Recommended books for students in Year 9 to Year 13

Prepared by

Sebastian Thomas
23/09/2021

Executive Summary

The purpose of this report is to inform English teachers in high school of the appropriate books to provide to their students in Year 9 to Year 13 to aid them in their course of study. It will first discuss the business case in more detail, highlighting the dataset used to obtain the list of books. From there, it will delve into the analysis conducted to the data obtained, providing insight into the machine learning model used and its accuracy in recommending the appropriate books. Finally, the report will conclude with the recommendations of books for each year level.

Table of Contents

Introduction

Reading literature enriches the way people express themselves and enhances their ability to reflect on themselves critically (Huemer et al., 2007). The association between developing cognitive skillsets in literature and reading the right level of books is extensive. Creating a learning environment that fits within the normal parameters of a student's age will ensure that they enjoy what they learn and drive success in their various learning paths (Gosalia, 2015). Current technology used to match the books to readers is inefficient. They lack "construct and theoretical validity" and are incredibly costly. In this scenario, machine learning can be utilised as a cost-effective solution that provides better results than the commercially available formulae. Due to their open-source nature, they are also easily upgradeable. The report will lay

out how data was analysed, the model it was fed into, and its output results in the following two sections.

Analysis

Preprocessing and vectorising data

Data provided by Kaggle's commonLit competition was provided to train the machine learning models. These files included the training set, the set of data to which the model was to fit, and the sample submission document. Before fitting the data to the different models, the training was preprocessed to provide the best data to work with. This process included removing unnecessary characters such as punctuation and commonly used words such as "the" and "a" that do not provide helpful insights in regards to the reading ease of the passages. The words in the extracts were also lemmatised, which simplifies several different forms of a word into its root form. For example, "singing", "sang", "sung" would all be simplified to the word "sing".

Once the passages were refined, they were fitted to a vectoriser known as the TF-IDF vectoriser. It is a means of transforming the preprocessed words into a meaningful representation of numbers that can be used to feed the machine learning algorithm (Chaudary, 2020).

Fitting the data to the model

The training data set was then split so that a subset of the data was used to train the model, and the rest were used to validate it. The data was then fit to various regression-based machine learning models to evaluate the reading ease of the required passages. These models included MLP regression, SVM, randomForest and Ridge. Our analysis found that the Ridge model outperformed the others in terms of accuracy for prediction. Thus, it was utilised to make predictions on the testing data provided by commonLit and submitted. This yielded a public score of 0.732, deeming it fit for use to determine the suitable books for each year level in New Zealand high schools.

Gathering suitable books for each year level with the model

Details of the top 100 books from the Gutenberg library were scraped to provide input to the machine learning model. These details included the book IDs, which Gutenberg supplied to each of its books, the names, and a paragraph from one of the chapters in the books. Like the training data set provided by commonLit, the passages from each book were put through the preprocessing stage and vectorised. They were then given as input to the Ridge machine learning model, which yielded the following results:

1) Year 9 (Figure 1)
    a. The book that could be provided to the most capable students would be Alice's Adventures in wonderland
    b. The book that could be suitable for students who are slower in the learning process is Grimm's Fairy Tales
2) Year 10 (Figure 2)

    a. The book that could be provided to the most capable students would be The Souls of Black Folk
    b. The book that could be suitable for students who are slower in the learning process is Adventures of Huckleberry Finn
3) Year 11 (Figure 3)
    a. The book that could be provided to the most capable students would be The Extraordinary Adventures of Arsene Lupin, Gentleman-Burglar
    b. The book that could be suitable for students who are slower in the learning process is Dracula

4) Year 12 (Figure 4)
    a. The book that could be provided to the most capable students would be Jane Eyre: An Autobiography
    b. The book that could be suitable for students who are slower in the learning process is The Life and Adventures of Robinson Crusoe
5) Year 13 (Figure 5)
    a. The book that could be provided to the most capable students would be The Republic
    b. The book that could be suitable for students who are slower in the learning process is The Works of Edgar Allan Poe Volume 2

Some of the predictions made here correlate with those deemed suitable through the Flesch-Kincaid algorithm. This includes the books such as The Souls of Black Folk and Huckleberry Finn for year ten students (Lit2Go, n.d.). However, the fact that only one paragraph from each of the books has been fit to the model somewhat deteriorates the quality of some of the predictions. In reality, it would be better to draw paragraphs from a larger pool of chapters than one, which would provide better insight into the reading ease of the book. However, since the model was trained on passages in which the median number of words were 70, the amount of input that the model could receive from the scraped data was limited to approximately 70.

Conclusion

This report has provided insight into how the analysis was conducted on the data provided, the machine learning models used to develop the predictions, and its accuracy in recommending books to students between Year 9 and Year 13. The processes involved were preprocessing the data provided by Kaggle's commonLit competition and testing various machine learning models. The Ridge machine learning model produced a few reasonable predictions despite the given limitations from the analysis above.

Appendices

Year 9
Grimms' Fairy Tales The Jungle Book

Little Women
Peter Pan
The Jungle
Heart of Darkness
Old Granny Fox
Great Expectations
The Importance of Being Earnest: A Trivial Comedy for Serious People
A Modest Proposal
A Doll's House : a play
Metamorphosis
Dubliners
Pride and Prejudice
The Adventures of Sherlock Holmes
Treasure Island
War and Peace
The Secret Garden
The Awakening, and Selected Short Stories
Alice's Adventures in Wonderland

Figure 1 - Table showing the recommended book for year 9

Year 10
Adventures of Huckleberry Finn
Ethan Frome
Complete Original Short Stories of Guy De Maupassant
The Great Gatsby
Wuthering Heights
Emma
Crime and Punishment
Anthem
The Strange Case of Dr. Jekyll and Mr. Hyde
The Happy Prince, and Other Tales
Siddhartha
The King James Version of the Bible
The Hound of the Baskervilles
The Time Machine
Anne of Green Gables
The Call of the Wild
The Romance of Lust: A classic Victorian erotic novel
Oliver Twist
The Adventures of Tom Sawyer, Complete
The Souls of Black Folk

Figure 2 - Table showing the apppropriate book for Year 10 students

Year 11

Dracula
The Slang Dictionary: Etymological, Historical and Andecdotal
The Count of Monte Cristo, Illustrated
Songs of Innocence, and Songs of Experience
An Index of The Divine Comedy by Dante
Narrative of the Captivity and Restoration of Mrs. Mary Rowlandson
Thus Spake Zarathustra: A Book for All and None
A Pickle for the Knowing Ones
Walden, and On The Duty Of Civil Disobedience
A Christmas Carol in Prose; Being a Ghost Story of Christmas
Carmilla
The Mysterious Affair at Styles
Autobiography of Benjamin Franklin
The Picture of Dorian Gray
David Copperfield
Anna Karenina
Sense and Sensibility
The Prince
Frankenstein; Or, The Modern Prometheus
The Extraordinary Adventures of Arsene Lupin, Gentleman-Burglar

Figure 3 - Table showing the appropriate books for Year 11

Year 12
The Life and Adventures of Robinson Crusoe
Candide
Around the World in Eighty Days
The Brothers Karamazov
Simple Sabotage Field Manual
Ulysses
Uncle Tom's Cabin
The War of the Worlds
The Wonderful Wizard of Oz
Moby Dick; Or, The Whale
The Yellow Wallpaper
Beyond Good and Evil
The Legend of Sleepy Hollow
The Interesting Narrative of the Life of Olaudah Equiano, Or Gustavus Vassa, The African
Beowulf: An Anglo-Saxon Epic Poem
The Odyssey
The American Diary of a Japanese Girl
The Prophet
Leviathan
Jane Eyre: An Autobiography

Figure 4 - Table showing the appropriate books for Year 12

Year 13
The Works of Edgar Allan Poe â€" Volume 2
Les MisÃ©rables
Common Sense
The Elements of Style
A Tale of Two Cities
Gulliver's Travels into Several Remote Nations of the World
The Scarlet Letter
Essays of Michel de Montaigne â€" Complete
Notes from the Underground
Narrative of the Life of Frederick Douglass, an American Slave
A Study in Scarlet
The Confessions of St. Augustine
The History of the Peloponnesian War
The Problems of Philosophy
The Kama Sutra of Vatsyayana
Don Quixote
The Iliad
Second Treatise of Government
The Philippine Islands, 1493-1898, Volume 33, 1519-1522
The Republic

Figure 5 - Table showing the appropriate books for Year 13

References

Chaudary, M. (2020, April 24). *TF-IDF Vectorizer scikit-learn* [Medium.com].

https://medium.com/@cmukesh8688/tf-idf-vectorizer-scikit-learn-dbc0244a911a

Gosalia, P. (2015, June 10). *The Importance of Age-Appropriate Education* [The Swaddle].

https://theswaddle.com/importance-age-appropriate-education/

Huemer, W., Gibson, J., & Pocci, L. (2007). Why Read Literature? The cognitive function of form. *A SENSE OF THE WORLD. ESSAYS ON FICTION, NARRATIVE AND KNOWLEDGE*, 233–245.

Lit2Go. (n.d.). *Flesch–Kincaid Grade Level 10*.

https://etc.usf.edu/lit2go/readability/flesch_kincaid_grade_level/10/