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, valga de redundancia es una expresión compuesta de diferentes caracteres que nos permiten llevar a cabo análisis de texto, estas expresiones nos permiten poder identificar cadenas de texto o bien también pueden ser utilizadas para realizar ciertas validaciones, por ejemplo, la validación de un número telefónico como vimos en clase.

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

Análisis en cadenas de texto: usamos expresiones regulares para buscar ciertas cadenas dentro de un texto e identificar patrones

Validaciones en formularios: Validar ciertas variables, que el formato que el usuario pone sea correcto a lo que pide el formulario, ejemplo: una contraseña para la banca virtual.

Buscadores de texto: nos permiten realizar búsquedas mas flexibles y facilitan la manipulación de texto

Extracción de datos: las expresiones regulares se utilizan para extraer información específica de documentos

1. Explique brevemente las 3 condiciones que establecen que una tabla se encuentra en formato ***tidy.*** (5 pts)
2. Cada variable forma una columna:

Cada variable en el conjunto de datos debe representarse en una columna separada. Esto significa que todas las mediciones o características deben ser organizadas de manera estructurada en columnas

1. Cada observación forma una fila:

Cada fila en la tabla debe representar una observación única o un caso individual.

1. Cada tipo de unidad observacional forma una tabla:

Los datos deben organizarse de manera que cada tipo de unidad observacional constituya su propia tabla.

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



Primero los años están siendo tratados como columnas y no como un valor de una variable año. Por lo que la tabla falla en uno de los conceptos básicos de tidy data. Para pasarla a un formato tidy podríamos agregar una nueva columna año con los valores (2008,2009 y 2010) y otra columna valor que contenga el numero respectivo, así formando un único registro.

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 otro caso, en la columna de Jugador están incluyendo la posición que tiene este, por lo que también esta tabla falla en cuanto a que cada variable tiene que ser una columna. Al crear una nueva columna que se llame Posición, la tabla quedaría en formato tidy y sería mucho más fácil de entender. Esto nos permitiría filtrar a jugadores por posición

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



No está en formato tidy ya que la tabla presenta el precio como rango, en vez de tener una columna que sea rango y que en esa columna el registro pueda tomar el valor del rango del precio en el que entra cada producto. Al igual que Urbano y Rural , se podría categorizar como una variable de “Area”.

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

La diferencia entre estas funciones es que la función duration mide el tiempo real que ha pasado tomando en cuenta eventos irregulares mientras que la función period registra los cambios en reloj sin tomar en cuenta las irregularidades.

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

Usaría la función de period cuando quiera intervalos de tiempo que siguen el calendario, mientras que la función de duration la usaría cuando necesite medir un intervalo de tiempo fijo.

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

Este concepto refiere a una situación en la que la probabilidad de que falten datos es independiente tanto de las variables observadas como de las no observadas. También lo podemos intepretar cómo la probabilidad de que un valor faltante no esté relacionado con el valor real que debería haber sido observado.

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

Podría ser una imputación por regresión

1. 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.

Usaría este método ya que nos permitiría utilizar la información que tengamos disponibles sin la necesidad de tronarme todas las observaciones que al menos tengan 1 faltante.

1. 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)
2. listwise deletion.
3. pairwise deletion.
4. outliers cap via standard deviation.
5. **outliers cap via percentile approach.**
6. min-max scaling.
7. ¿En qué contexto de Machine Learning se recomienda utilizar Min Max Scaling? (6 pts)

Este método de normalización es mejor usarlo cuando trabajamos con algoritmos sensibles a la escala de las variables, como aquellos que utilizan medidas de distancia o aquellos que son sensibles a la magnitud de los valores.

1. 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)

La técnica que deberías de utilizar es la Log-transformation, esta nos permitiría tener un approach muy cercano a una distribución normal

1. 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)

Necesitaríamos dos variables dummy así si las dos variables dummy nos dan un valor de 0 por ejemplo, sabemos que la observación sería de la tercer nivel de la variable categórica y no de los primeros dos.

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

En el contexto de machine learning , usamos el one hot encoding cuando tenemos variables categóricas, ya que este nos permite representar los niveles como variables numéricas

1. ¿Qué es un n-gram? (5 pts)

Es una secuencia de n elementos que pueden ser ya sea palabras, sílabas etc. Que nos permiten identificar patrones en los textos. Puede haber distintos tipos de n-gramas ya sea un unigrama , bigrama y así sucesivamente.

1. 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*

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)

Una herramienta como esta podría funcionar analizando varias cosas en los entregables de los alumnos, primero podría empezar con una comparación de las palabras más frecuentes en los entregables, al haber una gran cantidad de palabras iguales con gran frecuencia con el texto comparado, ya es un indicio de posible plagio. Una herramienta como esta me imagino que también analizaría la diversidad léxica del entregable y el documento a comparar, al no haber diversidad es otro indicio. Por otro lado, los n-grams también serían otra forma que esta herramienta analice los textos para ver plagio, los estudiantes al tener n-grams exactamente iguales podría ser otro indicio de plagio. Otra forma podría ser usando el método de TF-IDF. Este método asigna un peso a cada palabra en un documento en función de su frecuencia en ese documento (TF) y su inversa frecuencia en el conjunto de documentos (IDF). Palabras que son frecuentes en un documento, pero raras en el conjunto de documentos recibirán un peso más alto. Un valor muy bajo de TD-IDF nos indicaría que las palabras tienen una frecuencia muy alta pero también que son comunes en el resto del cuerpo que se esta analizando.

1. Utilizando el dataset de “Student Performance”, realice una presentación respondiendo alguna de las siguientes preguntas (10 pts)
   1. ¿Cuál es el efecto de la dieta del estudiante antes de la prueba?
   2. ¿Existe alguna diferencia entre grupo de estudiantes (gender/race) al estar previamente preparados?
   3. ¿Existe alguna relación entre los resultados de matemáticas, lectura y escritura para los diferentes grupos de estudiantes (gender/race)?

A screenshot of a graph

Description automatically generated