Práctica 1

OSCAR JAVIER BACHILLER SANDOVAL

Respuestas a los puntos indicados

1. Contexto. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.

El sitio Web seleccionado para realizar el procedimiento de Web Scraping corresponde con el espacio de la revista Journal of Learning Analytics, cuya dirección electrónica es https://learning-analytics.info/journals/index.php/JLA.

Esta revista es la publicación oficial de la Society for Learning Analytics Research (SoLAR), una sociedad que con base en lo consignado en su sitio Web https://solaresearch.org/, corresponde con una red interdisciplinaria de investigadores internacionales que están explorando el papel y el impacto del análisis de datos en la enseñanza, el aprendizaje, la capacitación y el desarrollo, promoviendo la publicación y difusión de la investigación en este campo.

Es un tema muy interesante que cada vez más toma mayor trascendencia y relevancia con el objeto de analizar y tomar decisiones con base en datos de un contexto educativo, de escenarios de mejora y transformación que pueden darse en el contexto de la enseñanza y el aprendizaje y de la necesidad asociada de promover la transferencia de conocimiento en esta área.

Así se llevo a cabo un procedimiento de Web Scraping, con el objeto de poder recopilar datos que no se encuentran en otros medios; la mayor parte de las publicaciones en esta área se encuentran por medio de artículos disponibles en consulta a través de los portales de revistas especializadas en la difusión de conocimiento en esta disciplina

Por medio del uso del lenguaje de programación Python y de su conjunto de librerías específicamente desarrolladas para este tipo de tareas, se realizó un análisis de la estructura del sitio web, se identificó aquellos datos de interés en cuanto a la publicación de artículos, en este ejercicio en un número de publicación determinado, con el objeto de identificar que temáticas, autores y que volumen de publicación se esta produciendo en este campo

2. Definir un título para el dataset. Elegir un título que sea descriptivo.

infoarticulospublicados_JOLA_scraping.csv

3. Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).

Acorde a uno de los números de la revista, se ha extraído el título de un artículo, sus autores, el enlace de dicho articulo y el resumen de este, todo es texto. Dada sus características, va a permitir su análisis para identificar patrones en el texto, que con base en nuestro propósito contribuya a la gestión del conocimiento, de manera particular en el escenario de la producción académica e investigativa que se publica en este tipo de revistas especializadas, se esperaría que, dada la rigurosidad de estos escenarios de publicación, el trabajo asociado de limpieza de datos sea mínimo.

4. Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente

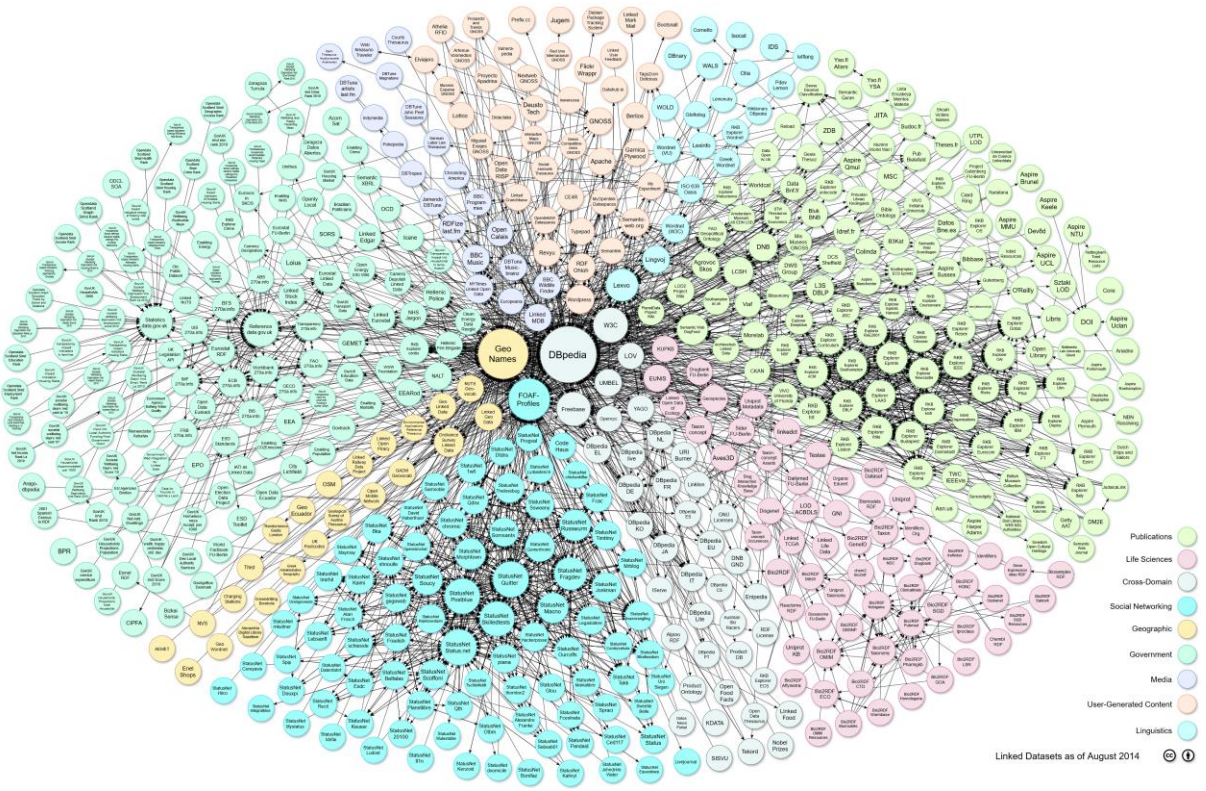

Imagen de dominio público

5. Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.

Título: corresponde con el titulo con el cual ha sido publicado el artículo en la revista.
Autor(es): corresponde al nombre del autor o a los nombres de los autores del artículo publicado en la revista.
Enlace: corresponde al enlace para acceder al contenido del artículo.
Resumen: corresponde con el abstract del articulo en donde se indica su intencionalidad

El tiempo de recolección de los datos es atemporal, dado corresponde con la fecha de publicación y el numero asociado a la revista, en este caso Vol 5 No 3 (2018): Special Section: LAK-18 Invited Papers.

Se ha hecho uso de un procedimiento de Web Scraping haciendo uso de BeautifulSoup, Requests y Pandas.

BeautifulSoup fue utilizado como un medio para extraer y navegar contenido HTML

Requests, para establecer comunicación con el sitio Web por medio de HTTP y solicitar de esta manera el conjunto de servicios asociados a dicho sitio

Pandas, para almacenar los resultados obtenidos del procedimiento de web scrapiong en un archivo csv a través de la conversión a un formato tabular

6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).

Para el desarrollo de este trabajo, se expresa un enorme agradecimiento a la Universidad de Athabasca, https://www.athabascau.ca/, quién es la propietaria de este sitio, al igual que a la Universidad Tecnológica de Sidney, https://www.uts.edu.au/ como principales promotoras de SoLAR. Se ha de hacer una mención especial en el sentido de que los artículos son puestos a disposición de la sociedad a través de una iniciativa de datos abiertos, dado son de acceso abierto (Open Access).

7. Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder.

Este escenario lo he escogido porque es el área sobre la cual estoy encaminando mis esfuerzos de aprendizaje, para apropiarme y poder desenvolverme en el campo del Learning Analytics

Dentro ya de un escenario de divulgación del conocimiento, es importante saber que es lo que se está publicando y quien lo está haciendo, para de esta manera poder comenzar a identificar el estado del arte o el estado de la cuestión frente al desarrollo de esta temática, donde ya se empiezan a articular productos de investigación como libros, pero principalmente la publicación de artículos en muchos casos resultados de investigación en torno a una temática en particular asociada

Es por ello por lo que considero importante conocer que se está produciendo, que es lo que este conjunto de datos puede permitir obtener como elemento de información inicialmente acotado al análisis de la publicación de un numero de la revista, pero que puede extenderse a las demás publicaciones y/o a otras revistas de similar propósito

Así entre las preguntas que podrían responderse están

¿Qué temas relacionados con learning analytics se están publicando?
¿En qué contexto se están llevando a cabo este tipo de publicaciones?
¿Quiénes son los autores que más volumen de publicación están llevando a cabo?
¿Qué redes de conocimiento se están creando a partir de la publicación conjunta de artículos?
¿Qué enfoque están abordando estas publicaciones?

Ello a partir de por ejemplo el análisis del texto por medio de categorías en los títulos el nombre de los autores y en los resúmenes de las publicaciones, a través por ejemplo de minería de textos o inclusive análisis de sentimientos para determinar condiciones de desarrollo de estas iniciativas, etc.

Se mencionaba anteriormente que una de sus posibles aplicaciones se asocia con la gestión del conocimiento. Analizando el texto va a ser posible identificar que expertos existen acorde a una temática particular, que escenarios y enfoques se suelen trabajar a partir de ello y quienes pueden ser a su vez pares temáticos, algo así como contar con una “página amarilla” de profesionales expertos por un campo determinado. En cuanto a la investigación como tal, también es posible establecer pautas de acción sobre el desarrollo de nuevos proyectos.

8. Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección:
    a. Released Under CC0: Public Domain License
    b. Released Under CC BY-NC-SA 4.0 License
    c. Released Under CC BY-SA 4.0 License
    d. Database released under Open Database License, individual contents under Database Contents License
    e. Other (specified above)
    f. Unknown License

La licencia seleccionada es “Released Under CC BY-NC-SA 4.0 License” con el animo de apoyo a la generación de una cultura de respeto por el trabajo de los demás y el propio. Por ello se solicita reconocimiento al autor, la imposibilidad de cualquier uso comercial y de compartir igual en caso de mejora a partir de lo desarrollado

9. Código. Adjuntar el código con el que se ha generado el dataset preferiblemente en Python o, alternativamente, en R.

Disponible en el espacio de GitHub https://github.com/obachiller/WebScraping/tree/master/Notebooks

10. Dataset. Presentar el dataset en formato CSV

Disponible en el espacio de GitHub https://github.com/obachiller/WebScraping/tree/master/Dataset

Contribuciones	**Firma**
Investigación previa	OJBS
Redacción de las respuestas	OJBS
Desarrollo código	OJBS

Recursos

Jarmul, K., & Lawson, R. (2017). Python Web Scraping. Packt Publishing Ltd. Second Edition.
Masip, D. El lenguaje Python. Editorial UOC.
Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.

Tutorial de Github https://guides.github.com/activities/hello-world .
Vanden Broucke, S., & Baesens, B. (2018). Practical web scraping for data science: best practices and examples with Python. Apress.