The Words are not Enough:

An Investigation into the Viability of Textual Complexity as a Feature for Recommendation Systems

Mees van Smaalen
M.F.vanSmaalen@student.tudelft.nl
Responsible Professor: M.S. Pera

I. Background

- For children it is crucial to develop good reading skills and making sure they enjoy reading at a young age helps tremendously
- Recommender Systems could be employed for finding good recommendations
- RS often use collaborative filtering which works suboptimal for children due to lack of user feedback
- Recommend on content features instead
- Textual complexity is a possible feature to base recommendations on

II. Research Question

Does the language used in books and their descriptions match the age of the children it's intended for?

III. Methods

Determine the best fitting readability formula and apply to all texts then evaluate the error terms per age bucket

Measure the average amount of sentences and variation of each text to discover trends

Compare each word in the text to Age of Acquisition data to explore differences in distributions

IV. Results

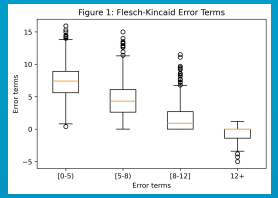


Figure 1: Flesch-Kincaid Error Terms

Age Bucket	Sentence Length Average	Sentence Length Variation
[0-5)	15.806	63.340
[5-8)	16.544	66.237
[8-12)	17.393	81.186
12+	17.830	77.651

Table 2: Mean of the Average and Variation of Book Description Sentence Lengths across Age Buckets

- For book descriptions it was found that the Flesch-Kincaid reading algorithm performed the best
- The algorithm becomes more accurate towards books intended for older children
- Average sentence length increases with age
- Sentence length variation increase with age
- Not substantially enough to draw strong conclusion
- Large amount of general language results in distributions skewing to lower ages
- Small overall increase in overall word difficulty in the older age buckets
- Difficulty seems to stagnate in the [8-12) age bucket

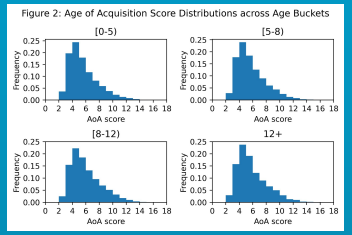


Figure 2: Age of Acquisition Score Distributions across Age Buckets

V. Description vs Full Text

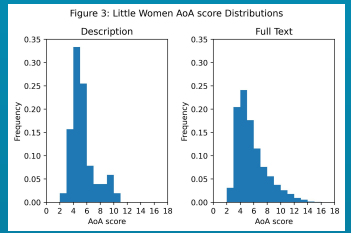
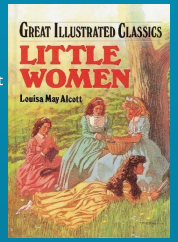


Figure 3: Little Women AoA score Distributions

- Full text has a more equal spread of word difficulty
- The full text naturally has more varied sentences
- The description is rated lower by Flesch-Kincaid but higher by Dale-Chall compared to the full text
- Results seem contradictory
- Description not that reflective of full text in this case

Title: Little Women
Author: Louisa May Alcott
Year: 1868

Age Bucket	Description [8-12)	Full Text 12+
Sentence length avg	18.4	21.95
Sentence length var	67.84	248.77
Flesch-Kincaid	11.4	13.3
Dale-Chall	16.37	12.5



VI. Conclusions

- Textual complexity might not be the optimal feature to base recommendations on
- High error terms in the lower age buckets, likely the descriptions for these books are not aimed at the young children but instead at their parents
- Increase in sentence length not substantial enough for reliable recommendation
- Due to the difficulty of words stagnating in the higher age buckets, recommending based on word difficulty becomes tricky
- Description text complexity might not accurately reflect book text complexity
- Books naturally have more varied sentences
- Accuracy dependent on readability formula
- In this example full text has more spread out distribution, in the full data set this is not seen and in general the full texts contain lower AoA scores due to a higher ratio of general language to subject specific language

VII. Limitations

- Data set was lacking in size
- Limited amount of age buckets
- Full texts were dated
- Amount of readability formulas applied could be increased
- Better filtering of general language

VIII. References

1. Milton, S., Batista, L., Allen, G., Gao, S., Ng, Y. D., & Pera, M.S. (2020). "Don't Judge a Book by its Cover": Exploring Book Traits Children Favor.
2. https://github.com/textstat/textstat
3. Allen, G., Milton, A., Landau Wright, L., Fails, J. A., Kennington, C., & Pera, M.S. (2022). "Supercalifragilisticexpialidocious: Why Using the "Right" Readability Formula in Children's Web Search Matters"

TU Delft