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)

Una expresión regular es un conjunto de caracteres que nos establecen un resultado esperado, es decir, nos permiten establecer que valores esperamos y cómo identificamos a ese conjunto de carateres.

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

* Verificación de constraseñas: en este caso, se usa la expresión para validar el “formato” permitido de la contraseña, por ejemplo: poner sólo 8 caracteres, con al menos un número y un símbolo (#$%&\_-) . Clave: 12Carlo$.
* Buscadores de texto: En este caso, buscamos palabras, frases, oraciones que contengan ciertas palabras, etc. La idea es establecer la expresion regular que buscamos encontrar o que contenga la cantidad de valores “especiales” que buscamos encontrar. A veces se usa para elminar palabras que son muy repetitivas que no deseamos analizar (el, por, que, ella, la, los, ect).
* Análisis léxico en un compiladores: en este caso, son las expresiones regulares, aquellos tokens que permite el lenguaje para su compilación, es decir, establecen palabras y hasta símbolos que se pueden manejar dentro del lenguaje.
* Códigos de estudiantes: en este caso, el carné para estudiantes se puede ver ya más detallado, porque puede ser una expresión regular que establezaque que primero debe de ir el año, luego iniciales del estudiante y por último el número de cuando fue insctrito. Esto permite un mejor manejo de los carnés.

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

* Las columas son variables: en este caso, buscamos que las columnas esten nombrados con las variables que se requieren en la tabla.
* Las filas deben de ser observaciones: al momento de ver la fila, estemos seguros que estamos teniendo una observación ordenada y clara de la tabla.
* Y la tabla se debe de componer de sólo una unidad observacional: en este caso, que todas las observaciones esten destinadas a un mismo “tipo de observación” es decir, que todas sean congruentes en qué tipo de información se ve por cada observación.

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



En este caso, esta poniendo en el nombre de la columna, valores, que en este caso son los años, pero no esta poniendo una etiqueta año, sino cada año. Por lo que hace falta una columna que contenga una cabal los valores que vemos en las intersecciones entre el país y el año.

El proceso para llegar a la tabla tidy, es: crear una columna donde este se llame: Año, donde sus valores en ella serán los que estaban anteriormente como encabezados de las columnas. Luego, se pondrían una columna que diga Valor: donde contiene los números que intersectan en la tabla anterior.

Country. |Año. |Valor

Guatemala|2008|5

…. Los demás 2008

Guatemala|2009|9

… así sucesivamente.

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



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



1. Sobre lubridate: Explique la diferencia entre las funciones period y las funciones duration. (5 pts)
2. ¿En qué contexto utilizaría una función period y en cúal utilizaría una función duration? (5 pts)
3. Explique el concepto de data Missing Completely at Random (MCAR). (6 pts)
4. Si logramos verificar que la data faltante es MCAR, ¿cuál imputación recomendaría utilizar? (5 pts)
5. 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.
6. 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)
7. listwise deletion.
8. pairwise deletion.
9. outliers cap via standard deviation.
10. outliers cap via percentile approach.
11. min-max scaling.
12. ¿En qué contexto de Machine Learning se recomienda utilizar Min Max Scaling? (6 pts)
13. 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)
14. 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)
15. ¿En cuál contexto utilizamos one hot encoding? (5 pts)
16. ¿Qué es un n-gram? (5 pts)
17. 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)