Autor: Carlos Herrera Carballo.

Asignatura: Tipología y ciclo de vida de los datos (Curso 2020-2021).

Práctica 1: Web Scraping

1. Contexto.

En esta actividad práctica, se ha implementado un scraper mediante el cual se pueden obtener dos conjuntos de datos de distinto tipo. Por un lado, el primer conjunto de datos generado recoge la información publicada, en su edición de enero de 2021, del "Ranking Webometrics of World Universities" para cada universidad del continente especificado. Por otro lado, el segundo conjunto de datos presenta esta misma información pero perteneciente a las publicaciones de dicho ranking entre los años 2012 y 2020, de modo que se tiene un registro histórico de datos para cada universidad. Además, en el segundo caso, únicamente se ha extraído la información de las primeras 100 universidades de cada ranking por región. Por tanto, puede extraerse la información actual o histórica de los ránquines de los cinco continentes: Europa, Asia, Oceanía, África y América (dividido en Norteamérica y Latinoamérica).

El "Ranking Webometrics of World Universities" es una iniciativa del Cybermetrics Lab, grupo de investigación perteneciente al Consejo Superior de Investigaciones Científicas (CSIC), que analiza cuantitativamente los contenidos de Internet y de la Web, particularmente, aquellos que tienen relación con los procesos de generación y comunicación académica del conocimiento científico. Así, cada seis meses se lleva a cabo la actualización y publicación de este ranking en la página web en la que se ha llevado a cabo la extracción de datos.

2. Definir un título para el dataset.

El título escogido para este dataset es el siguiente: "*Ranking universitario Webometrics"*.

3. Descripción del dataset.

Tal y como se ha descrito anteriormente, el primer conjunto de datos generado recoge la información para cada universidad del "Ranking Webometrics of World Universities" perteneciente a su edición de enero 2021 y, el segundo, esta misma información publicada en las ediciones del ranking desde 2012 a 2020 para las primeras 100 universidades de la región escogida . En concreto, las variables extraídas son: el nombre de la universidad, su página web, su posición en el ranking regional, su posición en el ranking mundial y sus valores calculados por Cybermetrics Lab para los indicadores de impacto, apertura y excelencia. En el caso del registro histórico de datos, se ha añadido también una variable que hace referencia al año de publicación del ranking en el que se obtiene la información.

4. *Representación gráfica.*

5. *Contenido.*

5.1 Descripción de las variables.

Para cada universidad, que constituye un registro en el conjunto de datos generado, se obtienen las siguientes variables:

- **Name**: nombre oficial de la universidad.
- **Web**: web oficial de la universidad.
- **Regional_ranking**: posición de la universidad en el ranking regional (en función del continente elegido).
- **World_ranking**: posición de la universidad en el ranking mundial.
- **Impact_score**: puntuación para el indicador de impacto.
- **Openness_score**: puntuación para el indicador de apertura.
- **Excellence_score**: puntuación para el indicador de excelencia.
- **Year**: año de publicación del ranking (sólo en el dataset histórico)

Respecto al tipo de estas variables, las puntuaciones de los distintos indicadores, el año, así como las posiciones de los ránquines son de tipo entero, mientras que el resto de variables son de tipo texto.

5.2 Descripción de los indicadores.

Las cifras publicadas en el ranking se denominan “ranks” y cuanto menor sea el valor de las mismas, mejor será el resultado de la universidad en este apartado. Así, en el ranking Webometrics podemos encontrar los siguientes ranks:

- **Presencia**: número total de páginas del dominio web principal de la institución. La fuente de este indicador es Google y tiene un peso del 5% en la ponderación final.

- **Impacto / Visibilidad**: número de redes externas que enlazan con las páginas web de la institución. En este caso, las fuentes son Ahrefs y Majestic y su peso en la ponderación es del 50%.

- **Apertura / Transparencia**: número de citas de los principales autores (concretamente de los primeros 210 autores). Para este indicador, la fuente de información es Google Scholar Citations y tiene un peso del 10% en la ponderación.

- **Excelencia**: número de artículos entre el 10% más citado en cada una de las 27 disciplinas de la base de datos. La fuente de información utilizada para este indicador es Scimago y su peso es del 35%.

Nota: Es necesario señalar que el indicador de presencia ha sido omitido en las últimas ediciones del ranking. En consecuencia, en esta actividad práctica los valores para este rank no se recogieron.

5.3 Periodo de tiempo de los datos y método de extracción.

Tal y como se ha descrito en apartados anteriores, los datos que corresponden a la primera edición de 2021 del "Ranking Webometrics of World Universities", han sido recolectados por Cybermetrics Lab en enero del mismo año. El periodo de validez de estos datos serán seis meses, dado que dicho ranking es actualizado transcurrido este tiempo. En este sentido, la próxima actualización de estos datos se llevará a cabo en julio de 2021.

En el caso de los datos históricos, el periodo de validez será de un año, ya que, una vez finalizado el año 2021, si se quiere tener los datos actualizados deberán extraerse nuevamente obteniéndose así el histórico desde el año 2012 hasta el año 2021.

Para llevar a cabo la extracción de los datos del ranking del año 2021, se ha implementado un script en lenguaje de programación Python a través del cual, por medio de técnicas de Web Scraping, se han obtenido los datos requeridos para este proyecto de las páginas HTML del código fuente de la dirección web oficial del "Ranking Webometrics of World Universities". Por otra parte, para la extracción de los datos históricos de los ránquines publicados desde 2012 hasta 2020, se ha empleado las mismas técnicas pero, dado que esta información no estaba en la misma página web, se ha realizado Web Scraping sobre Wayback Machine, un servicio y base de datos que contiene copias de una gran cantidad de páginas o sitios de Internet. Así, se han extraído los datos publicados más cercanos al mes de diciembre de cada año, asegurando de este modo tener los últimos datos actualizados en dicho año por Cybermetrics Lab en la página web del ranking.

Finalmente, es necesario remarcar que se ha elegido la versión inglesa de esta página web debido a que la versión española no estaba actualizada en el momento de llevar a cabo esta actividad práctica (Marzo y Abril de 2021).

6. Agradecimientos.

El Cybermetrics Lab, organismo adscrito al CSIC, viene desarrollando estudios cuantitativos focalizados en la web académica desde mediados de los noventa. Tras publicar un primer indicador en 1996, comenzó la recopilación de los primeros datos web de universidades europeas en 1999. Posteriormente, hacia el año 2003 y tras de la publicación del ranking de la Universidad Jiatong de Shanghai, el Ranking Académico de Universidades del Mundo (ARWU),

el grupo decidió adoptar las innovaciones introducidas en dicho ranking y comenzar el proyecto "Ranking Webometrics of World Universities", cuya primera edición fue publicada en 2004. En este sentido, el objetivo original del Ranking es promover la presencia académica en la web, apoyando las iniciativas de Open Access para incrementar significativamente la transferencia de conocimiento científico y cultural generado por las universidades a toda la Sociedad.

En la actualidad, queda patente la importancia de cuantificar la presencia y visibilidad web de las universidades y , con ello, poder estudiar diferentes aspectos que sean reflejo de su calidad como instituciones en el ámbito académico y, además, nos permita establecer comparaciones entre distintos centros de esta índole. Esto, se traduce en la necesidad de desarrollar indicadores web que nos ayuden a medir estas características y sus tendencias a lo largo de los años, tal y como exponen los autores Isidro F. Aguillo y Begoña Granadino (F. Aguillo & Granadino, 2006).

En la literatura publicada, se han llevado a cabo diversos estudios para analizar dicha presencia y visibilidad web de las universidades, un ejemplo de ello, es el estudio llevado a cabo por los autores Enrique Orduña-Malea, Jorge Serrano-Cobos, José-Antonio Ontalba-Ruipérez y Nuria Lloret-Romero en el que han estudiado la evolución del tamaño y visibilidad de los dominios web de las universidades públicas españolas desde enero hasta junio de 2009 (Orduña-Malea et al., 2010). Específicamente, también es relevante conocer la capacidad de las universidades para transmitir el conocimiento científico fruto de las investigaciones que se realizan en ellas. Por ello, el estudio de la presencia en redes sociales científicas, tales como ResearchGate o Academia.edu, de las universidades resulta vital para tal fin. En este sentido, también existen análisis de este hecho, entre ellos, el realizado por Cristina González-Díaz, Mar Iglesias-García y Lluís Codina (González-Díaz et al., 2015).

En conclusión, la búsqueda de datos confiables, sobre los que sustentar más estudios como los anteriormente citados, es clave para permitir a equipos de investigación seguir indagando en este campo de estudio y, con ello, orientar a las instituciones hacia la universalidad del conocimiento en la Web en un mundo cada vez más interconectado.

7. Inspiración.

El conjunto de datos obtenido en la presente actividad práctica permite llevar a cabo estudios de distinta índole. A continuación, se describen algunas de las posibilidades de estudio o cuestiones que pueden resolverse tras un análisis de los datos extraídos del ranking Webometrics.

En primer lugar, se puede llevar a cabo un estudio de la posición de las universidades de una región concreta en el ranking mundial, siendo esto un indicativo importante para detectar posibles déficits en cuanto a la generación y comunicación del conocimiento científico, concretamente, a través de la web en las instituciones de dicho continente. En este sentido, en aquel continente en el que se detecten bajas posiciones en el ranking mundial de sus universidades, se evidenciará la necesidad de remediar esta situación, lo que se traducirá en una mejora que repercutirá positivamente en la transmisión del conocimiento no solo en dicha región, sino también en otros territorios que se valgan del mismo de forma online.

En segundo lugar, también constituye una fuente de información valiosa para las propias universidades, debido a que pueden conocer las puntuaciones de los distintos indicadores de su "competencia", es decir, los puntos débiles y fuertes de otras instituciones (regionales e internacionales), así como los propios. En consecuencia, puede servir de ayuda en la toma de decisiones orientada a conseguir una mejora en aquellos aspectos que se consideren necesarios tras el análisis de estos datos. Con esto, se favorecerá una competencia entre instituciones que, nuevamente, conllevará un enriquecimiento mutuo del conjunto de universidades de la región.

En tercer lugar, podremos tener una colección de datos históricos que permita evaluar la tendencia de estos indicadores para las distintas universidades tras el paso del tiempo. Así, se podrán llevar a cabo estudios predictivos para aproximar el impacto global de las universidades de una región en el futuro, o hacer balance sobre la evolución o involución experimentada. Tanto los datos actualizados como los históricos, también pueden ser una gran fuente de información para llevar a cabo un trabajo de periodismo de datos.

Finalmente, en comparación con los estudios expuestos en el apartado anterior, el conjunto de datos generado presenta una mayor capacidad analítica debido a ciertos aspectos. Por un lado, para cada universidad se recoge un total de cinco indicadores (posición ranking mundial, posición ranking regional, puntuación de impacto, puntuación de apertura y puntuación de excelencia), en vez de únicamente dos indicadores como el tamaño web y el impacto empleados en el estudio del autor Orduña-Malea (Orduña-Malea et al., 2010). Además, la definición de los indicadores de impacto, apertura y excelencia se encuentran en un constante proceso de mejora por parte de los investigadores del Cybermetrics Lab, esto permite tener una medición más precisa y fiable de cada uno de estos aspectos para cada universidad estudiada. Por otro lado, en el estudio de la autora González-Díaz (González-Díaz et al., 2015) se recalca la imposibilidad de comparar la presencia web en distintas redes sociales académicas (Researchgate y Academia.edu) dado que cada una tenía su propio indicador de impacto definido, sin embargo, el presente conjunto de datos nos permite llevar a cabo estas comparaciones dado que sus indicadores son métricas estandarizadas, lo que permite solventar la problemática de no tener medidas normalizadas para cada universidad.

Bibliografía empleada en los apartados 6 y 7:

[1] AGUILLO, Isidro F.; GRANADINO, Begoña (2006). «Indicadores web para medir la presencia de las universidades en la Red». En: ROCA, Genís (coord.). La presencia de las universidades en la Red[monográfico en línea]. Revista de Universidad y Sociedad del Conocimiento(RUSC). Vol.3, n.°1. UOC. [Fecha de consulta: 25/03/2021]. http://www.uoc.edu/rusc/3/1/dt/esp/aguillo_granadino.pdf ISSN 1698-580X

[2] Orduña-Malea, E., Serrano-Cobos, J., Ontalba-Ruipérez, J. A., & Lloret-Romero, N. (2010). Presencia y visibilidad web de las universidades públicas españolas. Revista Española De Documentación Científica, 33(2), 246–278. https://doi.org/10.3989/redc.2010.2.740

[3] Díaz, Cristina & Iglesias-García, Mar & Codina, Lluís. (2015). Presencia de las universidades españolas en las redes sociales digitales científicas: caso de los estudios de comunicación. El Profesional de la Informacion. 24. 10.3145/epi.2015.sep.12.

8. Licencia.

El tipo de licencia Creative Commons escogida para la publicación de los posibles conjuntos de datos que pueden obtenerse con el scraper implementado es: **Released Under CC BY-NC-SA 4.0 License**. Esto se debe a que las cláusulas de este tipo de licencia se alinean, consecuentemente, con los objetivos del proyecto planteado. A continuación, haciendo uso de las definiciones expuestas por Creative Commons en su web, se describe las características de esta licencia:

- ***Attribution (BY)***: se debe otorgar el crédito correspondiente al propietario del conjunto de datos, proporcionar un enlace a la licencia e indicar si se realizaron cambios. Puede hacerse de cualquier manera razonable, pero no de ninguna manera que sugiera que el licenciante respalda al usuario del conjunto de datos o su uso.

- ***NonCommercial (NC):*** no puede utilizarse el material con fines comerciales**.**

- ***ShareAlike (SA):*** si se modifica, transforma o construye sobre el material, se debe distribuir las contribuciones del usuario bajo la misma licencia que el original.

De este modo, con la primera cláusula (BY) nos aseguraremos de que se reconozca el trabajo realizado en la extracción del conjunto original de datos, otorgándonos dicho mérito por ello siendo citados en posibles investigaciones en las que se haga uso de este material. Además, evitamos tener responsabilidad sobre el uso que otras personas puedan darle al conjunto de datos. Con la segunda cláusula (NC), preservamos uno de los objetivos fundamentales de Cybermetrics Lab y, también, del presente proyecto, como es promover el acceso abierto al conocimiento científico generado en las universidades de todo el mundo. Finalmente, la última cláusula (SA) nos permite seguir salvaguardando el anterior punto, es decir, el uso no comercial de estos datos, así como de cualquier modificación o construcción de los mismos.

9. Código.

El código con el que se ha generado los distintos datasets se encuentra publicado en el siguiente repositorio GitHub: https://github.com/cherreracar/Practica1_WebScraping

10. Dataset.

Tras generar todos los posibles conjuntos de datos que pueden obtenerse al ejecutar el scraper implementado, lo cual, supone un total de doce posibles archivos csv (seis para los ránquines actuales de 2021 y otros seis para el registro de datos históricos desde 2012 hasta 2020), se comprimieron en un archivo rar que, posteriormente, se publicó en Zenodo. De este modo, el **DOI** obtenido para este archivo ha sido el siguiente: **10.5281/zenodo.4663054** con su correspondiente url: **https://doi.org/10.5281/zenodo.4663054**

A continuación, adjuntamos imagen para verificar la publicación:

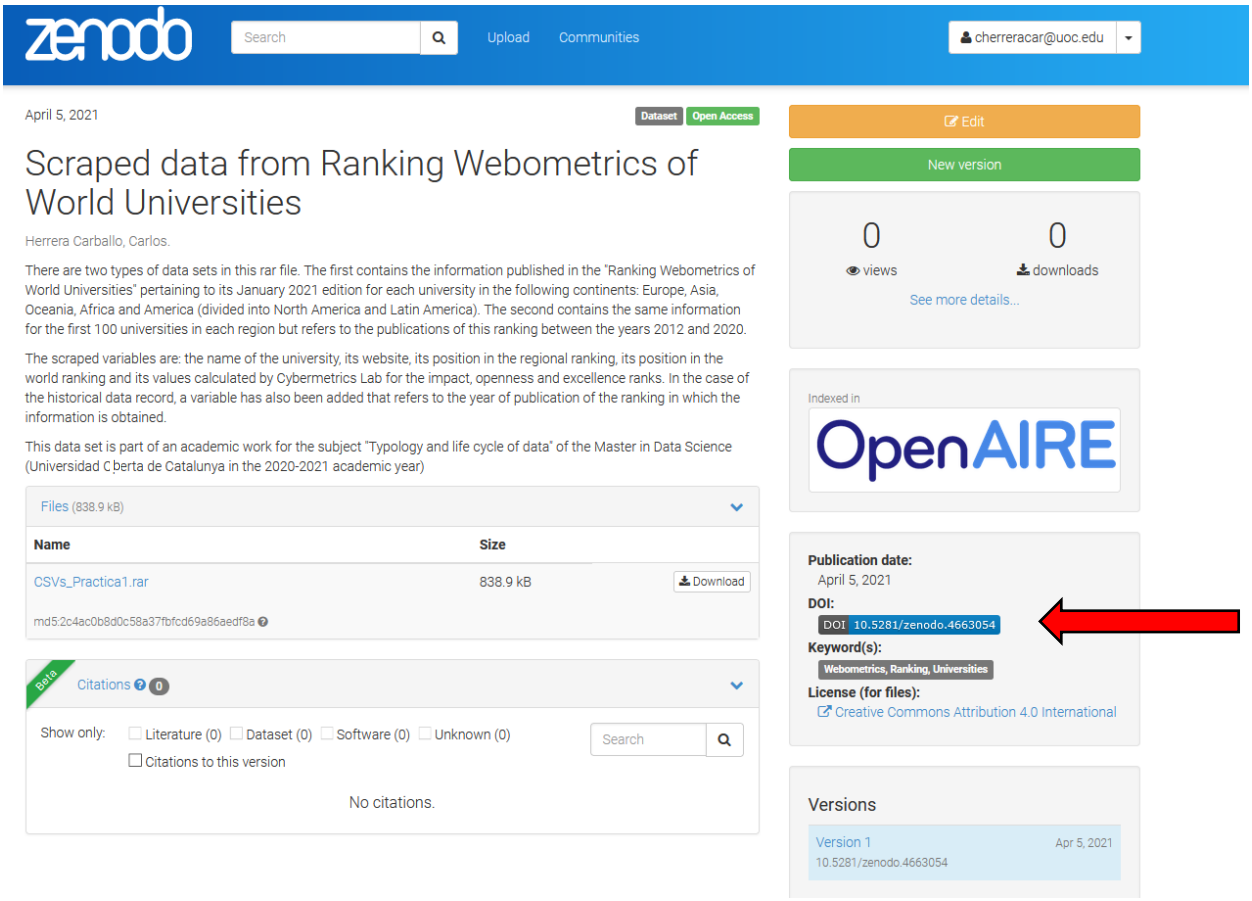

Tabla de contribuciones al trabajo

Contribuciones	Firma
Investigación previa	CHC
Redacción de las respuestas	CHC
Desarrollo código	CHC