**Práctica 1 (25% nota final)**

**INTEGRANTES: Gerard Alcalde y Guillem Rochina**

**NOVIEMBRE 2022**

1. **Contexto.** Explicar en qué contexto se ha recolectado la información. Explicar por qué el sitio web elegido proporciona dicha información. Indicar la dirección del sitio web.

En el presente proyecto, nuestro equipo ha convenido llevar a cabo un proyecto de scraping del metabuscador de reservas de alojamiento conocido como **Booking.com.**

Dicha plataforma proporciona información y permite hacer reservas en su plataforma en función de los criterios/preferencias que el usuario introduce en su buscador, siendo el número de reseñas positivas, la proximidad al destino indicado por el usuario y la disponibilidad en las fechas determinadas algunos de los criterios que más influencia tienen en la posición del ranking (explicado en el apartado de inspiración) entre los resultados de la búsqueda. Es decir, un alojamiento con más reseñas positivas, mejor nota y más próximo al lugar indicado por el usuario ocupará una posición en las primeras páginas del buscador.

No obstante, no solo dichos factores influyen en la posición que ocupan cada oferta de alojamiento. ¿Qué elementos son relevantes? ¿Existen algunas características que pueden ser mejoradas para escalar posiciones en los rankings? Estas son preguntas que todo manager de hotel se ha realizado alguna vez en su carrera, y a su vez nos ofrece un contexto sobre el que trabajar en el presente proyecto.

En concreto, la dirección del sitio web donde nuestro web scraper inicia su proceso es <https://www.booking.com>

1. **Título.** Definir un título que sea descriptivo para el dataset.

Ofertas de alojamiento vacacional en las principales ciudades españolas: Principales indicadores y características.

1. **Descripción del dataset.** Desarrollar una descripción breve del conjunto de datos que se ha extraído. Es necesario que esta descripción tenga sentido con el título elegido.

Tal y como se observa en el título, el dataset presenta los datos e indicadores más relevantes de cada uno de los hoteles encontrados en función de determinados criterios de búsqueda. En este dataset en concreto presentamos datos para distintas fechas, número de adultos, niños y habitaciones a fin de tener una muestra más amplia e informativa que la que nos daría una búsqueda con tan sólo unos parámetros fijos.

Dado que el enunciado de la práctica así lo indica, se ha optado por no realizar una limpieza del dataset resultante. Por tanto, encontramos una serie considerable de elementos que dan pie a un proyecto de limpieza bastante elaborado, por ejemplo, nos encontramos con algunas columnas conformadas por listas o diccionarios que a su vez presentan valores dentro de estos, así como datos faltantes en algunos registros del dataset que pueden ser objeto de inputación o eliminación en caso de ser oportuno.

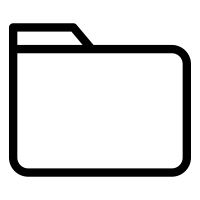
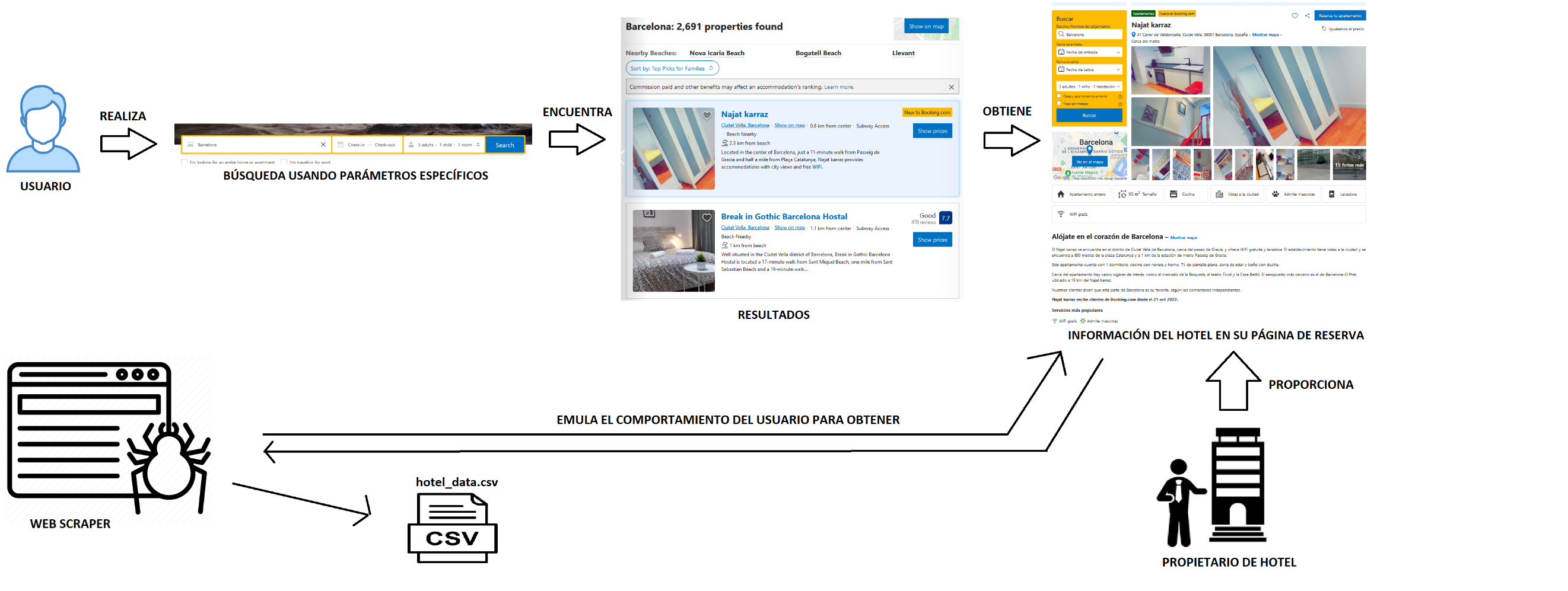
Finalmente, cabe destacar que el formato escogido para el fichero resultantes es un csv, pues facilita su tratamiento y compartición.

**RECUPERA**

1. **Representación gráfica.** Dibujar un esquema o diagrama que identifique el dataset visualmente y el proyecto elegido

**Figura 1: Representación gráfica del proyecto escogido**

**LA INFORMACIÓN SE DESTINA A UN CSV Y LAS IMÁGENES A UN DIRECTORIO**



**IMG**

**Hotel\_images**

Graphical user interface, text

Description automatically generated**Figura 2: Representación gráfica del dataset resulta**

Text

Description automatically generated

1. **Contenido.** Explicar los campos que incluye el dataset y el periodo de tiempo de los datos.

A continuación, se realiza una breve descripción de cada una de las columnas:

|  |  |  |
| --- | --- | --- |
| Columna | Tipo | Descripción |
| Name | **Str** | Nombre del hotel |
| City | **Str** | Nombre de la ciudad donde se realiza la búsqueda (parámetro de búsqueda inputado) |
| Check-in | **Str** | Fecha de entrada al hotel (parámetro de búsqueda inputado) |
| Check-out | **Str** | Fecha de salida del hotel (parámetro de búsqueda inputado) |
| Adults | **Int** | Número de adultos para los que se realiza la reserva (parámetro de búsqueda inputado) |
| Children | **Int** | Número de niños para los que se realiza la reserva (parámetro de búsqueda inputado) |
| Num\_rooms | **Int** | Número de habitaciones reservadas (parámetro de búsqueda inputado) |
| Address | **Str** | Dirección postal del hotel, incluyendo calle, número, zona de la ciudad, código postal, ciudad y país |
| Hotel\_coordinates | **Str** | Coordenadas del hotel formadas por longitud y latitud separadas por una coma |
| Hotel\_score | **Int** | Nota general que recibe el hotel por los usuarios |
| Hotel\_scores | **Dict** | Diccionario que recoge la puntuación del hotel para distintas categorías. Las categorías incluidas son: Staff, Facilities, Cleanliness, Comfort, Value for money, Location, Free WiFi. |
| Hotel\_description | **List** | Descripción aportada por el propietario del hotel. Cada elemento de la lista es una línea de la descripción en formato string |
| Features | **List** | Lista de las distintas características del hotel |
| Room\_data | **Dict** | Diccionario para cada una de las distintas habitaciones disponibles en el hotel los siguientes atributos:   * + room\_name (string): Nombre de la habitación.   + room\_price (string): Precio de la habitación.   + room\_capacity (string): Texto con la capacidad de la habitación indicando número de adultos y de niños.   + room\_options (list of strings): Lista de opciones del hotel asociadas a la habitación (cancelación gratuita, tipo de pensión, etc)   + room\_facilities (list of strings): Lista de características de la habitación.   + room\_bed\_type (string): Tipo de camas en la habitación. |
| Page\_count | **Int** | Posición en la que ha aparecido el hotel al buscar en booking |
| Current\_page | **Int** | Página en la que ha aparecido el hotel al buscar en booking. |
| In\_Page\_count | **Int** | Número de la posición en la que se encuentra el hotel dentro de la página de resultados |
| Search\_date | **Str** | Fecha en la que se realizó la búsqueda |

El objetivo de este dataset es obtener las características que hacen que un hotel esté mejor posicionado que otro. El posicionamiento de los hoteles suele ser algo estático y por ello con una sola lectura de la página web puede ser suficiente. No obstante, se podría realizar un scrapping con diferentes fechas que nos permitiese obtener hoteles que puedan no aparecer en la búsqueda debido a que no tengan disponibilidad en las fechas seleccionadas.

El dataset abarca los búsquedas para los meses de Diciembre, Marzo y Junio de la temporada 2022-2023 para añadir un componente temporal al análisis posterior de los datos.

1. **Propietario.** Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares. Justificar qué pasos se han seguido para actuar de acuerdo a los principios éticos y legales en el contexto del proyecto.

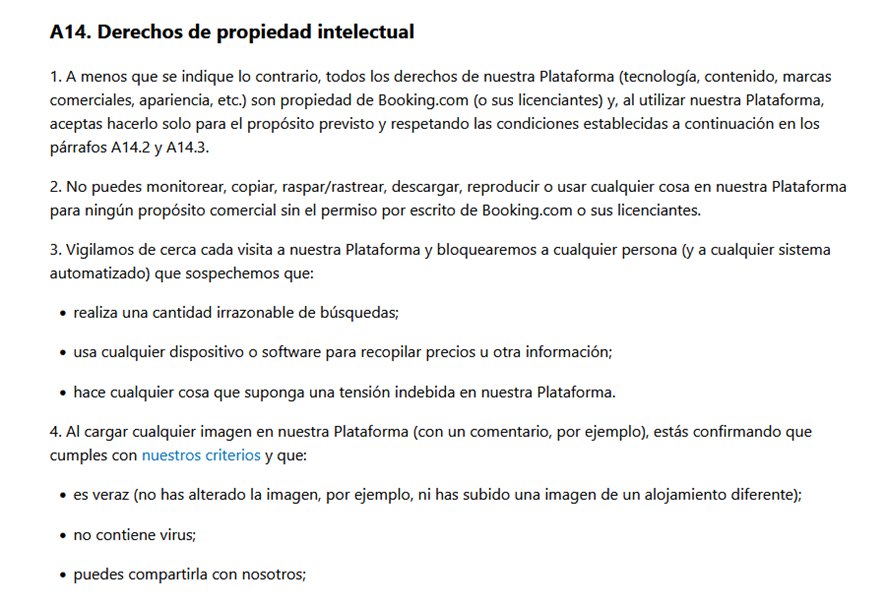
Como se ha presentado en el apartado de contexto, los datos han sido recopilados de la web de Booking.com estableciendo en su motor de búsqueda distintos parámetros, por lo que el propietario del conjunto de datos es **Booking Holdings**.

Como metabuscador, esta plataforma se encarga desde hace más de 20 años de agregar tarifas de habitaciones/apartamentos en un mismo sitio web a fin de ofrecer facilidades a los individuos en su búsqueda por un sitio donde alojarse durante sus viajes. Actualmente, miles de viajeros ocasionales o aquellos que viajan por negocios hacen uso de esta plataforma para reservar hoteles, hostales o incluso barcos, por lo que esta web supone una herramienta muy potente tanto para huéspedes como para las cadenas hoteles o propietarios que deciden hacer uso de sus propiedades como alojamientos vacacionales.

Scrapear la web de Booking.com no es una tarea que no se haya hecho nunca. De hecho, existen muchas páginas que ofrecen sus servicios de scraping para esta pagina como por ejemplo ***Bright Data*** *(*<https://cutt.ly/IMDMFbx>*).* Asimismo, existen muchos posts donde se ofrecen recomendaciones sobre como mejorar, como en la web de ***Hostaway*** (<https://www.hostaway.com/how-to-rank-1-on-booking-com/>). Sin embargo, no hemos sido capaces de encontrar ningún proyecto que junte las dos vertientes, scrapear los datos y analizarlos a fin de obtener recomendaciones sobre como posicionarse mejor en Booking.com. Por tanto, esta búsqueda se justifica como la herramienta que poner fin al “gap” entre estas dos actividades (el scrapeo web de Booking y las recomendaciones de páginas web especializadas en posicionamiento web de hoteles).

Finalmente, con el fin de seguir con los principios éticos y legales previamente se investigó el archivo: <https://www.booking.com/robots.txt> y la página de términos y condiciones de Booking.com: <https://www.booking.com/content/terms.es.html>, cuyo contenido de interés en nuestro caso se encuentra en el apartado A14.2 (véase Figura 3). En este se indica que no se puede scrapear ningún dato con fines comerciales sin permiso escrito. No obstante, dado que nuestro proyecto tan solo persigue fines académicos y no comerciales, consideramos que legalmente estamos cubiertos en este aspecto.

Asimismo, el apartado A14.3 indica la posibilidad de bloqueo a cualquier usuario que muestre indicios de hacer web scraping (usar software para recopilar precios o realizar una cantidad irrazonable de búsquedas). A fin de cubrirnos en este aspecto, se han incluido varios retardos a lo largo del código para no sobrecargar la página y se ha diseñado el proyecto de forma que las búsquedas que realizar el programa sean lo más fidedignas posibles a como actuaria un usuario real, además de cambiar el header del user agent.

Figura 3: Términos y condiciones de Booking.com

1. **Inspiración.** Explicar por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.

Aparecer en los primeros resultados de una búsqueda de un usuario por internet se ha convertido en los últimos años en una de las preocupaciones principales de cualquier negocio que busca hacerse un hueco en el mundo digital. Este fenómeno explica la proliferación de herramientas como el SEO y el SEM, mediante las cuales se intenta innovar (más allá del pago de tarifas para aparecer en las primeras páginas de búsqueda) a fin de lograr un mejor “posicionamiento” en buscadores como Google o Bing. El negocio hotelero no es una excepción en este caso.

No obstante, el uso de dichas herramientas no resulta útil si nos enfrentamos a un buscador dentro de una página web, como el caso que nos ocupa con Booking.com. ¿Qué hacer entonces? ¿Cómo mejorar el posicionamiento dentro de la web?

En la web del metabuscador de alojamientos encontramos que define a dichas posiciones entre los resultados como “ranking”, descrito por ellos mismos como el orden en el que se muestran los alojamientos disponibles en los resultados de búsqueda. En su blog “Booking.com Partner Hub” indican que los resultados son ordenados según la relevancia en base a las preferencias particulares de cada “cliente”, incluyendo en este caso también las “dinámicas del mercado”, el “rendimiento del alojamiento”, etc. Variable que poca información revela al interesado en que sus hoteles/alojamientos aparezcan en posiciones más “atractivas” para este.

En concreto, en el mismo post analizado se destaca la relevancia de revisar regularmente las “condiciones flexibles”, los comentarios, los precios externos, las promociones, la puntuación de la página del alojamiento… a fin de lograr escalar puestos en la clasificación.

Dado el impacto que puede tener en el negocio de un hotel el aparecer en posiciones más inmediatas del buscador, el presente proyecto pretende ser un paso inicial (el de recopilación de datos) que finalice realizando un análisis inferencial que destaque que factores tienen un mayor peso en el posicionamiento de un alojamiento en **Booking.com.**

No obstante, la herramienta desarrollada puede tener múltiples funcionalidades más allá de la presentada. A saber, puede ser utilizada por un usuario que busque el mejor momento en términos económicos para realizar un viaje a una ciudad determinada (si lo cruza con datos relacionados con el precio de vuelos, viajes en barco o el coste de cualquier otro modo de transporte) o, por otro lado, un hotel puede utilizarla para hacer llevar a cabo una comparación de sus servicios con respecto a la competencia (en este caso, dejamos a consideración del propio usuario determinar si los datos están siendo o no utilizados con fines comerciales, pues, como se ha visto en el apartado 6 está prohibido por las políticas de Booking.com.)

1. **Licencia.** Seleccionar una licencia adecuada para el dataset resultante y justificar el motivo de su elección.

A la hora de escoger una licencia para el dataset resultante se ha optado por una licencia ampliamente conocida como es la Creative Commons, ya que esta permitirá conocer los requisitos establecidos de una forma sencilla.

Creative Commons recoge distintos tipos de licencia basada en distintos parámetros como son la atribución, el uso a realizar de los datos o como se deben tratar estos. En este sentido se ha optado por una licencia: **Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)**, esta licencia estipula:

* Se pueden compartir los datos en cualquier medio o formato
* Se permite modificar los datos o trabajar sobre ellos, siempre y cuando se indique que esta modificación se ha llevado a cabo.
* Se debe indicar quien es el autor de los datos.
* Si se modifican o se trabaja sobre los datos, estos se deberán distribuir bajo la misma licencia.

Se ha optado por esta licencia ya que permite que cualquiera pueda acceder al dataset que hemos generado y utilizarlo con la finalidad deseada, pero con el compromiso de compartir sus modificaciones de forma pública y gratuita promoviendo así que aumente el contenido de calidad y gratuito en la red.

1. **Código.**

El código desarrollado en esta práctica se encuentra en el enlace del Github: <https://github.com/guiruha/WebScrapingUOC>

Concretamente, el código, desarrollado principalmente en Python, se encuentra en la carpeta ***src*** tal como indican las instrucciones de práctica, agrupados en dos ficheros:

* **Main.py:** script que se encarga de ejecutar el proceso diseñado en la clase ***BookingSpider.*** El código dentro de dicho script es bastante simple, pues se encarga de importar la clase, instanciarla y ejecutar las funciones principales.
* **BookingScraper.**py: script que concentra todo el desarrollo e implementación de la clase ***BookingSpider,*** el cual se encarga de realizar las búsquedas dentro del portal de Booking.com con los parámetros introducidos al ejecutar el script previamente descrito, recorrer toda la lista de resultados, obtener la información de la página de reserva de cada hotel encontrado y almacenarlo todo en un fichero csv y una carpeta donde guardar las imágenes.

Como añadido, se ha incluido un scrip en lenguaje ***bash, sampleDataCollection.sh****,* para automatizar el proceso de ejecución del ***Main.py*** con varios parámetros de búsqueda distintos. Dicho script activa el environment de conda (a fin de evitar problemas con los paquetes) y ejecuta ***Main.py*** con distintos parámetros.

1. **Dataset.** Publicar el dataset obtenido en formato CSV en Zenodo, incluyendo una breve descripción. Obtener y adjuntar el enlace del DOI del dataset (https://doi.org/…). El dataset también deberá incluirse en la carpeta **/dataset** del repositorio.

El ejemplo de dataset resultante del proceso previamente descrito se encuentra de nuevo el repositorio de Github nombrado. Se incluye tanto el csv obtenido como varias carpetas de ejemplo con imágenes de algunos hoteles scrapeados.

1. **Vídeo**. Realizar un breve vídeo explicativo de la práctica **(máximo 10 minutos)**, que deberá contar con la participación de los dos integrantes del grupo. En el vídeo se deberá realizar una presentación del proyecto, destacando los puntos más relevantes, tanto de las respuestas a los apartados como del código utilizado para extraer los datos. Indicar el enlace del vídeo (https://drive.google.com/…), que deberá ubicarse en el Google Drive de la UOC.

|  |  |
| --- | --- |
| **Contribuciones** | **Firma** |
| Investigación previa | Gerard Alcalde, Guillem Rochina |
| Redacción de las respuestas | Gerard Alcalde, Guillem Rochina |
| Desarrollo del código | Gerard Alcalde, Guillem Rochina |
| Participación en el vídeo | Gerard Alcalde, Guillem Rochina |