Universidad Francisco Marroquín
Data Wrangling 2020
Catedrático: Juan Carlos Girón
Auxiliar: José Josue

UFM UNIVERSIDAD FRANCISCO MARROQUÍN
VERITAS • LIBERTAS • JVSTITIA

Examen Final Data Wrangling 2020

Instrucciones

- Usted tiene el período de la clase para resolver el examen final.
- La entrega del final, al igual que las tareas, es por medio de su cuenta de GitHub, adjuntando el link en el portal de MiU.
- Pueden hacer uso del material del curso e internet (stack overflow, etc.). Sin embargo, si encontramos algún indicio de copia, se anulará el exámen para los estudiantes involucrados.

Serie Única: Conteste a las siguientes preguntas

1. ¿Qué es una expresión regular? (5 pts)
    a. Son una secuencia de caracteres que forman un patrón de búsqueda, considerador lenguajes formales. Son principalmente utilizados en operaciones de sustitución y busquedas de patrones en cadenas de caractres. En cs, proporcionan una forma flexible de buscar y reconocer cadenas de texto.

2. Enumere y explique brevemente cuatro aplicaciones prácticas en las cuales las expresiones regulares son utilizadas. (5 pts)
    a. Web scraping: Las expresiones regulares ayudan enormemente a procesar los datos textuales desordenados. Hay muchas aplicaciones diferentes, como buscar títulos de varios artículos de blogs.
    b. Analisis lexico en un compilador: El analizador léxico necesita escanear e identificar solo un conjunto finito de cadenas / token / lexemas válidos que pertenecen al idioma en cuestión. Busca el patrón definido por las reglas del idioma.Las expresiones regulares tienen la capacidad de expresar lenguajes finitos definiendo un patrón para cadenas finitas de símbolos.

c. Validacion de formatos de constraseñas: Se establece la cantidad y tipos de caracteres que debe contener la contrasena para ser considerada aceptada.
d. Validacion de correos: Se establece que que dominio debe contener, la cantidad de caracteres, etc.

3. Explique brevemente las 3 condiciones que establecen que una tabla se encuentra en formato ***tidy.*** (5 pts)
    a. Cada variable debe tener su propia columna
    b. Cada observación debe tener su propia fila
    c. Cada valor debe tener su propia celda

4. Diagnostique y explique por qué la siguiente tabla no está en formato ***tidy.*** Luego, explique cómo convertirla a formato ***tidy.*** (7 pts)

Country	2008	2009	2010
Guatemala	5	9	13
United States	9	13	23
Belgium	7	13	18
Argentina	9	18	28
France	7	13	24
United Kingdom	3	3	5
Germany	10	15	27
Poland	1	2	2

- La tabla no se encuentra en formato Tidy porque los nombres de las columnas no son nombres de variables sino que son valores de una variable. En este caso los nombre de las columnas 2008, 2009 y 2010 representan valores de la variable año, los valores en dichas columnas representan valores de la variable 'casos' y cada fila representa tres observaciones no solo una. Para transformarla a formato tidy tendria que crear nuevas columnas para asignarle variables de "año" y "casos" y respectivamente hacer una sumatoria de los casos por año para que sea reflejado en la columna de casos.

5. Diagnostique y explique por qué la siguiente tabla no está en formato ***tidy.*** Luego, explique cómo convertirla a formato ***tidy.*** (7 pts)
    a. La tabla no es considerada tidy porque se puede observar que la posición se encuentra a la par del nombre, por lo que debería tener una

variable aparte para establecer la posición del jugador ya que no se pueden tener dos variables en la misma columna. Para arreglarla se deberia crear una columna con la variable posición.

Equipo	**Jugador**
Real Madrid	Federico Valverde - Mediocentro
Juventus	Cristiano Ronaldo - Delantero
Barcelona	Frenkie De Jong - Mediocentro
Manchester United	Marcus Rashford - Delantero
Manchester City	Eric García - Defensa
Liverpool	Alisson - Portero
Atlético de Madrid	Joao Félix - Delantero
AC Milan	Sandro Tonali - Mediocentro
Roma	Pedro - Delantero
Inter de Milan	Achraf Hakimi - Defensa
Sevilla	Lucas Ocampos - Delantero
Valencia	Jose Luis Gayá - Defensa
PSG	Neymar - Delantero
Monaco	Cesc Fábregas - Mediocentro
Bayern Munich	Alphonso Davies - Defensa

6. Diagnostique y explique por qué la siguiente tabla no está en formato ***tidy.*** Luego, explique cómo convertirla a formato ***tidy.*** (7 pts)
    a. Las observaciones deberian al menos estar representadas en numeros binarios, y los precios no deberian ser variables sino que agregar una variable de precio e incluir ahí los precios.

Producto	Urabno	Rural	Q0 - Q50	Q50 - Q100	Q100 - Q500	Q500 +
Banano 12 und.	x		x			
Café molido 1 lb	x		x			
Televisión Samsung 32"		x				x
Carne Molida 5 lb		x		x		
Licuadora 1 lt	x				x	

7. Sobre lubridate: Explique la diferencia entre las funciones period y las funciones duration. (5 pts)
    a. Durations mide la cantidad exacta de tiempo entre dos momentos, funciona como un cronometro y es independiente a una fecha de inicio,

mientras que periods mide de forma precisa los tiempos del 'reloj' o 'calendario' sin tomar en cuenta años bisiestros o 'day light savings' independiente a una fecha de inicio (interpretración más humana del tiempo).

   i. Durations: "2021-02-27 06:00:00 UTC"
   ii. Periods: "2021-02-28"

8. ¿En qué contexto utilizaría una función period y en cúal utilizaría una función duration? (5 pts)
   Durations lo utilziaria cuando necesito saber tiempo exacto por ejemplo la duración exacta con segundos de un tiempo de aterrizaje mientras que period la utilizaria en un contexto donde la fecha no sea de gran importancia la exactitud, como por ejemplo calcular la fecha en la que se realizó cierta accion más humana (check in de hotel).

9. Explique el concepto de data Missing Completely at Random (MCAR). (6 pts)
   a. MCAR consiste en la propensión de que falte algún punto de datos es completamente aleatoria. Es decir, la probabilidad de que falten valores en una variable es la misma para todas las muestras, No existe una relación entre la ausencia de un dato y cualquier valor del dataset, perdido u observado.

10. Si logramos verificar que la data faltante es MCAR, ¿cuál imputación recomendaría utilizar? (5 pts)
    a. Imputación múltiple

11. Si estamos realizando el análisis de una encuesta en la cual tenemos información sobre 150 individuos y tenemos valores faltantes en diferentes variables de nuestra tabla, ¿cúal de los siguientes métodos utilizaría y por qué? (6 pts)
    a. listwise deletion.
    b. pairwise deletion.
    c. outliers cap via standard deviation.
    d. outliers cap via percentile approach.

12. Usted se encuentra realizando un modelo sobre la capacidad necesaria que necesita para atender la demanda de transporte de un producto

determinado. Se requiere que cumpla con el 90% de la demanda mensual. ¿Cúal de los siguientes métodos utilizaría para determinar con qué población de sus datos trabajar? (6 pts)

a. listwise deletion.
b. pairwise deletion.
c. outliers cap via standard deviation.
d. outliers cap via percentile approach.
e. min-max scaling.

13. ¿En qué contexto de Machine Learning se recomienda utilizar Min Max Scaling? (6 pts)
    a. Unicamente cuando los limites ya se encuentran definidos.

14. Si encuentra que la distribución de sus datos tiene un comportamiento exponencial, ¿cúal técnica de normalización utilizaría para transformar los datos a una distribución normal? (5 pts)

    a. Log Transformations

15. Si se tiene una variable categórica con tres niveles, cúantas variables dummy necesita para poder pasar la data a un modelo econométrico o de machine learning? (5 pts)

    a. 3

16. ¿En cuál contexto utilizamos one hot encoding? (5 pts)

    a. One hot encoding es utilizado para transformar variables categóricas a vectores binarios en categorías. Un contexto donde se tendría que utilizar one hot enconding sucede cuando la característica categórica no es ordinal, también cuando la cantidad de características categóricas es menor por lo que se puede aplicar one hot enconding de manera efectiva.

17. ¿Qué es un n-gram? (5 pts)
    a. Los n grams de texto se utilizan amplicamente en 'text mining' y en las tareas de procesamiento de lenguaje natural. Son un conjunto de

palabras que coexisten dentro de una ventana determinada (secuencia de palabras N), al calcular los n-grams normalmente se avanza una palabra.

18. Si quiero obtener como resultado las filas de la tabla A que no se encuentran en la tabla B, ¿cómo debería de completar la siguiente sentencia de SQL? (5 pts)

*SELECT * FROM A LEFT JOIN B ON A.KEY = B.KEY WHERE B.key IS NULL*

_____