“Inteligencia Artificial en la extracción de datos en la web para incrementar las ventas.”

“Artificial Intelligence in data extraction on the web to increase sales.”

Autor: Luis Fabrizzio Rios Ruiz,

Luisf.rios@edu.uag.mx

Co-autor: Zaira Ruth Zuviría López,

zzuviria@edu.uag.mx

“Inteligencia Artificial en la extracción de datos en la web para incrementar las ventas.”

“Artificial Intelligence in data extraction on the web to increase sales.”

Fecha: Agosto 2023
Date: August 2023
Universidad Autonoma de Guadalajara
Zapopan, Jalisco

Resumen

Los apartados de desarrollo web, e-commerce y ventas en línea es uno de los principales objetivos para uso de Inteligencias Artificiales. Una de las practicas más conocidas y que podrá ser mayormente beneficiada es el Web Scraping, que es básicamente el filtrado, manejo y consumo de datos de páginas web para obtener datos en crudo y poder procesarlos para toma de decisiones, automatización de procesos o testing de software. En esta investigación han utilizado algoritmos más robustos y adaptables para el web scraping. Estos algoritmos utilizan la clasificación de texto y modelos como redes neuronales convolucionales, redes neuronales residuales y LSTM para extraer datos de manera automatizada y precisa de sitios web en constante cambio. El resultado del análisis estudio destaca la importancia de la extracción de datos web en el contexto del comercio electrónico y sugiere que el sistema propuesto tiene el potencial de revolucionar el proceso de extracción de datos automatizado en el futuro.

Resumee

The sections of web development, e-commerce and online sales is one of the main objectives for the use of Artificial Intelligence. One of the best-known practices and one that can benefit the most is Web Scraping, which is basically the filtering, handling, and consumption of data from web pages to obtain raw data and be able to process it for decision-making, process automation, or test testing. software. In this research, more robust and adaptable algorithms have been used for web scraping. These algorithms use text classification and models such as convolutional neural networks, residual neural networks, and LSTM to automatically and accurately extract data from ever-changing websites. The result of the analysis highlights the importance of web data extraction in the context of e-commerce study and suggests that the proposed system has the potential to revolutionize the automated data extraction process in the future.

Palabras Clave

Inteligencia Artificial, Algoritmo, Internet, Comercio electronico, Procesamiento de datos, Automatizacion, Procesamiento de datos, Programacion informática, Lenguaje, Lenguaje de programación, Computadora

Keywords

Artificial Intelligence, Algorithms, Internet, E-commerce, Data processing, Automation, Data processing, Computer programming, Language, Computer language, Computer

Introducción

Los apartados de desarrollo web, e-commerce y ventas en línea es uno de los principales objetivos para uso de Inteligencias Artificiales. Una de las practicas más conocidas y que podrá ser mayormente beneficiada es el Web Scraping, que es básicamente el filtrado, manejo y consumo de datos de páginas web para obtener datos en crudo y poder procesarlos para toma de decisiones, automatización de procesos o testing de software.

Introduction

The sections of web development, e-commerce and online sales is one of the main objectives for the use of Artificial Intelligence. One of the best-known practices and one that can be most benefited is Web Scraping, which is basically the filtering, management, and consumption of data from web pages to obtain raw data and be able to process it for decision-making, process automation, or testing. software.

Metodologia

Para este estudio se realizaron múltiples investigaciones en el campo de Web Scraping, NLP, Machine Learning y composición de los documentos de HTML. Además de eso, se realizó una investigación a profundidad a partir de múltiples artículos en internet que tomaban o abarcaban este tema de forma parcial. Cabe mencionar que este tema o esta aplicación no tiene un gran campo de estudio, por lo que actualmente se encuentra en una etapa temprana, o, en otras palabras, es un tema nuevo que aún se encuentra en pañales y que evolucionara periódicamente.

Los temas por tratar fueron redactados tomando en cuenta las lecturas previas, se enviaron a revisión y, tras algunos ajustes se aplicó de forma directa por los autores de esta investigación. Los primeros datos a obtener corresponden a información general obtenida directamente de la investigación, que es en pocas palabras los algoritmos y modelos que se utilizaran a la hora de buscar una aplicación del concepto mencionado. Luego tenemos la etapa de aprendizaje o de generación del conocimiento, que es la etapa donde se aprende de los temas y como pueden relacionarse y aplicarse al mismo tiempo. Y la última que es la etapa de aplicación o conclusión, que es la etapa donde se obtiene una confirmación de en verdad este concepto puede ser aplicado de forma satisfactoria y con aplicaciones evidentes.

Methodology

For this study, multiple investigations were carried out in the field of Web Scraping, NLP, Machine Learning and composition of HTML documents. In addition to that, an in-depth investigation was carried out based on multiple articles on the internet that partially took or covered this topic. It is worth mentioning that this topic or this application does not have a large field of study, so it is currently in an early stage, or, in other words, it is a new topic that is still in its infancy and will evolve periodically.

The topics to be discussed were written taking into account the previous readings, they were sent for review and, after some adjustments, they were applied directly by the authors of this research. The first data to be obtained correspond to general information obtained directly from the investigation, which is, in a few words, the algorithms and models that will be used when looking for an application of the concept. Then we have the stage of learning or generation of knowledge, which is the stage where you learn about the topics and how they can be related and applied at the same time. And the last one is the application or conclusion stage, which is the stage where a confirmation is obtained that this concept can really be applied satisfactorily and with obvious applications.

Resultados

La mayoría de la información obtenida revela que los modelos de clasificación de Machine Learning y en específico el modelo de SVM (Support Vector Machine) pueden ser aplicados en un 96% de efectividad para poder hacer una extracción de datos de la web con Eb Scraping, esto en conjuntos con el filtrado de Lenguaje Natural y con un aprendizaje supervisado, dan pie a una aplicación que pueda adaptarse a todas las formas distintas de composición de los documentos HTML.

Como resultado del análisis de estudio y en base a todos los artículos visitados y referenciados, se destaca la importancia de la extracción de datos web en el contexto del comercio electrónico y sugiere que el sistema propuesto tiene el potencial de revolucionar el proceso de extracción de datos automatizado en el futuro.

Esto logra resaltar la importancia que tiene la IA en ámbitos de Desarrollo web y de compra y venta por internet, que desde hace unos años ha cobrado gran relevancia en el sector, por lo que esto podría dar pie a una nueva revolución y una nueva rama de la Inteligencia Artificial aplicada a algoritmos de extracción de datos de la web mediante modelos de clasificación.

Results

Most of the information obtained reveals that the Machine Learning classification models and specifically the SVM model (Support Vector Machine) can be applied with 96% effectiveness in order to extract data from the web with Eb Scraping, This, in conjunction with Natural Language filtering and supervised learning, gives rise to an application that can adapt to all the different forms of HTML document composition.

As a result of the study analysis and based on all the articles visited and referenced, the importance of web data extraction in the context of electronic commerce is highlighted and suggests that the proposed system has the potential to revolutionize the data extraction process. automated in the future.

This manages to highlight the importance of AI in the areas of web development and internet buying and selling, which for a few years has gained great relevance in the sector, so this could give rise to a new revolution and a new branch of Artificial Intelligence applied to data extraction algorithms from the web through classification models.

Conclusiones

En esta investigación se han analizado algoritmos robustos y adaptables que podrían usarse para darle al algoritmo de Web Scraping una funcionalidad más compleja, útil y automatizada. Estos algoritmos utilizan la clasificación de texto y modelos como SVM, modelos de clasificación y Análisis del Lenguaje Natural, para darle al algoritmo la capacidad de decidir y analizar múltiples documentos HTML en constante cambio. En conclusión, la capacidad de las Inteligencias Artificiales al pasar los años ha ido en aumento al igual que su uso. El ámbito de las compras y ventas se ha visto evolucionado por el uso de estas y con el pasar del tiempo, el uso de algoritmos para extracción y análisis de datos en la Web se verá en aumento, y abrirá un nuevo panorama para el comercio electrónico.

Conclusions

In this research we have analyzed robust and adaptable algorithms that could be used to give the Web Scraping algorithm a more complex, useful, and automated functionality. These algorithms use text classification and models such as SVM, classification models, and Natural Language Analysis, to give the algorithm the ability to decide and analyze multiple, constantly changing HTML documents. In conclusion, the capacity of Artificial Intelligence over the years has been increasing as well as its use. The field of purchases and sales has evolved due to the use of these and over time, the use of algorithms for data extraction and analysis on the Web will increase, and will open a new panorama for electronic commerce. .

Referencias bibliográficas/ Bibliographical references

Carle V. Web Scraping using Machine Learning [Internet]. DIVA. 2020. Available from: https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1468583&dswid=4980

Khder MA. Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application [Internet]. 2021 [cited 2023 Aug 15]. Available from: http://ijasca.zuj.edu.jo/PapersUploaded/2021.3.11.pdf

Rahman RU, Tomar DS. Threats of price scraping on e-commerce websites: attack model and its detection using neural network. [Internet]. 2020 [cited 2023 Aug 15]. Available from: https://link.springer.com/article/10.1007/s11416-020-00368-6#auth-Rizwan_Ur-Rahman

Gunawan R, Rahmatulloh A, Darmawan I, Firdaus F. Comparison of Web Scraping Techniques: Regular Expression, HTML DOM and Xpath. [Internet]. 2019 [cited 2023 Aug 15]. Available from: https://www.atlantis-press.com/proceedings/icoiese-18/55914830

Patnaik SK, Babu NC. Trends in web data extraction using machine learning. [Internet]. 2021 [cited 2023 Aug 15]. Available from: https://content.iospress.com/articles/web-intelligence/web210465

Salgado AF. A web scraping framework for stock price modelling using deep learning methods. [Internet]. 2019 [cited 2023 Aug 15]. Available from: https://upcommons.upc.edu/handle/2117/178104

Idris AY, Bamoallem R, Mohamad Hatta MHA. Web scraping and regression analysis based on machine learning for COVID-19 with rapid software platform. [Internet]. 2022 [cited 2023 Aug 15]. Available from: https://www.researchgate.net/publication/361367966_Web_Scraping_and_Regression_Analysis_based_on_Machine_Learning_for_COVID-19_with_Rapid_Software_Platform

Prehanto DR, Indriyanti AD, Prismana IGLE, Permadi GS, Prastyo EHA. Implementation of Web Scraping on News Sites Using the Supervised Learning Method. [Internet]. 2021 [cited 2023 Aug 15]. Available from: https://web.s.ebscohost.com/abstract?site=ehost&scope=site&jrnl=13053515&AN=150153249&h=QpgJz6IK%2bzGn1uhqFKa1W9VQRunx1qqz97eCevQLP9825whrDIXO%2f81Bp0zagr0izXaUbwBDfKN6CiXIOFExpw%3d%3d&crl=c&resultLocal=ErrCrlNoResults&resultNs=Ehost&crlhashurl=login.aspx%3fdirect%3dtrue%26profile%3dehost%26scope%3dsite%26authtype%3dcrawler%26jrnl%3d13053515%26AN%3d150153249

Selvy PT, Anitha M, Vishnu Varthan LR, Sethupathi P, Adharsh SP. Intelligent Web Data Extraction System for E-commerce. [Internet]. 2022 [cited 2023 Aug 15]. Available from: https://www.publishoa.com/index.php/journal/article/view/545

Sanchez de La Fuente CF. Machine and Deep Learning models for house price prediction in United States of America and Portugal. [Internet]. 2022 [cited 2023 Aug 15]. Available from: https://repositorio.iscte-iul.pt/handle/10071/27005

Egger R, Kroner M, Stöckl A. Web Scraping. In: Egger R, editor. Applied Data Science in Tourism. Tourism on the Verge. Springer, Cham; 2022. p. Chapter 5. Available from: https://doi.org/10.1007/978-3-030-88389-8_5

Coronado-Guerrero AE, Mares-Barrientos BJ, Hernandez-Varela J, Mora-Herrera JE, Pantoja-Guitierrez M, Lopez-Chernyshov PE, Díaz-Pacheco A. Recolección y análisis de datos mediante técnicas de IA aplicadas al sector turístico. JÓVENES EN LA CIENCIA. 2023;21:1–10. Available from: https://www.jovenesenlaciencia.ugto.mx/index.php/jovenesenlaciencia/article/view/4117

Chauhan VK, Dahiya K, Sharma A. Problem formulations and solvers in linear SVM: a review. Artif Intell Rev. 2019;52:803–855. https://doi.org/10.1007/s10462-018-9614-6

Wang Z-q, Sun X, Zhang D-x, Li X. An Optimal SVM-Based Text Classification Algorithm. In: Proceedings of the 2006 International Conference on Machine Learning and Cybernetics; 2006. Dalian, China. p. 1378-1381. DOI:10.1109/ICMLC.2006.258708.

Semblanza del autor

Mi nombre es Luis Fabrizzio Rios Ruiz, tengo 20 años y estudio la carrera de Ing. En Software en la Universidad Autónoma de Guadalajara (UAG). Soy una persona estoica, de mente abierta, respetuosa, profesional, entregada, en constante búsqueda de conocimiento y superación. Tengo especial interés en proyectos de automatizacion de redes, de desarrollo de back-end, front-end,desarrollo móvil, web scraping, desarrollo de API's, desarrollo de algoritmos de Inteligencia Artificial,. Tambien estoy interesado en aprender y certificarme sobre Arquitecturas de Software, Certificaciones de Redes de Cisco y Servicios en la Nube. Como objetivo de vida, me gustaría ser arquitecto de software.

Portrait of the author

My name is Luis Fabrizzio Rios Ruiz, I am 20 years old, I am studying Software Engineering at the Universidad Autonoma de Guadalajara (UAG). I am a stoic person, open-minded, respectful, professional, devoted, in constant search for knowledge and self-improvement. I have special interest in network automation projects, back-end development, front-end, mobile development, web scraping, API's development, development of Artificial Intelligence algorithms. I am also interested in learning and getting certified on Software Architectures, Cisco Network Certifications and Cloud Services. As a life goal, I would like to be a software architect.