Project Deliverable 1

1. Dataset
   https://www.kaggle.com/zygmunt/goodbooks-10k/kernels

   The dataset I will use is kaggle goodbooks-10k to access book descriptions and book ids. Goodreads has many ids such as goodreads_book_id and best_book_id which point to most popular book edition, while word_id refers to the book in the abstract sense.

   (bag of words concept; best for analyzing text, for comparing book descriptions and seeing how they're similar, way to represent passage of text as vector)

2. Methodology
   a. Data processing
      Load data from csv files using pandas for data processing with columns book_id, best_book_id, author, and other relevant information. Then, use TfidVectorizer function from scikit-learn to transform text to feature vectors and cosine similarity to calculate a numeric value that denotes similarity between books.

   b. Machine learning model
      I want to predict/estimate one or few books that are similar to a book input to generate book recommendations. I can use the bag of words model which can count word occurrences and generate document vectors to obtain how closely related two books are. I can write scripts that look up books in a database to read book descriptions and compare it with other books. If appropriate, I can also use sentiment analysis to retrieve better recommendation outputs.

   c. Final conceptualization
      I want to try integrating this model into a simple landing-page webapp to create a book recommendation generator.