Obtención de una base de datos de perfiles de investigadores en Google Scholar basado en web scraping

Obtaining a database of researcher's profiles in Google Scholar based on web scraping

Claudio Isaac Soriano-Burgos[1], Juan Antonio Bautista[1], Misael López-Ramírez[1]
[1]Departamento de Estudios Mutidisciplinarios, División de Ingenierías, Campus Irapuato-Salamanca, Universidad de Guanajuato
ci.sorianoburgos@ugto.mx, ja.bautista@ugto.mx, lopez.misael@ugto.mx

Resumen

Los sistemas de información se han colocado como una solución a los problemas de manejo y gestión de datos en todas las áreas del conocimiento debido a su capacidad de captura, almacenamiento y procesamiento de los datos, con la finalidad de agilizar procesos administrativos o alcanzar conocimiento nuevo a partir del procesamiento de los datos. El uso de sistemas de información es cada vez más común, por lo que es necesario contar con sistemas que se adapten a las necesidades de los usuarios, por ejemplo, aportar a las actividades de vinculación e investigación de la comunidad científica. El sistema propuesto realizará la obtención automática de datos de publicaciones de investigadores adscritos a la Universidad de Guanajuato que cuenten con un perfil en Google Scholar, por medio de *web scraping*.

Palabras clave: *Web scraping*; google scholar.

Introducción

El uso de *web scraping* se ha presentado como una herramienta para realizar la recopilación de datos de varias páginas web sin la necesidad de abrir cada una de ellas de forma manual. Esta forma de recopilación es una oportunidad para realizar la extracción de datos en donde se requiera la consulta de varias páginas web. El mecanismo de *web scraping* (Figura 1) consiste en enviar solicitudes a un servidor. La respuesta de una solicitud exitosa se envía en forma de un documento en HTML y posteriormente, se analiza el documento buscando patrones en donde se encuentren las características que se quieran extraer y almacenar. En este proyecto se propone el uso de Google Scholar como fuente de generación de bases de datos, con la información de las publicaciones de investigadores como autores, resumen de artículos, revista y año de publicación, de manera automática por medio de *web scraping*.

Figura 1. Web scraping

Metodología Propuesta

El mecanismo de *web scraping* se puede separar en dos etapas: la navegación en las páginas de Internet y la extracción de datos de las páginas a las que se tuvo acceso. En la etapa de navegación, se dirigirá al enlace en Google Scholar y en la etapa de extracción de datos se identificarán patrones en la estructura HTML de las páginas web consultadas. Este proceso se realizará utilizando el lenguaje Python por medio de las librerías Requests, Scrapy y Beautiful Soup.

Figura 2. Obtención del set de la base de datos

Las primeras solicitudes al servidor (Figura 2) se realizarán para capturar los enlaces a cada uno de los perfiles de los investigadores adscritos a la Universidad de Guanajuato, almacenando los enlaces en un archivo CSV. Posteriormente, se utilizará el archivo generado para realizar una segunda etapa de solicitudes para consultar cada uno de los enlaces y extraer los datos de las publicaciones de todos los usuarios encontrados, almacenándolos un nuevo archivo CSV. Para la generación de la base de datos, se utilizarán los archivos resultantes.

Resultados y conclusiones

El primer conjunto de datos obtenido fue el de los enlaces de todos los investigadores adscritos a la Universidad de Guanajuato del sitio https://scholar.google.es/citations?view_op=view_org&org=7919973288022170384&hl=es&oi=io. Se realizaron 59 solicitudes get exitosas al servidor utilizando la librería Requests sin especificar un tiempo entre solicitudes. Como resultado de esta primera etapa, se almacenaron 587 enlaces correspondientes a los perfiles de los investigadores en donde se encuentran los enlaces a sus publicaciones.

En la segunda etapa de solicitudes, se utilizó Scrapy y Beautiful soup, debido a que, en algunos casos, no se pudo observar un patrón definido cuando faltaba algún dato. En esta etapa, se realizaron 5237 solicitudes al servidor, especificando un tiempo aleatorio entre 1 y 3 segundos entre consultas. Como resultado de esta segunda etapa de solicitudes (Figura 3), se generó un archivo CSV con los datos de año de publicación, título, autores, revista, resumen de la publicación y el enlace a la revista de publicación de cada uno de los investigadores.

V±o	TV▼tulo	Autores	Revista	DescripciV2n	l
2022	Metastable vacua from torsion and machine learning	Cesar Damian, Oscar Loaiza-Brito	arXiv prepi	By implementing an error function on a Machine Learning algorithm we look fo	h
2021	Strings: Geometry and Symmetries for Phenomenology	Nana Cabo-Bizet, Anamarƒ±a Font, Oscar Loaiza-Brito, Hans I	0	String phenomenology is the study of the possible connections between the ma	h
2021	Inflationary Implications of the Covariant Entropy Bound and the Swan	Dibya Chakraborty, Cesar Damian, Alberto Gonzv"lez Bernal, (	Universe	We present a proposal to relate the de Sitter conjecture (dSC) with the time de	h
2021	Quantum implications of the covariant entropy bound in string inflatio	Dibya Chakraborty, Cesar Damian, Alberto Gonzalez Bernala,	arXiv e-pri	We present a proposal to trace the De Sitter Conjecture back to quantum aspec	h
2020	Testing swampland conjectures with machine learning	Nana Cabo Bizet, Cesar Damian, Oscar Loaiza-Brito, Damiv"n	The Europ	We consider Type IIB compactifications on an isotropic torus	h
2020	Fat inflatons, large turns and the CΣ-problem	Dibya Chakraborty, Roberta Chiovoloni, Oscar Loaiza-Brito, G	Journal of	It is commonly believed that a successful period of inflation driven by a single o	h
2019	Leaving the Swampland: Non-geometric fluxes and the Distance Conje	Nana Cabo Bizet, Cesar Damian, Oscar Loaiza-Brito, Damian I	Journal of	We study a Type IIB isotropic toroidal compactification with non-geometric flux	h
2019	Some remarks on the dS conjecture, fluxes and K-theory in IIB toroidal	Cesar Damian, Oscar Loaiza-Brito	arXiv prepi	In this note we present a description on the implications on the refined dS Swa	h
2019	Two,Âêfield axion inflation and the swampland constraint in the flux,Â	Cesar Damian, Oscar Loaiza,ÂêBrito	Fortschritt		h
2017	Meromorphic flux compactification	Damian Cesar, Oscar Loaiza-Brito	Journal of	We present exact solutions of four-dimensional Einstein,Âõs equations related	h
2017	Meromorphic flux compactification	Cesar Damian, Oscar Loaiza-Brito	Journal of	We present exact solutions of four-dimensional Einstein,Âõs equations related	h
2016	Mirror quintic vacua: hierarchies and inflation	Nana Cabo Bizet, Oscar Loaiza-Brito, Ivonne Zavala	Journal of	We study the moduli space of type IIB string theory flux compactifications on th	h
2015	The Closed String Tachyon and its relationship with the evolution of th	Celia Escamilla-Rivera, G Garcia Jimenez, O Loaiza-Brito, O C	0	We present a cosmological landscape where the classical closed string tachyon	h
2014	Half-flat quantum hair	Hugo Garcia-Compean, Oscar Loaiza-Brito, Aldo Martinez-Me	Physical Rı	By wrapping D 3-branes over 3-cycles on a half-flat manifold, we construct an e	h
2013	More stable de Sitter vacua from S-dual nongeometric fluxes	Cesar Damian, Oscar Loaiza-Brito	Physical Rı	Stable vacua obtained from isotropic tori compactification might not be fully sti	h
2013	Exotic orientifolds in non-geometric flux cosmology	Cesar Damian, Oscar Loaiza-Brito	AIP Confer	We report on the existence of a stable de Sitter vacum in Type IIB non-geometr	h
2013	Slow-roll inflation in non-geometric flux compactification	Cesar Damian, Luis R Dv▼az-Barrv2n, Oscar Loaiza-Brito, Mig	Journal of	By implementing a genetic algorithm we search for stable vacua in Type IIB nor	h
2013	Closed string tachyon: inflation and cosmological collapse	Celia Escamilla-Rivera, Gerardo Garcv▼a-Jimv©nez, Oscar Lo	Classical a	Starting with a compactification of critical bosonic string theory on an internal s	h
2012	Towards a K-theory description of quantum hair	H Garcv▼a-Compev"n, O Loaiza-Brito	AIP Confer	The first steps towards a proposal for a description of the quantum hair in 4D st	h
2012	Robinson-Bertotti solution from flux compactification	O Loaiza-Brito, L Vazquez-Mercado	Journal of	We present a 10 dimensional supergravity compactification threaded with a sin	h

Figura 3. Datos de las publicaciones de un usuario obtenidos de Google Scholar por web scraping

Web scraping se ha presentado como una herramienta efectiva para la extracción de las publicaciones de artículos de los investigadores adscritos a la Universidad de Guanajuato que cuentan con un perfil en Google Scholar. De acuerdo con el archivo robots.txt, el sitio permite la extracción automática de los perfiles de usuario, sin embargo, la extracción automática permanece limitada a un determinado número de solicitudes al servidor. De las librerías utilizadas en Python, Requests reportó el mayor número de solicitudes con código 200 en la respuesta del servidor, mientras que Scrapy presentó la menor cantidad. En el proceso de identificación de patrones y extracción de características, Beautiful soup Como trabajo futuro, estos datos servirán para la implementación de un sistema generador de grupos de interés utilizando un algoritmo de aprendizaje no supervisado, implementándolo en una aplicación web que permita ayudar a la colaboración científica intra e interinstitucional en un campo disciplinar en específico.

Referencias

AlMarzouq, M., AlZaidan, A., & AlDallal, J. (2020). Mining GitHub for research and education: challenges and opportunities. International Journal of Web Information Systems, 16(4).

Henrys, K. (2021). Importance of Web Scraping in E-Commerce and E-Marketing. SSRN Electronic Journal, January.

Idris, A. Y. (2022). Web Scraping and Regression Analysis based on Machine Learning for COVID-19 with Rapid Software Platform. 3(1).

Pratiba, D., Abhay, M. S., Dua, A., Shanbhag, G. K., Bhandari, N., & Singh, U. (2018). Web Scraping and Data Acquisition Using Google Scholar. Proceedings 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions, CSITSS 2018.

Rahmatulloh, A., & Gunawan, R. (2020). Web Scraping with HTML DOM Method for Data Collection of Scientific Articles from Google Scholar. Indonesian Journal of Information Systems, 2(2).