Ranking Reading Levels

Megan Sorel.
Data Scientist at McGraw-Hill

Recommending Better Reading Material

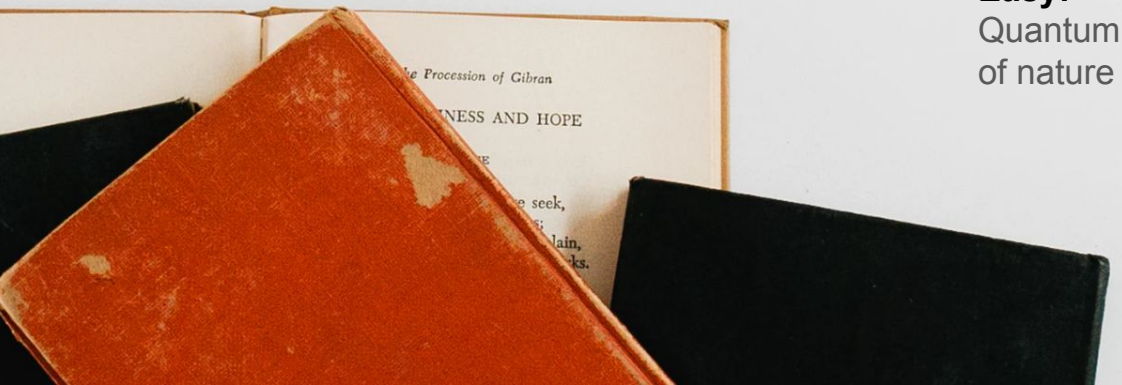
Increase reading accessibility and motivation through tailored recommendations for schools and independent readers.

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.

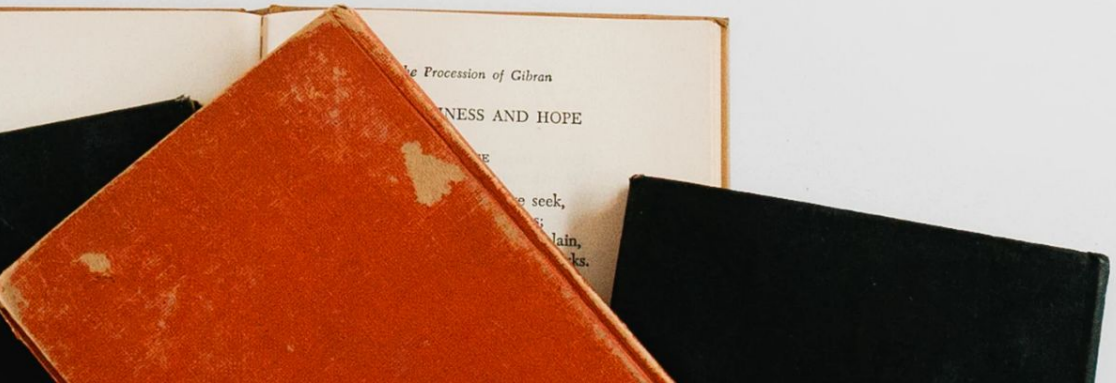
The Goal: Accurately Rank These Sentences

A

When you think of dinosaurs and where they lived, what do you picture?

B

The commutator is peculiar, consisting of only three segments of a copper ring, while in the simplest of other continuous current generators several times that number exist, and frequently 120 segments are to be found.



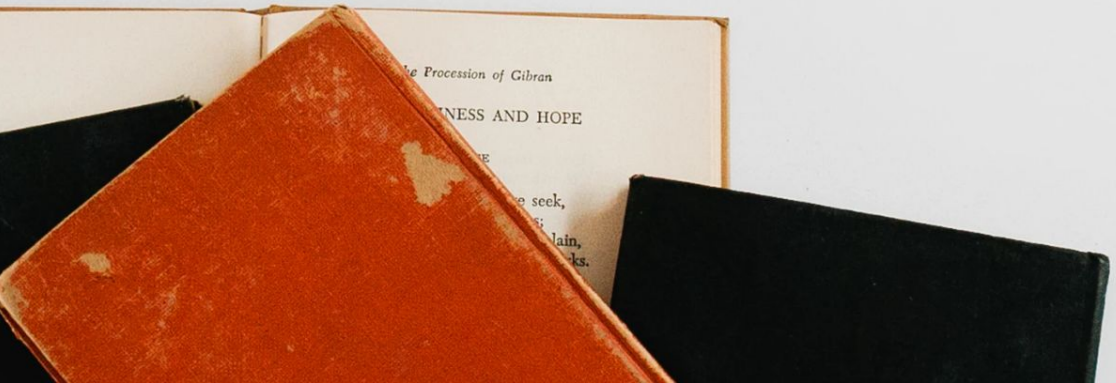
What Makes These Sentences Difficult?

A

When you think of dinosaurs and where they lived, what do you picture?

B

The commutator is peculiar, consisting of only three segments of a copper ring, while in the simplest of other continuous current generators several times that number exist, and frequently 120 segments are to be found.



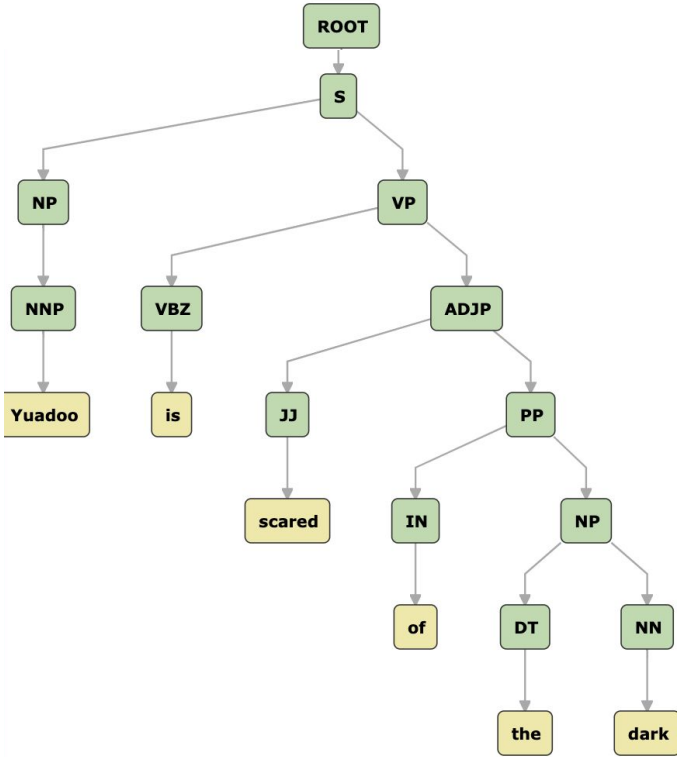
Word Familiarity

“It is sad but true. Autumn is often called the sad time of the year, and it is the sad time.”

“...In retaliation for the "Boston Tea Party," Parliament closed the port of Boston and virtually abolished provincial self-government”

Grammatical Complexity

Measured by ‘Syntax Trees’ and their depth.

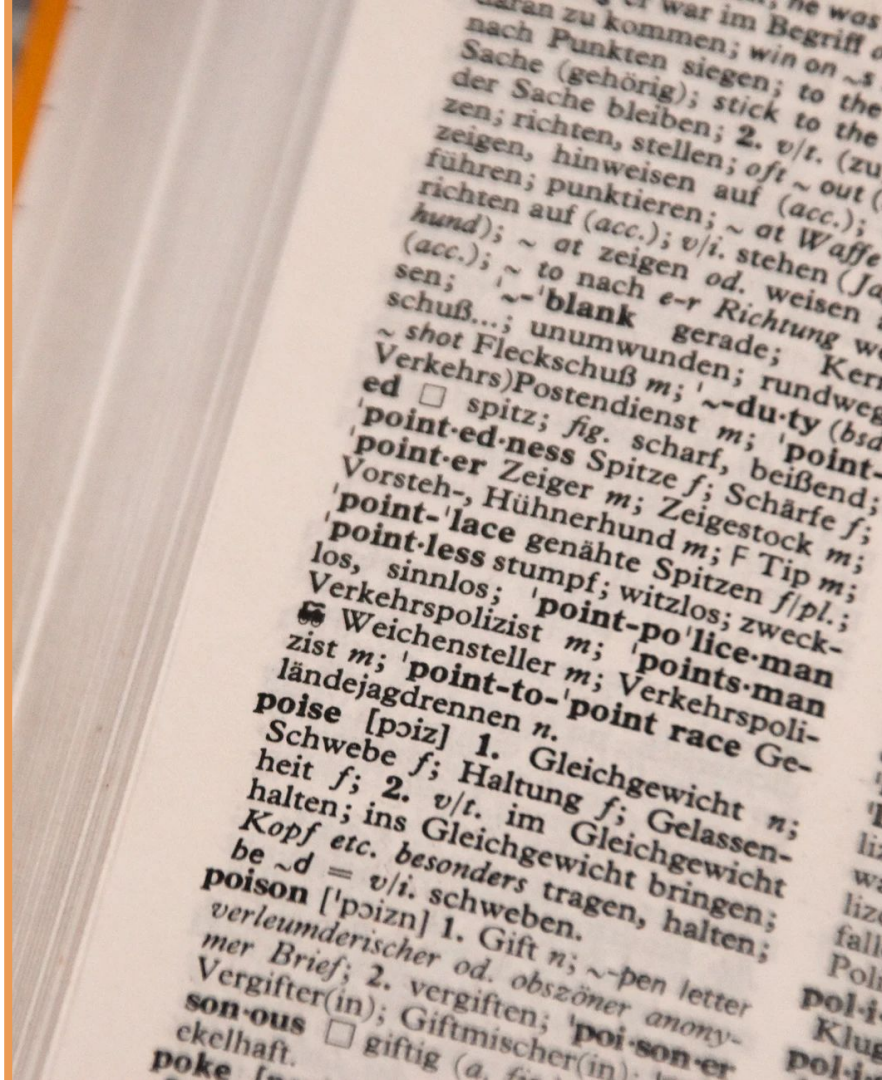


Types of Words

Third-person pronoun use:

‘the weak man must become strong of himself; he must, by his own efforts, develop the strength which he admires in another. None but himself can alter his condition.’

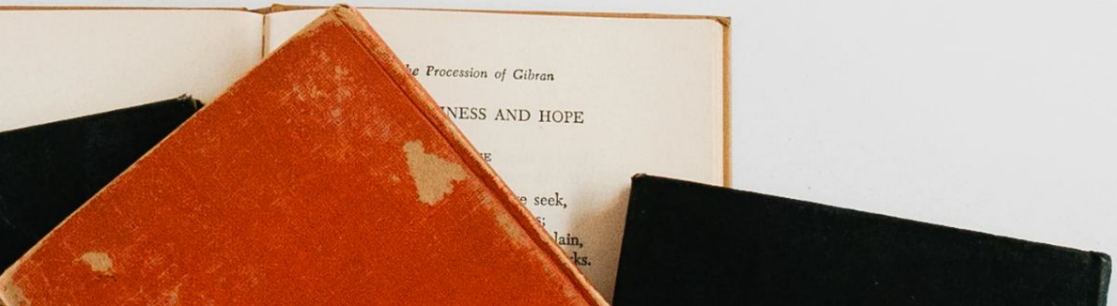
Types of conjunctions, parts of speech, verb tenses.



Predicting Readability

Readability scores were calculated by polls where people rated the text.

	Baseline	Old Model	New Model
Percent of Difficulty Accounted for in Model		16%	44%
Mean Squared Error	1.06	.92	.60



Model Improvements

Pronoun co-reference

Nominalization

Bigram word-frequency

Split infinitives

Heavy Noun Phrase shifting

Sentence predictability and ‘Garden Path’ sentences

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