Memoria del proyecto

El presente documento refleja los puntos principales de la realización de la práctica 1 de web scraping de la asignatura *Tipología y ciclo de vida de los datos* del Máster en Ciencia de Datos de la Universitat Oberta de Catalunya, la cual tiene como objetivo general la recolección de datos de una página web.

1. Contexto

El contexto de esta práctica se basa en el estudio y el análisis del estado de la situación actual del alquiler, ya que, como se ha vivido en los últimos años existe una gran demanda ligado a su correspondiente aumento de precios. Para esto se ha trabajado con la web Fotocasa a través de la herramienta Selenium y centrándonos exclusivamente en Madrid.

A pesar de centrarnos solamente en la web Fotocasa todo el código desarrollado se ha enfocado para extrapolar la metodología aplicada a otras web con similar contexto.

2. Título

A partir del contexto anterior se ha establecido el nombre ***Madrid Rent Prices*** para el conjunto de datos generado.

3. Descripción del dataset

Como se puede intuir, cada elemento del conjunto representa una casa/piso de alquiler en Fotocasa y sus correspondientes atributos los cuales se describirán en las siguientes secciones. Estos elementos se han obtenido en diversos días durante el mes de abril de este año (2022) para tener una variabilidad de los cambios ocurridos en determinados periodos de tiempo.

4. Representación gráfica

5. Contenido

El dataset se encuentra formado por las siguientes características:

- ***zipCode***:
- ***buildingSubtype***:
- ***buildingType***:
- ***clientAlias***:
- ***clientId***:
- ***clientTypeId***:
- ***dateDiff***:
- ***dateUnit***:
- ***dateOriginalDiff***:
- ***dateOriginalUnit***:
- ***dateOriginalTimestamp***:
- ***description***:
- ***_id***:
- ***isDiscarded***:
- ***isHighlighted***:
- ***isPackAdvancePriority***:
- ***isPackBasicPriority***:
- ***isPackMinimalPriority***:
- ***isPackPremiumPriority***:
- ***isMsAdvance***:

- ***isNew***:
- ***isNewConstruction***:
- ***hasOpenHouse***:
- ***isOpportunity***:
- ***isTrackedPhone***:
- ***isTop***:
- ***minPrice***:
- ***multimedia***:
- ***otherFeaturesCount***:
- ***periodicityId***:
- ***price***:
- ***promotionId***:
- ***promotionUrl***:
- ***promotionTitle***:
- ***promotionTypologiesCounter***:
- ***realEstateAdId***:
- ***reducedPrice***:
- ***subtypeId***:
- ***transactionTypeId***:
- ***typeId***:

6. Agradecimientos

Específicamente el conjunto de datos obtenido se ha generado, como se comento en secciones anteriores, de la plataforma de venta y alquiler de viviendas Fotocasa. Gracias al enfoque de servicios que presenta nos ha permitido definir diversas características para obtener todos los datos mediante técnicas de scraping y conseguir obtener el conjunto de datos aquí presente. Siguiendo los modelos que presentan este tipo de plataformas para anonimizar su localización dentro de los datos recolectados se han obviado algunos como podrían ser cualquier aspecto con el que se pueda obtener la dirección real del establecimiento. Todo con el fin de que esto no pueda generar ningún problema al proceso de alquiler o al actual propietario. A parte de este tipo de información el resto de atributos se consideran dentro de los principios éticos y legales del contexto del proyecto.

7. Inspiración

El fin de trabajar con estos datos es por el potencial que presentan. De similar forma existen otros conjuntos de datos los cuales tienen unas similares características pero están enfocados en una mayor medida a el precio de venta de determinadas viviendas, podemos ver un ejemplo en el conjunto de datos *House Prices - Advanced Regression Techniques* de Kaggle. En este, como su título indica se busca aplicar técnicas de regresión para predecir los precios de una vivienda según sus características.

En nuestro caso, no sólo se busca poder responder a estas cuestiones, si no que se planean otras como podrían ser la predicción del tiempo que una vivienda estará en alquiler (por esto el motivo de registrar los datos existentes en diferentes fechas). Con esto se podría optimizar el tiempo de publicación de los anuncios, destacando los parámetros que influyen en mayor medida para que un alojamiento sea alquilado en un determinado periodo de tiempo.

8. Licencia

Dentro de las licencias establecidas para su publicación se le ha asignado la **CC BY-SA 4.0 License**. El motivo de aplicar este tipo de licencia es que permite el uso de la obra, incluyendo en esto su modificación pero destacando que la autoría original tiene que estar presente definiendo así cualquier cambio o transformación realizada en el conjunto de datos. Además, por el potencial que observamos en los datos para su uso comercial permitimos así su uso en este contexto.

9. Código

El código para generar el correspondiente conjunto de datos se ha realizado mediante el lenguaje de programación Python. Este se encuentra disponible en el siguiente repositorio de Github. En el README principal de dicho repositorio se describen detalladamente las carpetas y códigos existentes que lo forman, además de las diversas formas de ejecución desarrolladas y otras notas importantes al trabajar con este proyecto.

10. Dataset

Finalmente, el resultado final del dataset se puede encontrar en Zenodo a través del siguiente [enlace del DOI]().

11. Contribuciones

En este apartado se reflejan las contribuciones realizadas por cada uno de los autores en las diversas tareas realizadas. Las iniciales en la sección de firma representan la confirmación por parte del autor de su participación en el apartado correspondiente.

Contribución	Firma
Investigación previa	
Redacción de las respuestas	
Desarrollo del código:	
- User-agent aleatorio	
- Pruebas con proxys	
- Descarga de imágenes	
- Obtención de características	
Dockerización de la aplicación	
Base de datos MongoDB Atlas	