

ISPC-TSIT 4.0-2023- PROYECTO INTEGRADOR

Informe de Proyecto: Análisis de Reseñas de Hoteles en Córdoba a través de Web Scraping en TripAdvisor

El presente informe detalla el proceso integral de nuestro proyecto de web scraping en TripAdvisor, orientado al análisis cualitativo de reseñas de hoteles en Córdoba, Argentina. Desde la definición de objetivos hasta la aplicación de metodologías ágiles y consideraciones éticas, este proyecto fue ejecutado con precisión y ética profesional.

Definición de Objetivos y Alcances

El proyecto se inició con la clara definición de objetivos. Establecimos metas específicas, desde la recopilación de datos hasta su análisis, con un enfoque en contribuir al desarrollo de estrategias de marketing y adquirir habilidades de investigación y programación.

Metodologías Ágiles y Trabajo en Pareja (Pair)

La elección de metodologías ágiles y la implementación del trabajo en pareja fueron elementos clave para el éxito y eficiencia de nuestro proyecto de web scraping en TripAdvisor. Este enfoque se diseñó con la flexibilidad y la adaptabilidad como principios fundamentales, permitiendo una respuesta efectiva a los desafíos en constante evolución que surgieron durante el desarrollo del proyecto.

1. Iteraciones Rápidas y Adaptabilidad:

- La metodología ágil se centró en la entrega iterativa de resultados, lo que nos permitió ajustar y mejorar continuamente nuestro enfoque a medida que surgían nuevos desafíos o se refinaban los objetivos. Las iteraciones rápidas fueron esenciales para mantenernos ágiles y receptivos a los cambios en los requisitos del proyecto.

2. Colaboración Constante y Comunicación Efectiva:

- El trabajo en pareja fomentó una colaboración constante entre los miembros del equipo. La comunicación fluida y directa facilitó la resolución rápida de problemas y la toma de decisiones conjunta. La retroalimentación inmediata entre los compañeros mejoró la calidad del código y aceleró el proceso de desarrollo.

3. División de Tareas y Especialización:

- La asignación de tareas específicas a cada miembro del equipo durante el trabajo en pareja permitió una mayor especialización. Cada miembro pudo enfocarse en áreas de su experiencia, maximizando así la eficiencia y la calidad del trabajo realizado. Esta división de tareas también facilitó la resolución de desafíos técnicos específicos.

4. Trello como Herramienta de Gestión:

- Trello se convirtió en una pieza central de nuestra gestión de proyectos. La herramienta proporcionó una estructura organizativa visual que permitió un seguimiento claro del progreso, la asignación de tareas y la identificación de posibles cuellos de botella. La interfaz intuitiva de Trello contribuyó a un flujo de trabajo ordenado y eficiente.

5. Flexibilidad para Abordar Desafíos Técnicos:

- Los desafíos técnicos inherentes al web scraping, como la detección de bots y cambios en la estructura de la página, exigieron un enfoque flexible. La metodología ágil y el trabajo en pareja nos proporcionaron la flexibilidad necesaria para adaptarnos a estos desafíos de manera proactiva y eficiente.

6. Mejora Continua y Lecciones Aprendidas:

- La metodología ágil fomentó una cultura de mejora continua. Después de cada iteración, revisamos y analizamos los resultados obtenidos, identificando áreas de mejora y ajustando nuestra estrategia en consecuencia. Esta mentalidad de aprendizaje continuo contribuyó al desarrollo académico del equipo.

En resumen, la combinación de metodologías ágiles y trabajo en pareja no solo fue instrumental para la ejecución exitosa del proyecto, sino que también sentó las bases para prácticas de desarrollo ágil y colaborativo en futuros proyectos. La agilidad, la comunicación efectiva y la adaptabilidad fueron elementos clave que contribuyeron al logro de nuestros objetivos de manera eficiente y ética.

Desarrollo y Desafíos Técnicos

Durante el desarrollo del proyecto, nos enfrentamos a diversos desafíos técnicos que requerían soluciones creativas y adaptativas. A continuación, se detallan algunos de los principales desafíos y las estrategias utilizadas para superarlos:

1. Detección de Bots:

- ☐ Desafío: TripAdvisor implementa medidas de seguridad para detectar bots, lo que podría obstaculizar el proceso de scraping.
- ☐ Ejemplo: Al intentar acceder a múltiples páginas de reseñas, nos encontramos con bloqueos automáticos.
- ☐ Solución: Se implementó el uso del encabezado "User-Agent" para simular un navegador real. Este enfoque evitó la detección como bot, permitiendo la extracción continua de datos.

2. Estructura de la Página Cambiante:

- ☐ Desafío: TripAdvisor puede modificar la estructura de su sitio web, afectando nuestras selecciones y búsquedas de elementos.
- ☐ Ejemplo: Cambios en la clasificación de elementos HTML que estábamos extrayendo.
- ☐ Solución: Implementamos técnicas de scraping dinámico, ajustando nuestro script para adaptarse a cambios en la estructura. Esto incluyó la identificación flexible de elementos mediante clases y etiquetas específicas.

3. Paginación y Límites:

- ☐ Desafío: TripAdvisor puede limitar la cantidad de resultados por página y las páginas accesibles, restringiendo la cantidad de datos extraíbles.
- ☐ Ejemplo: Se limitaba la cantidad de reseñas mostradas en una sola página.
- ☐ Solución: Se implementó la paginación y bucles para recopilar resultados de múltiples páginas. Al no funcionar esto, se optó por navegar una a una las páginas para la extracción.

4. Duplicación de Datos:

- ☐ Desafío: Algunas páginas web pueden mostrar resultados duplicados, lo que lleva a la duplicación de datos en la extracción.
- ☐ Ejemplo: Repetición de reseñas en diferentes secciones de la página.
- ☐ Solución: Se incorporaron lógicas específicas para eliminar duplicados durante el proceso de extracción y almacenamiento de datos, garantizando la integridad y precisión de la información recopilada.

5. Calificaciones de Viajeros:

- ☐ Desafío: Obtener calificaciones de viajeros puede ser desafiante debido a la falta de una estructura numérica clara.
- ☐ Ejemplo: Las calificaciones estaban representadas visualmente y no de forma numérica en la página.
- ☐ Solución: Implementamos un mapeo de las etiquetas de calificación a valores numéricos, permitiendo una interpretación cuantitativa de las opiniones de los usuarios.

Consideraciones Adicionales sobre el Uso de Selenium:

Dado que en el equipo tuvimos un compañero que por cuestiones técnicas se le imposibilitó, luego de iniciado el proyecto e instaladas las librerías; dado una actualización en su sistema, se optó porque el utilizara Selenium para:

- Verificar si la página permite la extracción de ciertos datos antes de que el equipo cree un script eficiente en BeautifulSoup y Requests.
- Se utilizó como un explorador controlado para evaluar la estructura de la página y confirmar la viabilidad de la extracción.

Estos desafíos y soluciones demuestran la naturaleza dinámica y compleja del web scraping, así como la necesidad de enfoques flexibles y adaptativos para garantizar el éxito del proyecto. La combinación de herramientas como Selenium y estrategias específicas nos permitió superar obstáculos y obtener datos valiosos de manera ética y eficiente.

Consideraciones Éticas y Legales

El cumplimiento de regulaciones éticas y legales fue una prioridad. Documentamos exhaustivamente el proceso, asegurando la transparencia y el respeto por los términos de servicio de TripAdvisor. La atención a la privacidad y la integridad de los datos fue fundamental para garantizar un enfoque ético en todo momento.

Análisis de Datos y Hallazgos

El análisis de datos proporcionó una visión detallada de la experiencia del usuario en hoteles cordobeses. Identificamos aspectos positivos y áreas de mejora, generando recomendaciones específicas. La categorización y evaluación cuantitativa y cualitativa nos permitieron identificar patrones y tendencias relevantes.

Análisis Detallado de Datos y Resultados:

Durante el proceso de extracción y análisis de datos en TripAdvisor, se realizó una revisión minuciosa de las opiniones de los usuarios, proporcionando insights valiosos sobre las experiencias de los huéspedes en diversos hoteles de la región. A continuación, se resumen los hallazgos clave:

Aspectos Positivos Destacados:

Atención Excepcional: Hoteles como Azur Real Hotel Boutique & SPA y Amérian Executive Córdoba Hotel fueron consistentemente elogiados por ofrecer una atención excepcional a sus huéspedes.

Calidad de Servicios: Muchos establecimientos recibieron elogios por la calidad de sus servicios, ubicación estratégica y excelentes instalaciones.

Áreas de Mejora Identificadas:

Desafíos con la Conectividad WiFi: Se identificaron problemas de conectividad WiFi en varios hoteles, incluyendo Windsor Hotel y NH Urbano, lo que podría afectar la satisfacción del huésped.

Comentarios Negativos sobre Ruido y Limpieza: Algunos lugares recibieron críticas negativas relacionadas con el ruido y la limpieza, aspectos cruciales para la experiencia del huésped.

Recomendaciones:

Mejora de Servicios de WiFi: Se sugiere a los hoteles con problemas de conectividad abordar y mejorar sus servicios de WiFi para garantizar una experiencia sin contratiempos para los huéspedes.

Medidas Correctivas para Ruido y Limpieza: Aquellos con comentarios negativos sobre el ruido y la limpieza podrían considerar la implementación de medidas correctivas, mejorando así la experiencia general del huésped.

Este análisis detallado no solo destaca los aspectos positivos y áreas de mejora en la oferta hotelera de Córdoba, sino que también proporciona recomendaciones específicas para abordar los desafíos identificados, contribuyendo así a la mejora continua de la calidad de los servicios en la industria hotelera local.

Lecciones Aprendidas y Desarrollo Académico

Este proyecto no solo se trató de la extracción de datos, sino también del crecimiento académico. Aprendimos a aplicar el conocimiento deontológico, desarrollamos habilidades de programación y exploramos técnicas avanzadas de evaluación de mercado. Las lecciones sobre flexibilidad metodológica y adaptabilidad fueron invaluable.

Conclusiones y Recomendaciones

Este proyecto no solo fue un ejercicio técnico en web scraping, sino también un esfuerzo integral para comprender la percepción del usuario y contribuir al desarrollo de estrategias de marketing informadas. La combinación de habilidades técnicas, ética profesional y análisis de datos meticuloso destaca la capacidad del equipo para abordar proyectos complejos de manera efectiva. El informe final sirve como testimonio de nuestro compromiso con la excelencia en la investigación y el respeto por los principios éticos en la ciencia de datos.

Desde una perspectiva ética, el proyecto se llevó a cabo con integridad y respeto. Se recomienda mantener la actualización constante del script para enfrentar posibles cambios en el sitio web y ampliar el alcance del estudio a otras ciudades para obtener una visión más completa del turismo en la región.

Reflexiones Finales y Continuidad del Proyecto

Este proyecto no solo logró sus objetivos técnicos, sino que también destacó la importancia de la ética en la ciencia de datos y el web scraping. La combinación de habilidades técnicas, metodologías ágiles y consideraciones éticas sentó las bases para futuros proyectos de investigación y análisis de datos.

El análisis de datos no solo cumple con los objetivos de la investigación, sino que también proporciona una base sólida para futuros estudios y mejoras en la oferta turística de Córdoba. La continuidad del proyecto podría incluir una expansión geográfica para obtener una visión más completa de la región y la actualización periódica del análisis para reflejar cambios en las preferencias y expectativas de los huéspedes.