado y limpieza de datos con funciones de cadena

1. Lee el fichero proporcionado.
2. Cuenta el número de líneas del fichero.
3. Cuenta el número de veces que aparece el artículo «el» en el fichero (nota: elemento contiene los caracteres «el» pero no debe contar puesto que no contiene a el artículo «el».
4. Elimina las tildes del fichero y convierte todas las letras a minúscula.
5. Crea un fichero de texto con el nombre «resultado.txt» que contendrá en la primera línea el número de líneas del fichero, en la segunda línea el número de veces que aparece el artículo «el», y a partir de la tercera línea el texto original todo en minúsculas y sin tilde.

Tercera parte: Procesado de datos con Pandas y Numpy

El objetivo del tercer ejercicio es procesar unos registros de datos sobre municipios españoles donde almacenamos su provincia, su código postal y su población. Para realizar la práctica sigue los siguientes pasos, empleando Pandas y Numpy tanto como sea posible:Carga en el entorno el «poblacionMunicipios.csv» proporcionado por el profesor respetando las tildes y caracteres españoles.

1. Carga en el entorno el «poblacionMunicipios.csv» proporcionado por el profesor respetando las tildes y caracteres españoles.
2. Elimina las filas correspondientes a municipios sin población.
3. Imprime el total de la población de todos los municipios.
4. Crea un DataFrame de Pandas donde la primera columna sean todas las provincias que aparecen en el archivo, la segunda columna el número total de habitantes por provincia, la tercera la desviación típica en el número total de habitantes por provincia y la cuarta el número total de municipios por provincia.
5. Carga el archivo «CP_Municipios.csv».
6. Empleando el código de provincia y el código de municipio, cruza los dos datasets comentados hasta el momento. El resultado final debe ser un DataFrame.
7. Genera un CSV llamado «faltan.csv» con todos los códigos postales que estén en «CP_Municipios.csv» pero que no tengan en «poblacionMunicipios.csv» ningún municipio con el que poder cruzar. Si todo cruza correctamente el archivo «faltan.csv» estará vacío. Se puede decir que «faltan.csv» son los códigos postales que se han perdido en el cruce a la hora de considerar la población.
8. Agrupa la información por código postal. Se quiere un DataFrame con las siguientes columnas:
    a. Código postal.
    b. Número de municipios que tienen dicho código postal asignado.
    c. Población: se calculará como la suma de la población de todos los municipios que incluyen a dicho código postal.
    d. Provincia: a la que está asignada el código postal.

Se realizará en un único notebook con el código y comentarios asociados.