Universidad Rey Juan Carlos

Práctica 1: Recopilación, estructuración y análisis de datos

El objetivo de esta primera práctica es el de profundizar en algunos de los conceptos vistos en las sesiones de teoría. La situación que se plantea es la siguiente: acabas de entrar en una gran compañía que proporciona servicio de consumo de series y películas bajo demanda. Tu labor dentro de la compañía será analizar algunos procesos del **modelo de negocio** y desarrollar un pequeño **sistema de información**.

PRIMERA TAREA:

En la empresa tienen definidos una gran variedad de procesos, pero para esta práctica nos interesa el proceso mediante el cual se decide si se va a producir una serie, o película, o no.

Este proceso de negocio por el cual se evalúa la producción de una serie o película comienza cuando el guionista de la posible serie, o película, manda al servicio de producción la sinopsis de la película, o el guion de los tres primeros episodios (en caso de tratarse de una serie). En ese momento, ese email se reenvía al director de producción que se encargará de leer dicho guion y decidir si continua adelante o no. Si el director no da su visto bueno, el servicio de producción contestará al email del cliente con su negativa. En caso contrario, será el propio director el que mandará el guion, o sinopsis, a los siguientes departamentos, o servicios, para que realicen las siguientes tareas:

1. Servicio de guionistas. La compañía tiene un servicio de personas encargadas de evaluar cada guion recibido. En este punto se evalúan varios aspectos: 1) si la historia es atractiva y engancharía a los espectadores, y 2) si el guion contiene sorpresas o giros en la historia, y 3) si la duración de la propuesta se corresponde con la dirección del guion. En caso de que se satisfagan estas tres condiciones, el servicio de guionistas daría su visto bueno.
2. El segundo departamento que recibe el guion será el de marketing y tiene como tarea proponer un presupuesto de toda la campaña de marketing para promocionar esa nueva serie o película. En concreto, el presupuesto dependerá de si se hará una estrategia de marketing a nivel nacional o internacional.
3. Por último, el departamento de producción evaluará los costes que tendría la compañía en producir dicha serie, o película. Este coste, dependerá de si se va a rodar por completo en ubicaciones extranjeras, si se va a rodar por completo en España, o si se va grabará usando localizaciones nacionales o internacionales. Dependiendo de esto, el coste de producción será diferente.

Cuando estos tres departamentos hayan terminado sus correspondientes tareas le mandarán su conclusión al director. Si el director recibe un “Sí” por parte de los guionistas, calculará el precio final que le cobrará al cliente teniendo en cuenta los costes que le han dicho tanto marketing como el departamento de producción.

Finalmente, este precio final se lo comunicará al servicio de producción para que contesten al cliente con el precio final.

La primera tarea dentro de esta compañía será la de **modelar**, usando **BPMN** y **UML**, este proceso de negocio que se acaba de describir.

SEGUNDA TAREA:

La segunda tarea que te han encomendado es la de desarrollar, una primera versión de un **sistema de información gerencial** o MIS (*Management Information System*).

Para ello, se te proporciona un archivo CSV extraído de la actual base de datos de películas y series. Tu labor será la de desarrollar, usando el lenguaje de programación con el que te sientas cómodo y una base de datos MySQL, este pequeño MIS que genere la información que se detalla en los diferentes ejercicios de este enunciado (mostrados a continuación). Además, se pondrá en práctica el uso de librerías pensadas para realizar un análisis de datos que permita generar información interesante para un potencial usuario final, cliente o, simplemente, alguien interesado en recibir la información que se obtiene tras el filtrado y análisis de los datos proporcionados.

Para aquellos alumnos que no tengan una preferencia clara sobre el lenguaje de programación a utilizar se les recomienda utilizar Python en un cuaderno de Jupyter junto con la librería Pandas para el tratamiento y manipulación de los datos.

Es importante que el alumnado realice esta primera práctica con éxito y de forma adecuada, ya que la segunda práctica se basará en gran medida en los resultados, o avances alcanzados en esta práctica. La práctica se realizará **en grupos de hasta 4 personas**.

CONJUNTO DE DATOS:

Para esta práctica, utilizaremos un dataset obtenido de Kaggle, una web que contiene miles de dataset para realizar diferentes tareas, entre ellas análisis de datos, aplicación de técnicas de machine learning, etc. En concreto, utilizaremos un dataset llamado *netflix_titles.csv,* que contiene **12 columnas** correspondientes a información de títulos (películas y series) contenidas en Netflix. El fichero está disponible como un adjunto en aula virtual.

EJERCICIOS A REALIZAR:

Ejercicio 1 [2 puntos]

En este primer ejercicio, el grupo deberá desarrollar el modelado del proceso de negocio descrito anteriormente (PRIMERA TAREA) usado las dos notaciones vistas en teoría: *Business Process Modeling Notation* **(1 punto)** y *Unified Modeling Language* **(1 punto)**

Ejercicio 2 [1 punto]

El objetivo de esta tarea será el de desarrollar un sencillo sistema de información. Para ello partimos de los datos contenidos en el archivo CSV. Debemos diseñar las tablas en la base de datos y desarrollar los códigos necesarios para leer los datos del fichero CSV y almacenarlos en la base de datos.

Después, será necesario leer los datos desde la BBDD (usando diferentes consultas) y se almacenarán los resultados apropiadamente en una tabla para poder manipularlos.

En este ejercicio, para el correcto desarrollo del sistema MIS, será necesario calcular los siguientes valores:

a) Número de muestras (valores distintos de *missing*).
b) Media y desviación estándar de la columna duración.
c) Valor mínimo y valor máximo de las columnas duración y año.

Ejercicio 3 [2.5 puntos]

Hay datos que nos interesa analizar basándonos en agrupaciones, para darle un sentido a nuestro análisis en base a esa agrupación. De una manera más específica, vamos a trabajar con las siguientes agrupaciones:

a) Por tipo de contenido (serie o película)
b) Número de temporadas en series: estableceremos dos rangos diferentes, el primero aquellos contenidos que tengan 1 o 2 temporadas, y aquellos que tengan más de 2.
c) Duración en películas: estableceremos dos rangos diferentes, el primero aquellos contenidos que tengan una duración de más de 90 minutos, y aquellos que tengan una duración inferior.

En este caso deberemos calcular la siguiente información para la variable *Duración*:

a) Número de observaciones
b) Número de valores ausentes (*missing*)
c) Mediana
d) Media
e) Varianza
f) Valores máximo y mínimo

Ejercicio 4 [2 puntos]

Una vez que tenemos nuestros datos divididos y estructurados, debemos simular la interacción de los usuarios con la plataforma. Para ello, se debe crear una tabla nueva con un número razonable de usuarios (pueden ser entre 20 y 30), cuyos datos almacenados sean:

a) Id de usuario
b) Nombre de usuario
c) Fecha de último inicio de sesión

Hay que tener en cuenta que cada usuario habrá visto al menos una película o serie, y que existirán usuarios que hayan visto más de una, lo que debe contemplarse, generando una tabla extra para registrar el/los visionado/s, que contenga:

d) Id del visionado
e) Id de la Película/Serie visionada (**ha de ser el mismo que el de la tabla original para poder operar con ambas tablas en un futuro**)
f) Id del usuario que la ha visionado (**ha de ser el mismo que el de la tabla anterior para poder operar con ambas tablas en un futuro**)
g) Puntuación otorgada al contenido (película/serie)

Estos datos pueden rellenarse de forma aleatoria, pero hay que tener especial cuidado en no generar duplicados u otras cuestiones que puedan ensuciar el conjunto de datos.

Ejercicio 5 [2.5 puntos]

Por último, se programarán las diferentes funciones del MIS. En concreto, se deben generar gráficos sencillos para obtener los siguientes datos:

a) Mostrar las 10 películas con más visionados, representadas en un gráfico de barras.
b) Mostrar las 10 series con más visionados, representadas en un gráfico de barras.
c) Mostrar la media de visionados de las películas de más de 90 minutos frente a las de menos de 90 minutos.
d) Mostrar la media de visionados de las series de más de 2 temporadas frente a las de 1 ó 2 temporadas.

NORMAS DE ENTREGA:

La entrega de la práctica consistirá en un archivo comprimido con los siguientes ficheros:

a) Archivo txt con los nombres y apellidos de los integrantes del grupo
b) Código fuente en el lenguaje de programación elegido
c) Archivo db con la base de datos creada por los alumnos
d) Memoria en formato PDF en la que se muestren los ejercicios resueltos
e) Las dos imágenes (pdf, o jpg) que contengan los diagramas del modelo de negocio
f) Cualquier otro fichero de texto plano con instrucciones de compilación/ejecución que los alumnos consideren necesarias

La fecha límite para entregar esta práctica será el 24 de marzo a las 23:55 y se realizará por la plataforma Aula Virtual.

PESO DE LA PRÁCTICA:

La evaluación de esta práctica supondrá un 20% de la nota total de la asignatura.