Capstone Project – Sprint 0

- **The Problem area**: What is your area of interest? Within it, what challenges or opportunities could your project address?

  The focus of my project is to develop an advanced book recommendation system that addresses the challenges faced by readers who struggle to find suitable titles after finishing a book, as well as new readers unsure of where to start. As someone who has recently embraced the joys of reading, I often find myself feeling a sense of emptiness upon completing a captivating novel, coupled with a strong desire to discover another book that resonates with my interests. Unfortunately, my attempts to search online for comparable titles often lead to overwhelming choices and unsatisfactory results.

  To tackle this issue, the project will create a user-centric recommendation system that utilizes data-driven techniques and algorithms to provide personalized suggestions based on individual reading preferences and past experiences. By leveraging existing datasets of books, genres, and user ratings, this system aims to enhance the reading experience for all users, fostering a deeper connection between readers and literature

- **The User**: Who experiences these problems? How would they benefit from the outcomes of your project?

  Many readers, particularly those who have recently developed a passion for literature or are seeking to cultivate a reading habit, encounter significant challenges in finding titles that align with their tastes. Once a reader discovers a book, genre, or writing style they enjoy, the natural inclination is to seek out similar works. However, with billions of books available and countless reviews reflecting a wide range of preferences, this search can quickly become overwhelming. This project aims to develop a personalized book recommendation system that enables users to input a book or author they have enjoyed and receive tailored suggestions for comparable titles. By streamlining the discovery process, this system will enhance the reading experience, foster deeper engagement with literature, and assist individuals in efficiently navigating the vast array of options to identify books they are likely to appreciate.

- **The Big Idea**: How can machine learning bring solutions to these areas? Research how other people have approached the problem previously. Refer to the "Intro to Capstone" slides on synapse for an overview of different machine learning approaches.

  Machine learning significantly enhances book recommendation systems by utilizing algorithms to analyze user preferences and book characteristics, thereby facilitating the discovery of books that readers are likely to enjoy. Key methodologies include collaborative filtering, which recommends titles based on the preferences of similar users, and content-based filtering, which focuses on the specific attributes of the books, such as genre and author. These methods can be combined into hybrid approaches to improve recommendation accuracy. Advanced techniques, such as matrix factorization and deep learning, enable the identification of complex patterns in user preferences and book features. Furthermore, the user feedback allows the system to continuously learn and refine its recommendations over time. Overall, machine learning provides more personalized suggestions and tailored reading experiences, leveraging data-driven insights to meet the diverse interests of readers.

- **The Impact**: What societal or business value do you anticipate your project to add? If possible, try to quantify the scale of the problem (in dollars, in CO2, in time spent, ...)

  This project aims to encourage individuals to embrace reading and remain engaged with new books, making it easier for them to discover titles that resonate with their interests, whether in fiction or non-fiction. By fostering a deeper connection to literature, the project has the potential to enhance knowledge acquisition and promote social engagement. Additionally, the recommendation system will enable online bookstores to provide more relevant book suggestions, allowing users to better gauge their interests and preferences, thereby making businesses more customer focused. Ultimately, this initiative will reduce the time spent searching for suitable books, allowing readers to dedicate more time to reading and gaining knowledge, while also providing a valuable escape from excessive screen time for both youth and adults.

- **The Data**: Identify several possible datasets in this subject area and describe them at a high level. Include references.

  Goodreads-10k Dataset: The Goodreads-10k dataset, sourced from Kaggle, comprises several interconnected files that provide a rich overview of books and user interactions on the platform. The primary file, **books.csv**, contains comprehensive metadata for each book, including unique identifiers, author names, publication years, average ratings, total ratings, reviews, and genre classifications. The **ratings.csv** file captures extensive user ratings linked to specific books, allowing for insights into user preferences and engagement. Additionally, the **book_tags.csv** file records user-generated tags assigned to books along with their frequencies, while the **tags.csv** file translates these tag IDs into their corresponding names. Collectively, these datasets offer a valuable resource for analyzing book popularity, user behavior, and categorization.

  Reference: https://github.com/zygmuntz/goodbooks-10k

  Backup Datasets

  Book Recommendation Dataset: The **books.csv** dataset provides a detailed catalog of books identified by their respective ISBNs, with invalid entries already filtered out. It includes essential content-based information such as book titles, authors, publication years, and publishers, sourced from Amazon Web Services. The **ratings.csv** file contains book rating information, with ratings expressed on a scale from 1 to 10 (indicating higher appreciation) or as implicit ratings marked by zero. Lastly, the **users.csv** file includes anonymized user IDs mapped to integers, along with demographic data such as location and age, although some fields may contain NULL values if the information is unavailable. Together, these datasets offer a comprehensive overview of books, their ratings, and user demographics.

  Reference: https://www.kaggle.com/datasets/arashnic/book-recommendation-dataset/data?select=Users.csv

  Amazon Books Review Dataset: The Amazon Books Review Dataset, sourced from Kaggle, contains information similar to the Goodreads dataset. This dataset includes two primary CSV files: book ratings and book data. The book data file features essential attributes such as the book title, author, publication date, and rating count (average rating).

In contrast, the ratings dataset provides more detailed information, including user ID, title, price, review score, review summary, and the text of the review.

This dataset is relatively recent, having been updated two years ago, and it includes reviews sourced from Goodreads. The comprehensive nature of this dataset makes it a valuable resource for analyzing reader feedback and trends in book ratings

Reference: https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews

- **The Alternative**: In a few sentences, summarise a problem in an alternative subject area that also interests you

Another intriguing area to explore is the prediction of housing market prices in Canada. This project aims to develop a model that estimates the selling price of homes by utilizing a dataset that identifies comparable properties and employs machine learning algorithms to enhance price prediction accuracy. By analyzing various features such as location, size, amenities, and historical sales data, the model will provide insights into how much a house in Canada is likely to sell for. This predictive capability will not only assist potential buyers in making informed decisions but also aid sellers in setting competitive prices. Ultimately, this initiative seeks to leverage data-driven approaches to improve transparency and efficiency in the Canadian housing market, benefiting all stakeholders involved.