

Especialización en Ciencia de Datos

Natalia Boretto Ramírez

IT ACADEMY



"Extracción y Análisis de Información de Hoteles en la Costa Dorada utilizando Web Scraping en Booking.com"

Resumen

El objetivo principal del trabajo final de la especialización de Data Science fue crear un dataset utilizando técnicas de web scraping con la herramienta Selenium para la página web de Booking.com. La búsqueda se centró en alojamientos para dos adultos y dos niños en la Costa Dorada de España durante la primera semana de julio de 2023.

El proyecto se desarrolló utilizando Python y las librerías Selenium para recopilar información de diferentes alojamientos disponibles en la página web. Se recopilaron datos como el nombre del alojamiento, la ubicación, las tarifas, las comodidades y las valoraciones de los clientes.

Posteriormente, se llevó a cabo un proceso de limpieza y

preprocesamiento de los datos para asegurar la calidad del dataset y se realizó un análisis exploratorio de los mismos para identificar patrones y tendencias relevantes. Se aplicaron modelos de regresión para predecir el precio de los alojamientos en función las características y valoraciones.

Con el objetivo de implementar casi todas las herramientas estudiadas en el Bootcamp, se realizó web scraping con BeautifulSoup en la web de Booking según la búsqueda principal, en uno de los alojamientos de la muestra, obteniendo así los comentarios de los clientes. A esta información se la incluyó en un nuevo dataset y se le realizaron los procedimientos de Análisis de Sentimientos.

En resumen, este trabajo final no solo demostró la utilidad del web scraping para la creación de datasets, su aplicación en proyectos de análisis y modelado de datos, sino que también ilustró cómo aplicar técnicas de modelado predictivo, como la regresión, para obtener información

valiosa para la toma de decisiones en la industria turística.

Introducción

El turismo es una industria clave en todo el mundo, y España es un destino turístico muy popular en Europa. La Costa Dorada, situada en la provincia de Tarragona, es conocida por sus impresionantes playas y su clima mediterráneo. En este proyecto, se presenta un dataset creado a través del web scraping con Selenium en la página web de Booking.com, con el objetivo de buscar alojamientos disponibles en la Costa Dorada para una familia de dos adultos y dos niños durante una semana en julio de 2023.

El web scraping es una técnica automatizada que permite recopilar datos de sitios web, y en este caso, Booking.com se convierte en una fuente de datos valiosa para obtener información sobre alojamientos y servicios turísticos en todo el mundo. La creación de un dataset a partir del web scraping de Booking permite obtener información detallada sobre los alojamientos, servicios, precios y disponibilidad en diferentes destinos turísticos, lo que puede ser muy útil para diversas aplicaciones, como la

planificación de viajes, la investigación de mercado y el análisis de la industria turística.

Además, al utilizar herramientas de análisis de datos, es posible extraer información relevante y valiosa, como tendencias de precios, popularidad de destinos turísticos y preferencias de los usuarios. Todo esto puede ser muy útil para tomar decisiones informadas y estratégicas en el sector turístico o en cualquier otra industria relacionada. En resumen, la creación de un dataset a partir del web scraping de Booking.com es una tarea interesante y valiosa para obtener información relevante y útil sobre la industria turística y otros campos relacionados, lo que puede ayudar a tomar decisiones informadas y estratégicas en el futuro.

Antecedentes

En los últimos años, el uso de técnicas de web scraping se ha vuelto cada vez más común en diferentes campos, incluyendo la industria turística. Con el aumento de la disponibilidad de datos en línea, se ha vuelto más fácil y eficiente recopilar información valiosa y relevante para tomar decisiones informadas en este sector.

En particular, el uso de web scraping en la industria turística permite obtener información sobre los precios, la disponibilidad de alojamiento y la popularidad de diferentes destinos turísticos. Esto puede ser muy útil para la toma de decisiones, como la planificación de viajes, la investigación de mercado y la elaboración de estrategias comerciales.

Además, la aplicación de modelos de análisis de datos en los datos recopilados a través del web scraping puede proporcionar información aún más valiosa, como patrones y tendencias en el comportamiento del consumidor y la demanda de servicios turísticos.

Por lo tanto, la creación de un dataset a partir del web scraping de Booking.com para obtener información relevante y valiosa sobre la industria turística es una tarea importante y útil para diferentes aplicaciones.

Metodología

1-El día 23 de marzo de 2023 se llevó a cabo el Web scraping utilizando Selenium para automatizar el proceso de recolección de datos de la página web Booking.com, para una búsqueda de alojamientos para dos

adultos y dos niños en la Costa Dorada de España durante una semana en el mes de julio de 2023.

2-Creación del dataset: Los datos recolectados con Selenium son almacenados en un archivo CSV para su posterior análisis. El dataset creado consta de 1078 registros, con información sobre el nombre del alojamiento, precio de la estadia, ubicación, tipo de habitaciones, puntaje de los clientes, comodidades.

3-Limpieza del dataset.

4-Análisis exploratorio de datos y visualizaciones: Se utiliza la librería Pandas para analizar los datos recolectados y entender mejor sus características. En esta etapa, se identificaron patrones y tendencias importantes en los datos.

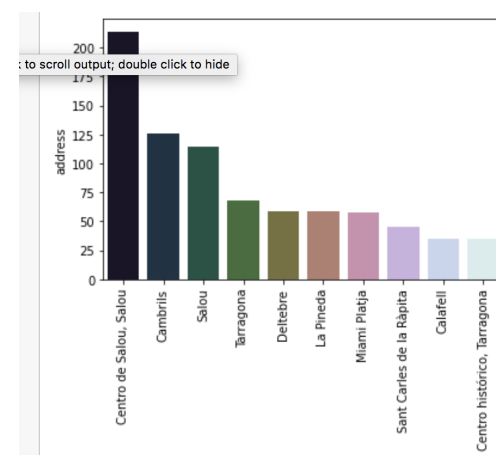


Figura: Las 10 ciudades con más oferta de alojamientos.

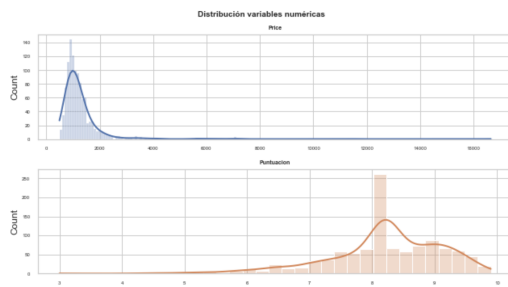


Figura: Gráfico de distribución para cada variable numérica.

Se utilizaron diferentes librerías como Matplotlib, Seaborn y Folium para crear visualizaciones que permitan entender mejor los datos recolectados. Geolocalización de los alojamientos con Folium:

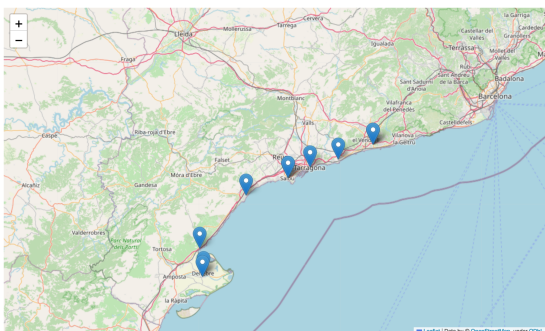


Figura: Top 10 de los alojamientos mejor puntuados por los clientes y su visualización en el mapa.

5-Preprocesado: Se realizan técnicas de preprocesamiento de datos.

6-División entre Features y Target. Se dividieron los datos disponibles en un conjunto de entrenamiento y un conjunto de test, aplicando un 80%-20%.

7-Modelos de regresión: Se utilizaron diferentes modelos de regresión de la librería Scikit-learn, como Regresión lineal, Random Forest y Support

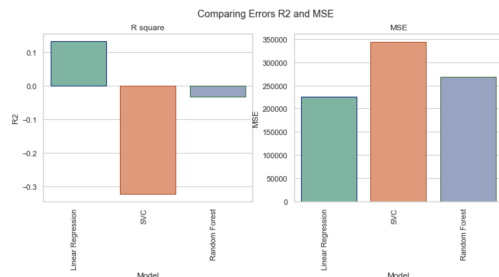
Vector Machine, para predecir variables de interés basadas en los datos recolectados. La regresión lineal múltiple predice una variable continua usando una ecuación lineal. El bosque aleatorio construye varios árboles de decisión y los combina para mejorar la precisión de la predicción. Las SVM buscan el hiperplano que mejor separa los datos en diferentes clases.

7.1-Evaluación y mejores parámetros: Se evaluaron los modelos de regresión utilizando diferentes métricas, R2 Score y Mean Squared Error (MSE). Además, se buscan los mejores parámetros para cada modelo para mejorar su precisión.

También se aplicó cross-validation que es una técnica utilizada para evaluar el rendimiento de un modelo de aprendizaje automático y prevenir el sobreajuste.

8-Resultados y conclusiones: después de aplicar diferentes modelos de regresión, ninguno de ellos generó buenas predicciones. Esto sugiere que el problema de modelización es complejo y puede requerir la inclusión de más variables, o la exploración de diferentes técnicas de modelización para mejorar la precisión de las predicciones.

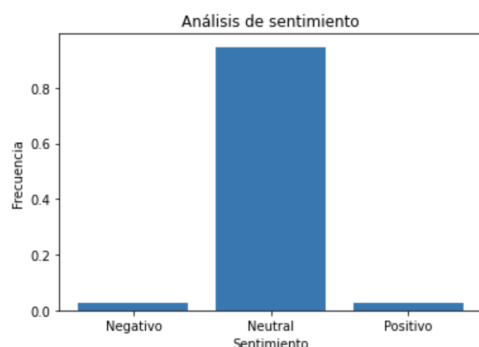
	Model	R2	MSE
0	Linear Regression	0.132540	225479.899949
1	SVC	-0.323194	343939.386792
2	Random Forest	-0.032532	268387.197191



9- Por último se realizó el día 7 de abril de 2023 Web scraping con BeautifulSoup de los comentarios de un alojamiento de la muestra analizada (ya que algunos alojamientos cuentan con más de 1.500 comentarios de usuarios), con el objetivo de realizar Minería de texto y Análisis de Sentimientos, para determinar si los comentarios son positivos, negativos o neutros.

{'neg': 0.025, 'neu': 0.947, 'pos': 0.028, 'compound': 0.15'}

Figura: Resultado de Análisis de sentimiento.



Conclusión

En conclusión, el proceso de obtención de datos mediante web scraping con Selenium, seguido de la limpieza y análisis de los mismos, ha resultado en una valiosa fuente de información para este trabajo final. La capacidad de extraer datos de la web de Booking.com de manera automatizada y eficiente ha permitido obtener una gran cantidad de datos que han sido sometidos a un proceso de limpieza y preparación para su análisis.

Durante el análisis, se han identificado patrones y tendencias en los datos, lo que ha proporcionado una mayor comprensión del fenómeno estudiado. Este análisis también ha permitido la generación de gráficos y visualizaciones que facilitan la interpretación de los resultados y la comunicación de hallazgos clave.

De la aplicación de los algoritmos de los modelos elegidos se ha encontrado que estos no han sido los más óptimos para el conjunto de datos en cuestión.

Es importante considerar la posibilidad de que algunas variables pueden no tener un efecto lineal en la variable de interés y, por lo tanto, se deban incluir variables más complejas o usar

técnicas de transformación de datos para identificar mejor estas. Se debe tener en cuenta que los modelos de regresión lineal no siempre son los más adecuados para todas las situaciones y se deben explorar otras alternativas para lograr resultados más precisos y efectivos.

Por último, tras realizar el análisis de sentimiento sobre los comentarios de los clientes del alojamiento analizado, podemos concluir que, en general, el tono emocional del texto es neutro, aunque se puede observar una leve presencia de sentimientos tanto positivos como negativos. Esta variabilidad puede deberse a la naturaleza subjetiva de las opiniones de los clientes, las cuales se ven influenciadas por sus experiencias personales y preferencias individuales.

Referencias

<https://www.analyticslane.com/>

<https://www.freecodecamp.org/>

<https://www.datacamp.com/tutorial/text-analytics-beginners-nltk>

<https://machinelearningmastery.com/hyperparameter-optimization-with-random-search-and-grid-search/>

<https://towardsdatascience.com/hypothesis-testing-explained-as-simply-as-possible-6e0a256293cf>

<https://machinelearningmastery.com/statistical-hypothesis-tests-in-python-cheat-sheet>

Enlaces a los archivos de

Jupyter Notebook:

[Web scraping -creación del Dataset](#)

[Análisis Exploratorio- Modelos](#)

[Análisis de sentimientos](#)