

**APRENDIENDO A EDUCAR: información para la mejora**

Dirección de Innovación Educativa-Coordinación del Modelo Educativo

Vol. I, No.1, 20 de Febrero de 2020,

Como parte del Telescopio Educativo, se reseñan documentos, que se consideran relevantes y pertinentes para mejorar el trabajo educativo.

Las reseñas se

1. Título: **Web Scraping using Machine Learning**
2. **Autor**: Victor Carle
3. **Año**: 2020. **País**: Suecia. **Publicación**: Universidad Tecnológica Real de Estocolmo (KTH)
4. **Palabras clave**: web scraping, aprendizaje automático, clasificación de texto, máquina de vectores de soporte, HTML, preprocesamiento de datos y patrones.
5. **Línea de investigación educativa**:
6. **Objeto estudio**: Inteligencia Artificial
7. **Fundamentos teóricos**:

Esta tesis se basa en dos áreas principales de estudio: Web Scraping y aprendizaje automático. En cuanto al web scraping, se discuten los conceptos básicos de HTML y cómo se puede utilizar para extraer datos de un sitio web. También se aborda el problema de la variabilidad en el código HTML y cómo esto puede afectar la capacidad de un raspador web para recuperar datos precisos y estructurados. En cuanto al aprendizaje automático, se discuten los conceptos básicos de la clasificación de texto y el uso de máquinas de vectores de soporte (SVM) para entrenar modelos que puedan reconocer patrones en el código HTML. También se aborda el preprocesamiento de datos, que es una técnica utilizada para limpiar y transformar los datos antes del entrenamiento del modelo. Además, se discuten otros conceptos teóricos relevantes, como la selección del conjunto de datos, la evaluación del modelo y las limitaciones del aprendizaje automático en general. En resumen, los fundamentos teóricos están diseñados para proporcionar una comprensión sólida tanto del web scraping como del aprendizaje automático, lo que permite al autor desarrollar un algoritmo robusto que pueda superar las limitaciones comunes asociadas con estas técnicas.

1. **Metodología:**

Esta tesis se divide en tres fases principales: recopilación de datos, preprocesamiento de datos y entrenamiento del modelo. En la primera fase, se recopilan datos de sitios web seleccionados utilizando técnicas de web scraping. Se utilizan herramientas como Beautiful Soup y Selenium para extraer datos no estructurados del sitio web y convertirlos en un formato estructurado. En la segunda fase, se realiza el preprocesamiento de los datos. Esto implica limpiar los datos y transformarlos en un formato que pueda ser utilizado por el modelo de aprendizaje automático. Se utilizan técnicas como la eliminación de palabras vacías, la lematización y la tokenización para preparar los datos para el entrenamiento del modelo. En la tercera fase, se entrena el modelo utilizando máquinas de vectores de soporte (SVM). Se utiliza una técnica llamada clasificación binaria para entrenar el modelo para reconocer patrones específicos en el código HTML. El modelo se evalúa utilizando técnicas como validación cruzada y matriz de confusión para determinar su precisión y capacidad para generalizar a nuevos conjuntos de datos.

1. **Resultados principales**:

Los resultados de esta tesis indican que el algoritmo desarrollado es capaz de recuperar datos precisos y estructurados de sitios web seleccionados, incluso si el código HTML subyacente cambia. Se logró una precisión del 90% en la clasificación binaria utilizando máquinas de vectores de soporte (SVM). Además, se encontró que el preprocesamiento de datos es una parte crítica del proceso y puede afectar significativamente la precisión del modelo. La eliminación de palabras vacías y la lematización fueron técnicas particularmente efectivas para mejorar la precisión del modelo. También se encontró que la selección cuidadosa del conjunto de datos es importante para garantizar que el modelo sea capaz de generalizar a nuevos conjuntos de datos. La evaluación del modelo utilizando técnicas como validación cruzada y matriz de confusión fue útil para determinar su capacidad para generalizar a nuevos conjuntos de datos.

1. **Contacto:**

Victor Carle - [**eva.agius@ub.uu.se**](mailto:eva.agius@ub.uu.se)

1. **Comentarios**:

* Considero que esta tesis es muy útil como base para investigación sobre la aplicación de técnicas de Web Scraping con Inteligencia Artificial.
* Puede ser enfocado a temas de e-commerce.
* Existe poca información sobre web scraping con inteligencia artificial por lo que esta tesis puede ser la base de una nueva rama de la Inteligencia Artificial.

1. **Elaboró**: Luis Fabrizzio Rios Ruiz

Referencia:

**Carle, V. (2020). *Web Scraping using Machine Learning*. DIVA.** <https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1468583&dswid=4980>