

Práctica 1- Tipología y ciclo de vida de los datos

Autores

- David Fernández Álvarez
- Sara Robisco Cavite

1. Contexto

El contexto de esta práctica es la obtención de las detecciones de ondas gravitacionales por parte de LIGO, VIRGO y KAGRA, tanto aquellas confirmadas y publicadas como las descartadas.

Para ello obtendremos los datos de la web del consorcio LIGO, VIRGO y KAGRA. En esta web no sólo exponen información sobre los diferentes observatorios y sus actividades, sino que además muestran y comparten sus datos de manera pública. El enlace es <https://www.gw-openscience.org/> (<https://www.gw-openscience.org/>).

Hemos escogido esta web porque nos ha parecido un conjunto de datos original dentro de una web lo suficientemente compleja para la elaboración de la práctica.

2. Título del dataset

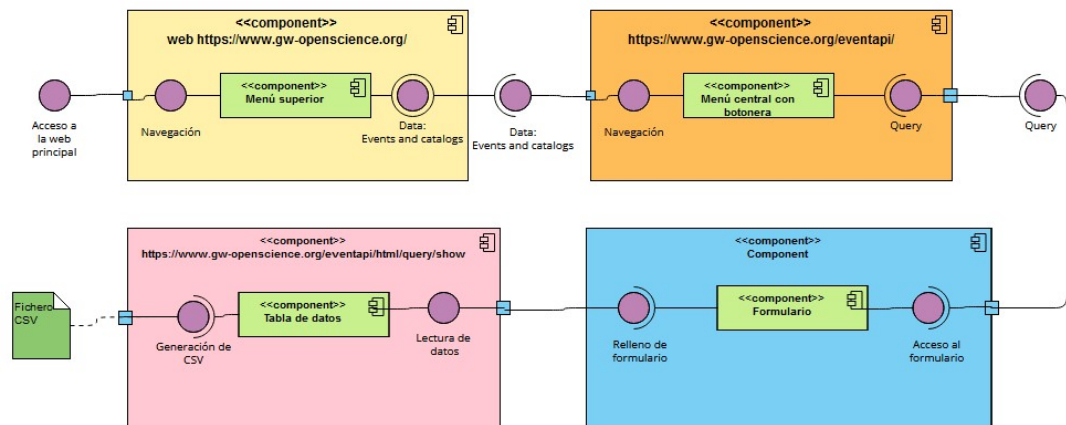
Hemos decidido denominar al dataset `detecciones_ondas_gravitacionales.csv`

3. Descripción del dataset

El conjunto de datos extraído son los datos de las detecciones de ondas gravitacionales, tanto aquellas confirmadas como aquellas que se han asociado a ruido y se han descartado. Nuestra aplicación recoge aquellas detecciones dentro del intervalo de fechas que seleccionemos con la intención de ir las añadiendo a una base de datos sin necesidad de traerse todos los datos. Esto también lo hacemos porque la carga de los datos es lenta para la tabla completa, por lo que es preferible ir descargando los últimos datos periódicamente e insertarlos en una base de datos mediante herramientas como Pentaho.

4. Representación gráfica

La representación visual de la generación del dataset es la siguiente:



5. Contenido

Los campos de nuestro dataset son:

- **Name:** identificador de la detección.
- **Version:** versión de la detección. Se revisan periódicamente.
- **Release.**
- **GPS:** fecha y hora de la detección en formato GPS.
- **Mass_1:** masa del primer objeto en masas solares.
- **Mass_1_upper:** valor máximo del rango de error de la masa del primer objeto.
- **Mass_1_lower:** valor mínimo del rango de error de la masa del primer objeto.
- **Mass_2:** masa del segundo objeto en masas solares.
- **Mass_2_upper:** valor máximo del rango de error de la masa del segundo objeto.
- **Mass_2_lower:** valor mínimo del rango de error de la masa del segundo objeto.
- **Network_snr:** ratio señal/ruido en la red.
- **Network_snr_upper:** valor máximo del rango de error del ratio señal/ruido en la red.
- **Network_snr_lower:** valor mínimo del rango de error del ratio señal/ruido en la red.
- **Distance:** distancia a la que se ha producido la colisión en Megapársecs.
- **Distance_upper:** valor máximo del rango de error de la distancia.
- **Distance_lower:** valor mínimo del rango de error de la distancia.
- **chi_eff:** correlación de campo z de las fusiones de agujeros negros binarios.
- **chi_eff_upper:** valor máximo del rango de error de la correlación de campo.
- **chi_eff_lower:** valor mínimo del rango de error de la correlación de campo.
- **Total_mass:** masa total de ambos cuerpos. Medida en masas solares.
- **Total_mass_upper:** valor máximo del rango de error de la masa total.
- **Total_mass_lower:** valor mínimo del rango de error de la masa total.
- **Chirp_mass:** masa efectiva de un sistema binario. Medida en masas solares.
- **Chirp_mass_upper:** valor máximo del rango de error de la masa efectiva.
- **Chirp_mass_lower:** valor mínimo del rango de error de la masa efectiva.
- **Detector_Frame_Chirp_Mass:** marco del detector de la masa efectiva. Medida en masas solares.
- **Detector_Frame_Chirp_mass_upper:** valor máximo del rango de error del marco del detector de la masa efectiva.
- **Detector_Frame_Chirp_mass_lower:** valor mínimo del rango de error del marco del detector de la masa efectiva.

- **Redshift**: corrimiento al rojo, marca la velocidad a la que se alejan de nosotros.
- **Redshift_upper**: valor máximo del rango de error del corrimiento al rojo.
- **Redshift_lower**: valor mínimo del rango de error del corrimiento al rojo.
- **False_Alarm_Rate**: tasa de falsa alarma. La medida es años elevado a -1.
- **P_astro**: probabilidad de que el evento tenga un origen astrofísico.
- **Final_mass**: masa final del objeto resultante tras la colisión. Medida en masas solares.
- **Final_mass_upper**: valor máximo del rango de error de la masa final.
- **Final_mass_lower**: valor mínimo del rango de error de la masa final.

6. Propietario

Para obtener el propietario de la web hemos hecho un script de python, para poder ejecutarlo hemos instalado la librería python-whois. A continuación se muestra el script y los datos del propietario de la web:

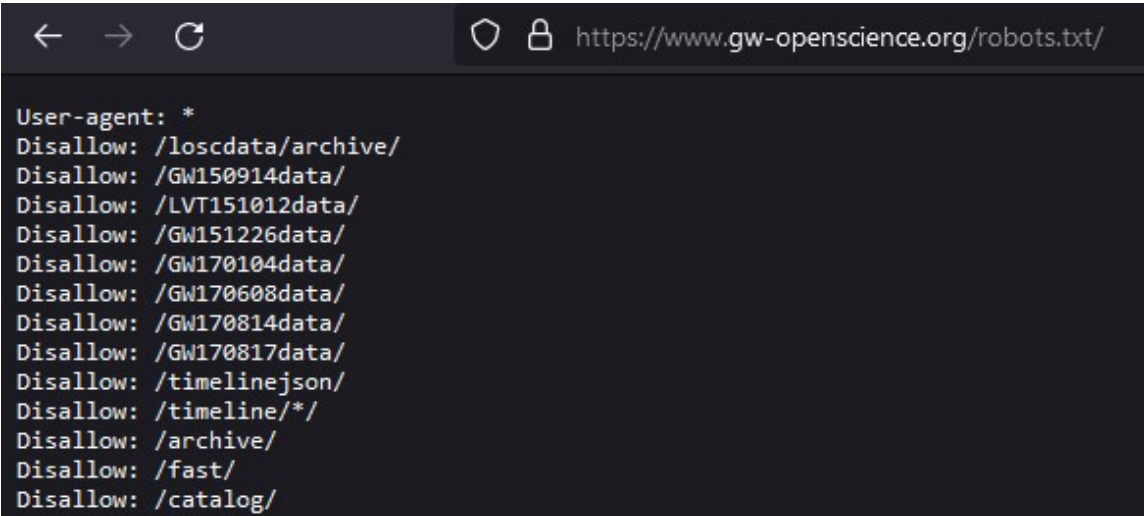
```
In [1]: ▶ import whois
w = whois.whois('https://www.gw-openscience.org')
print(w)

{
  "domain_name": "gw-openscience.org",
  "registrar": "Domain.com, LLC",
  "whois_server": "http://whois.domain.com",
  "referral_url": null,
  "updated_date": "2022-06-12 19:25:09",
  "creation_date": "2017-09-04 21:11:08",
  "expiration_date": "2027-09-04 21:11:08",
  "name_servers": [
    "ligo.ligo.caltech.edu",
    "mercutio.ni.caltech.edu",
    "tepid.ni.caltech.edu"
  ],
  "status": [
    "clientTransferProhibited https://icann.org/epp#clientTransferProhibited",
    "clientUpdateProhibited https://icann.org/epp#clientUpdateProhibited"
  ],
  "emails": "compliance@domain-inc.net",
  "dnssec": "unsigned",
  "name": "REDACTED FOR PRIVACY",
  "org": "Domain Privacy Service FBO Registrant.",
  "address": "REDACTED FOR PRIVACY",
  "city": "REDACTED FOR PRIVACY",
  "state": "FL",
  "registrant_postal_code": "REDACTED FOR PRIVACY",
  "country": "US"
}
```

Observamos que el propietario de la web es "Domain Privacy Service FBO Registrant.", esto suena a que no quieren mostrar abiertamente los datos del propietario. Sabemos que están en Estados Unidos y por los nombres de los servidores podemos deducir que realmente el dueño es el Caltech.

Al observar lo celosos que son de su privacidad nos dio miedo que no nos dejaran hacer

webscraping, así que lo primero que hicimos fue consultar su archivo Robots.txt, situado en <https://www.gw-openscience.org/robots.txt/> (<https://www.gw-openscience.org/robots.txt/>). Veamos su contenido:



```
User-agent: *
Disallow: /loscdata/archive/
Disallow: /GW150914data/
Disallow: /LVT151012data/
Disallow: /GW151226data/
Disallow: /GW170104data/
Disallow: /GW170608data/
Disallow: /GW170814data/
Disallow: /GW170817data/
Disallow: /timelinejson/
Disallow: /timeline/*/
Disallow: /archive/
Disallow: /fast/
Disallow: /catalog/
```

Por lo que vemos sólo hay restricciones a ciertas páginas, evitando acceder a ellas no incumpliremos sus restricciones y actuaremos de manera ética. Además hemos insertado en el código retardos de tiempo para evitar saturar el servidor con nuestra actividad. Para evitar saturar su web con consultas, nuestra idea es hacer una extracción de datos semanal, con los datos de la última semana cada vez. De este modo no saturaremos su servidor, además no se detectan ondas gravitacionales a diario, por lo que el intervalo semanal es lo mejor para todos.

Si fuésemos a realizar descargas masivas de una forma asídua, o si hubiese existido alguna restricción que nos hubiera impedido obtener los datos, nos hubiéramos puesto en contacto con el grupo que trabaja con la colaboración Virgo en la Universidad de Valencia. Explicándoles la necesidad de los datos estamos seguros de que nos ayudarían a obtenerlos.

7. Inspiración

Lo que nos ha hecho tomar la decisión de obtener este conjunto de datos es conocer más sobre las detecciones de ondas gravitacionales. Cada vez que el consorcio formado por LIGO, VIRGO y KAGRA publican una detección, ésta se convierte en un acontecimiento en medios y nos surgen muchas preguntas:

- ¿Qué intervalos de masas de objetos son los más detectados?
- ¿Hay periodos del año donde haya más probabilidad de detecciones? Si es así ¿De qué región del espacio provienen?
- ¿Qué hace que una señal se considere buena o se descarte?
- ¿Cuáles son las detecciones más cercanas? ¿Y las más lejanas?

Estos son ejemplos de preguntas a las que nos gustaría dar respuesta. Lo positivo es que los datos que necesitamos para ello están en partes de la web de acceso permitido, por lo tanto es viable obtenerlos.

8. Licencia

Hemos decidido compartir el dataset bajo la licencia CC BY 4.0. Esta licencia consiste en:

- Permisos:
 - Permite compartir y distribuir los datos en cualquier medio.
 - Permite remezclar y transformar los datos.
- Restricciones:
 - Debes dar crédito a los autores, proveer un enlace a la licencia e indicar si has hecho cambios sobre los datos originales.

Se pueden conocer los detalles de esta licencia aquí: <https://creativecommons.org/licenses/by/4.0/> (<https://creativecommons.org/licenses/by/4.0/>).

El motivo es que la licencia de los datos originales de la web es CC BY 4.0 y creemos que es ético que los compartamos con el mismo tipo de licencia. Además queremos que todo el mundo pueda usar nuestro dataset, pudiendo compartirlo y adaptar los datos a sus necesidades (coger las columnas que quieran y descartar las que no, por ejemplo). La idea es que siempre que lo usen den atribución al consorcio LIGO, VIRGO y KAGRA, pues son el autentico origen de datos.

9. Código

El código de nuestro proyecto se encuentra en el siguiente enlace: <https://github.com/SRobiscoUOC/Practica1> (<https://github.com/SRobiscoUOC/Practica1>). Se ha dividido la aplicación en un programa principal, llamado scraper.py, y otro archivo con las funciones requeridas para el correcto funcionamiento de la aplicación, llamado scraper_functions.py. Veamos su contenido:

- **Scraper.py:** contiene el programa principal que recibe como parámetros Fecha_inicio y Fecha_fin. Con estos parámetros invoca a la función execute_form, situada en scraper_functions.py, y rellena el fichero csv con el dataset resultante.
- **scraper_functions.py:** contiene la función execute_form que accede a la web <https://www.gw-openscience.org/> (<https://www.gw-openscience.org/>) empleando Selenium, esta función va navegando por los menús de la web hasta encontrar el formulario para generar la tabla de datos. Una vez ahí, rellena el formulario y pulsa el botón commit generando la tabla. Podría parecer que al generar la tabla sólo habría que recorrerla pero no es así: al trabajar con la pantalla de Chrome minimizada (hemos usado Chrome debido a que firefox no permite hacer webscrapping en sus últimas versiones) no mostraba la totalidad de las columnas. Esto ha hecho que, tras generar la tabla esperemos unos segundos y pulsemos un botón que abre el desplegable del selector de columnas, una vez abierto hemos tomado los elementos del selector y, mediante una función que hemos hecho para comprobar si el elemento estaba sin seleccionar y seleccionarlo en ese caso, hemos dejado todos marcados, quedando la tabla completa y lista para recorrerla. Otro problema encontrado han sido los datos: había datos con espacios que hemos tenido que tratar, pero lo peor ha sido la presencia de subíndices y superíndices que ha habido que tratar, dando lugar a nuevas columnas con los márgenes de error. Ha sido todo un reto al que ha habido que añadir que la mayor parte de los elementos de la web carecían de id. Otro problema con el que nos hemos topado ha sido la lentitud: el sistema tarda unos 17 minutos en general el documento debido a la complejidad de la tabla.

Además de las dificultades descritas, debemos añadir que al usar Chromedriver con Selenium nos encontramos unos extraños errores de lectura de USB que no tenían nada que ver con nuestro código. Tras buscar documentación lo solucionamos añadiendo el

siguiente código en la declaración de nuestro navegador:

```
In [ ]:  options = webdriver.ChromeOptions()
         options.add_experimental_option('excludeSwitches', ['enable-logging'])
         driver = webdriver.Chrome(options=options)
```

Con esto el programa se ejecuta de manera fluida sin lanzar errores. Es muy importante que lo lancemos desde la carpeta source, pues tiene todo direccionado a ese directorio raíz. Para ejecutar la aplicación lo haremos del siguiente modo: *python scraper.py --Fecha_inicio YYYYMMdd --Fecha_fin YYYYMMdd*. Un ejemplo de ejecución sería la siguiente:

```
python scraper.py --Fecha_inicio 20150101 --Fecha_fin 20221105
```

Con estas fechas aseguramos tomar desde la primera detección de ondas gravitacionales, que tuvo lugar el día 14/09/2015. Damos un margen porque queremos obtener también las detecciones descartadas.

En la carpeta source hemos añadido el archivo requirements.txt con las librerías requeridas tanto para la elaboración de la práctica como para la creación de este documento. Su contenido es el siguiente:

appdirs==1.4.4

async-generator==1.10

brotlipy==0.7.0

builtwith==1.3.4

certifi @ file:///C:/b/abs_ac29jvt43w/croot/certifi_1665076682579/work/certifi

cfffi @ file:///C:/Windows/Temp/abs_6808y9x40v/croots/recipe/cfffi_1659598653989/work

click==8.1.3

colorama @ file:///C:/Windows/TEMP/abs_9439aeb1-0254-449a-96f7-33ab5eb17fc8apleb4yn/croots/recipe/colorama_1657009099097/work

cryptography @ file:///C:/b/abs_36x9ifdcl4/croot/cryptography_1665612655344/work

exceptiongroup==1.0.1

fastprogress @ file:///tmp/build/80754af9/fastprogress_1634699150215/work

future==0.18.2

h11==0.14.0

idna @ file:///C:/b/abs_bdhbebrioa/croot/idna_1666125572046/work

importlib-metadata==5.0.0

itsdangerous==2.1.2

Jinja2==3.0.3

mkl-fft==1.3.1
mkl-random @ file:///C:/ci/mkl_random_1626186184308/work
mkl-service==2.4.0
numpy @ file:///C:/b/abs_53f_dbvhzc/croot/numpy_and_numpy_base_1665773185489/work
outcome==1.2.0
pycparser @ file:///tmp/build/80754af9/pycparser_1636541352034/work
pyee==8.2.2
pyOpenSSL @ file:///opt/conda/conda-bld/pyopenssl_1643788558760/work
pypeteer==1.0.2
PySocks @ file:///C:/ci/pysocks_1605307512533/work
python-whois==0.8.0
selenium==4.6.0
six @ file:///tmp/build/80754af9/six_1644875935023/work
sniffio==1.3.0
sortedcontainers==2.4.0
tqdm @ file:///C:/b/abs_0axbz66qik/croots/recipe/tqdm_1664392691071/work
trio==0.22.0
trio-websocket==0.9.2
urllib3 @ file:///C:/b/abs_a8_3vfznn_/croot/urllib3_1666298943664/work
websockets==10.4
Werkzeug==2.2.2
win-inet-pton @ file:///C:/ci/win_inet_pton_1605306162074/work
wincertstore==0.2
wsproto==1.2.0

10. Dataset

El dataset se ha publicado en Zenodo con el DOI: 10.5281/zenodo.7308240

DOI 10.5281/zenodo.7308240

<https://doi.org/10.5281/zenodo.7308240>

El contenido del dataset se corresponde con los datos de las detecciones de ondas gravitacionales obtenidas desde el 01 de enero de 2015 hasta el 05 de noviembre de 2022.

En total son 119 filas que contiene tanto los datos de aquellas detecciones confirmadas como ondas gravitacionales generadas por la colisión de dos objetos interestelares, como aquellas detecciones descartadas. A continuación se muestra una previsualización de los datos en excel:

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
name	version	release	cpu	mass_1	mass_1_upper	mass_1_lower	mass_2	mass_2_upper	mass_2_lower	network_size	network_size_upper	network_size_lower	distance	distance_upper	distance_lower	chi_sq	chi_sq_upper
GW200322_091133	v1	GWTC-3-confident	12689035113	34	48	-18	140	168	-87	60	17	-12	3600	7000	-2000	24	43
GW200316_215756	v1	GWTC-3-confident	12684310941	131	102	-29	78	19	-29	103	4	-7	1120	470	-440	13	27
GW200311_115853	v1	GWTC-3-confident	12679631513	342	64	-38	277	41	-59	178	2	-2	1170	280	-400	-2	16
GW200311_103121	v1	GWTC-3-marginal	12679578997								92						
GW200308_173609	v1	GWTC-3-confident	12677241877	364	112	-96	138	72	-33	71	5	-5	5400	2700	-2600	65	17
GW200306_093714	v1	GWTC-3-confident	12675226521	283	171	-77	148	65	-64	78	4	-6	2100	1700	-1100	32	28
GW200302_015811	v1	GWTC-3-confident	12671490995	378	87	-85	200	81	-57	108	3	-4	1480	1020	-700	1	25
GW200225_060421	v1	GWTC-3-confident	12666458795	193	50	-30	140	28	-55	125	3	-4	1150	510	-530	-12	17
GW200224_222234	v1	GWTC-3-confident	12666181724	400	89	-45	325	50	-72	200	2	-2	1710	490	-640	10	15
GW200220_124850	v1	GWTC-3-confident	12662381481	389	141	-86	279	92	-90	85	3	-5	4000	2800	-2200	-7	27
GW200220_061928	v1	GWTC-3-confident	12662147867	87	40	-23	61	26	-25	72	4	-7	6000	4800	-3100	6	40
200219_201407	v1	GWTC-3-marginal	12661784659								136						
GW200215_094415	v1	GWTC-3-confident	12661400731	375	101	-69	279	74	-84	107	3	-5	3400	1700	-1500	-8	23
GW200216_220804	v1	GWTC-3-confident	12659261028	51	22	-13	30	14	-16	81	4	-5	3800	3000	-2000	10	34
200214_224526	v2	GWTC-3-marginal	12657555445								131						
GW200210_092254	v1	GWTC-3-confident	12653617929	241	75	-46	283	47	-42	84	5	-7	940	430	-340	2	22
GW200209_085452	v1	GWTC-3-confident	12653737101	356	105	-68	271	78	-78	96	4	-5	3400	1900	-1800	-12	24
GW200208_222617	v1	GWTC-3-confident	12652359959	51	104	-30	123	90	-57	74	14	-12	4100	4400	-1900	45	43
GW200208_130117	v1	GWTC-3-confident	12652020959	378	92	-62	274	61	-74	108	3	-4	2230	1000	-850	-7	22
GW200202_154313	v1	GWTC-3-confident	12646934115	101	35	-14	73	11	-17	108	2	-4	410	150	-160	4	13
GW200201_203549	v1	GWTC-3-marginal	12642404670								90						
GW200129_065458	v1	GWTC-3-confident	12643161164	345	99	-32	289	34	-93	268	2	-2	900	290	-380	11	11

11. Vídeo

El vídeo explicativo se encuentra en el siguiente enlace

12. Tabla de contribuciones

A continuación se muestra la tabla con las contribuciones de las personas del equipo:

Contribuciones	Firma
Investigación previa	D.F.A., S.R.C.
Redacción de las respuestas	D.F.A., S.R.C.
Desarrollo del código	D.F.A., S.R.C.
Participación en el vídeo	D.F.A., S.R.C.