Laboratorio Módulo 2 - REALIZAR EL PROCESO ETL PARTIENDO DE UN DATASET Y DATOS DE WIKIDATA Y EL MUSEO DEL PRADO

Juan Sánchez de Corta

Contenido

1. Fase 1: Carga de los datos

1.1. Análisis previo

- Uri:
    - La imagen de “Guacamayo” no es correcta ya que en su lugar se encuentra la imagen de “Lobo”.
    - El link de wikidata de “Lobo” es incorrecto ya que en vez de aparecer https://www.wikidata.org/wiki/Q18498 aparece Q18498.
    - La imagen de “Nutria” es incorrecta ya que en su dirección aparece el protocolo htttps en vez de https.
    - El link de wikidata de “Perro” es incorrecto ya que en vez de aparecer https://www.wikidata.org/wiki/Q20717272 aparece Q20717272.
    - La imagen de "Pavo” no es correcta ya que en su lugar aparece la imagen de “Oso”.
    - El link de wikidata de “Rana” debería ser https://www.wikidata.org/wiki/Q59237, pero en su lugar aparece https://www.wikidata.org/wiki/Q11788, que pertenece a “Oso”.
    - El link de wikidata de “Búfalo” no es correcto, ya que aparece https://www.wikidata.org/wiki/Q40435, que pertenece a una ciudad. En su lugar debería aparecer https://www.wikidata.org/wiki/Q42710.
    - El link de wikidata de “Ciervo” era https://www.wikidata.org/wiki/Q37548859, correspondiente a Ciervo en español. En su lugar lo he cambiado por https://www.wikidata.org/wiki/Q29838690, correspondiente a deer.
- Correspondencia de campos:
    - El campo Interés de “Rana” no contiene una cadena que represente su interés, sino que representa una uri a wikidata.
- Comprobar nombres comunes:
    - “Águila” no contiene nombre.
    - “Búho” tiene el nombre común ‘Strigiformes’ que he sustituido por ‘Bubo’.
    - “Caballo” tiene el nombre común ‘Equus ferus caballus’ que he sustituido por ‘horse’.
    - “Conejo” tiene el nombre común ‘Oryctolagus cuniculus’ que he sustituido por ‘rabbit’.
    - “Lagartija” tiene el nombre común ‘Lacertidae’ que he sustituido por ‘lizard’.
    - “Nutria” tiene el nombre común ‘Lutrinae’ que he sustituido por ‘otter’.
    - “Pavo” tiene el nombre común ‘Meleagris gallopavo domesticus’ que he sustituido por ‘domestic turkey’.
    - “Rana” tiene el nombre común ‘rana’ que he sustituido por ‘Northern Red-legged Frog’.
    - “Ardilla” tiene el nombre común ‘Sciurus vulgaris’ por ‘tree squirrell’.
- Comprobar descripciones:
    - “Ardilla” contiene la descripción de “Águila”.
    - “Águila” contiene la descripción de “Ardilla”.
    - “Búho” no contiene su propia descripción, sino una descripción genérica de aves rapaces.
    - “Ciervo” contiene la descripción de ‘Cervidus’, la cual es muy genérica, en vez de contener la suya propia.
    - “Gato” contiene la descripción de “Gaviota”.
    - “Gaviota” contiene la descripción de “Gato”.

- Inconsistencias internas:
    - “Águila” no contiene velocidad.
    - Se nombra a “Asno” como “Burro” y viceversa.
    - “Gaviota” no tiene velocidad.
    - “Guacamayo” no tiene velocidad.
    - “Nutria” contiene 2 campos de velocidad.
    - “Perdiz” no tiene velocidad.
    - “Rana” no tiene velocidad.
    - “Pavo” contiene una velocidad de 160 km/h, la cual he sustituido por 80 km/h.
    - “Asno” contiene 2 campos de interés.
    - “Camaleón” contiene una velocidad de 100 km/h, la cual he sustituido por 2 km/h.
    - “Rana” tiene en el campo interés una uri a wikidata.
    - El campo interés de “Corzo” tiene el valor de ‘Altoo’, el cual habría que sustituir por ‘Alto’.

1.2. Número y Valor de Propiedades Distintas

Para cada propiedad, contaremos los valores distintos y revisaremos si hay errores en los datos:

- descripción: 20 valores distintos
- imagen: 20 valores distintos
- velocidad: 17 valores distintos el valor de "Perro" tiene una velocidad en millas/hora que debe ser convertido a km/h si se quiere consistencia)
- link wikidata: 18 valores distintos
- nombre común: 18 valores distintos
- interés: 4 valores distintos (Alto, Medio, Bajo, Intermedio)

1.3. Identificar el Animal con Más Información

Cuento el número de campos con información completa para cada animal:

- Ardilla: 6 campos (descripción, imagen, velocidad, link wikidata, nombre común, interés)
- Águila: 5 campos (descripción, imagen, velocidad, link wikidata, nombre común, interés está incorrecto)
- Asno: 5 campos (imagen, velocidad, link wikidata, nombre común, descripción)
- Búho: 5 campos (imagen, velocidad, link wikidata, nombre común, descripción)
- Búfalo: 5 campos (descripción, imagen, velocidad, link wikidata, nombre común)
- Caballo: 5 campos (descripción, imagen, velocidad, link wikidata, nombre común)
- Caracol: 5 campos (descripción, imagen, velocidad, link wikidata, nombre común)
- Camaleón: 5 campos (descripción, imagen, velocidad, link wikidata, nombre común)
- Conejo: 6 campos (descripción, imagen, velocidad, link wikidata, nombre común, interés)
- Ciervo: 5 campos (descripción, imagen, velocidad, link wikidata, nombre común)
- Corzo: 5 campos (descripción, imagen, velocidad, link wikidata, nombre común, interés)
- Gamo: 5 campos (descripción, imagen, velocidad, link wikidata, nombre común)
- Gato: 5 campos (descripción, imagen, velocidad, link wikidata, nombre común)
- Gaviota: 5 campos (descripción, imagen, velocidad, link wikidata, nombre común)
- Guacamayo: 5 campos (descripción, imagen, velocidad, link wikidata, nombre común)

- Lobo: 5 campos (descripción, imagen, velocidad, link wikidata, nombre común)
- Lagartija: 5 campos (descripción, imagen, velocidad, link wikidata, nombre común)
- Nutria: 5 campos (descripción, imagen, velocidad, link wikidata, nombre común)
- Perro: 5 campos (descripción, imagen, velocidad, link wikidata, nombre común)
- Perdiz: 5 campos (descripción, imagen, velocidad, link wikidata, nombre común)
- Pavo: 5 campos (descripción, imagen, velocidad, link wikidata, nombre común)
- Oso: 5 campos (descripción, imagen, velocidad, link wikidata, nombre común)
- Rana: 5 campos (descripción, imagen, velocidad, link wikidata, nombre común, interés está en formato incorrecto)

Ardilla y Conejo tienen 6 campos con información completa.

1.4. Multiplicidad de Cada Propiedad

- descripción: 20 valores distintos
- imagen: 20 valores distintos
- velocidad: 17 valores distintos (revisar conversiones)
- link wikidata: 18 valores distintos
- nombre común: 18 valores distintos
- interés: 5 valores distintos

1.5. Propiedades que Deben Tener un Valor Único

Propiedades que deben ser únicas para cada animal:

- Descripción
- Imagen
- Link wikidata
- Nombre común

1.6. Propiedades que Deberían Ser Obligatorias

Propiedades que deberían ser obligatorias, es decir, casi siempre presentes:

- descripción: Presente en todos los casos.
- imagen: Presente en todos los casos.
- velocidad: presente en todos los animales y requiere consistencia.
- link wikidata: Presente en todos los casos.
- nombre común: Presente en todos los casos.
- interés: Faltante para algunos animales y puede ser inconsistente.

1.7. Rango de Aceptabilidad para Cada Propiedad

- descripción: Texto largo, puede tener variaciones en longitud.
- imagen: URL válida de una imagen.
- velocidad: Debe estar en km/h (conversión necesaria).
- link wikidata: Debe ser una URL válida de Wikidata.
- nombre común: Texto corto, debe estar estandarizado.
- interés: Debe estar entre los valores esperados (Alto, Medio, Bajo, Intermedio).

1.8. Variables Susceptibles de Ser Categóricas

- interés: Es categórica (Alto, Medio, Bajo, Intermedio).

Análisis de Categorías de interés:

- Alto: 6 animales
- Medio: 2 animales
- Bajo: 2 animales
- Intermedio: 1 animal

En resumen, hay algunos problemas de consistencia en el dataset, especialmente con la propiedad velocidad e interés. También hay propiedades que deben ser obligatorias para asegurar la integridad de la información. La mayoría de los campos están bien definidos y tienen datos consistentes.