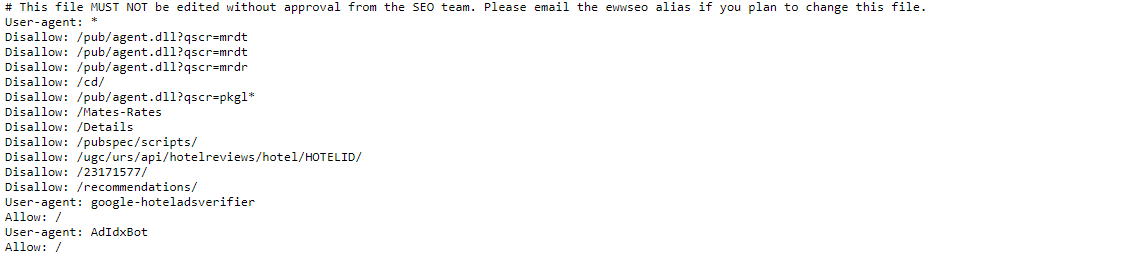
# CONTEXTO

Los datos recolectados en éste proyecto han sido extraídos del sitio web [www.expedia.es](http://www.expedia.es) con fines académicos. El objetivo académico de la extracción de datos es avanzar en el conocimiento de las técnicas de *web scraping*.

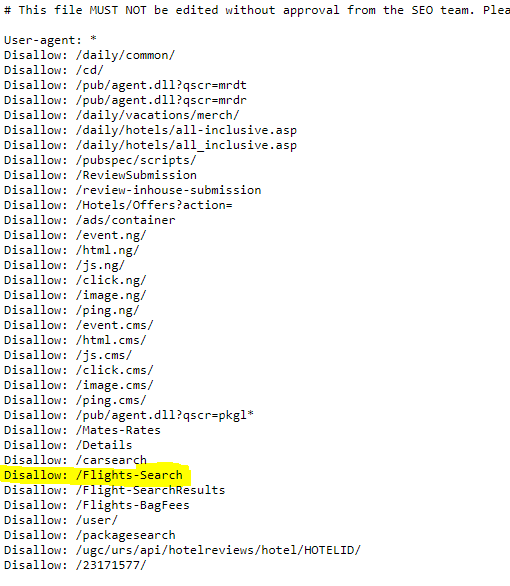
Dado que el enunciado de la práctica daba libertad para escoger el sitio web deseado, hemos decidido extraer información de [www.expedia.es](http://www.expedia.es) dado que es una agencia de viajes que ofrece servicio completo de agencia de viajes como pueden ser: la reserva de billetes de avión, hotel, alquiler de vehículos, cruceros, paquetes vacacionales y varios parques de atracciones, y todo ello a través de internet.

Esta elección estaba supeditada al análisis de las restricciones y características del sitio web para evaluar la legalidad y complejidad de la misma. En este análisis, el primer paso fue acceder al fichero robots.txt del dominio [www.expedia.es](http://www.expedia.es) y este fue el resultado:



Comprobamos que la ruta [www.expedia.es/flights-search](http://www.expedia.es/flights-search) no estaba restringida.

Curiosamente, el fichero robots.txt de la versión internacional de Expedia es muy diferente. En el mismo, se establece una regla de Disallow para la ruta /Flights-Search:



Por lo tanto, decidimos realizar el scraping sobre la versión española ([www.expedia.es](http://www.expedia.es)).

En éste proyecto nos vamos a centrar en trabajar en los **vuelos de ida únicamente**, dado que realizar la extracción de las demás opciones no aportaría contenido académico al trabajo, y sí redundancia de repetir una misma operación. No queremos decir que no fuese interesante si el fin fuese comercial, pero dado que no es el caso, creemos que no resulta de interés académico.

El objetivo de ésta extracción iría enfocado a poder aplicar un modelo en el que se pueda predecir el mejor momento del año para poder comprar un cierto billete de avión. Para ello es necesario automatizar la extracción de ésta web que combina ofertas de diferentes compañías aéreas, y con dicha información almacenada entrenar un modelo que aprenda de cómo se han comportado los precios en el pasado para poder predecir los del futuro.

# TÍTULO DEL DATASET

Título del *data set*: **Precios de un vuelo concreto por mes.**

Siglas: **PVCM**

Título de los archivos: **PVCM-MM-AAAA.csv**

# DESCRIPCIÓN DEL DATASET

Dado que el objetivo que nos hemos fijado es el de saber cuál es el mejor **mes** para ir a un determinado sitio de **nuestra lista de países por visitar** saliendo desde nuestra lista de **aeropuertos cercanos**, se ha decidido extraer información el día 15 de cada mes del año que combine todas las salidas y llegadas de los siguientes 12 meses.

Por lo tanto, el usuario del hipotético sistema de predicción tendría que indicar el mes en el que quiere viajar, y el país al que quiere viajar, y el sistema le indicaría los mejores meses en orden descendente para comprar el billete a ese sitio en concreto y ese mes en concreto.

En nuestro *data set*, por acotar la información almacenada, vamos a fijar la siguiente lista de países por visitar, y lista de aeropuertos cercanos:

* Lista de países por visitar:
  + Aeropuerto de Roma (ROM)
  + Aeropuerto de Nueva York (NYC)
  + Aeropuerto de Londres (LHR-Heathrow)
  + Aeropuerto del Cairo (CAI)
* Lista de aeropuertos cercanos:
  + Aeropuerto de Bilbao (BIO)
  + Aeropuerto de San Sebastián (EAS)
  + Aeropuerto de Vitoria (VIT)
  + Aeropuerto de Pamplona (PNA)

Los datos extraídos cada mes se almacenarán en un archivo .csv diferente, con un nombre diferente como ya se ha comentado en el apartado *Descripción del Dataset*. Un ejemplo de archivo en un mes concreto sería**PVCM-04-2019.csv**

VALIDEZ DE LOS DATOS (¿IMPERECEDEROS?)

# REPRESENTACIÓN GRÁFICA

PENDIENTE DE OBTENER EL PVCM-04-2019.csv

# CONTENIDO

En la siguiente tabla se exponen los diferentes atributos que vamos a extraer o añadir a nuestro data set:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **NOMBRE COLUMNA** | **TIPO DE**  **DATOS** | **ACEPTA**  **NULOS** | **FORMATO** | **VALOR POR**  **DEFECTO** |
| date\_flight | Fecha (sin tiempo) | NO | dd/mm/aaaa | - |
| name\_departure\_airport | xxxx | NO | - | - |
| name\_arrival\_airport | xxxx | NO | - | - |
| name\_airline | xxxx | NO | - | - |
| duration | xxxx | NO | Xd Xh Xm | - |
| time\_departure | xxxx | NO | hh:mm am/pm | - |
| time\_arrival | xxxx | NO | hh:mm am/pm | - |
| number\_stops | xxxx | NO | (X stop/s) / (Nonstop) | - |
| name\_layovers | xxxx | SÍ | - | NaN |
| prices | xxxx | NO | - | - |

Nótese que no se ha añadido la columna *date\_extraction* dado que cada **PVCM-MM-AAAA.csv** hace referencia a un mes diferente de extracción. El motivo de realizar esto así es que en el caso de que no acotemos y queramos guardar todos los datos de todas las combinaciones de países que pueda haber en el sitio web (lo cual sería una aplicación más cerca de la realidad, y menos académica) cada mes ocuparía una extensión de información muy grande, por lo que lo más óptimo para analizar el intervalo de tiempo deseado sería guardarlo en archivos por meses y luego cargar únicamente los archivos deseados para entrenar al modelo.

PENDIENTE DE SEGUIR COMPLETANDO LA TABLA

# AGRADECIMIENTOS

PENDIENTE

# INSPIRACIÓN

PENDIENTE

# LICENCIA

Hemos seleccionado la licencia *Open Database License (ODbL)* para nuestro dataset. Esto quiere decir que los usuarios pueden compartir sus datos con libertad y sin temor a los derechos de autor o cuestiones de propiedad. Cualquier puede hacer uso libre de los datos contenidos en este dataset sin miedo a ninguna infracción de derechos de autor y se permite que utilicen los datos para añadirlos a sus bases de datos. El objetivo de esta licencia es el de comprar y compartir información y datos de forma fácil y libre.

Estas son las condiciones que define la licencia *Open Database License*:

* Atribución: La referencia a la base de datos se debe incluir en todo momento en cualquier emisión pública con información del dataset 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 de los otros usuarios bajo la Licencia Abierta de Bases de Datos.
* Mantener abierta: la redistribución de cualquier versión o de cualquier parte de la base de datos debe estar restringida. Ninguna sección de la base de datos puede estar detrás de un PayWall.

Como ya hemos adelanta en el contexto, el objetivo sería predecir el mes del año en el que resulta más interesante comprar un billete a un destino concreto en un mes concreto del año. Para ello el tipo de información recabada resulta de vital importancia para entrenar modelos de minería de datos como los que siguen:

HABLAR MÁS A FONDO DE LOS MODELOS DE MINERÍA QUE SERÍAN INTERESANTES IMPLEMENTAR.

# CÓDIGO

PEGAR AQUÍ EL CÓDIGO DE GIT TAMBIÉN

# DATASET

CAPTURA DE PANTALLA DE PARTE DEL CSV GENERADO.

# CONTRIBUCIONES DE INTEGRANTES

|  |  |
| --- | --- |
| Contribuciones | Firma |
| Investigación previa | B., J.A. |
| Redacción de las respuestas | B., J.A. |
| Desarrollo código | B., J.A. |

*COPIA DE SEGURIDAD DE ROBOTS.TXT*

[www.expedia.es/robots.txt](http://www.expedia.es/robots.txt)

<https://www.expedia.es/Flights-Search?flight-type=on&starDate=03%2F10%2F2019&mode=search&trip=oneway&leg1=from%3ABilbao%2C+Espa%C3%B1a+%28BIO-Todos+los+aeropuertos%29%2Cto%3ALondres%2C+Reino+Unido+%28LHR-Heathrow%29%2Cdeparture%3A03%2F10%2F2019TANYT&passengers=children%3A0%2Cadults%3A1%2Cseniors%3A0%2Cinfantinlap%3AY&fareCalendarPrice=83>

# This file MUST NOT be edited without approval from the SEO team. Please email the ewwseo alias if you plan to change this file.

User-agent: \*

Disallow: /pub/agent.dll?qscr=mrdt

Disallow: /pub/agent.dll?qscr=mrdt

Disallow: /pub/agent.dll?qscr=mrdr

Disallow: /cd/

Disallow: /pub/agent.dll?qscr=pkgl\*

Disallow: /Mates-Rates

Disallow: /Details

Disallow: /pubspec/scripts/

Disallow: /ugc/urs/api/hotelreviews/hotel/HOTELID/

Disallow: /23171577/

Disallow: /recommendations/

User-agent: google-hoteladsverifier

Allow: /

User-agent: AdIdxBot

Allow: /

[www.expedia.com/robots.txt](http://www.expedia.com/robots.txt)

# This file MUST NOT be edited without approval from the SEO team. Please email the seoamers alias if you plan to change this file.

User-agent: \*

Disallow: /daily/common/

Disallow: /cd/

Disallow: /pub/agent.dll?qscr=mrdt

Disallow: /pub/agent.dll?qscr=mrdr

Disallow: /daily/vacations/merch/

Disallow: /daily/hotels/all-inclusive.asp

Disallow: /daily/hotels/all\_inclusive.asp

Disallow: /pubspec/scripts/

Disallow: /ReviewSubmission

Disallow: /review-inhouse-submission

Disallow: /Hotels/Offers?action=

Disallow: /ads/container

Disallow: /event.ng/

Disallow: /html.ng/

Disallow: /js.ng/

Disallow: /click.ng/

Disallow: /image.ng/

Disallow: /ping.ng/

Disallow: /event.cms/

Disallow: /html.cms/

Disallow: /js.cms/

Disallow: /click.cms/

Disallow: /image.cms/

Disallow: /ping.cms/

Disallow: /pub/agent.dll?qscr=pkgl\*

Disallow: /Mates-Rates

Disallow: /Details

Disallow: /carsearch

**Disallow: /Flights-Search**

Disallow: /Flight-SearchResults

Disallow: /Flights-BagFees

Disallow: /user/

Disallow: /packagesearch

Disallow: /ugc/urs/api/hotelreviews/hotel/HOTELID/

Disallow: /23171577/

# <Checkout>

Disallow: /Confirmation-Flight\*

Disallow: /HotelCheckout

Disallow: /Checkout/\*

# </Checkout>

Allow: /MobileHotel/Webapp/SearchResults

Allow: /MobileHotel/Webapp/HotelOffers

# disallowing multi-filters

# Theme+Star

Disallow: /1Star-\*-Hotels-\*.s10-0-d\*-t\*.Travel-Guide-Filter-Hotels

Disallow: /2Star-\*-Hotels-\*.s20-0-d\*-t\*.Travel-Guide-Filter-Hotels

Disallow: /3Star-\*-Hotels-\*.s30-0-d\*-t\*.Travel-Guide-Filter-Hotels

Disallow: /4Star-\*-Hotels-\*.s40-0-d\*-t\*.Travel-Guide-Filter-Hotels

Disallow: /5Star-\*-Hotels-\*.s50-0-d\*-t\*.Travel-Guide-Filter-Hotels

Disallow: /1Star-\*-Hotels-\*.s10-n\*-t\*.Travel-Guide-Filter-Hotels

Disallow: /2Star-\*-Hotels-\*.s20-n\*-t\*.Travel-Guide-Filter-Hotels

Disallow: /3Star-\*-Hotels-\*.s30-n\*-t\*.Travel-Guide-Filter-Hotels

Disallow: /4Star-\*-Hotels-\*.s40-n\*-t\*.Travel-Guide-Filter-Hotels

Disallow: /5Star-\*-Hotels-\*.s50-n\*-t\*.Travel-Guide-Filter-Hotels

# Theme Brand

Disallow: /\*-Hotels-\*.0-0-d\*-t\*-\*b\*.Travel-Guide-Filter-Hotels

Disallow: /\*-Hotels-\*.0-n\*-t\*-\*b\*.Travel-Guide-Filter-Hotels

# Star Brand

Disallow: /1Star-\*-Hotels-\*.s10-0-d\*--b\*.Travel-Guide-Filter-Hotels

Disallow: /2Star-\*-Hotels-\*.s20-0-d\*--b\*.Travel-Guide-Filter-Hotels

Disallow: /3Star-\*-Hotels-\*.s30-0-d\*--b\*.Travel-Guide-Filter-Hotels

Disallow: /4Star-\*-Hotels-\*.s40-0-d\*--b\*.Travel-Guide-Filter-Hotels

Disallow: /5Star-\*-Hotels-\*.s50-0-d\*--b\*.Travel-Guide-Filter-Hotels

Disallow: /1Star-\*-Hotels-\*.s10-n\*--b\*.Travel-Guide-Filter-Hotels

Disallow: /2Star-\*-Hotels-\*.s20-n\*--b\*.Travel-Guide-Filter-Hotels

Disallow: /3Star-\*-Hotels-\*.s30-n\*--b\*.Travel-Guide-Filter-Hotels

Disallow: /4Star-\*-Hotels-\*.s40-n\*--b\*.Travel-Guide-Filter-Hotels

Disallow: /5Star-\*-Hotels-\*.s50-n\*--b\*.Travel-Guide-Filter-Hotels

# Theme Star Brand

Disallow: /1Star-\*-Hotels-\*.s10-0-d\*-t\*-b\*.Travel-Guide-Filter-Hotels

Disallow: /2Star-\*-Hotels-\*.s20-0-d\*-t\*-b\*.Travel-Guide-Filter-Hotels

Disallow: /3Star-\*-Hotels-\*.s30-0-d\*-t\*-b\*.Travel-Guide-Filter-Hotels

Disallow: /4Star-\*-Hotels-\*.s40-0-d\*-t\*-b\*.Travel-Guide-Filter-Hotels

Disallow: /5Star-\*-Hotels-\*.s50-0-d\*-t\*-b\*.Travel-Guide-Filter-Hotels

Disallow: /1Star-\*-Hotels-\*.s10-n\*-t\*-b\*.Travel-Guide-Filter-Hotels

Disallow: /2Star-\*-Hotels-\*.s20-n\*-t\*-b\*.Travel-Guide-Filter-Hotels

Disallow: /3Star-\*-Hotels-\*.s30-n\*-t\*-b\*.Travel-Guide-Filter-Hotels

Disallow: /4Star-\*-Hotels-\*.s40-n\*-t\*-b\*.Travel-Guide-Filter-Hotels

Disallow: /5Star-\*-Hotels-\*.s50-n\*-t\*-b\*.Travel-Guide-Filter-Hotels

Disallow: /recommendations/

User-agent: AhrefsBot

Disallow: /pictures/

User-agent: google-hoteladsverifier

Allow: /

User-agent: AdIdxBot

Allow: /