Universidad Nacional Autónoma de México
Facultad de Ciencias

Profesores:
Dra. Amparo López Gaona
M. I. Gerardo Avilés Rosas

Laboratorista:
José Luis Vázquez Lázaro

1. Objetivo

Al finalizar esta práctica el alumno sera capaz de:

- Entender la naturaleza de los datos en un entorno de la vida real, siendo esta la de inconsistencias, redundancias, valores ausentes y anómalos.
- Entender la naturaleza de la de Extracción, Limpieza y Transformación de los Procesos ETL.
- Conocer software de Procesamiento y Limpieza de datos (**DataCleaner**).

2. Marco teórico

La **Limpieza de Datos** (en inglés ***Data Cleaning***) es el acto de descubrimiento y corrección ó eliminación de registros de datos erróneos de una fuente de datos (por lo general, de BDR). El proceso de limpieza de datos permite identificar datos incompletos, inconsistentes, ruidosos (incorrectos), no pertinentes, etc. y luego sustituir, modificar o eliminar estos "datos sucios" ("***Data Duty***"). Después de la limpieza, la fuente de datos podrá ser compatible con otras fuentes de la misma naturaleza.

Los datos incompletos, inconsistentes y/o ruidosos en un conjunto de datos pueden haber sido causado por: definiciones de diccionario de datos diferentes, errores de entrada del usuario y/o corrupción en la transmisión o el almacenaje de estos. El proceso de limpieza de datos tiene como finalidad alcanzar datos de calidad.

3. Instrucciones

- Descargar el archivo `nuevosVendedores.cvs`.
- Descargar **DataCleaner**.
- Descargar **Spoon**.

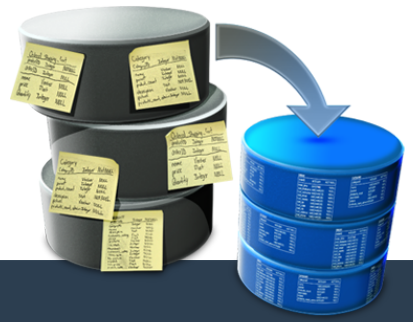
Elaborado por José Luis Vázquez Lázaro / 06 de febrero de 2018.

4. Caso de uso

La Agencia “Continental” ha contratado nuevos empleados en cada una de sus sucursales, por lo que es necesario registrarlos en su BD. El Área de Sistemas de la empresa (en la cual tú trabajas) ha recibido un archivo `.cvs` con la información de los nuevos vendedores. Sin embargo, antes de vaciar estos datos en la BD será necesario analizarlos, pues no se tiene certeza de su calidad.

5. Actividades

1. Con ayuda de **DataCleaner**, analiza cada campo del archivo proporcionado para conocer:
   - *Significado del campo*.
   - *Tipo*: **numérico** o **nominal**.
   - *Número de registros*.
   - *Número de valores nulos*.
   - Para campos numéricos obtener:
     - *Máximo y mínimo valor.*
     - *Media*.
     - *Mediana*.
     - *Desviación estándar.*
   - Para campos nominales:
     - *Moda (sólo si los valores del dominio pueden repetirse en varios registros)*.
     - *Máximo número de caracteres*.
2. Después de este análisis sobre los campos, aplica, si es necesario y con ayuda de **Spoon**, la o las técnicas de limpieza que consideres apropiadas para obtener datos de calidad. Estas pueden ser:
   - *Normalización /Estandarización*.
   - *Valores Perdidos*.
   - *Valores Atípicos*.



6. Desarrollo

El desarrollo de la práctica será exclusivamente en el horario de laboratorio.

7. Entregables

Deberás enviar un archivo `.zip`, con nombre `<número de cuenta>_practica2`, que contenga lo siguiente:

- Un archivo `.pdf` que contenga:
  - El análisis de cada uno de los campos, como se indica en 5.1.
  - ¿Cuáles campos necesitaron limpieza? y ¿por qué?
  - ¿Cuáles técnicas utilizaste para llevar a cabo la limpieza? y ¿por qué?
- Los archivos generados en **DataCleaner** que necesitaste para el análisis y los archivos generados en **Spoon** que necesitaste para la limpieza de los datos.
- El archivo `.cvs` resultante de la limpieza

a la dirección de correo `luis_lazaro@ciencias.unam.mx` con el asunto `[A&MD2018-2]Practica02`.