Ranking Reading Levels

Megan Sorel.
Data Scientist at McGraw-Hill

Recommending Better Reading Material

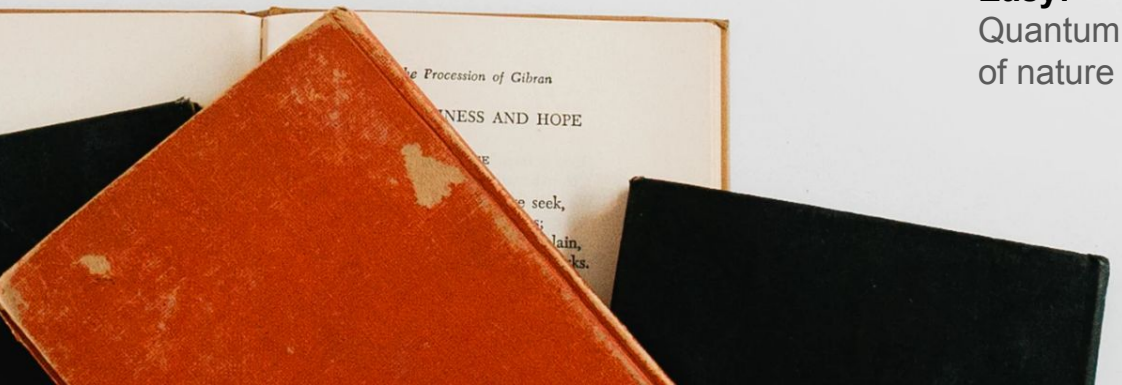
Increase reading accessibility and motivation through tailored recommendations for schools and independent readers.

The Problem: Current Models

Number of syllables in words

Difficult:
Celebration, librarian, hippopotamus

Easy:
Abscond, bane, curt



Length of the sentence

Difficult:
I am reading right now and reading is fun because I can learn many things and I like to learn, so I will read as many books as I can.

Easy:
Quantum mechanics describes the physical properties of nature at the scale of atoms and subatomic particles.

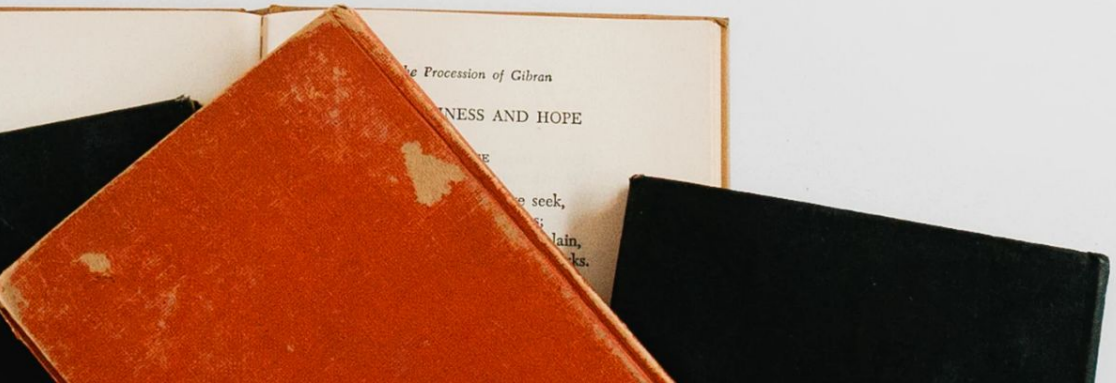
The Goal: Readability based on Linguistic Features

Model using

- Semantics
- Syntax
- Morphology

Readability Target:

- Readability ratings scored multiple
  - Range: -3.67 - 1.71
- Standard deviation of scores
  - Median: 0.48



Semantics

Word familiarity
- How frequently words occur in English

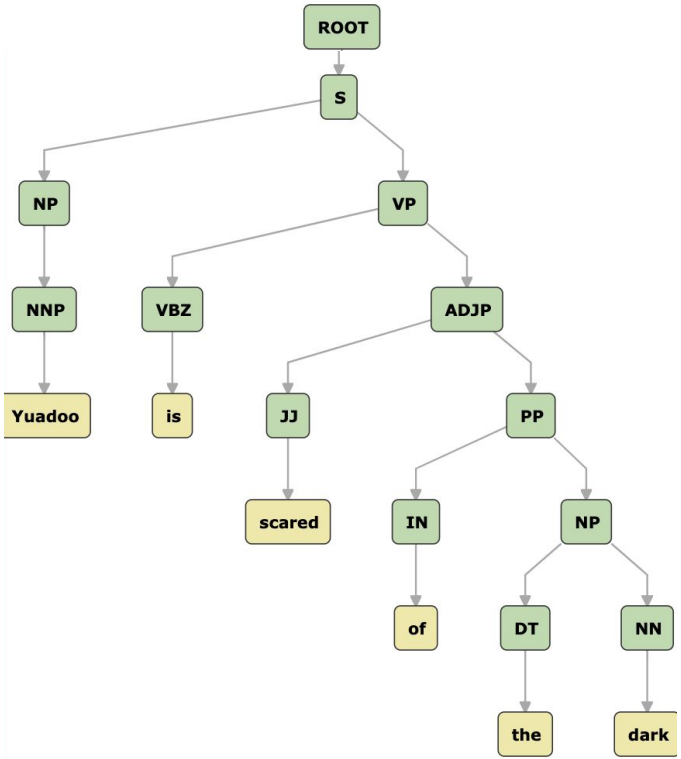
Lexical Diversity
- How many unique words occur in the text

Lexical Diversity * Word familiarity
- The interaction of the word difficulty and number of words

Topics
- War, Nature, Science, Outdoors and Family

Syntax

‘Yuadoo is scared of the dark’



The number and length of subtrees in a text as a measure of constituent complexity

The number of nodes that started with unique types of words to see how deep and diverse the tree was

Morphology

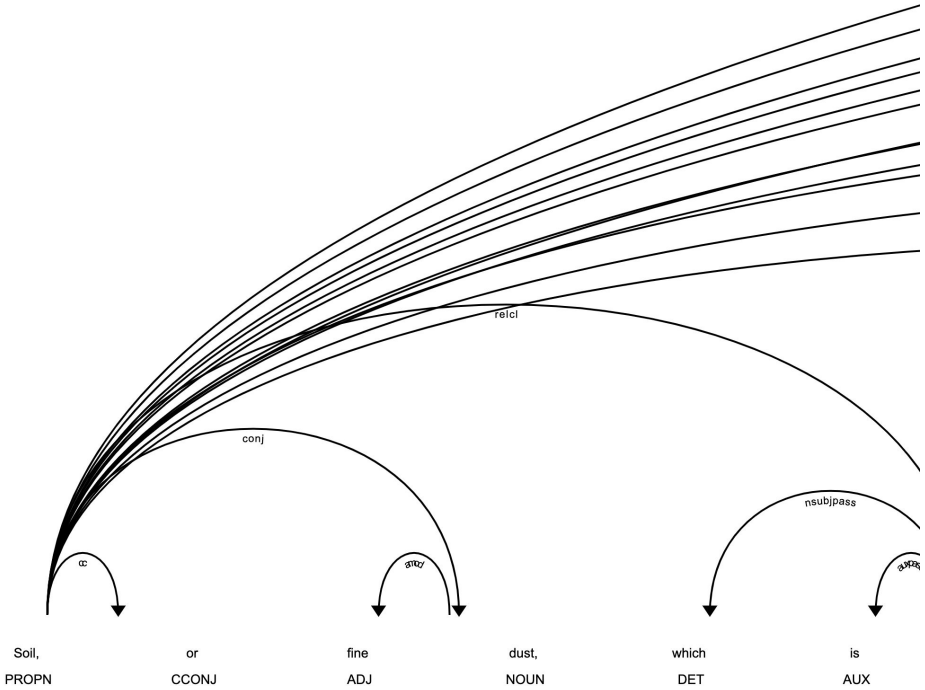
Third-person pronoun use

Counts of unique fine grained parts of speech in the sentence

Number of dependencies and distance between the word and its children

The use of the perfect tense

Example of the start of a dependency-complex sentence



Models Fitted with PCA Features

Ridge was chosen due to its low MSE and overfitting

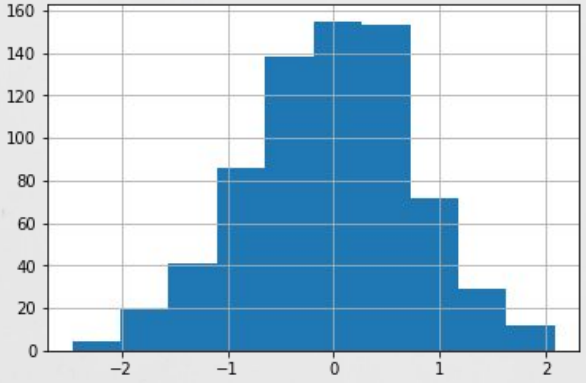
	Regularization	MSE	Train R2	Test R2
Ridge Regression	**Alpha**: 100 from gridsearch	.595	.46	.43
GradientBoostingRegressor	Default	.605	.64	.45
Keras Seq. Neural Network	**L2**: 0.1 **EarlyStopping**: mindelta: 0, patience: 10	.671	NA	NA

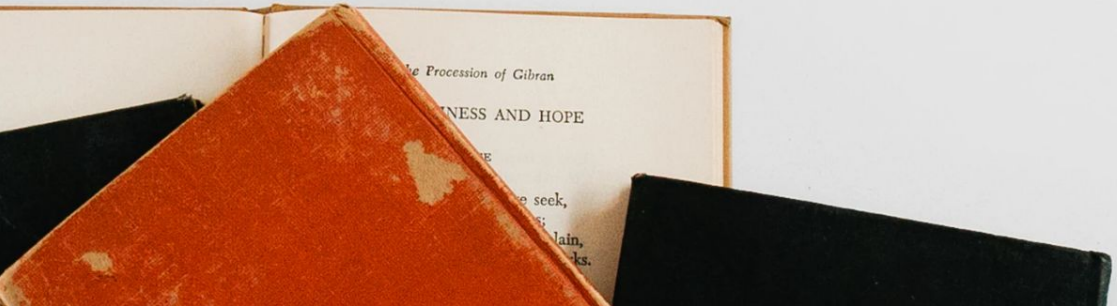
Production Model

Ridge after PCA and an alpha of 100

	Baseline	Old Model	New Model
R2		16%	44%
MSE	1.06	.92	.60

Residuals





Model Improvements

Pronoun co-reference with Neuralcoref

Nominalization with

Bigram word-frequency

Split infinitives

Heavy Noun Phrase shifting

Sentence predictability and ‘Garden Path’ sentences using Shift-Reduce-Parsers and Sentence similarity models

Future Applications

Create interactive EDA that allows selection of text based on features

Model for the target's standard deviation to find features that cause the highest variation in readability ranking

Make ESL version of model for adult learners

Resources

https://unsplash.com/

https://corenlp.run/

https://pypi.org/project/readability/

https://www.kaggle.com/c/commonlitreadabilityprize