Análisis y Curación de datos
Trabajo Práctico I

Consignas:

El objetivo en este punto es preparar los datos que serán el input de modelos de aprendizaje automático (ML).

a - Realizar el merge entre los 2 datasets que tenemos.
b - Determinar la cantidad de categorías aparecen en los resultados con menor frecuencia y agruparlas creando una nueva categoría.
c - Imputar valores para las columnas que contengan variables numéricas. Graficar la distribución antes y después de la imputación. Concluir al respecto.
d - Realizar los encodings necesarios sobre las variables categóricas: directory, device, category, type y title. Podrían utilizar One Hot Encoder de sklearn.
e - Se podría determinar la distancia entre el código postal donde se realiza la búsqueda y el código postal de los primeros 3 resultados (position 1, 2 y 3)? Se podría determinar la distancia entre el código postal donde se realiza la búsqueda y el código postal de los últimos 3 resultados (position 18, 19 y 20)? De ser posible determinarla: crear dos nuevas columnas que tengan la distancia promedio para los 3 primeros resultados y otra columna que tenga la distancia promedio para los últimos 3 resultados .
f - Determinar cuáles columnas son candidatas a ser desechadas por no aportar valor.
g - Realizar una reducción de dimensiones utilizando PCA. Concluir al respecto.

Consideraciones a tomar en el proceso:

- Dejar bien documentadas las decisiones que se tomaron en los puntos anteriores

Análisis del Contenido

- Determinar si todas las variables tienen el tipo apropiado.
- Analizar las features con tipo Objeto.
- Cuando sea conveniente llevar todo a minúsculas y quitar signos de puntuación.
- Eliminar palabras muy frecuentes que no nos dicen nada (‘de’, ‘en’, ‘con’, ‘para’, ‘la’, ‘el’, ‘&’, etc.) en la columna title.

Entregables

Los entregables de este práctico consisten en:

- Luego de pasar por todos los puntos de las consignas, almacenar en un nuevo archivo los datos resultantes.
- Elaborar un script que contenga una función (o varias) para curar nuevos datos con la misma estructura.

Guía para realizar un checklist por feature:
https://dimewiki.worldbank.org/Checklist:_Data_Cleaning