Project Proposal Data Mining

By Novella Alvina and Steven Tasmin

Reading is one of the most common leisure activities. Every reader has their own preferred choice of book genres. Oftentimes, after reading a good book, we usually wonder if there are similar books that existed. This is what attracted us to build a recommendation system for books. We plan to use data from https://www.gutenberg.org/ and possibly https://www.goodreads.com/. Gutenberg can provide us with the text from the book that we can analyse for similarities using Jaccard algorithms with the addition of the book details like authors, published year, publisher, etc and especially the genres or tags. Goodreads is useful if we possibly want to cross-check books for similar tags (genres) with the books we are examining. Hence, the structure of the data we plan to mine includes ISBN, title, author, publisher, published year, genre and the book text/synopsis itself. In general, when we are searching for similar books for a particular book, our plan is to use clustering algorithms to group the books in the database with similar tags or genres as this particular book. Then, with this result, use Jaccard to analyse the text in the book and look for the highest similarity. This can also be applied to the book details. We believe this is something very useful and relatable to implement in the real life. It significantly assists readers to find and move on to the next book of their choice effectively.