Artículo Original Pág. 35 - 46 ISSN 2415-2323

ILUMATE VOLUMEN 12

EXPLORACIÓN DE DATOS Y SISTEMA DE RECOMENDACIÓN DE LIBROS

Data Exploration & Book's System Recommendation

Lozada Gómez José Adolfo
jalgoz95@gmail.com
Instituto de Investigación en Ciencia y Tecnología,
Universidad La Salle-Bolivia

Arana Sánches Esmeralda Yennifer
aranaesmeralda7@gmail.com
Instituto de Investigación en Ciencia y Tecnología,
Universidad La Salle-Bolivia

Resumen

La presente investigación se realizó con la intención de aplicar los conocimientos adquiridos del área de Aprendizaje Automático en las publicaciones de libros en la red. El Objetivo fue analizar las propiedades de cada libro (libros más leídos, autores más consultados y calificación promedio de un autor o libro) y las relaciones que pueden tener dentro de un conjunto de datos. En base al análisis, añadir un sistema de recomendación para el uso de los usuarios. La metodóloga empleada fue cuantitativa de diseño de investigación no experimental tipo de estudio correlacional. Los resultados fueron, que utilizando el Machine Learning se pudo analizar los datos y así

poder brindar un sistema de recomendación.

Palabras claves

Análisis de datos, Relación de datos, Kmeans Clustering, KNN.

Abstract

This research was carried out with the intention of applying the knowledge acquired from the area of Machine Learning in the publication of books on the web. The objective was to analyze the properties of each book (most read books, most consulted authors and average rating of an author or book) and the relationships they may have within a data set. Based on the analysis, add a recommendation system for user use. The methodology used was quantitative non-experimental research design type of correlational study. The results were that using Machine Learning it was possible to analyze the data and thus be able to provide a recommendation system.

Key words

Data analysis, Data relations, Kmeans Clustering, KNN.

Introducción

La principal idea del proyecto es analizar y reconocer las propiedades más importantes con las que cuenta cada libro del dataset, encontrar las relaciones existentes dentro, en este caso información sobre los libros existentes, explorarlo y verificar que los datos sean correctos y estén libres de anomalías para su próxima ejecución en un sistema de recomendación.

Para la solución del problema, se debe reconocer el tamaño y las columnas del dataset para una noción básica de su estructura, utilizando la librería de NumPy. A continuación, seguir con la exploración de los datos usando las

gráficas de la librería Seaborn. Para utilizar los datos los libros se dividirán en grupos usando los métodos de Kmeans y Neighbors de la librería Sklearn que nos ayudarán a identificar al tipo de preferencia según las búsquedas y gustos del usuario.

Referentes Conceptuales

El aprendizaje automático es un paradigma que puede referirse al aprendizaje de experiencias pasadas (que en este caso son datos previos) para mejorar el rendimiento futuro. El único enfoque de este campo son los métodos de aprendizaje automático. El aprendizaje se refiere a modificación o mejora del algoritmo basado en "experiencias" pasadas automáticamente sin ninguna ayuda externa de humano. (Das & Narayan Behera, 2017).

El aprendizaje automático es la ciencia de enseñar a las computadoras a hacer predicciones basadas en datos. En un nivel básico, el aprendizaje automático implica dar a una computadora un conjunto de datos y pedirle que haga una predicción. Al principio, la computadora tendrá muchas predicciones incorrectas, sin embargo, en el transcurso de miles de predicciones, la computadora actualizará su algoritmo para hacer mejores predicciones (Norman, 2017, pág. 60).

Tipos de aprendizaje según (Cambronero, 2006):

> **Aprendizaje inductivo:** Creamos modelos de conceptos a partir de generalizar ejemplos simples. Buscamos patrones comunes que expliquen los ejemplos. Se basa en el razonamiento inductivo: Obtiene conclusiones generales de información especifica. El conocimiento

obtenido es nuevo. No preserva la verdad (nuevo conocimiento puede invalidar lo obtenido). No tiene una base teórica bien fundamentada.

Aprendizaje analítico o deductivo: Aplicamos la deducción para obtener descripciones generales a partir de un ejemplo de concepto y su explicación. Se basa en el razonamiento deductivo: Obtiene conocimiento mediante el uso de mecanismos bien establecidos. Este conocimiento no es nuevo (ya está presente implícitamente). Nuevo conocimiento no invalida el ya obtenido. Se fundamenta en la lógica matemática.

Aprendizaje analógico: Busca soluciones a problemas nuevos y encontrar similaridades con problemas ya conocidos y adaptando sus soluciones.

Aprendizaje genético: Aplica algoritmos inspirados en la teoría de la evolución para encontrar descripciones generales a conjuntos de ejemplos. Aprendizaje conexionista: Busca descripciones generales mediante el uso de la capacidad de adaptación de redes de neuronas artificiales.

Uno de los objetivos que persigue el Machine Learning es clasificar automáticamente los objetos o datos. En base a esto podríamos mencionar según (APD, 2019) a tres clases de algoritmos:

Aprendizaje supervisado

En el aprendizaje supervisado, la máquina se enseña con el ejemplo. De este modo, el operador proporciona al algoritmo de aprendizaje automático un conjunto de datos conocidos que incluye las entradas y salidas deseadas, y el algoritmo debe encontrar un método para determinar cómo llegar a esas entradas y salidas.

Mientras el operador conoce las respuestas correctas al problema, el algoritmo identifica patrones en los datos, aprende de las observaciones y hace predicciones. El algoritmo realiza predicciones y es corregido por el operador, y este proceso sigue hasta que el algoritmo alcanza un alto nivel de precisión y rendimiento.

Aprendizaje sin supervisión

Aquí, el algoritmo de aprendizaje automático estudia los datos para identificar patrones. No hay una clave de respuesta o un operador humano para proporcionar instrucción. En cambio, la máquina determina las correlaciones y las relaciones mediante el análisis de los datos disponibles.

En un proceso de aprendizaje no supervisado, se deja que el algoritmo de aprendizaje automático interprete grandes conjuntos de datos y dirija esos datos en consecuencia. Así, el algoritmo intenta organizar esos datos de alguna manera para describir su estructura. Esto podría significar la necesidad de agrupar los datos en grupos u organizarlos de manera que se vean más organizados.

A medida que evalúa más datos, su capacidad para tomar decisiones

sobre los mismos mejora gradualmente y se vuelve más refinada.

Aprendizaje por refuerzo:

El aprendizaje por refuerzo se centra en los procesos de aprendizajes reglamentados, en los que se proporcionan algoritmos de aprendizaje automáticos con un conjunto de acciones, parámetros y valores finales.

Al definir las reglas, el algoritmo de aprendizaje automático intenta explorar diferentes opciones y posibilidades, monitorizando y evaluando cada resultado para determinar cuál es el óptimo.

Para este caso se utilizó el algoritmo de clasificación no supervisada **k-means**. A continuación, algunos conceptos de las librerías que se utilizaron.

NumPy, o Python numérico, es una biblioteca basada en Python para cálculos matemáticos y matrices de procesamiento. Python no admite estructuras de datos en más de una dimensión, y los contenedores como listas, tuplas y diccionarios son unidimensionales. Los contenedores y tipos de datos incorporados en Python no se pueden reestructurar en más de una dimensión y tampoco se prestan a cálculos complejos. Estos inconvenientes son limitaciones para algunas de las tareas involucradas al analizar datos y construir modelos, lo que hace que los arreglos sean una estructura de datos vital. (Rajagopalan, 2020).

Pandas es una librería que proporciona estructuras de datos y herramientas de análisis de datos de alto rendimiento para el lenguaje de programación Python (Vidal, 2019).

La librería Sklearn sirve para ilustrar el manejo practico de los algoritmos de aprendizaje automático, además implementa numerosos algoritmos de aprendizaje automático, así como diversos procedimientos relacionados y, junto con la característica sencillez de Python, proporciona un entorno rápido y cómodo para su uso (Jose Manuel Robles, 2020).

El algoritmo K-means, creado por MacQueen en 1967 es el algoritmo de clustering más conocido y utilizado ya que es de muy simple aplicación y eficaz. Sigue un procedimiento simple de clasificación de un conjunto de objetos en un determinado número K de clústeres, K determinado a priori (Cambronero, 2006).

El análisis de clúster o Clustering agrupa un conjunto de objetos de datos en clústers o grupos de manera que en cada grupo los objetos sean similares entre sí y disimiles de los objetos de otros grupos. En la actualidad, existen distintas técnicas de agrupamiento que permiten cumplir con esta tarea. En búsqueda de un algoritmo más natural se hizo uso del concepto de densidad atómica de los elementos como base para generar uno nuevo. El algoritmo propuesto tiene como ventajas, poseer un método concreto de selección de centroides, además de tener mejores agrupamientos que otros algoritmos basados en centroides como k-means y k-medoids (Medina, 2016).

El algoritmo consta de tres pasos según (Laruta, 2020):

1. **Inicialización:** Una vez hayamos seleccionados el número de grupos, k, estableceremos k centroides en el espacio de los datos pudiendo escogerlos de forma aleatoria.

2. **Asignación objetos a los centroides:** Cada objeto de los

datos deberá ser asignado al centroide más cercano que se encuentre.

3. **Actualización centroides:** Se debe actualizar la posición que tiene cada centroide tomando como nuevo centroide la posición promedio de los objetos pertenecientes a dicho grupo.

La asignación de objetos a los centroides y la actualización de centroides va a ser repetitiva hasta que los centroides no se muevan o caso contrario se muevan por debajo de una distancia umbral en cada paso.

Este algoritmo se lo podría definir como una función de optimización de la suma de las distancias de cada objeto al centroide de su clúster.

Tiene varias ventajas entre las principales están que es un método rápido y sencillo, sin embargo, para ello se requieres decidir el valor de K y el resultado final que se obtenga dependerá mucho de donde estén iniciados los centroides.

Metodología

La metodóloga empleada fue cuantitativa de diseño de investigación no experimental tipo de estudio correlacional.

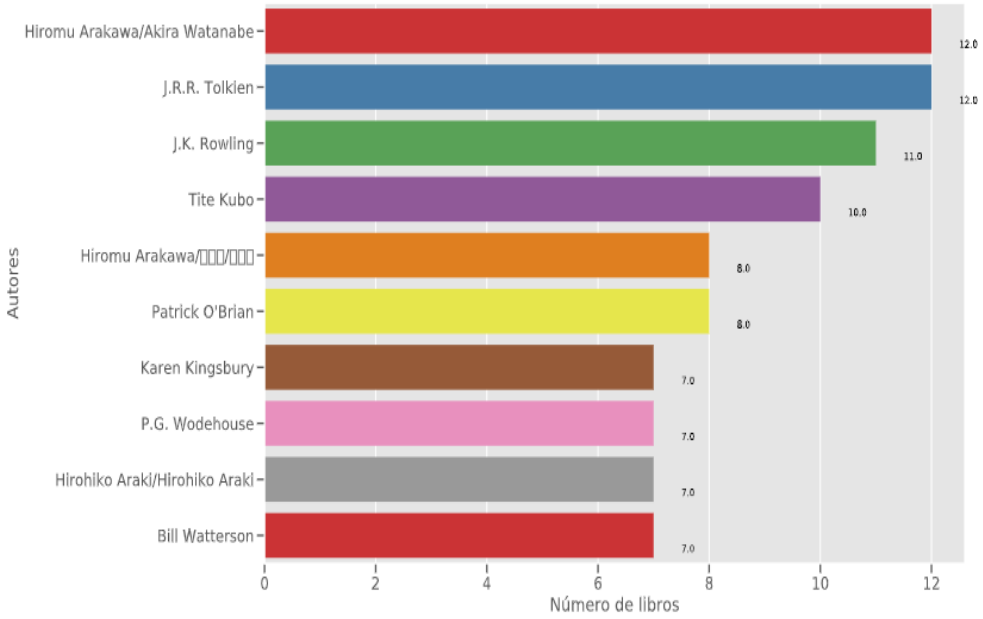
Los instrumentos utilizados fueron:

- Lenguaje de programación Python y las librerías Numpy, Pandas, Seaborn y Sklearn.
- Editor de código Visual Studio Code

Resultados y discusión:

Como resultado se pudo encontrar algunas ambigüedades en la relación de datos, como ser la calificación promedio y el número de reseñas, debe haber persistencia al cambio de datos en las columnas y tendremos como resultado un dataset que esté libre de anomalías, así como una exploración gráfica e informativa en la división de grupos en los libros, y la buena ejecución en el sistema de recomendación.

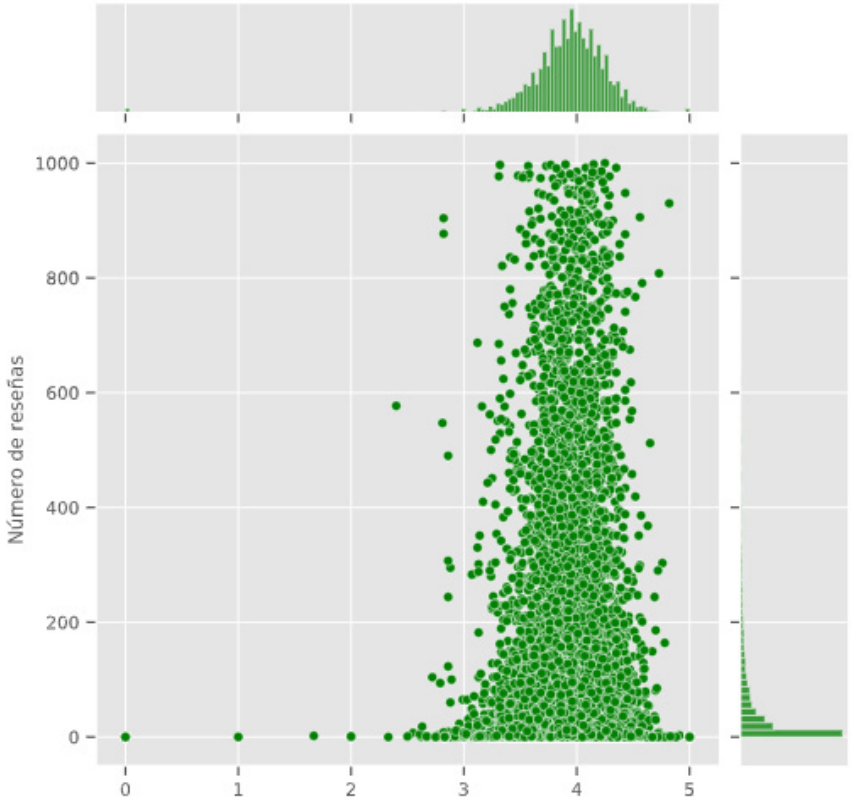
Fig 1. Cantidad de libros



Fuente Elaboración propia (VS Code, 2020)

La figura 1 muestra la cantidad de libros que existen por autor en el software que se hizo correr.

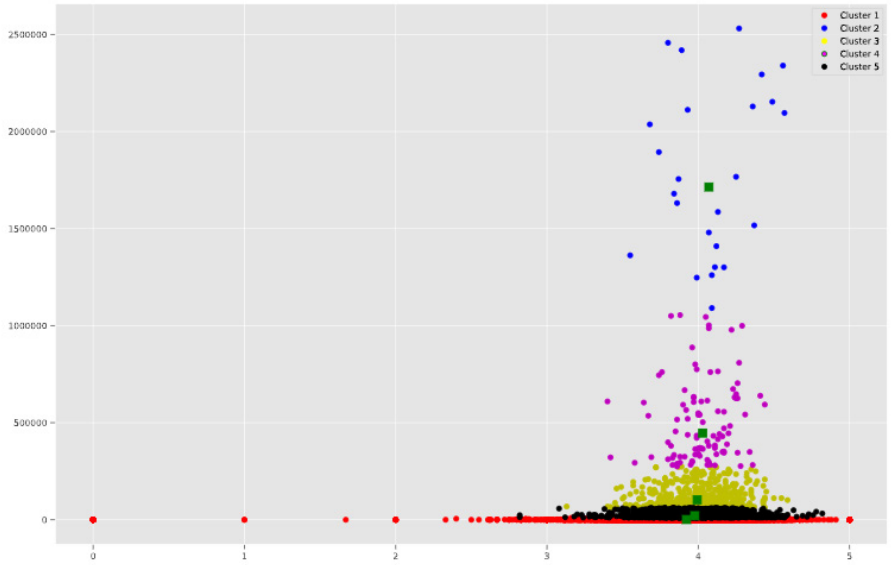
Fig 2. Número de reseñas y calificación promedio



Fuente Elaboración propia (VS Code, 2020)

No sé puede afirmar o negar si existe una relación entre el número de reseñas y la calificación promedio

Fig 3. División en clústeres del dataset



Fuente Elaboración propia (VS Code, 2020)

Conclusiones

Se logra ver que todas las herramientas que ofrece python al momento de un análisis de los datos como también para los diversos métodos de aprendizaje automático son potentes, fáciles de usar y entender, obteniendo una gran información sobre el conjunto de datos ya trabajados, dando a los desarrolladores poder manipular y usarlos a un beneficio final, en este caso para un sistema de recomendación.

Referencias

- APD, R. (04 de 04 de 2019). apd. Obtenido de apd: https://www.apd.es/algoritmos-del-machine-learning/#:~:text=Este%20tipo%20de%20algoritmos%20por,de%20caracter%C3%ADsticas%2C%20utilizando%20la%20probabilidad.

- Cambronero, C. G. (2006). ALGORITMO DE APRENDIJAZE KNN & KMEANS. 7.
- Das, K., & Narayan Behera, R. (2 de 2017). www.ijircce.com. Obtenido de www.ijircce.com: https://www.researchgate.net/profile/Rabi-Behera-2/publication/316273553_A_Survey_on_Machine_Learning_Concept_Algorithms_and_Applications/links/59017eb94585156502a094fd/A-Survey-on-Machine-Learning-Concept-Algorithms-and-Applications.pdf
- Jose Manuel Robles, R. C. (2020). Big data para cientificos sociales. Madrid: RALI, S.A.
- Medina, O. M. (2016). SEDICI. Retrieved from http://sedici.unlp.edu.ar/handle/10915/56757
- Moya, R. (29 de 05 de 2016). Jarroba. Obtenido de Jarroba: https://jarroba.com/k-means-python-scikit-learn-ejemplos/
- Norman, A. T. (2017). Aprendizaje Automática en Acción. Tektime .
- Rajagopalan, G. (23 de 12 de 2020). SpringerLink. Obtenido de SpringerLink: https://link.springer.com/chapter/10.1007/978-1-4842-6399-0_5#citeas
- Vidal, J. H. (2019). Aprendizaje automático en el diseño de un detector de estrés a partir de señales biomédicas. Universidad de La Laguna.

Artículo recibido: 23-09-2020
Artículo aceptado: 05-11-2020