|  |
| --- |
| **ACTIVIDAD 2.3 – Ejercicio Teórico**  **Anteproyecto de minería de datos web** |
| Imagina que decides montar un sistema de minería de datos web basado en un motor predictivo de Machine Learning.  Este es un ejercicio teórico. Explica en una **extensión máxima de 1 página** (aproximado) tu idea de proyecto, intentando contestar en la medida de lo posible a cuantas puedas de las siguientes preguntas. No es necesario que escribas nada de código, pero si lo ves necesario puedes comentar algunas de las tecnologías que usarías también. |
| Nombre Alumno: David Burgos; Demostenes Garcia; Edwin Sisalema |
| * ¿Cuál es el objetivo de tu proyecto? * ¿Qué tipo de minería web usarás: de contenido, de utilización, ¿de estructura? * ¿Cuáles serían tus fuentes de datos? ¿Qué web o conjunto de webs usarías para extraer datos? * ¿Usarías una web de un dominio propio o de tu empresa, o usarías webs externas? * ¿Qué tipo de datos quieres minar: numéricos, textos, ¿imágenes? * Señala la web concreta y pon ejemplos concretos de datos en esa web querrías minar * ¿Prevés algún problema en la adquisición de los datos? * ¿Con qué frecuencia tendrías que hacer el minado de datos? * ¿Qué cantidad de datos esperas manejar para crear tu dataset de entrenamiento? * ¿Y una vez esté el sistema en funcionamiento? * ¿Qué tipo de sistema predictivo (de Machine Learning) propones? * ¿Qué cosa te interesa predecir? * ¿Vas a usar sistemas de clustering, clasificación, regresión, análisis de textos? * ¿Cuáles crees que son las métricas adecuadas para validar tu sistema de minería web? * Más allá de este ejercicio teórico, ¿cómo de viable crees que sería tu propuesta de sistema en la vida real? |
| **Propuesta de Projecto**  A continuación, describe tu proyecto:  El objetivo principal de este proyecto es desarrollar un sistema predictivo basado en Machine Learning para analizar las preferencias cinematográficas de los últimos 30 años y utilizar estos patrones para ayudar a las empresas cinematográficas a producir películas que se alineen con los intereses del público. Esto les permite tomar decisiones más informadas en la producción de nuevas películas.  Para lograr esto, se empleará una estrategia de minería de contenido, que incluirá el análisis detallado de información relacionada con películas, como géneros, directores, actores, críticas y otros elementos influyentes. Las fuentes de datos podrían ser páginas web de reseñas de películas y bases de datos de películas como IMDb y Rotten Tomatoes, así como datos de redes sociales como Twitter. Además, se explorará la posibilidad de realizar encuestas en línea y la creación de un sitio web específico para recopilar datos personalizados.  Se anticipa la necesidad de trabajar con datos tanto textuales como numéricos. Esto implica analizar reseñas, sinopsis, opiniones y datos cuantitativos como clasificaciones e ingresos de taquilla. La minería se llevará a cabo en webs externas para obtener datos públicos o en este caso información internar debido a la creación del sitio web para recopilar datos personalizados.  Algunos problemas que podríamos enfrentar durante la adquisición de datos podrían ser el manejo de grandes volúmenes de datos, la limpieza de los datos y la posible necesidad de lidiar con las limitaciones de la API o el bloqueo de IP durante la extracción de datos. La frecuencia del minado de datos podría variar, pero una actualización mensual o trimestral podría ser un buen punto de partida, ya que se busca capturar y adaptarse a las tendencias en la industria cinematográfica, que está en constante cambio.  Para el conjunto de datos de entrenamiento, podríamos comenzar con unas pocas decenas de miles de películas. Una vez que el sistema esté en funcionamiento, podríamos estar manejando datos de cientos de miles de películas. Podríamos utilizar un sistema de clasificación de Machine Learning para predecir el éxito de una película basándonos en ciertos atributos, como el género, los actores y el director. También, el análisis de texto será fundamental para entender las preferencias y comentarios de los usuarios.  Las predicciones deseadas se centrarán en predecir la recepción y el éxito de una película, proporcionando información clave para respaldar la toma de decisiones en la producción cinematográfica. Las métricas que podríamos utilizar para validar nuestro modelo podrían ser la precisión, la sensibilidad, la especificidad y el área bajo la curva ROC.  Este proyecto es bastante viable en la vida real. Las empresas cinematográficas ya están utilizando análisis de datos similares para tomar decisiones sobre la producción de películas. Sin embargo, la implementación de este proyecto requeriría una inversión significativa en términos de tiempo y recursos. |