**Visualización de la Información**

**Tarea #5** 4/marzo/2022

**Antecedentes**

Sea un conjunto de vectores $X = \{\boldsymbol{x}_i | \boldsymbol{x}_i \in \mathbb{R}^n \text{ e } 1 \leq i \leq N\}$, donde $\boldsymbol{x}_i = \left(x_{i,1}, x_{i,2}, \ldots, x_{i,n}\right)^{\mathrm{T}}$. Una distancia $d(\boldsymbol{a}, \boldsymbol{b})$ entre dos vectores $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$ se puede considerar como la longitud de la trayectoria que une a ambos vectores $\boldsymbol{a}$ y $\boldsymbol{b}$. Existen varias definiciones de distancia:

Distancia de Manhattan (norma $L_1$): $d(\boldsymbol{a}, \boldsymbol{b}) = \sum_{i=1}^{n} |a_i - b_i|$.

Distancia Euclideana (norma $L_2$): $d(\boldsymbol{a}, \boldsymbol{b}) = \left(\sum_{i=1}^{n} (a_i - b_i)^2\right)^{1/2}$.

Distancia Minkowski: $d(\boldsymbol{a}, \boldsymbol{b}) = \left(\sum_{i=1}^{n} |a_i - b_i|^p\right)^{1/p}$.

**Misión**

**a) Filtrado de Datos**: Haga la lectura del archivo “titanic3.csv" y responda las siguientes preguntas: *i*) ¿Cuántos cuerpos fueron encontrados?, *ii*) ¿Cuántos de ellos fueron hombres mayores a cuarenta años?, *iii*) ¿Cuantas mujeres desaparecieron entre las edades de 15 a 35 años?, *iv*) ¿Cuantos hombres mayores a 20 años sobrevivieron? y *v*) ¿Cuantas mujeres menores a 25 años sobrevivieron?.

Además, Genere una copia del conjunto de datos y rellene los datos faltantes (NA's) con un valor de 0 en el caso de datos numéricos usados como identificador la palabra “desconocido" en el caso de datos tipo cadena de caracteres y en el caso de variables numéricas use el promedio de los valores de esa columna (p.ej., la edad y la tarifa).

Finalmente, de los campos “*age*” y “*fare*” agregue columnas al conjunto de datos qué contengan los valores normalizados. Elija la normalización tipo $\frac{x_i - \overline{x}}{\sigma}$ para el caso de que la variable tenga una distribución normal y utilice la normalización tipo $\frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$ en cualquier otro caso.

**b) Distancias**: Utilizando el archivo “movies.csv” construya una o varias funciones qué permitan calcular una matriz de distancias para los datos numéricos en el *dataFrame*. La función debe permitir construir la matriz de distancia usando las distancias de Manhattan, Euclideana y de Minkowski (para *p* igual a 3).

Una matriz de distancia es una matriz cuadrada qué contiene las distancias entre los elementos de un conjunto (medidas un par a la vez).

Compare sus resultados con los que se obtienen por medio del método **distance_matrix** de **scipy.spatial**.

Además, Usando los métodos “*dendrogram*” y “*linkage*” construya un diagrama en forma de árbol (dendrograma) para el conjunto de datos en “movies.csv”. Repita el proceso ahora usando algún esquema de normalización del rango de los datos.

*i*) ¿Que diferencias puede encontrar en los resultados previos?.

*ii*) ¿En qué casos resulta importante llevar a cabo un proceso de normalización del rango de datos?.

*iii*) Consulte los diferentes tipos de distancias que se pueden usar como parámetro en el método “*linkage*”, ¿qué características de los datos se podría basar uno para elegir una determinada distancia?.

**Requisitos**

a) Crear su propio código.

**Datos**

Archivos titanic3.csv y movies.csv.

**Puntos Extras**

a) Se otorgará un punto extra si el programa genera la proyección de los dos puntos de vista de la cámara en la misma ventana, todas las vistas arregladas de manera horizontal. Esto significa que se divide el dispositivo en dos salidas de forma horizontal y que se *renderizan* dos cámaras al mismo tiempo. La selección entre perspectiva y ortogonal debe ocurrir en las dos vistas de manera simultanea.

**Entregables**

Es obligatorio entregar:

a) Código fuente.
b) Reporte que explique:
    a. Conceptos usados.
    b. Estructura de su código.
    c. Instrucciones de ejecución y utilización de su programa. En particular los mecanismos implementados para cambiar los puntos de vista y el tipo de proyección.
    d. Explicar razonadamente las elecciones que han tomado y sus resultados.
c) Se sugiere incluir algunas imágenes de su programa funcionando.

**Fecha de Entrega**

11/marzo/2022

**PD**. Cualquier duda o asunto no descrito en este documento se puede consultar por correo electrónico.