Universidad Francisco Marroquín

Data Wrangling

Catedrático: Juan Carlos Girón

Auxiliar: José Josue

Examen Final Data Wrangling

**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 examen para los estudiantes involucrados.

**Serie Única: Conteste a las siguientes preguntas**

1. ¿Qué es una expresión regular? (5 pts)

Es una expresión formal para representar una cadena de caracteres por medio de sus características. En una expresión regular podemos definir las características o el formato que debe tener un texto que vayamos a utilizar. Por ejemplo, como que inicie con tres números y que tenga una o muchas letras y termine con un numeral. Estas características definidas definidas en las expresiones regulares s

1. Enumere y explique brevemente cuatro aplicaciones prácticas en las cuales las expresiones regulares son utilizadas. (5 pts)
   1. Validación entradas en un programa, como contraseñas o textos en formularios. Por ejemplo, que la contraseña contenga una mayúscula o que el correo esté bien escrito (con @ antes del donminio).
   2. Tokenización de conversaciones para su análisis. Con esto se busca descifrar el desarrollo la conversación a lo largo del tiempo y cómo se han desarrollado la relación de quienes participan.
   3. Compilación de programas de un lenguaje a otro. La tokenización del código por medio de expresiones regulares es la primera etapa de la compilación. Que significa pasar un código de un lenguaje a otro manteniendo su funcionamiento.
   4. Análisis de reportes financieros para conocer si el resultado es bueno o malo. Esta aplicación está hecha para ahorrar el tiempo que se dedica a entender el estado financiero de una empresa a partir de unos análisis de texto. El problema es que en los reportes financieros se trata de esconder el estado de la empresa entre mucho texto.
2. Explique brevemente las 3 condiciones que establecen que una tabla se encuentra en formato ***tidy.*** (5 pts)
   1. Que cada variable debe estar representada en una columna. Las variables son todos los valores que pertenecen a la misma unidad. Por ejemplo, el color, en una tabla de automóviles.
   2. Cada observación debe ser una fila. Las observaciones son los valores de un mismo elemento u objetivo medido a través de varias unidades. Por ejemplo un automóvil, con modelo, año, color y placa.
   3. Y cada tabla debe estar compuesta por una sola unidad observacional. Un ejemplo de esto es que si tenemos una tabla con las canciones y el ranking semanal (dos unidades observacionales), se deben separar en dos tablas. Una solo para las canciones (su nombre y otra información) y otra con el ranking de las canciones de cada semana.
3. Diagnostique y explique por qué la siguiente tabla no está en formato ***tidy.*** Luego, explique cómo convertirla a formato ***tidy*.** (7 pts)



Porblema: Los nombres de las columnas 2008, 2008 y 2010 son valores y no etiquetas de variables.

Pasos:

* En un nuevo dataframe crearía las columnas: country, year y value.
* Pasaría los años que anteriormente se encuentran como variables o columnas a la columna año. Agregaría el país a cada año.
* Se generaría una combinación de país por cada año.
* Después añadiría el resto de los valores a la columna value, donde se intercepte el año y el país de la tabla anterior.

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



Problema: existen múltiples variables en una columna. El nombre de jugador y su posición están en la misma variable.

Pasos:

* Crear una nueva columna llamada **posición**.
* Hacer un split o una separación por “ - “ de la columna jugador.
* Asignar el primer elemento del split a la columna jugador (donde iría el nombre)
* Asignar el segundo elemento del split en la columna **posición**.

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



Problema: Los nombres de las columnas Urbano, Rural y el resto de columnas de rangos de dinero son valores y no etiquetas de variables.

Pasos:

* Crear las columnas área y rango.
* Colocar “urbano” en la columna **área** done la x esté marcada en la columna urbano y “rural” en la columna **área** done la x esté marcada en la columna rural.
* Haría lo mismo en la columna rango, pondría el nombre del nombre de la columna que esté marcada con “x” en la columna **rango**.

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

* Period: el periodo toma las fechas en términos relativos, en una compresión más parecida a la humana. Por ejemplo hacer una cita para dentro de un mes, aunque la duración de los meses no siempre sea la misma.
* Duration: Se utiliza para calcular la duración exacta de un evento en términos reales. Por ejemplo, que la llegada de un paquete sea en 30 días, aquí se toma independientemente de la duración.

1. ¿En qué contexto utilizaría una función period y en cúal utilizaría una función duration? (5 pts)
   1. Period:
   2. Duration: utilizaría esta función en el caso en que necesite predecir en evento exacto, por ejemplo el paso de un cometa, por tengo que calcular la fecha sin importar la percepción.
2. Explique el concepto de data Missing Completely at Random (MCAR). (6 pts)
3. Si logramos verificar que la data faltante es MCAR, ¿cuál imputación recomendaría utilizar? (5 pts)
4. 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)
   1. listwise deletion.
   2. pairwise deletion.
   3. outliers cap via standard deviation.
   4. outliers cap via percentile approach.
5. 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)
6. listwise deletion.
7. pairwise deletion.
8. outliers cap via standard deviation.
9. outliers cap via percentile approach.
10. min-max scaling.
11. ¿En qué contexto de Machine Learning se recomienda utilizar Min Max Scaling? (6 pts)
12. 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)
13. Si se tiene una variable categórica con tres niveles, cúantas variables dummy necesita para poder pasar la data a un modelo econométrico y de machine learning? (5 pts)
14. ¿En cuál contexto utilizamos one hot encoding? (5 pts)
15. ¿Qué es un n-gram? (5 pts)
16. 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 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

1. Actualmente la UFM implementó la herramienta Turnitin, utilizada para detectar plagio en los entregables de los alumnos. Explique, basado en los conceptos visto en clase, el funcionamiento de este tipo de herramientas que analizan texto. (10 pts)