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

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)

   Una expresión regular es una serie de parámetros o características que se establecen para manejar texto. Estas pueden armarse de distintas formas, como regex o librerías como stingr, y contienen características deseadas por el usuario, las cuales se utilizarán para generar, borrar, encontrar o modificar texto de forma estandarizada.

2. Enumere y explique brevemente cuatro aplicaciones prácticas en las cuales las expresiones regulares son utilizadas. (5 pts)

   Para extraer opiniones de una base de datos de comentarios, es decir saber si las opiniones son positivas o negativas.
   Para detectar errores de escritura dentro de un texto grande.
   Para generar automáticamente correos electrónicos incorporando fragmentos del nombre de una persona.
   Para filtrar búsquedas dentro de una base de datos.
   Para la generación de textos predictivos.

3. Explique brevemente las 3 condiciones que establecen que una tabla se encuentra en formato ***tidy.*** (5 pts)

   Cada fila debe ser una observación. Es decir que tiene que ser única, la misma observación no puede tener dos filas distintas.

   Cada columna debe ser una variable, es decir una característica de las observaciones.

   Cada celda debe contener únicamente un dato. Es decir que las variables o columnas no pueden tener más de una dimensión.

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 la serie de tiempo está dispersa como variables. Esto provoca que se acumulen múltiples ocurrencias en las celdas. Para convertirla a formato tidy, necesitamos generar una observación para cada ocurrencia. Pasaremos a tener una cantidad de filas igual al total de ocurrencias registradas en toda la tabla, y Country pasaría a ser una variable, al igual que el año. EJ:

   Ocurrencia | Pais | Año

   Oc1 |Guatemala | 2009

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

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

No está en formato tidy porque la variable Jugador tiene un doble registro, es decir que tiene 2 dimensiones. Adicionalmente, ya que el registro único es Jugador, cada jugador debe ser una observación, y no el equipo, ya que el equipo puede llegar a repetirse si hay varios jugadores del mismo equipo. Cada jugador pasará a ser una observación, y su posición se separará en una nueva variable. El equipo también será una variable. Ej:

Jugador | Equipo | Posición
Federico Valverde | Real Madrid | Mediocentro

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

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	

La tabla no está en formato tidy porque utiliza un sistema de selección de variables para llevar su registro. Este caso es bastante fácil de solucionar, ya que solamente se necesita generar una nueva variable que reemplace a las anteriores, la cual indique directamente el valor deseado. Ej.

Producto	Localidad	Precio
Banano 12 und.	Urbano	Q0 - 50

7. Sobre lubridate: Explique la diferencia entre las funciones period y las funciones duration. (5 pts)

   La funcion period recibe como input una cantidad de tiempo, y la unidad que deseada para el output, y genera un objeto periodo de tiempo. duration por su parte recibe un input similar, con la diferencia de que la cantidad de tiempo es acumulativa, y su output es en segundos.

8. ¿En qué contexto utilizaría una función period y en cúal utilizaría una función duration? (5 pts)

   Utilizaria la funcion period cuando vaya a trabajar con un periodo ya determinado de tiempo, el cual deba estar manejando constantemente (como un tiempo de espera el cual haya que sumar constantemente a horas de llegada por ejemplo) ya que tendría un objeto periodo de tiempo el cual puedo fácilmente sumar a cualquier unidad de tiempo o hora.

   La función period no se puede sumar a unidades de tiempo, de modo que no serviría para lo mencionado anteriormente, así que la utiizaría exclusivamente para medir cuánto duró un suceso, con una hora de inicio y una hora de finalización.

9. Explique el concepto de data Missing Completely at Random (MCAR). (6 pts)

Se refiere al caso en que los datos faltantes no son faltantes por ningún motivo en concreto, sino que los datos faltantes son faltantes de forma completamente aleatoria. Es decir que no hubo un motivo de paridad entre valors

faltantes. Esto quiere decir que no tendría porque haber ningun tipo de sesgo en los datos faltantes, y sus valores deberían comportarse igual que para el resto de registros.

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

Si se puede verificar que la data faltante es MCAR, quiere decir que los datos faltantes no tendrán ningún sesgo, y se comportan igual que el resto de los datos. En este caso recomendaría una imputación de la media, ya que esta es la que mejor refleja el comportamiento en datos numéricos, y la imputación tendra un impacto menor en la fidelidad de los datos en términos estadísticos (media y desviación estandar). Si es una variable categórica sería imputación de la moda sectorizada.

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.

    Utilizaría outliers cap via standard dev. ya que al tener una cantidad grande de observaciones, la distribucion tiende a normalizarse y la desviacion estandar se vuelve un buen criterio

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.

Utilizaría outliers cap via percentil approach, fijado al 90%. Utilizaría este método para reemplazar los valores que estén fuera del percentil 90 con el máximo, de esta forma estaría cumpliendo el objetivo de cumplir con el 90% de la demanda, y estas observaciones atípicas seguirán reflejando un registro de valor alto en la data.

13. ¿En qué contexto de Machine Learning se recomienda utilizar Min Max Scaling? (6 pts)

    La estandarización Min-Max scaling se utiliza en ciertos modelos de machine learning para garantizar el buen funcionamiento del modelo.

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)

    Transforma los valores a la version normalizada con valores Z segun la distribución normal. Esto debido a que una transformación min-max scaling le asignaria una distribución más lineal a los datos, ya que los valores solo van entre 0 y 1

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)

    Si la variable tiene 3 niveles, se necesitaría dos dummies para incluirla en un modelo, una variable que sea 1 para el nivel A, y otra que sea 1 para el nivel B. El nivel C estará cubierto ya que será 0 en ambas variables, e incluir esta 3ra dummy sería redundante.

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

    Cuando tenemos variables categóricas y queremos analizar el efecto que tiene cada una de las categorias. Utilizamos one hot encoding para poder

analizarlas de forma numérica en modelos econométricos como regresiones, o modelos de machine learning.

17. ¿Qué es un n-gram? (5 pts)

    Un n-gram es una representación gráfica de la distribución de variables de texto. Este muestra, generalmente por tamaño, la distribucion de la cantidad de ocurrencias registradas para cada 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 ___ JOIN B ON A.KEY = B.KEY _____*

Select * form A LEFT JOIN B ON A.KEY = B.KEY WHERE B.key = null