Tipología y ciclo de vida de los datos.
Práctica 1
Yaiza Santana Santana

1. Título del dataset.
2. Subtítulo del dataset.
3. Agregad una imagen que identifique vuestro dataset visualmente.
4. Contexto. ¿Cuál es la materia del conjunto de datos?
5. Contenido. ¿Qué campos incluye? ¿Cuál es el período de tiempo de los datos y cómo se ha recogido?
6. Agradecimientos. ¿Quién es el propietario del conjunto de los datos? Incluir citas de investigación o análisis anteriores.
7. Inspiración. ¿Por qué es interesante este conjunto de datos? ¿Qué preguntas le gustaría responder la comunidad?
8. Licencia. Seleccionad una de las licencias y decid porqué la habéis seleccionado.
Anotaciones.
Bibliografía.

1. Título del dataset.

Para esta práctica se ha elegido el siguiente título para el *dataset*:

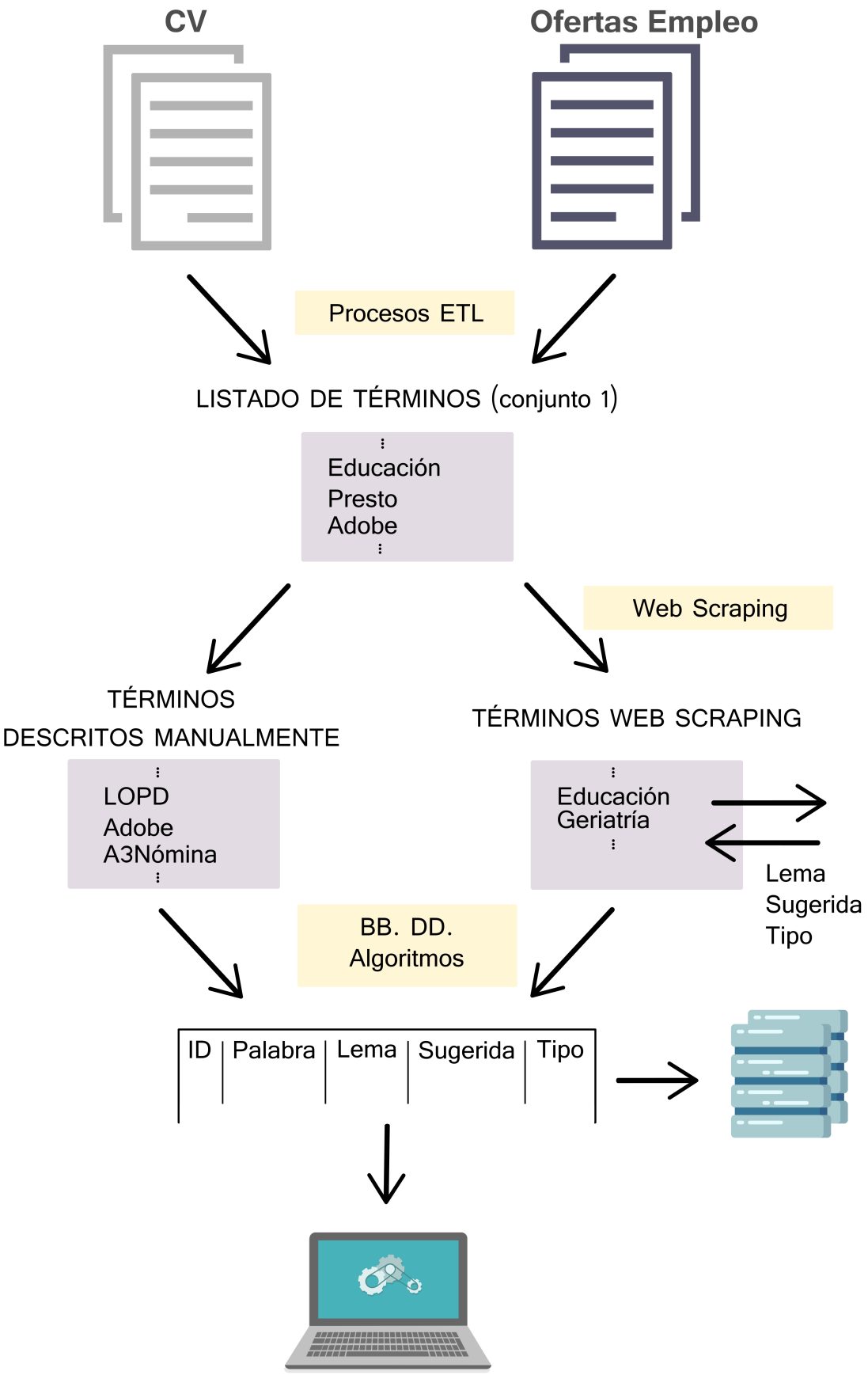
“Diccionario de empleo en Canarias”

2. Subtítulo del dataset.

Competencias demandadas en un portal de empleo, para detección de candidatos válidos para las ofertas de empleo publicadas en dicho portal mediante tratamiento de la información no estructurada de dichos CV's.

En esta etapa nos centraremos sólo en el estudio del diccionario de empleo referido a los CV's almacenados en bases de datos propias.

3. Agregad una imagen que identifique vuestro *dataset* visualmente.



4. Contexto. ¿Cuál es la materia del conjunto de datos?

En ocasiones resulta un trabajo tedioso para el demandante de empleo encontrar un puesto deseado para él en función de sus competencias, y para la empresa encontrar un buen candidato para el puesto ofertado.

Aunque existan datos estructurados en las bases de datos de este portal, como por ejemplo, la titulación del usuario que ayuda a cruzar la oferta de empleo con él si ésta ha sido registrada en la oferta, puede ser que la empresa haya pedido de manera no estructurada (como texto en el campo de descripción del puesto o tareas del puesto) alguna competencia que el usuario ha introducido también de manera no estructurada en el campo de formación, experiencia o conocimientos.

Pongamos un ejemplo, si el usuario ha introducido como formación “Geriatría” (en forma de texto) y la titulación no ha sido introducida, y en una oferta de trabajo se pide “Experiencia en Geriatría” y titulación en Enfermería, si no tratamos este texto para hacer el *maching*, el sistema no podrá sugerir esta vinculación entre demandante y oferta con sólo cruzar titulaciones desde una consulta a la base de datos.

Si conseguimos realizar el tratamiento de texto de todos esos campos mencionados y extraemos las competencias pedidas y su vocabulario lematizado, podremos analizar tanto las ofertas como los CV's para proponer al usuario ofertas de empleo más ajustadas a su perfil, y para proponer mejores candidatos para las ofertas de empleo.

Por lo tanto, hablaremos de dos conjuntos de datos con respecto a los CV's:

- Conjunto 1: La fuente de origen para el *Web Scraping* pedido: diccionario de palabras extraídas de los CV's.
- Conjunto 2: El conjunto de datos obtenido tras el *Web Scraping:* diccionario enriquecido con lema y tipo de la palabra para el estudio de las competencias y reducción del diccionario.

Para conseguir las palabras que conforman el diccionario que aparecen en dichos CV's, debemos llevar a cabo un proceso de extracción, transformación y carga (proceso ETL), y posteriormente, acceder a la página de la RAE para extraer el lema de la palabra y el tipo de palabra a través de *Javascript Scraping* (*Web Scraping* de contenido dinámico).

En el siguiente gráfico se explica el *Workflow* llevado a cabo:

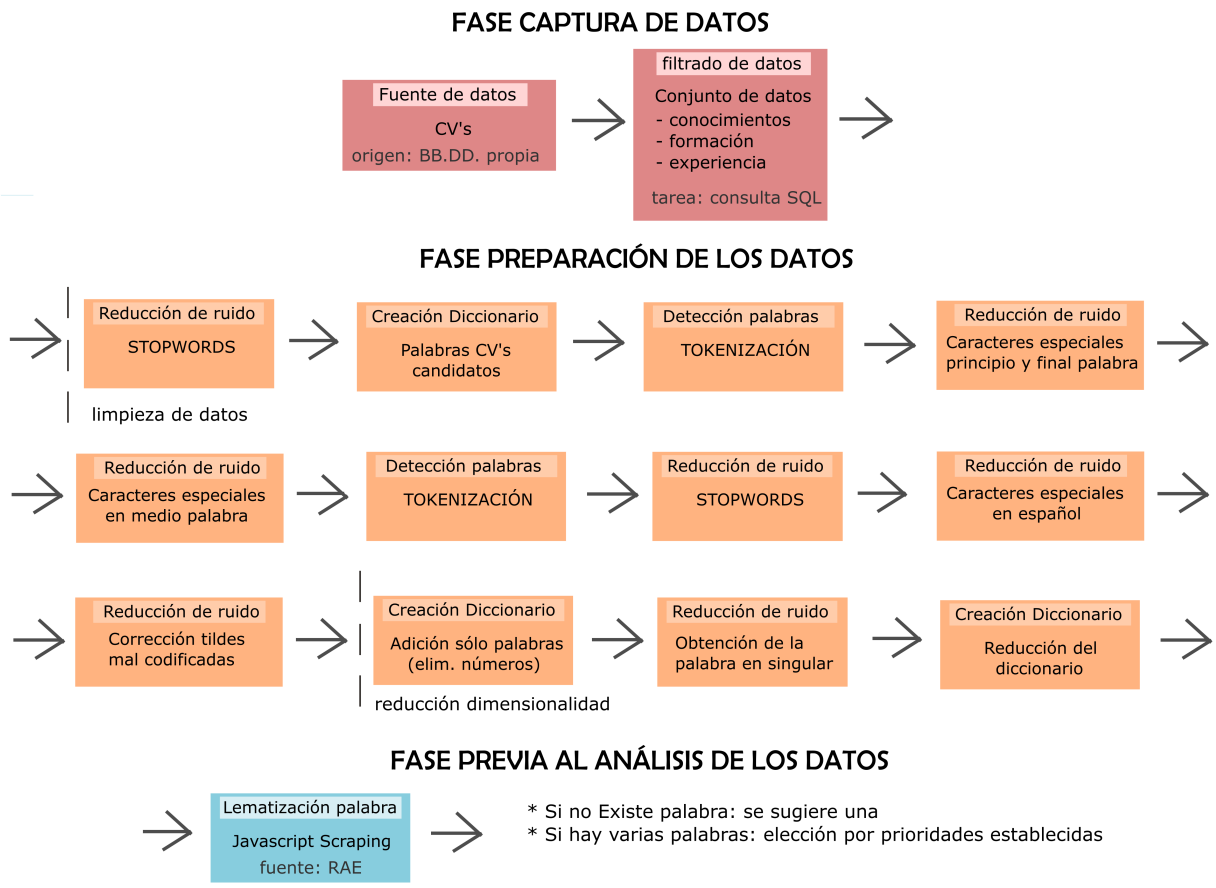


Figura 1. *Workflow* para la obtención del diccionario final enriquecido.

5. Contenido. ¿Qué campos incluye? ¿Cuál es el período de tiempo de los datos y cómo se ha recogido?

El diccionario del conjunto 1 y 2 obtenido es **diacrónico**, delimitado en el tiempo y en el espacio:

- En el tiempo: se hizo la creación de diccionarios a través del estudio de los CV's de los usuarios que accedieron al portal de empleo en el mes de octubre o actualizaron su CV's en el mismo mes.
- En el espacio: estos CV's son en gran mayoría de residentes en las Islas Canarias.

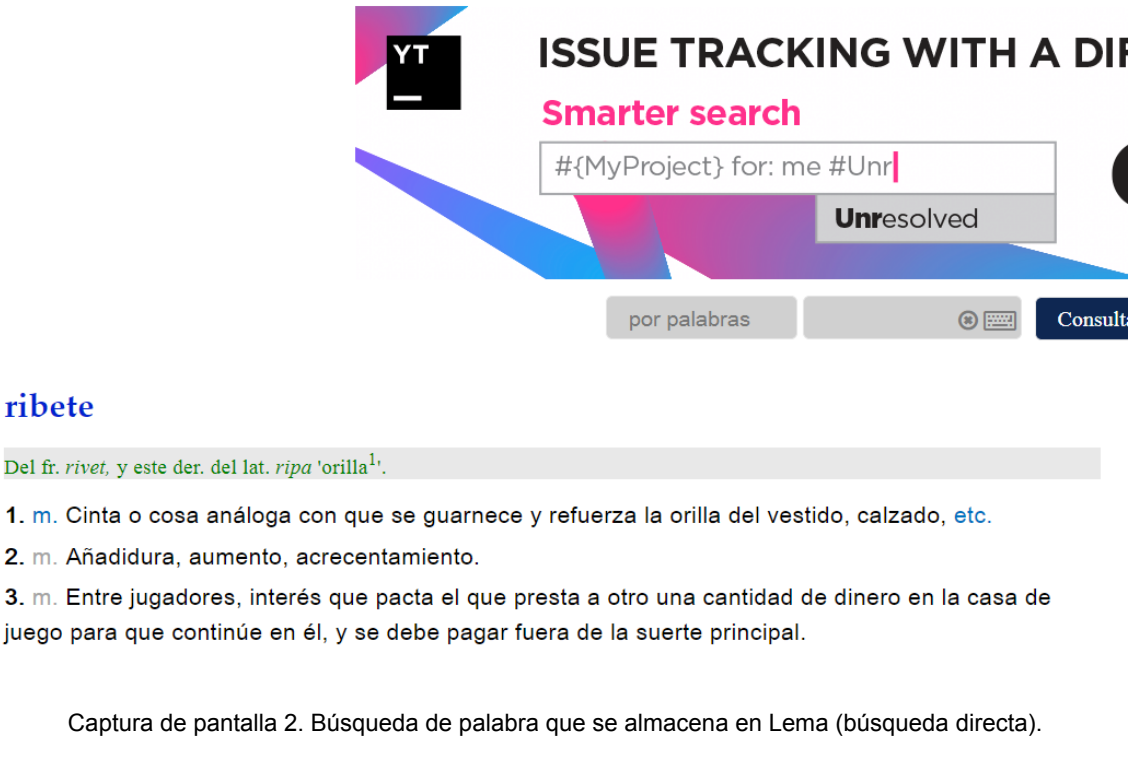
El conjunto 1 se ha recogido con consultas SQL a la base de datos propia de manera agregada. Posteriormente se ha tratado estos datos como se indicó en el apartado anterior. La tabla de esta base de datos que almacenará el conjunto 1 tiene los siguientes campos (columnas):

- ID de Palabra: Identificador único de palabra.
- Palabra: palabras que se obtuvieron tras los procesos ETL descritos anteriormente (último cajón naranja).
- Término: indica si es terminología de empleo que no podrá detectar la RAE, ya que son abreviaciones, software, etc.
- Descripción: descripción corta de ese termino de empleo que no detecta la RAE.

El conjunto 2 se ha conseguido enriqueciendo el conjunto 1 con *Javascript Scraping.* La base de datos que almacenará el conjunto 2 tendrá:

- ID de Palabra: Identificador único de palabra (referencia al ID del conjunto 1).
- Palabra: palabras que se obtuvieron tras los procesos ETL descritos anteriormente (último cajón naranja).
- Lema: lema de la palabra.
- Sugerida: Palabra sugerida por la RAE en el caso de varias propuestas, por preferencia:
  1. Nombre masculino (m.)
  2. Nombre femenino (f.)
  3. Adjetivo (adj.)
  4. Transitivo (tr.)
  5. Intransitivo (intr.)

Captura de pantalla 1. Búsqueda de palabra con dos sugerencias.



Captura de pantalla 2. Búsqueda de palabra que se almacena en Lema (búsqueda directa).

La palabra *kele* no está registrada en el Diccionario. Las entradas que se muestran a continuación podrían estar relacionadas:

- criar
- queche
- quedar
- quejar
- quemar
- queque
- querer[1]

Captura de pantalla 3. Búsqueda de palabra que no se busca ni lema ni sugerida por tener más de tres palabras sugeridas.

6. Agradecimientos. ¿Quién es el propietario del conjunto de los datos? Incluir citas de investigación o análisis anteriores.

La empresa que gestiona el portal de empleo es la propietaria de los datos almacenados en las bases de datos propias (conjunto 1), los cuales son recopilados de los usuarios del portal, quienes han dado permiso expreso para el tratamiento de éstos en beneficio de ellos, siempre salvaguardando la seguridad e integridad de los datos.

Cabe destacar que, gracias a la RAE, se pudo obtener el lema y tipo de las palabras del diccionario obtenido en el conjunto 2, para la mejor comprensión y reducción de los datos.

El conjunto 2 será de uso público debido a su gran interés social para la comunidad canaria.

7. Inspiración. ¿Por qué es interesante este conjunto de datos? ¿Qué preguntas le gustaría responder la comunidad?

Me he inspirado en este conjunto de datos para ayudar a mejorar la búsqueda de empleo a la comunidad joven canaria en dicho portal de empleo, la más castigada por el desempleo juvenil de toda España.

Se pretende ofrecer respuestas a los usuarios en su búsqueda de empleo, y ofrecer más CV's a las empresas para mejorar la empleabilidad.

8. Licencia. Seleccionad una de las licencias y decid porqué la habéis seleccionado.

Finalmente he decidido seleccionar la licencia **Released Under CC0: Public Domain License** [1], puesto que son las licencias más usadas en el ámbito de la investigación, para renunciar a todos sus derechos de autor y derechos relacionados en sus trabajos en la medida máxima permitida por la ley. Los datos del conjunto 2 serán datos de uso público, para así poder compartirlos con la comunidad de interés.

Anotaciones.

Se ha consultado a la RAE las primeras 1700 palabras del conjunto 1 de datos obtenido (total de 5698 palabras). Puede suceder que mientras se realiza el Web Scraping se pierda la conexión de datos, y por lo tanto el script desarrollado en Python dejará de funcionar correctamente. Para no perder todos los resultados obtenidos hasta que eso sucediera, el script se debe ejecutar en grupos de datos más pequeños, por ejemplo, de 50 en 50.

Bibliografía.

[1] https://creativecommons.org/share-your-work/public-domain/cc0/