Tipología y ciclo de vida de los datos.

Práctica 1

Curso 2018/19 - Primavera

NOMBRES: Daniel Mato Regueira e Iago Veiras Lens

AULA:1

CONTEXTO

Desde hace unos años, BlaBlaCar se ha postulado como una de las alternativas de transporte más populares para viajes de corta y media distancia. El funcionamiento de la web es sencillo. Por un lado, conductores particulares publican los viajes que van a realizar en las próximas fechas, especificando las características de este (origen, destino, paradas, duración, precio, etc.). Por otro lado, los viajeros buscan plazas libres de estos viajes y las reservan a través de la propia web de BlaBlaCar, ejerciendo esta última de intermediaria y aseguradora del contrato entre ambas partes.

Para obtener los datos de los viajes publicados en BlaBlaCar, prepararemos un crawler que navegará por la página de búsqueda de BlaBlaCar con unos parámetros definidos y los extraerá directamente, ya que la información más importante y general de los viajes está disponible en esa página.

TÍTULO DEL DATASET

El título del dataset es ***blablacar_extraccion_viajes.csv***.

DESCRIPCIÓN DEL DATASET

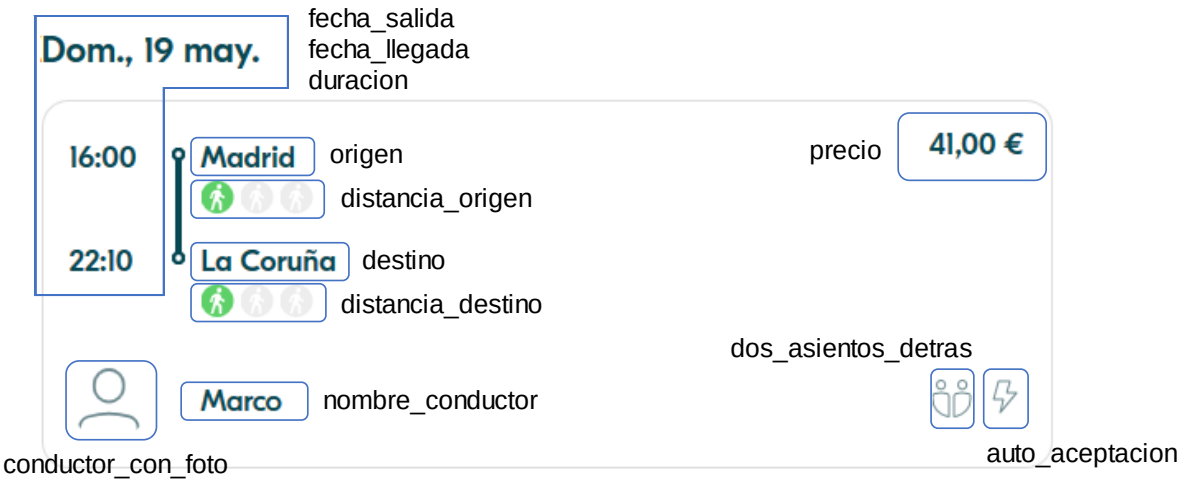
La extracción realizada de BlaBlaCar mediante técnicas de *web scraping* ha tenido los siguientes parámetros de entrada:

- **Plazas disponibles** = 1
- **Localidad de salida** = Madrid
- **Fecha de salida** = 17/04/2019
- **Localidad de llegada** = Cualquiera
- **Límite de páginas escaneadas** = 200

Por lo tanto, podemos decir que el contenido del dataset actual contiene hasta 2000 viajes (como máximo) con salida desde Madrid (o una localidad cercana) el 17/04/2019, con al menos una plaza disponible y con destino cualquier localidad. Cabe destacar que estos viajes eran los que cumplían estas condiciones el día de la ejecución de la consulta (15/04/2019).

REPRESENTACIÓN GRÁFICA

Para representar gráficamente el conjunto de datos, nos valdremos de la visualización que se puede encontrar en la web que descargamos. En el siguiente gráfico se muestra una tarjeta de viaje de la página de BlaBlaCar tras una búsqueda como la simulada con el código adjunto. En ella está indicadas el origen de cada una de las variables que se encuentran en el dataset, a excepción de "*nombre*" y "*url*", que se obtienen de las características de la propia tarjeta del viaje.



CONTENIDO

En este dataset se encuentran las siguientes 14 variables:

- **nombre**: nombre del viaje en BlaBlaCar. Forma parte de la última parte del campo *url*.
- **url**: dirección web del viaje en BlaBlaCar.
- **nombre_conductor**: nombre de pila del conductor del viaje.
- **origen**: nombre de la localidad de partida del viaje. En este dataset está fijada en Madrid y localidades cercanas.
- **fecha_salida**: fecha y hora de salida del viaje desde la localidad origen.
- **destino**: nombre de la localidad de destino del viaje.
- **fecha_llegada**: fecha y hora de llegada del viaje a la localidad destino.
- **duracion**: duración del viaje, calculada a partir de los campos *fecha_salida* y *fecha_llegada*.
- **precio**: precio total del viaje (en euros).

- **distancia_origen**: distancia de la localidad de salida del viaje respecto al punto fijado en la búsqueda. Admite tres categorías diferentes en función de la distancia en km de las dos localidades.
- **distancia_destino**: distancia de la localidad de salida del viaje respecto al punto fijado en la búsqueda, aunque en este caso, al no haber especificado destino, no está informada. Admite tres categorías diferentes en función de la distancia en km de las dos localidades.
- **auto_aceptacion**: característica del viaje que permite realizar la reserva de este sin tener que esperar a la confirmación del conductor.
- **dos_asientos_atras**: característica del viaje que indica si el conductor utilizará únicamente dos plazas en los asientos traseros, dejando la tercera libre.
- **conductor_con_foto**: característica del conductor del viaje que indica si tiene cargada una foto en BlaBlaCar.

El contenido del dataset ha sido extraído de la página web de BlaBlaCar, realizando una búsqueda según los criterios del apartado anterior mediante técnicas de *web scraping* con el código indicado en el repositorio de GitHub. La validez de los datos es bastante efímera, ya que suele haber bastante variabilidad en la oferta disponible en BlaBlaCar de un día para otro.

AGRADECIMIENTOS

El propietario de este conjunto de datos es la propia empresa **BlaBlaCar**, ya que los conductores que publican sus viajes en la web ceden sus datos a la empresa en cuestión para poder utilizar sus servicios.

Aunque no nos hemos basado en ningún estudio anterior desarrollado sobre la información de esta página, sí hemos utilizado como base la documentación disponible de la API pública que tiene BlaBlaCar en su web (https://dev.blablacar.com/docs/versions/1.0/resources/trips) para saber qué tipo de parámetros podíamos pasar a la consulta de búsqueda de viajes.

INSPIRACIÓN

Debido a la cantidad y variedad de oferta de la web, estudiar y analizar los viajes publicados en cada momento en BlaBlaCar puede servir de interés a las dos partes implicadas.

Desde el punto de vista del conductor, este análisis le puede permitir ajustar los detalles de su viaje para maximizar las posibilidades de reserva de sus plazas, y por lo tanto de minimizar los gastos de su viaje.

Si nos ponemos en la piel de los viajeros, tener esta visión global de todo lo que está ofertado en la web le puede permitir escoger la mejor opción para su viaje (e incluso adaptar sus viajes a lo que está ofertado en la página).

LICENCIA

Para el dataset obtenido hemos escogido la licencia abierta de bases de datos o **ODbL**. Esta es una licencia que permite a los usuarios compartir datos con libertad y sin temor a la infracción de los derechos de autor. Las razones por las cuales hemos escogido esta licencia son sus condiciones:

- **Atribución**: la *ODbL* dicta que en la referencia se debe incluir en todo momento una emisión pública con la información contenida en el mismo y conclusiones derivadas.
- **Compartir Igual**: las adaptaciones, revisiones o adiciones a los contenidos o estructura de la organización de la base de datos también deben estar a libre disposición bajo la Licencia Abierta de Bases de Datos.
- **Mantener abierta**: Las secciones de la base de datos no pueden ser protegidos detrás de un muro de pago.

CÓDIGO

El código fuente para la extracción de los datos de la página web de BlaBlaCar está disponible en el mismo repositorio de GitHub (https://github.com/iveirasuoc/BlaBlaCrawler/tree/master/src) en el que se encuentra este fichero.

DATASET

El dataset generado a partir del código anteriormente mencionado está también disponible en el propio repositorio de GitHub (https://github.com/iveirasuoc/BlaBlaCrawler/tree/master/csv). Tal y como se especificaba en un apartado anterior, este conjunto de datos hace referencia a los viajes disponibles durante el día 15/04/2019 en BlaBlaCar (día en el que lanzó la consulta para extraer los datos), con salida de Madrid el día 17/04/2019 y al menos una plaza disponible.

CONTRIBUCIONES AL TRABAJO

Para la realización de este trabajo, los dos integrantes del grupo han participado en las tres tareas de igual manera. Esto queda indicado en la siguiente tabla.

Contribuciones	Firma
Investigación previa	DMR, IVL
Redacción de las respuestas	DMR, IVL
Desarrollo código	DMR, IVL