

Seminario de investigación – 20 de noviembre de 2024

Introducción al web scraping y análisis de sentimiento para el tratamiento de datos estructurados turísticos

Seminario 054. Departamental. Campus de Vicálvaro (Madrid)
15:30 – 19:30

Impartido por: Sofía Blanco Moreno, Universidad de León

Organizado por: Laura Fuentes Moraleda, Titular de Universidad, URJC

Antecedentes y justificación

El *web scraping* es una técnica invaluable para recabar datos de manera eficiente y sistemática en línea, permitiendo a los investigadores acceder a una amplia gama de información relevante para su trabajo. Al dominar esta técnica, los investigadores pueden optimizar sus procesos de recolección de datos, obtener información de manera oportuna y precisa, y avanzar en su investigación de manera más efectiva, contribuyendo así al desarrollo del conocimiento en el ámbito del desarrollo social sostenible.

El *web scraping* se ha convertido en una herramienta crucial tanto en la investigación académica como en las empresas de investigación de mercados, permitiendo la extracción de grandes cantidades de datos de sitios web y plataformas *online*. Las empresas están aprovechando el *web scraping* para recopilar datos de plataformas como TripAdvisor, Booking y Google Reviews para obtener información más profunda sobre el comportamiento, las preferencias y las tendencias del mercado de los consumidores. Es necesario que los investigadores tengan la capacidad de acceder a grandes conjuntos de datos directamente desde fuentes en línea. La creciente importancia de la toma de decisiones basada en datos hace que la capacidad de recopilar y procesar eficientemente dichos datos sea una habilidad esencial tanto para los investigadores como para los profesionales.

Una vez recopilados los datos, el siguiente paso implica la limpieza y el procesamiento, transformando los datos sin procesar en formatos utilizables. Este paso es esencial para garantizar la calidad y la precisión de los datos, tanto en la investigación como en las aplicaciones industriales. Con la cada vez mayor disponibilidad de datos no estructurados de fuentes *online*, comprender cómo procesar adecuadamente estos conjuntos de datos es vital para extraer información significativa. Además, esta preparación sienta las bases para análisis más avanzados, como el análisis de sentimientos y el modelado de temas, donde herramientas como MeaningCloud y Azure pueden evaluar automáticamente el tono emocional y los temas clave en los datos. Estas herramientas ofrecen versiones gratuitas y de pago, lo que las hace accesibles a investigadores y empresas de todos los tamaños.

El análisis de sentimientos es cada vez más reconocido por sus aplicaciones tanto en la investigación como en los negocios. Desde la comprensión de las opiniones de los clientes hasta el análisis de conversaciones en las redes sociales a gran escala, el análisis de sentimientos ofrece información fundamental sobre la percepción pública. Tanto los investigadores como las empresas están utilizando esta técnica para dar sentido a grandes cantidades de datos textuales no estructurados. Este seminario presenta una oportunidad para dotar a los estudiantes e investigadores de las habilidades no solo para recopilar datos, y también para aplicar técnicas de análisis avanzadas. A través de este seminario, los participantes adquirirán experiencia práctica en el manejo de conjuntos de datos del mundo real y la realización de un análisis de sentimientos integral para extraer información procesable, lo que lo hace muy relevante y oportuno en el mundo actual impulsado por los datos.

El seminario lo impartirá Sofía Blanco Moreno, profesora experta en Marketing y Turismo con especialización en Marketing Digital, y tecnologías como *web scraping* e inteligencia artificial (*machine learning* y *deep learning*) aplicadas en Turismo y Hotelería. Es profesora en la Universidad de León, tras haber desarrollado su carrera profesional como experta en Marketing en empresas internacionales como Meliá Hoteles y Grupo Oblanca. Ha recibido varios premios de investigación, como el de mejor comunicación en los congresos AIRSI y AIM 2023, y los premios de transferencia de conocimiento I UAM-ASSECO Business Case Award 2022 y Prueba de Concepto dentro del Plan de Transferencia de Conocimiento Universidad-Empresa para la plataforma Photo Data Tour Analytics (<https://mktingphotodatatour.unileon.es/>).

Entre sus publicaciones más recientes destacan:

- What do you do or with whom? Understanding happiness with the tourism experience: an AI approach applied to Instagram in Journal of Humanities and Social Sciences Communications.
- Big data in tourism marketing: past research and future opportunities in Spanish Journal of Marketing – ESIC.
- Women's happiness and brand content marketing in Journal of Management Decision.
- Social media impact of tourism managers: a decision tree approach in happiness, social marketing and sustainability in Journal of Management Development.
- Unconscious and Conscious Aspects of Healthy Food Consumption: A Neuromarketing and Artificial Intelligence Approach in Global Challenges for a Sustainable Society.
- Comparison and positioning of NGOs aimed at children from the perspective of social marketing on Twitter in Journal of Humanities and Social Sciences Communications.
- Social marketing and happiness in employment. Evidences from Glassdoor in BMC Psychology.

Objetivos del seminario

Explicación en el aula por un lado del manejo de herramientas de *web scraping* para la descarga de datos web, y por otro, del análisis de sentimiento de *reviews* y de *topics* a los datos obtenidos de internet. Se desarrollará un caso práctico en el que los alumnos deberán descargar datos de sitios web, y posteriormente se limpiará la base de datos y se estructurará. Posteriormente se analizarán los datos descargados de sitios webs con diferentes herramientas libres.

Resultados esperados y *networking* para los miembros del Centro

- Manejo de herramientas de *web scraping* para la descarga de datos.
- Estructuración y limpieza de los datos para su tratamiento posterior.
- Manejo de herramientas de análisis de sentimiento y de *topics*.
- Análisis estadístico de los datos cuantificados.

Programa de trabajo

- Explicación sobre qué es *web scraping*, cómo se está usando en investigación y en las empresas, y principales herramientas libres y de pago. Uso de dos herramientas de *web scraping* para descarga de datos de diferentes plataformas, por ejemplo, TripAdvisor, Booking, Civitatis, Google Reviews, etc. (1h).
- Limpieza y tratamiento de las bases de datos descargadas (1h).
- Explicación sobre qué es el análisis de sentimiento, tipos y aplicaciones. Cómo se está usando en investigación y en las empresas, y principales herramientas libres y de pago. Uso de dos herramientas de *web scraping* de análisis de sentimiento y de *topics*, como MeaningCloud, Azure o NoCode (1h).
- Análisis estadístico de los resultados del análisis de sentimiento (1h).