**CommonLit Readability Model – A text readability algorithm**

Prepared by

Raymond Luo

17/09/2021

**Contents**

3 – Executive Summary

3 – Introduction

3 – Limitations

4 – Model

4 – Results

5 – Conclusion

6 – Appendix A

**Executive Summary**

CommonLit aims to develop an algorithm which will help identify a text’s relative difficulty. This report summarises the approach which has been chosen. A set of text recommendations are also provided to help high school teachers to obtain an outline of which texts are suitable for their students to read. The texts come from a number of popular texts and are to be assigned to year groups between Year 9 and 13.

This purpose of this report is to highlight the mechanisms of this model, the limitations that such a model may have, and provide some text recommendations targeted towards high school students of differing year levels. The model was built for the purpose of providing a more effective way of evaluating the difficulty of texts. It is important to note that such a model will have inherent limitations, and text recommendations should be taken only as a guideline.

**Introduction**

CommonLit is a non-profit education technology organization serving over 20 million teachers and students with free digital reading and writing lessons for grades 3-12. They have requested for a machine learning algorithm which will aid in identifying the appropriate reading level of a passage of text to improve upon existing solutions. Currently available readability methods such as the Flesch-Kincaid Grade Level utilize weak and limited solutions, such as character/syllables per word or number of words per sentence, to signify the difficulty of texts. This report will provide a non-technical explanation of the methodologies in our machine learning model, touch on the limitations which the model possesses, as well as using our model to deduce a list of various text recommendations for different high school year levels.

**Limitations**

Any machine generated predictions will have inherent limitations. Some of the limitations which belong to this model are described below:

* The model may encounter some degree of trouble in providing a holistic score for entire texts. The model was trained on excerpts of around 100-250 words, but texts typically have far more words. Thus, the perceived ‘difficulty’ of each excerpt within the same text will likely be subject to some variation.
* The difficulty of texts especially in a high school setting, lie in the symbolic meanings of the text and is not solely influenced by the complexity of its vocabulary or how difficult a sentence is to read. Such a model is not very capable of picking up on such factors, due to the scope of which it has been trained on.

**Model**

The model utilizes a machine learning algorithm to learn to interpret the text and assign a numerical score to the text which signifies its readability. The model was trained on a given dataset of around 2800 text excerpts, with each excerpt being accompanied by its readability score. Through this training, the model will learn to identify key components of a passage of text and use these key features to determine an appropriate score.

A particularly powerful type of machine learning model known as a transformer has been employed. Such a model is more capable of modelling and remembering long-term dependencies within the texts, which is one of the reasons why it is effective at handling text inputs.

Feeding through the model involves the following steps:

1. Encoding the input text to a sequence of numbers
2. Passing the sequence of numbers through the model
3. Receiving the readability score from the model

**Results**

Excerpts from a set of 100 texts sourced from the Project Gutenberg website have been extracted, with a corresponding readability score calculated for each text. 10 sets of scores are provided for different excerpts from each text, with the average of all the scores calculated. The year level which the scores correlated to are defined below

|  |  |
| --- | --- |
| **Boundary Scores** *(square bracket denotes inclusive, round bracket denotes non-inclusive)* | |
| **Year Level** | **Boundaries** |
| Year 9- | 0.4+ |
| Year 9 | [-0.1, 0.4) |
| Year 10 | [-0.6, -0.1) |
| Year 11 | [-1.1, -0.6) |
| Year 12 | [-1.6, -1.1) |
| Year 13 | [-2.1, -1.6) |
| Year 13+ | -2.1- |

Year 9- and Year 13+ which are typically too easy for Year 9 students or too difficult for Year 13 students.

For a complete list of texts, with their associated readability score and model-determined year level, refer to **Appendix A**.

**Conclusion**

Machine learning can serve as an extremely powerful tool for language processing tasks such as readability testing. Employing machine learning allows the difficulty of various published works to be more accurately and efficiently scored, which is the main objective of creating this model. Some limitations of this model include its limited ability to analyse an entire text holistically, being limited to shorter passages or paragraphs. The model has been evaluated on 100 free e-books found on the Project Gutenberg website. The recommended year level for each of these texts have been provided in **Appendix A.**

**Appendix A**

|  |  |  |  |
| --- | --- | --- | --- |
| **Book** | **Author** | **Readability Score** | **Year Level** |
| The Wonderful Wizard of Oz | L. Frank Baum | **0.63606053** | Year 9- |
| Alice's Adventures in Wonderland | Lewis Carroll | **0.196855619** | Year 9 |
| The Happy Prince, and Other Tales | Oscar Wilde | **0.167637179** | Year 9 |
| Metamorphosis | Franz Kafka | **0.156419808** | Year 9 |
| Grimms' Fairy Tales | Jacob Grimm and Wilhelm Grimm | **0.133783128** | Year 9 |
| Old Granny Fox | Thornton W. Burgess | **0.128404216** | Year 9 |
| The Picture of Dorian Gray | Oscar Wilde | **0.061843601** | Year 9 |
| Anne of Green Gables | L. M. Montgomery | **-0.061287848** | Year 9 |
| The Adventures of Tom Sawyer, Complete | Mark Twain | **-0.196100439** | Year 10 |
| Peter Pan | J. M. Barrie | **-0.209679965** | Year 10 |
| Ethan Frome | Edith Wharton | **-0.22970996** | Year 10 |
| The Secret Garden | Frances Hodgson Burnett | **-0.255123299** | Year 10 |
| Dubliners | James Joyce | **-0.264040528** | Year 10 |
| The Great Gatsby | F. Scott Fitzgerald | **-0.322872348** | Year 10 |
| The Awakening, and Selected Short Stories | Kate Chopin | **-0.363434941** | Year 10 |
| The Adventures of Sherlock Holmes | Arthur Conan Doyle | **-0.373268525** | Year 10 |
| Little Women | Louisa May Alcott | **-0.387094964** | Year 10 |
| Anna Karenina | graf Leo Tolstoy | **-0.417576922** | Year 10 |
| Uncle Tom's Cabin | Harriet Beecher Stowe | **-0.480540673** | Year 10 |
| A Doll's House : a play | Henrik Ibsen | **-0.493843267** | Year 10 |
| Complete Original Short Stories of Guy De Maupassant | Guy de Maupassant | **-0.494295222** | Year 10 |
| The Jungle Book | Rudyard Kipling | **-0.496704648** | Year 10 |
| The Jungle | Upton Sinclair | **-0.506126039** | Year 10 |
| Adventures of Huckleberry Finn | Mark Twain | **-0.524061074** | Year 10 |
| The Yellow Wallpaper | Charlotte Perkins Gilman | **-0.525446022** | Year 10 |
| Wuthering Heights | Emily Brontë | **-0.550154527** | Year 10 |
| The Importance of Being Earnest: A Trivial Comedy for Serious People | Oscar Wilde | **-0.576073557** | Year 10 |
| Around the World in Eighty Days | Jules Verne | **-0.584880865** | Year 10 |
| A Study in Scarlet | Arthur Conan Doyle | **-0.595739232** | Year 10 |
| The Mysterious Affair at Styles | Agatha Christie | **-0.602806037** | Year 11 |
| Heart of Darkness | Joseph Conrad | **-0.605003121** | Year 11 |
| The Hound of the Baskervilles | Arthur Conan Doyle | **-0.612269917** | Year 11 |
| The Time Machine | H. G. Wells | **-0.613919303** | Year 11 |
| The Romance of Lust: A classic Victorian erotic novel | Anonymous | **-0.646312862** | Year 11 |
| Notes from the Underground | Fyodor Dostoyevsky | **-0.65000442** | Year 11 |
| A Tale of Two Cities | Charles Dickens | **-0.651247323** | Year 11 |
| Oliver Twist | Charles Dickens | **-0.655171454** | Year 11 |
| The Brothers Karamazov | Fyodor Dostoyevsky | **-0.66709577** | Year 11 |
| The American Diary of a Japanese Girl | Yoné Noguchi | **-0.672054987** | Year 11 |
| Crime and Punishment | Fyodor Dostoyevsky | **-0.676806971** | Year 11 |
| The Strange Case of Dr. Jekyll and Mr. Hyde | Robert Louis Stevenson | **-0.677584054** | Year 11 |
| Jane Eyre: An Autobiography | Charlotte Brontë | **-0.684091347** | Year 11 |
| Anthem | Ayn Rand | **-0.696143775** | Year 11 |
| The Call of the Wild | Jack London | **-0.70751268** | Year 11 |
| The Extraordinary Adventures of Arsene Lupin, Gentleman-Burglar | Maurice Leblanc | **-0.721378857** | Year 11 |
| Emma | Jane Austen | **-0.724689782** | Year 11 |
| A Christmas Carol in Prose; Being a Ghost Story of Christmas | Charles Dickens | **-0.726220938** | Year 11 |
| Narrative of the Captivity and Restoration of Mrs. Mary Rowlandson | Rowlandson | **-0.765783188** | Year 11 |
| Dracula | Bram Stoker | **-0.7749448** | Year 11 |
| Siddhartha | Hermann Hesse | **-0.805895925** | Year 11 |
| Treasure Island | Robert Louis Stevenson | **-0.821271659** | Year 11 |
| Sense and Sensibility | Jane Austen | **-0.871286494** | Year 11 |
| War and Peace | graf Leo Tolstoy | **-0.883074788** | Year 11 |
| The Count of Monte Cristo, Illustrated | Alexandre Dumas | **-0.923199788** | Year 11 |
| The War of the Worlds | H. G. Wells | **-0.941998056** | Year 11 |
| The Scarlet Letter | Nathaniel Hawthorne | **-0.942776176** | Year 11 |
| David Copperfield | Charles Dickens | **-0.947949196** | Year 11 |
| Persuasion | Jane Austen | **-0.963612074** | Year 11 |
| Narrative of the Life of Frederick Douglass, an American Slave | Douglass | **-0.999351599** | Year 11 |
| Les Misérables | Victor Hugo | **-1.042196175** | Year 11 |
| Frankenstein; Or, The Modern Prometheus | Mary Wollstonecraft Shelley | **-1.202142094** | Year 12 |
| Great Expectations | Charles Dickens | **-1.2104478** | Year 12 |
| Pride and Prejudice | Jane Austen | **-1.223281686** | Year 12 |
| The Art of War | active 6th century B.C. Sunzi | **-1.298505425** | Year 12 |
| The Odyssey | Homer | **-1.334930742** | Year 12 |
| Songs of Innocence, and Songs of Experience | William Blake | **-1.50390453** | Year 12 |
| Don Quixote | Miguel de Cervantes Saavedra | **-1.54133786** | Year 12 |
| The Souls of Black Folk | W. E. B. Du Bois | **-1.559904707** | Year 12 |
| The Kama Sutra of Vatsyayana | Vatsyayana | **-1.604768163** | Year 13 |
| The Prophet | Kahlil Gibran | **-1.607291502** | Year 13 |
| Moby Dick; Or, The Whale | Herman Melville | **-1.624414438** | Year 13 |
| Candide | Voltaire | **-1.631320238** | Year 13 |
| The Legend of Sleepy Hollow | Washington Irving | **-1.653975618** | Year 13 |
| The Republic | Plato | **-1.711748397** | Year 13 |
| Walden, and On The Duty Of Civil Disobedience | Henry David Thoreau | **-1.784447134** | Year 13 |
| The King James Version of the Bible |  | **-1.935586149** | Year 13 |
| The Problems of Philosophy | Bertrand Russell | **-1.960501313** | Year 13 |
| A Pickle for the Knowing Ones | Timothy Dexter | **-2.023770034** | Year 13 |
| The Elements of Style | William Strunk | **-2.039673162** | Year 13 |
| Ulysses | James Joyce | **-2.066578794** | Year 13 |
| Gulliver's Travels into Several Remote Nations of the World | Jonathan Swift | **-2.106035364** | Year 13+ |
| Beyond Good and Evil | Friedrich Wilhelm Nietzsche | **-2.283282995** | Year 13+ |
| A Modest Proposal | Jonathan Swift | **-2.31021204** | Year 13+ |
| The Interesting Narrative of the Life of Olaudah Equiano, Or Gustavus Vassa, The African |  | **-2.388278782** | Year 13+ |
| The Prince | Niccolò Machiavelli | **-2.398843741** | Year 13+ |
| Essays of Michel de Montaigne — Complete | Michel de Montaigne | **-2.426245248** | Year 13+ |
| Autobiography of Benjamin Franklin | Benjamin Franklin | **-2.43955847** | Year 13+ |
| Thus Spake Zarathustra: A Book for All and None | Friedrich Wilhelm Nietzsche | **-2.472623962** | Year 13+ |
| Common Sense | Thomas Paine | **-2.483844769** | Year 13+ |
| The Confessions of St. Augustine | Bishop of Hippo Saint Augustine | **-2.668735242** | Year 13+ |
| The History of the Peloponnesian War | Thucydides | **-2.699634862** | Year 13+ |
| Second Treatise of Government | John Locke | **-2.711859** | Year 13+ |
| The Iliad | Homer | **-2.771131575** | Year 13+ |
| Meditations | Emperor of Rome Marcus Aurelius | **-2.777133274** | Year 13+ |
| Leviathan | Thomas Hobbes | **-2.956110883** | Year 13+ |
| Beowulf: An Anglo-Saxon Epic Poem | J. Lesslie Hall | **-3.239355731** | Year 13+ |
| An Index of The Divine Comedy | Dante | **-3.26967175** | Year 13+ |