Universidad Francisco Marroquín

Data Wrangling

Catedrático: Juan Carlos Girón

Auxiliar: José Josue

Examen Final Data Wrangling

Sergio Cuevas 20190216

**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)

La “unidad” más básica de una expresión regular es el “carácter”, partiendo de esto podemos decir que las expresiones regulares nos ayudan para poder encontrar los patrones que puede haber en un “string”. Las necesitamos para poder extraer o analizar partes específicas de un texto en base a estos “patrones” que pueden llegar a tener las palabras.

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

* Pueden ser muy utilices para la limpieza de datos. Es común que a veces la gente se confunda o no quiera llenar algún dato, estos “tipos” solo dificultan el análisis. Ejemplo si ponen un punto (.) en lugar de su edad, podremos reemplazar ese punto por algún otro carácter o dejarlo vacío, que no ayude para reconocerlo como NA o de la forma que nos convenga.
* Para “minar chats”, buscar palabras específicas, ver su recurrencia y analizar en base a estos resultados.
* Crear algoritmos de clasificación. Ej. Un análisis de sentimiento de como twittea la gente sobre una marca. En base a las palabras que usan o la secuencias, se podría clasificar si el tweet es positivo o negativo o incluso neutral y tomar decisiones de negocio en base a eso.
* Cuando solo necesitamos extraer cierta palabra o patrón de caracteres dentro de un texto, porque es lo único que nos interesa para analizar.

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

* Cada variable es una columna: Habla de una misma característica de las distintas observaciones. Ej. Estamos hablando de la altura (variable) de todas las personas encuestadas (observaciones).
* Cada observación es una fila: Se podría resumir como “una tabla tidy crece hacia abajo”. Cada observación es una fila se refiere a que cuando se “observa” o se tiene una nueva fila, significa que esa unidad observacional tiene sus todas sus características expresadas en columnas. Ej. La observación 14 son todos los datos/características (columnas) obtenidas de la encuesta de la persona #14.
* La tabla esta compuesta por una unidad observacional: Es la unidad que nos llama la atención estudiar. Ej. Si estamos tomando datos de investigación a perros, humanos o de resultados de examen, etc.

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



Primero, cada variable no es una columna. Se tiene como columna “Country”, “2008”, “2009”, “2010”. Significa que la verdadera variable debe de ser “Year” y 2008, 2009,2010 son los datos que puede tomar la observación en dicha variable.

Asumiendo que la unidad observacional si es país (Ej. Estamos viendo la cantidad de premios novel por país al año), la tabla en forma tidy tuviera que queda con solo 3 columnas: Country, Year y #premios\_Nobel (para el ejemplo) y 24 filas, que serían la observación de cuantos premios nobel gano cada país en cada año en específico.

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 la data no esta en formato tidy porque tiene múltiples variables en una columna, tiene nombre y posición en la misma.

Para convertir la data a tidy, primero crearía una columna que se llamara “Posición” y extraería (expresiones regulares) la posición para meter los datos en esta nueva columna. Luego, dependería de que es lo que necesite o de que trate la tabla, pero definiría que mi unidad observacional es el Jugador. La tabla quedaría con 3 columnas: Jugador, Equipo y Posición; y 15 filas.

\*No se mencionó, pero cabe aclarar que el guion y esos espacios en el medio no los necesitamos y que deberán ser eliminado también.

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



Porque se tienen múltiples variables que están en múltiples columnas. Cada variable debería de ser solo una columna.

Eso significa que, si queremos que la data esta tidy, se debería de hacer una columna llamada “área” que será de clasificación si se está en el área Urbana y Rural; luego, otra columna llamada Rango\_Precio, que podrá tomar los valores que se pueden observar en la tabla de arriba de la columna 4 a la 7. De esa forma nuestra tabla tendrá 3 columnas: Producto, Área y Rango\_Precio; y filas tendrá 5.

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

Duración es aquella que representa específicamente lo que dice, sin ningún “ajuste humano”, 1 mes exacto. En cambio, Period sería 1 mes con el “ajuste humano” y no de la forma en que la computadora lo ve. Cabe mencionar que duration por detrás es el número exacto, pero expresado en segundos.

Es más fácil entenderlo con un ejemplo: si estamos en 31 de enero y agregamos 1 mes con en forma de **duration** agregará exactamente 30 días y terminará en algún día de marzo. En cambio, si se hace con **period** y agregamos un mes, el resultado sería 28 de febrero; como si una persona dijera “nos vemos el otro mes”.

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

Si lo que necesito es ir agregando en busca del final de mes de cada mes específico usaría **period**, no me tendría que preocupar en pensar cuantos días tiene cada mes y pensar cuanto hay que agregarle. En cambio, **duration** lo utilizaría cuando es de vital importancia hablar en términos exactos, la duración de un proceso, agregarle el tiempo exacto a algún dato que estoy tomando, etc.

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

Quiere decir que no hay ninguna correlación o relación comprobable entre un dato faltante y algún otro dato faltante u observado. En otras palabras, la razón por la que la información que no está es porque es totalmente aleatoria en relación con los otros datos. Ej. Esa persona específicamente no quería responder la pregunta o la persona no habla el idioma entonces no entendió la pregunta y contesto más la encuesta/no contestó. Son totalmente aleatorios porque son específicos de cada situación.

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

Todo dependerá del tipo de dato faltante y el comportamiento de la variable. Si es categórica, será por medio de la moda; Si es numérica dependerá si la distribución está centrada en la media, por lo cual sería imputación por media; o si la distribución está sesgada podría ser mejor idea usar la mediana.

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.

Porque temenos muy pocos datos, significaría que el listwise nos dejaría con muchos menos. Y los métodos de outliers cap, podría como que sesgar mucho nuestros datos por la misma razón de que son muy pocos y nos están diciendo que se tienen varios valores faltantes.

Pairwise sirve porque no se eliminan, dependiendo de las columnas que necesitemos, aunque tiene desventajas también.

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.

Porque me piden un percentil específico entonces de esa forma puedo “apuntar” al 90% exacto.

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

Cuando la distribución de los datos que tenemos no se ajusta bien a la distribución de una curva normal y se conocen bien los limites superior e inferior.

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 mejor opción sería una transformación “Log”. Por la forma en la que se comportan los datos.

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)

Si se usa one hot encoding serían 3 variables dummy, una por cada nivel. Aunque vale la pena mencionar que en ciertos algoritmos, como el de regresión lineal se podrían tener solo 2 porque se tiene el “caso base” que sería tomado en cuenta en el intercepto.

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

Porque es la forma en que el algoritmo puede “entender” que es lo que provoca ese dato en ausencia o en presencia. De esa forma se usan para “fittear” el modelo. De no ser así, y tener 1, 2 o 3 como categoría, por ejemplo, el modelo interpretaría que el comportamiento cambia dependiendo si el numero aumento o disminuye, no entendería que es una categoría presente o ausente.

1. ¿Qué es un n-gram? (5 pts)

Algoritmo que busca un patrón de 2 elementos, si paso A entonces pondrá B. Porque se está hablando de específicamente una secuencia de palabras. Ej. Si puso “quiero”, puede ser “café”, “te”, etc.

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* ***WHERE B.KEY IS NULL****\_\_\_\_\_*

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)

Como fue explicado con anterioridad, considero que ha de utilizar cierto patrones que hay en los conjuntos de caracteres, por medio de REGEX y luego utilizar un tipo de “skipgrams” para poder encontrar esas relaciones que hay entre el texto que se está comparando con la tarea real del alumno y ver si la estructura y las palabras utilizadas son lo suficientemente comparables como para decir que es copia.