Tobias Kroll

Prof. Deokgun Park

May 4, 2021

Readability Scorer: Project Proposal

The Project

I'll be designing and implementing a model that scores text excerpts; readability. I will be using a k-nearest-neighbors approach in order to score novel excerpts. I will explore ways to vectorize the documents as I proceed. The goal is to compete (and in fact do so 3 months after this project is due) in the Kaggle competition here: https://www.kaggle.com/c/commonlitreadabilityprize.

Why

Readability of books and documents is a crucial metric which helps to bring the right book to the right reader. Traditional tools 'are based on weak proxies of text decoding'. While there do exist modern, machine learning based approaches for this problem, they are often proprietary, and thus not transparent to the public. In fact, the goal of the competition, and thus my project, is to fix this problem. Additionally, for the more materialistic and personal side, the prize pool for the competition is rather large.

Dataset

It's just the one for this Kaggle competition:
https://www.kaggle.com/c/commonlitreadabilityprize/data

Similar Systems

There are already several published notebooks for the competition:
https://www.kaggle.com/hannes82/commonlit-readability-roberta-inference

https://www.kaggle.com/leighplt/transformers-pytorch

https://www.kaggle.com/shahrukh0603/commomlit

to just post three. While it is unlikely that a kNN approach beats those using deep neural networks, it should significantly beat their training time, and hopefully not be entirely too poor a system prediction-wise.