Web de recomendaciones de películas (basada en MovieLens)

Objetivos

Trabajar en **parejas** para desarrollar una página web **original** que, utilizando Filtrado Colaborativo *User-User* e *Item-Item,* muestre recomendaciones personalizadas de películas.

Todo el trabajo realizado por los estudiantes debe ser original, es decir, el código y la documentación deben haber sido realizados por los dos miembros de la pareja, y no proceder de otras fuentes y/o grupos de trabajo.

Descripción del trabajo a realizar

Para lograr el objetivo descrito, descompondremos el trabajo entre varios elementos y tareas:

1. Fuentes de datos

Se trabajará con datos procedentes de MovieLens (movielens.org). En concreto con el conjunto de datos (dataset) recomendado para propósitos educativos y de desarrollo. Seleccionaremos inicialmente el dataset completo (full), caracterizado por:

- 27 millones de valoraciones
- 280.000 usuarios
- 58.000 películas
- URL: https://grouplens.org/datasets/movielens/latest/

La fuente de datos está compuesta por múltiples ficheros, peor para nuestro propósito debería bastarnos con los siguientes archivos:

- movies.csv
- ratings.csv
- tags.csv
- links.csv

Para conocer los detalles de los datos y formatos de cada fichero, es necesario estudiar el fichero "README.txt" incluido en el ZIP descargado: ml-latest.zip.

2. ETL [Tarea 1]

Una vez analizados los ficheros, los datos necesarios deberán cargarse en una base de datos, con la que trabajará la web.

El proceso por el cual se transforman y cargan los datos desde los ficheros hasta la BD es libre, pero debe hacerse de manera **automática**. Es decir, debe implementarse

un programa (Java, PHP, Python, etc.) que procese el fichero y cargue la información en una BD en **MySQL** o **MariaDB**.

Este proceso debe ser **fácilmente reproducible** por parte del profesor de cara a la evaluación de la actividad con un conjunto diferente de datos. Resultará por tanto muy importante su correcta y detallada documentación. Por eso, se recomienda trabajar con el entorno **XAMPP**.

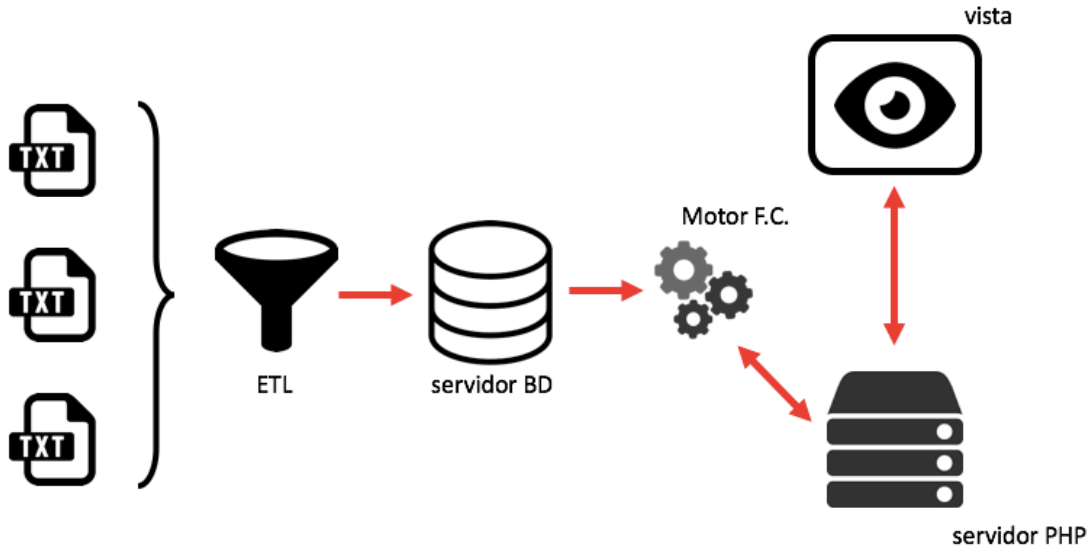

Figura 1. Esquema de la aplicación

3. Web de recomendaciones

El sistema a desarrollar será por tanto una **página web dinámica PHP con conexión a una base de datos** MySQL/MariaDB, por lo que habrá que gestionar una arquitectura de 3 capas (Vista-Servidor-Datos) que puede organizarse fácilmente con el citado XAMPP.

4. Recomendaciones User-User [Tarea 2]

Nuestra web deberá mostrar recomendaciones personalizadas en base al filtrado colaborativo basado en usuarios. Para ello se debe calcular la similitud entre usuarios utilizando la expresión de **Pearson**, y posteriormente tendremos que seleccionar un vecindario para hacer las predicciones.

Los/las estudiantes podrán seleccionar el criterio que estimen oportuno para configurar dicho vecindario. Este debe estar **bien documentado**, pero inicialmente se puede tomar como criterio base la utilización de aquellos usuarios con una similitud superior a cero (>0).

Lo ideal sería que dentro de la propia web se pudiese elegir ese umbral **[Tarea 3]**

Tras esto, se presentarán los siguientes datos para el usuario U seleccionado:

4.1. Ofrecer el **ranking de las N películas** con mejor predicción para el usuario U. Inicialmente N=5.
Idealmente debería ser también un parámetro configurable desde dentro de la propia página. **[Tarea 3]**

4.2. Ofrecer la **predicción** para una película P seleccionada de entre las que no ha visto el usuario U.
Idealmente la película debería poderse buscar con un buscador en función de su título.

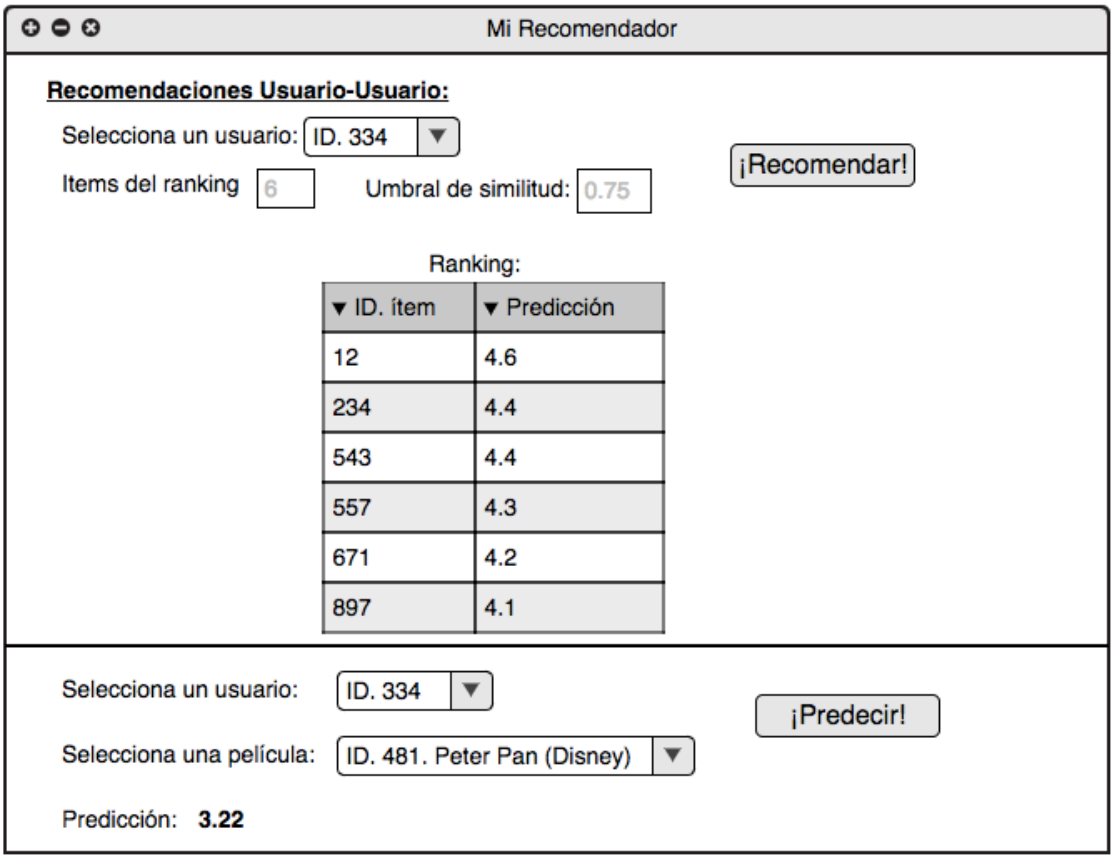

Figura 2. Posible vista para las recomendaciones Usuario- Usuario o Ítem-Ítem

5. Recomendaciones Item-Item [Tarea 5]

La web podrá mostrar también recomendaciones personalizadas basadas en el filtrado colaborativo entre ítems. Para ello deberá, igual que antes, calcularse la similitud entre ítems, y posteriormente las predicciones en base al vecindario seleccionado. Para el cálculo de la similitud se utilizará la expresión del **coseno ajustado.**

Las indicaciones anteriores para la selección del vecindario aplican igualmente para este tipo de recomendaciones.

De nuevo, lo ideal sería que dentro de la propia web se pudiese elegir ese umbral **[Tarea 3]**

Igualmente, el objetivo es ofrecer dos resultados:

5.1. El **ranking de las N películas** con mejor predicción para el usuario U (inicialmente N=5).
Como antes, lo ideal debería ser que N fuese un parámetro configurable desde la propia web. **[Tarea 3]**

5.2. La **predicción** para una película P seleccionada de entre las que no ha visto el usuario U.
Idealmente la película debería poderse buscar con un buscador en función de su título.

6. Gestión de un nuevo usuario [Tarea 6]

Como última tarea, los/las estudiantes podrán optar por incluir un "usuario 0", a través del cual se podrían valorar películas. Estas valoraciones se registrarían en la BD y por tanto permitirían ofrecer:

6.1. Un **ranking de las N películas** con mejor predicción para el usuario 0 (inicialmente N=5).
Como antes, lo ideal debería ser que N fuese un parámetro configurable desde la propia web. **[Tarea 3]**

6.2. La **predicción** para una película P seleccionada de entre las que no ha visto el usuario 0.
Idealmente la película debería poderse buscar con un buscador en función de su título.

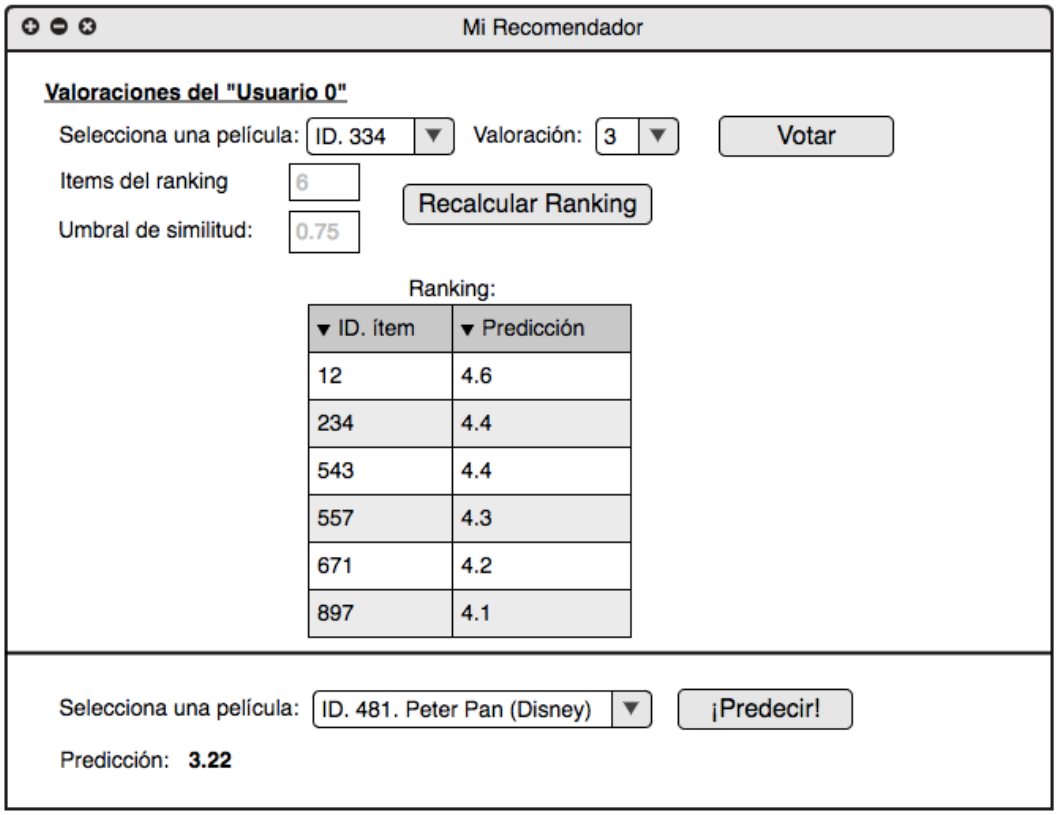

Figura 3. Posible vista para la gestión de votos y visualización de resultados del "Usuario 0".

Normas y forma de entrega

Ambos miembros de la pareja deberán adjuntar a través del Campus Virtual el resultado de su trabajo. La entrega se realizará a través de un único fichero comprimido en formato **ZIP** (no RAR). Deben seguirse las siguientes indicaciones:

Nombre del fichero: Apellido1.Nombre1_Apellido2.Nombre2.SSII.Práctica.**ZIP**
Contenido del fichero:

1. **Capeta Documentación:**
   - Fichero PDF con todos los detalles de la actividad.
   - Nombre del Fichero: Apellido1.Nombre1_Apellido2.Nombre2.SSII.Práctica.**PDF**
   - Debe incluir, al menos, los siguientes apartados:
     - Portada
     - Índice
     - Descripción de la base de datos
       - Definición de las tablas y sus atributos (nombres de columnas, tipos, etc.).
       - Datos de conexión con la BD (servidor, usuario, contraseña, etc.).
       - Scripts de la creación de las tablas.
       - Script de carga de datos.
     - Descripción del proceso de carga de datos.
     - Recomendaciones User-User:
       - Descripción y justificación de las funciones utilizadas, junto con sus seudocódigos.
     - Recomendaciones Item-Item:
       - Descripción y justificación de las funciones utilizadas, junto con sus seudocódigos.
     - Descripción del portal web:
       - URL de acceso (si existe).
       - Mapa de la página web desarrollada.
       - Breve manual de usuario.
     - Manual de instalación:
       - Describir los pasos a seguir para replicar en un entorno local el sistema creado por los/las estudiantes. **Este apartado es vital para la corrección de la actividad**.
       - Incluir:
         - Scripts para replicar la BD.
         - Usuarios y contraseñas.
         - Etc.
     - Conclusiones/comentarios.
     - Bibliografía

2. **Carpeta BD:**
   - Fichero txt (extensión .sql) con el script de creación de la BD.

- Fichero txt (extensión .sql) con el script de carga de datos en la BD.

3. **Carpeta Código:**
   - Código fuente de los diferentes ficheros necesarios para el funcionamiento de la página web.

Calificación

La calificación de la actividad dependerá de los siguientes aspectos:

- Claridad y calidad de la documentación.
- Claridad y calidad del código fuente.
- Claridad y calidad del manual de instalación.
- Trabajo en equipo.
- Entrega en tiempo y forma según las normas de entrega.
- Grado de consecución de las funcionalidades especificadas en la tabla de tareas que figura a continuación:

Tarea 1: ETL

Tarea 2: Web dinámica con recomendaciones User-User

Tarea 3: Umbrales configurables:

- Poder seleccionar el umbral de similitud al elegir los usuarios que formarán el vecindario.
- Poder elegir cuántos ítems se mostrarán como máximo al ofrecer el ranking de películas (N configurable)

Tarea 4: Cruzar datos de usuarios: utilizar la información contenida en otros ficheros de MovieLens para hacer la web más amigable y usable, mostrando otros datos, como los títulos de las películas, por ejemplo.

Tarea 5: Web dinámica con recomendaciones Ítem-Ítem

Tarea 6: Gestionar el "usuario 0", la inserción de sus valoraciones y sus recomendaciones personalizadas

Calificación máxima	Tarea 1	Tarea 2	Tarea 3	Tarea 4	Tarea 5	Tarea 6
5 puntos	•	•				
6 puntos	•	•	•			
7 puntos	•	•	•	•		
8 puntos	•	•	•	•	•	
10 puntos	•	•	•	•	•	•