1. Introducción

Este documento presenta la solución a la PRAC1 de la asignatura Tipología y ciclo de vida de los Datos del Máster de Ciencia de Datos. Los alumnos participantes en esta práctica son:

Gregorio Andrés García Menéndez (gagarcia)

Manuel Enrique Gómez Montero (mnlgmontero)

2. Solución

En esta práctica se pedía realizar un modelo de Web Scraping y presentar tanto el proceso como el resultado final. Siendo este documento un resumen de todo lo pedido, el código del programa y así como la explicación del mismo se encuentran en el repositorio de Github creado para la ocasión (https://github.com/megmontero/MovieScrapper).

A continuación presentamos los distintos puntos explicados detalladamente según se pedía en el enunciado de la práctica.

1. Contexto. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.

Hoy día, la inteligencia artificial y el uso de algoritmos de esta rama para tomar decisiones en contextos basados en experiencia de usuario es inmenso. Podemos encontrar casos como la publicidad basada en navegación de usuario que utiliza Google e inserta en cada página que visitamos adscrita a su red de publicidad, o también recomendaciones de productos en empresas de venta online como Amazon, en la que también basan las experiencias pasadas de los usuarios en cuanto a ventas y navegación para mandar publicidad y recomendaciones dirigidas.

La información recogida por nosotros y el dataset final viene dado por una necesidad similar. Teniendo en cuenta que estos sistemas funcionan muy bien al contar con muchísimos datos de múltiples usuarios, nuestro dataset recoge en varias visiones todos los datos de la plataforma colaborativa *IMDB*. Esta plataforma contiene prácticamente todas las películas, series y demás contenido audiovisual con toda la información pertinente (desde actores hasta taquilla según que casos), y también *ratings* y *reviews* de usuario sobre todo ese contenido.

Con toda esta gran fuente de datos, nuestro dataset se convierte en el pilar fundamental en el que basar un **recomendador de contenido audiovisual**. Y otro punto adicional e interesante, es que tal y como recogemos la información y en relación a la *forma* que hemos dado a nuestro dataset, el recomendador podría ser tanto *content-to-content,* es decir, recomendar contenido a un usuario según material relacionado o similar a dicho contenido, como *content-to-person*, esto es basar las recomendaciones en contenido de usuarios con gustos similares.

Así, con nuestro scraper recolectamos información sobre películas: información técnica propia de la película, personas involucradas en la película (actores, directores, etc) y usuarios de la plataforma que han visto y votado la película.

2. Definir un título para el dataset. Elegir un título que sea descriptivo

El título elegido para el dataset es: ***IMDB Movies, Persons and Users Dataset***.

3. 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).

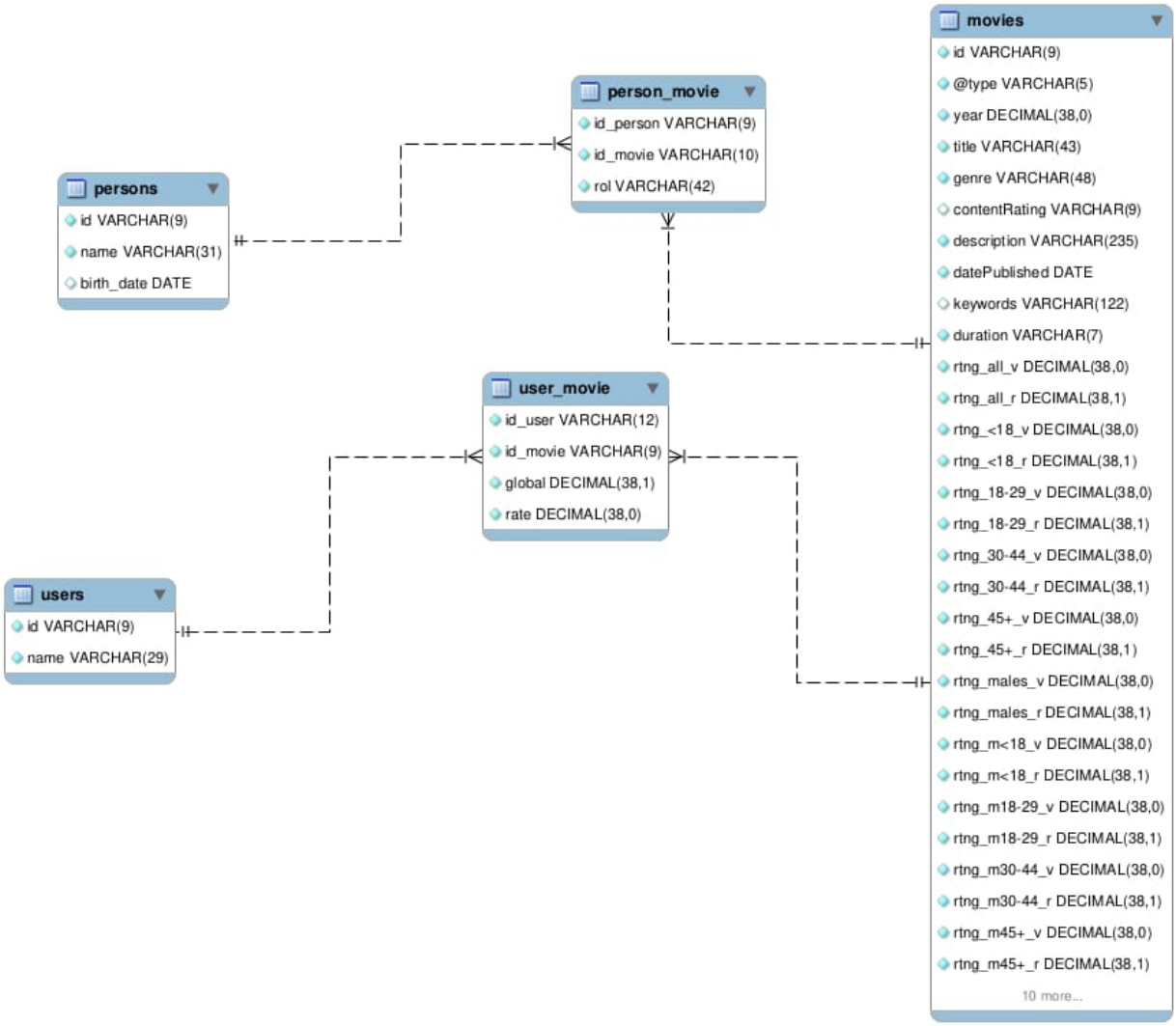
El dataset ha sido diseñado con una filosofía **NoSQL** mediante una BBDD documental, por simplicidad para la práctica se ha usado **UnQLite** ya que es una BBDD autocontenida que no requiere servidor; en un futuro podría usarse una BBDD más compleja que nos dé un mejor rendimiento y escalabilidad como puede ser MongoDB.

El objetivo de usar una BBDD documental es, por un lado, huir de la rigidez de un modelo relacional, teniendo más flexibilidad usando json que tuplas y, por otro, tener disponibles los datos tal y como queremos consumirlos. Los agregados (o perspectivas desde las que queremos acceder a los datos) creados son:

- **Movies**: Colección de películas. Cada película tiene propiedades generales(título, género, año...), personas que participan(actores, directores y productores), rating demográfico por edad y sexo y usuarios que han realizado reviews de las películas.
- **Persons**: Colección de personas y las películas en las que han participado con distintos roles.
- **Users**: Colección de usuarios con los ratings que han realizado de películas.

4. Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente

La explicación del apartado anterior se resume en el siguiente esquema gráfico de carácter general:



Hay que tener en cuenta que para nuestro enfoque es más adecuado tener en mente un esquema No Relacional. La imagen anterior corresponde a la explicación del dataset desde un punto de vista relacional.

A continuación se encuentran tres ejemplos, uno de cada visión de base de datos explicado:

Ejemplo de ***Movies***:

Modelo de datos IMDB Scrapper

Ejemplo documento colección *Movie*


```
{
  @type: "Movie",
  name: "The Haunting of Sharon Tate",
  + genre: ["Drama","Horror","Thriller"],
  contentRating: "R",
  description: "The Haunting of Sharon Tate is a movie starring
    Hilary Duff, Jonathan Bennett, and Lydia Hearst. Pregnant with
    director Roman Polanski's child and awaiting his return from
    Europe, 26-year-old Hollywood actress Sharon Tate becomes...",
  datePublished: "2019-04-05",
  keywords: "actress,year 1969,killer,charles manson,mass
    murder",
  duration: "PT1H34M",
  title: "The Haunting of Sharon Tate",
  year: 2019,
  id: "tt7976208",
  + actors: [{"name":"Hilary Duff","id":"nm0240381"},{"nam… ],
  + creators: [{"name":"Daniel Farrands","id":"nm0268107"}],
  + directors: [{"name":"Daniel Farrands","id":"nm0268107"}],
  + rating: [{"group":"all","rating":3.1,"votes":279},{"gr… ],
  + reviews: [{"user":{"name":"incredingo-37769","id":"ur73… ],
  __id: 0
}
```


Ejemplo de ***Persons***:

Ejemplo documento colección *Persons*


```
{
  id: "nm0230826",
  name: "Monica Dolan",
  birth_date: "1969-03-15",
  + Actress: [{"name":"National Theatre Live: All About Eve… ],
  + Self: [{"name":"Sunday Brunch","id":"tt2326935"},{"n… ],
  + Archive footage: [{"name":"Cormoran Strike","id":"tt4276618"}]
}
```

Ejemplo de ***Users***:

Ejemplo documento colección *Users*

A pesar de que nuestro modelo de datos para el caso de uso elegido es no relacional y para ello el formato JSON era ideal, también se ha realizado el formato CSV para cubrir todas las necesidades de futuros proyectos basados en este dataset.

5. Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.

Nuestro dataset incluye las tres colecciones mencionadas: *movies, persons* y *users.* La información recogida en cada una de ellas es la siguiente:

- **Movies**: cada registro corresponde a una película distinta. Los atributos recogidos de cada película son:
  - *@type:* Tipo del documento dentro de la base de datos (preparado por si queremos ampliar a más contenidos audiovisuales en un futuro)
  - *name:* Nombre de la película en IMDB
  - *genre*: etiquetas que describen el género de la película
  - *contentRating:* clasificación que indica el tipo de audiencia de la película
  - *description*: resumen descriptivo del argumento e información de la película
  - *datePublished*: año de publicación de la película
  - *keywords*: palabras clave o *tags* que identifican a la película
  - *duration*: duración de la película en formato ISO
  - *title*: Título oficial de la película
  - *year*: año de la película
  - *id*: ID único de la película dentro de IMDB y de nuestro dataset
  - *actors*: Actores y actrices involucrados en la película
  - *creators*: Creadores involucrados en la película
  - *directors*: Directores que participaron en la película
  - *rating*: información general de valoración de los usuarios a nivel demográfico
  - *reviews*: reseñas realizadas por los usuarios que contiene el *id* del usuario, el nombre y su puntuación

(ampliable en un futuro a recoger también el texto de la reseña de cara a realizar más analítica, como análisis de sentimientos o similar)

- **Persons**: Todas las figuras involucradas en las películas que recogemos:
  - *id*: ID único de la persona dentro de IMDB y de nuestro dataset
  - *name*: nombre de la persona
  - *birth_date*: fecha de nacimiento de la persona
  - Información sobre los proyectos en los que ha participado: esta parte del dataset es variable, ya que recogemos todas las personas involucradas en cada película, pero cada persona, además de haber participado en esa película, puede haber estado involucrada en numerosos tipos de proyectos: de actor, de director, en el departamento musical, etc. Esta parte es muy variopinta, y hace que no haya unos campos fijos, aunque sí que se llaman igual dentro de la plataforma y del dataset para dos personas distintas que han participado en roles iguales. Como queremos que el dataset sea completo y extensible, recogemos absolutamente todos los proyectos en los que la persona aparece y el role en cada uno de ellos. Ejemplos de categorías añadidas aquí son:
    - *Director*: la persona ha dirigido un contenido
    - *Actor:* la persona ha participado en el casting de un contenido
    - *Soundtrack*: la persona ha participado en la banda sonora de un contenido
    - *Self*: la persona ha salido en representación de sí misma en algún contenido

  Así, la lista se extiende hasta un número de categorías mucho mayor.
- **Users**: Los usuarios de la plataforma IMDB que han votado las películas recogidas, y todo lo que han votado y qué puntuación han dado al contenido de la plataforma:
  - *id:* ID único en IMDB y en nuestro dataset
  - *name*: alias del usuario en la plataforma
  - *ratings*: lista con todos los contenidos que ha votado y la puntuación que ha dado a los mismos

6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).

Agradecimientos al propietario de los datos extraídos es la plataforma, IMDB por disponer de los datos para nuestros fines académicos aquí presentados

7. Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder.

Como ya hemos mencionado anteriormente, este conjunto de datos tiene tres visiones diferenciadas y a la vez relacionadas: películas con todos los usuarios que han votado dicha película así como sus puntuaciones y todas las figuras involucradas en cada película; figuras involucradas en cada película y toda la información de proyectos en los que ha participado; y todos los usuarios de la plataforma y puntuaciones que han dado a cada contenido de la plataforma.

Toda la información disponible es ideal para construir nuestro recomendador basado en contenido y en usuarios, de tal forma que los casos de uso son muchos. Por ejemplo:

– Recomendación de películas similares en base a la película que estoy viendo *(¿Qué películas me recomiendas si me interesa esta película?)*:

– Recomendación de películas cuyas personas involucradas sean similares (*¿Qué películas me recomiendas si he visto muchas películas de esta/s persona/s?)*:

– Recomendación de películas basada en usuarios con gustos parecidos (*¿Qué películas me recomiendas que no haya visto y que gente similar a mí sí?)*:

Así, el potencial de casos de uso y utilidad de los datos es de gran valor y puede ser extendido a cualquier tipo de situación concreta (por ejemplo, si además de recomendaciones como las anteriores queremos tener una recomendación personalizada basada en una película nueva cuya promoción va a ser realizada a gran escala, si se va a realizar el relanzamiento de una serie o figura concretos, etc)

8. Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección:

- ○ ***Released Under CC0: Public Domain License***
- ○ ***Released Under CC BY-NC-SA 4.0 License***
- ○ ***Released Under CC BY-SA 4.0 License***
- ○ ***Database released under Open Database License, individual contents under Database Contents License***
- ○ ***Other (specified above)***
- ○ ***Unknown License***

En cuanto al licenciamiento del dataset, lo primero ha sido referirse al apartado de la propia web, que dice:

https://help.imdb.com/article/imdb/general-information/can-i-use-imdb-data-in-my-software/G5JTRESSHJBBHTGX#

Can I use IMDb data in my software?
Limited non-commercial use of IMDb data is allowed, provided the following conditions are met:

You agree to all the terms of our copyright/conditions of use statement.
Please also note that IMDb reserves the right to withdraw permission to use the data at any time at our discretion.
The data must be taken only from the datasets made available (see IMDb Contributor Datasets . You may not use data mining, robots, screen scraping, or similar online data gathering and extraction tools on our website. If the information/data you want is not present in our datasets, it means it's not available for non-commercial usage.
The data can only be used for personal and non-commercial use and must not be altered/republished/resold/repurposed to create any kind of online/offline database of movie information (except for individual personal use). Please refer to the copyright/license information enclosed in each file for further instructions and limitations on allowed usage.
You must acknowledge the source of the data by including the following statement:
Information courtesy of
IMDb
(http://www.imdb.com).
Used with permission.

We reserve the right to deny any individual request for any reason.

Commercial use
Looking for information on licensing IMDb content for commercial use? See here

---

Como nuestro uso es académico y no comercial, vamos a elegir una licencia abierta haciendo mención a IMDB tal y como especifica la web.

Según hemos estudiado los licenciamientos, en el caso de dataset funciona de forma distinta. Podríamos distinguir dos casos:

1. El creador de los datos es la propia entidad que licencia el dataset (ejemplo: un fotógrafo que crea un dataset de fotos sacadas por él sobre las cuales tiene todos los derechos)
2. El dataset viene de datos ya creados o pertenecientes a otra entidad o entidades (nuestro caso)

En el primer caso, el dataset puede estar acotado bajo la licencia que el creador de los datos considere adecuado.

En el segundo caso, en el que nuestro dataset recae, el licenciamiento viene más ligado a la forma del dataset y lo que aporta que a los datos en sí.
Esto es: si nos limitamos a descargar una base de datos de forma prácticamente literal, podemos estar infrigiendo alguna ley respecto a lo indicado por el creador de los datos originales.
Sin embargo, si nuestro dataset no es una copia exacta de los datos en cuanto a estructura y aporta algo más, el licenciamiento puede ser elegido acorde al origen de datos, pero dándole una licencia propia.

En nuestro dataset no nos hemos limitado a descargar los datos de forma literal: hemos creado un dataset propio con tres visiones distintas que, si alguien se plantea hacer uso de los datos para un caso de uso concreto, la utilidad que este dataset aporta es única y no puede ser dada por los datos originales en la forma en la que están (dentro de la web de imdb).
Por lo tanto, elegimos una licencia de CDLA-Sharing-1.0, haciendo siempre mención a IMDB tal y como piden en su web. Elegimos esta licencia debido al carácter abierto y de *copy-left* y que a ello permitiría cual uso, modificación o manipulación del dataset para fines iguales a los nuestros.

El contenido de la licencia, además de poder verla en la página de proyecto de Github, puede consultarse en el siguiente enlace: https://cdla.io/sharing-1-0/

9. Código. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.

El código ha sido desarrollado utilizando *Python 3.6*, en entornos Windows y Linux, bajo los que también ha sido completamente probados.

Principalmente se han usado las siguientes librerías:

- **Requests**: Utilizado para extraer el código HTML en páginas estáticas.
- **Selenium**: Utilizado para extraer el código HTML y navegar en páginas dinámicas.
- **Beautiful Soup**: Parser utilizado para extraer información del código HTML.

Todo el código ha sido realizado siguiendo las normativas oficiales de estilo PEP8 (https://www.python.org/dev/peps/pep-0008/) y verificado usando el módulo pycodestyle oficial para este propósito.

Para consultar el código completo, qué estructura tiene y ver en detalle todos los archivos, se debe visitar el proyecto de Github correspondiente a esta práctica:

https://github.com/megmontero/MovieScrapper

10. Dataset. Presentar el dataset en formato CSV

Como ya hemos expuesto, este dataset se aprovecha bien de una estructura de datos *NoSQL*, de ahí nuestra elección de UnQLite como base de datos de almacenamiento, pero también puede interesar tener un formato de datos relacional de cara a futuros casos de uso. Así, nuestro código también permite la exportación a CSV de los datos que el scraper va obteniendo de IMDB.

En este caso, en lugar de tener tres colecciones (movies, persons, users), tenemos cinco ficheros distintos (*tablas en un modelo relacional)* que guardan los datos de las tres colecciones y la interacción entre ellas. Esta naturaleza puede verse en los ficheros CSV de muestra que hay en la carpeta *dataset* dentro del proyecto de Github. Una muestra del dataset es la siguiente:

- ***movies.csv***: guarda la información propia completa de cada película

```
"id";"@type";"year";"title";"genre";"contentRating";"description";"datePublished";"keywords";"duration";"rtr
"tt4154756";"Movie";2018;"Avengers: Infinity War";"Action,Adventure,Sci-Fi";"PG-13";"Avengers: Infinity War
"tt1825683";"Movie";2018;"Black Panther";"Action,Adventure,Sci-Fi";"PG-13";"Black Panther is a movie starrir
"tt5463162";"Movie";2018;"Deadpool 2";"Action,Adventure,Comedy,Sci-Fi";"R";"Deadpool 2 is a movie starring P
"tt1727824";"Movie";2018;"Bohemian Rhapsody";"Biography,Drama,Music";"PG-13";"Bohemian Rhapsody is a movie s
"tt1677720";"Movie";2018;"Ready Player One";"Action,Adventure,Sci-Fi";"PG-13";"Ready Player One is a movie s
"tt6644200";"Movie";2018;"A Quiet Place";"Drama,Horror,Mystery,Sci-Fi,Thriller";"PG-13";"A Quiet Place is a
"tt1270797";"Movie";2018;"Venom";"Action,Sci-Fi,Thriller";"PG-13";"Venom is a movie starring Tom Hardy, Mich
"tt1517451";"Movie";2018;"A Star Is Born";"Drama,Music,Romance";"R";"A Star Is Born is a movie starring Lady
"tt1477834";"Movie";2018;"Aquaman";"Action,Adventure,Fantasy,Sci-Fi";"PG-13";"Aquaman is a movie starring Ja
"tt4912910";"Movie";2018;"Mission: Impossible - Fallout";"Action,Adventure,Thriller";"PG-13";"Mission: Impos
"tt2798920";"Movie";2018;"Annihilation";"Adventure,Drama,Horror,Mystery,Sci-Fi,Thriller";"R";"Annihilation i
"tt5095030";"Movie";2018;"Ant-Man and the Wasp";"Action,Adventure,Comedy,Sci-Fi";"PG-13";"Ant-Man and the Wa
"tt3778644";"Movie";2018;"Solo: A Star Wars Story";"Action,Adventure,Fantasy,Sci-Fi";"PG-13";"Solo: A Star W
"tt4154664";"Movie";2019;"Captain Marvel";"Action,Adventure,Sci-Fi";"PG-13";"Captain Marvel is a movie starr
"tt4881806";"Movie";2018;"Jurassic World: Fallen Kingdom";"Action,Adventure,Sci-Fi";"PG-13";"Jurassic World:
```

- ***persons.csv***: guarda la información propia completa de cada persona (figura) participante en películas:

```
"id";"name";"birth_date"
"nm0000375";"Robert Downey Jr.";"1965-04-04"
"nm1165110";"Chris Hemsworth";"1983-08-11"
"nm0749263";"Mark Ruffalo";"1967-11-22"
"nm0262635";"Chris Evans";"1981-06-13"
"nm1321655";"Christopher Markus";None
"nm1321656";"Stephen McFeely";None
"nm0498278";"Stan Lee";"1922-12-28"
```

- ***users.csv***: guarda la información propia completa de cada usuario de IMDB votante de las películas:

```
"id";"name"
"ur7108599";"kjames-26542"
"ur4849028";"shawneofthedead"
"ur1210420";"garethvk"
"ur5730393";"kevintgeisler"
"ur7766152";"blparker-31738"
"ur5004201";"mto10"
"ur8878081";"BiiivAL"
"ur8698769";"oliverdimitri"
"ur5627071";"eden-rabatsch"
"ur3518490";"hawkins_saints_rock"
"ur7325313";"milleniumlogan"
"ur6778625";"pjgs200"
"ur8324652";"upashnafuentes"
```

- ***person_movie.csv***: guarda la relación entre personas y películas, esto es: en qué películas ha participado cada persona y bajo qué tipo de cargo o rol:

```
"id_person";"id_movie";"rol"
"nm0000375";"tt6534532";"Actor"
"nm0000375";"tt2461172";"Actor"
"nm0000375";"tt4154796";"Actor"
"nm0000375";"tt2250912";"Actor"
"nm0000375";"tt3498820";"Actor"
"nm0000375";"tt1872194";"Actor"
"nm0000375";"tt1300854";"Actor"
"nm0000375";"tt1515091";"Actor"
```

- **user_*movie.csv***: guarda la relación entre los usuarios y las películas, es decir, las películas que cada usuario ha votado y qué puntuaciones ha dado:

```
"id_user";"id_movie";"global";"rate"
"ur7108599";"tt4154756";8.5;10.0
"ur4849028";"tt4154756";8.5;10.0
b'ur1210420';"tt4154756";8.5;10.0
"ur5730393";"tt4154756";8.5;10.0
"ur7766152";"tt4154756";8.5;10.0
"ur5004201";"tt4154756";8.5;10.0
"ur8878081";"tt4154756";8.5;10.0
"ur8698769";"tt4154756";8.5;10.0
"ur5627071";"tt4154756";8.5;10.0
```

3. Contribuciones

Contribuciones	Firma
Investigación previa	G. A G. M, M. E. G. M.
Redacción de las respuestas	G. A G. M, M. E. G. M.
Desarrollo	G. A G. M, M. E. G. M.