



Classifying Text by Readability

By **Nathaniel Green** (New York's premier education consultant and reader of some books)



Agenda

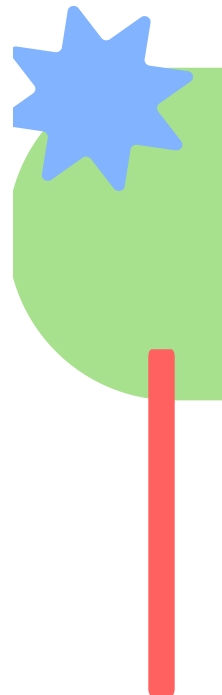
- 1. Introduce the Big Problem**
 - 2. Contextualize my data and target**
 - 3. Assess my models**
 - 4. Make Recommendation**
 - 5. Discuss Next Steps**
- 

The Problem:

Stakeholder: Teachers in the NYC DOE

The Problem: Teachers, especially science and social studies teachers, are always looking for timely, grade appropriate texts for their students. Assessing the academic rigor of a text can be a difficult task for teachers, despite it being absolutely essential for student learning and engagement. While expensive sites like NEWSELA exist, their texts are limited in their scope.

The Solution: Create a model that can categorize the readability (academic rigor) of any text into three Categories: "Hard", "Medium", and "Easy"



Can YOU tell the difference between these two texts??

“While we learn, the cells of the brain (called neurons) connect to each other by reaching out their tiny arms (called axons) or even by growing new arms.”

8th Grade

“Easy”



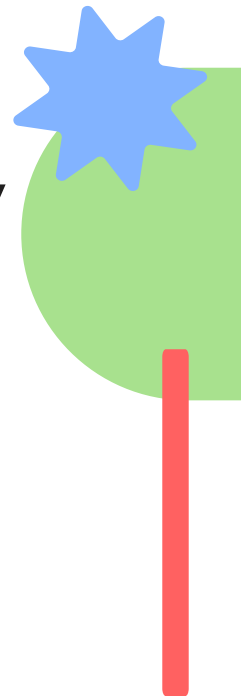
“The brain is constantly trying to find better ways to deliver and deal with information by creating or removing connections between neurons.”

College

“Hard”



**If you could distinguish between
those texts, you did better than my
model!**

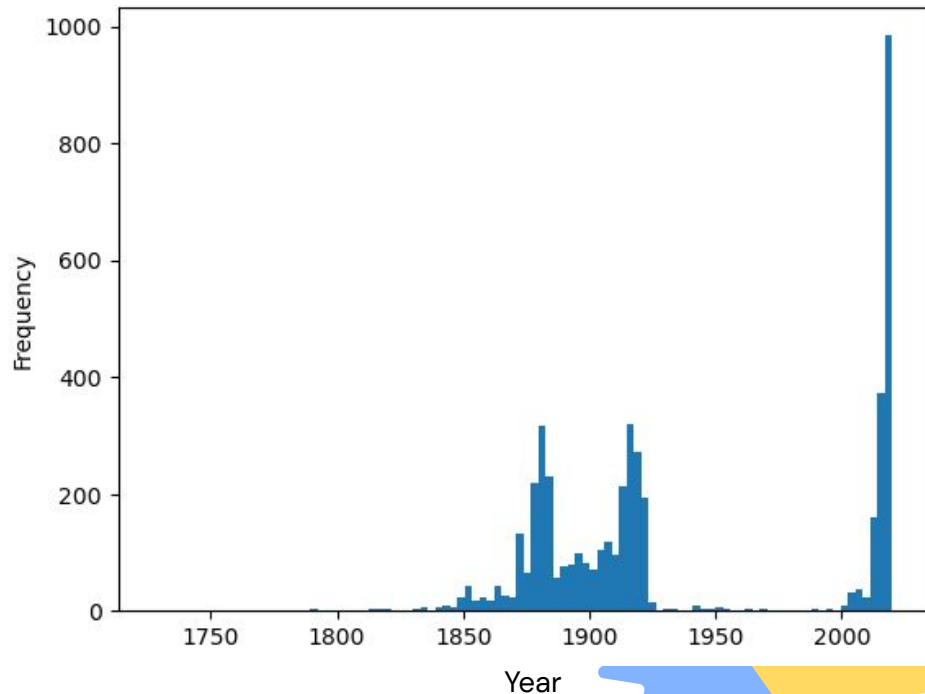


My Data: The Text

The text corpus I used was assembled by Georgia State University and consists of around 5,000 mostly public domain texts from a variety of sources including:

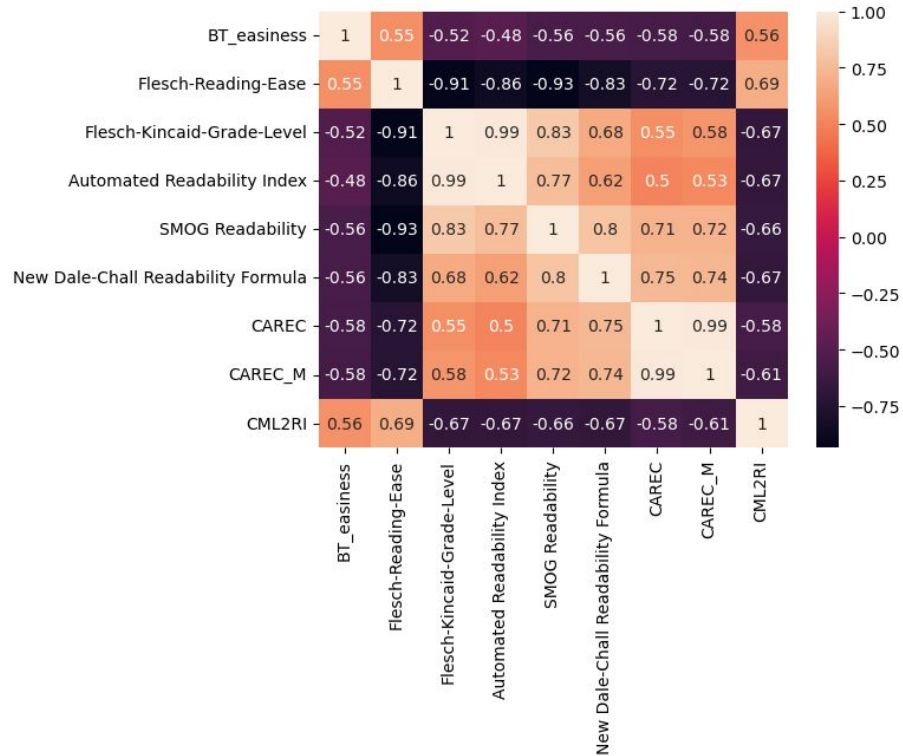
1. CommonLit
2. Wikipedia
3. USHistory.org
4. Dozen of open digital libraries

Note: All texts were filtered for being school appropriate!




My Data: The Target

I chose to measure the readability (academic rigor) using **Flesch Reading Ease** because it is the most widely used measure of text readability, used by most schools in the United States and was highly correlated to other known measures of readability.

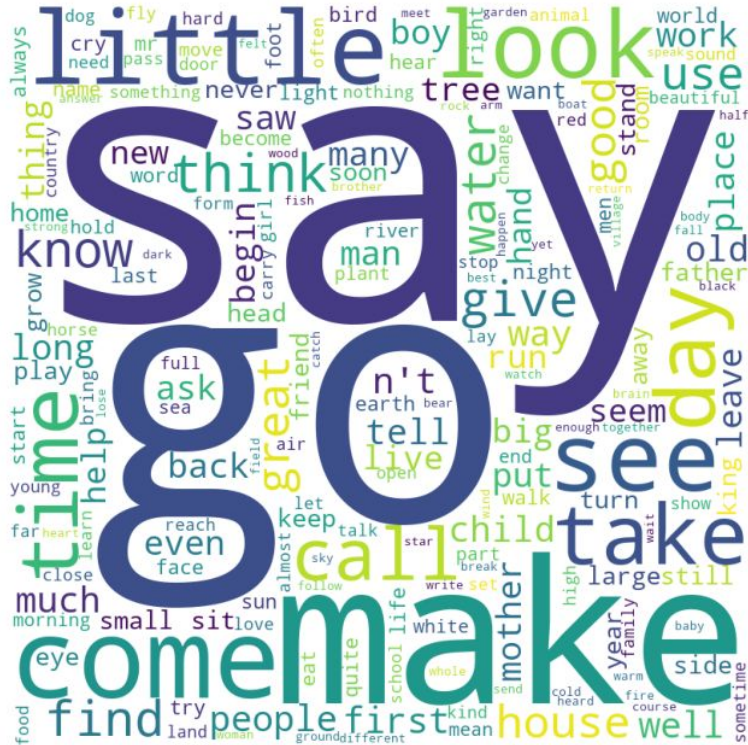


My Data: The Target

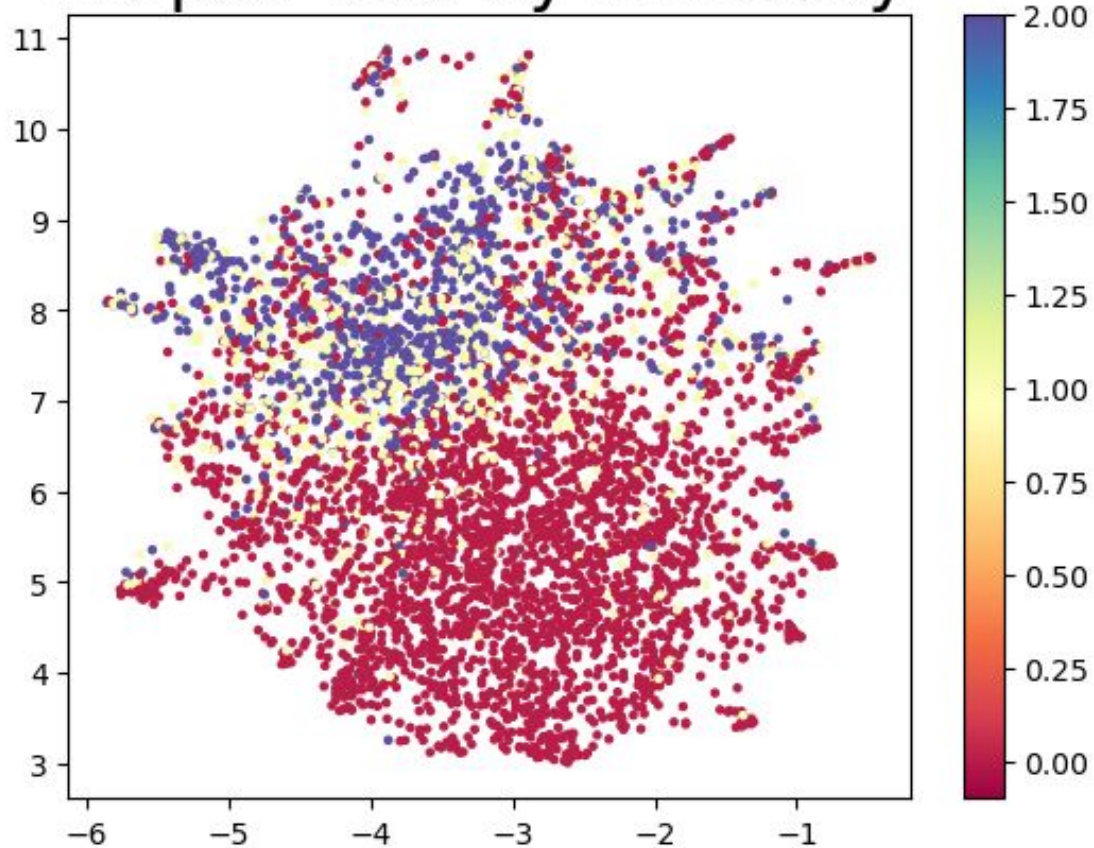


“Easy”(0)	“Medium”(1)	“Hard”(2)
8th Grade and Lower	High School	College
-30 to 50	50 to 60	60 to 120

^ Very Small Range

[illegible]

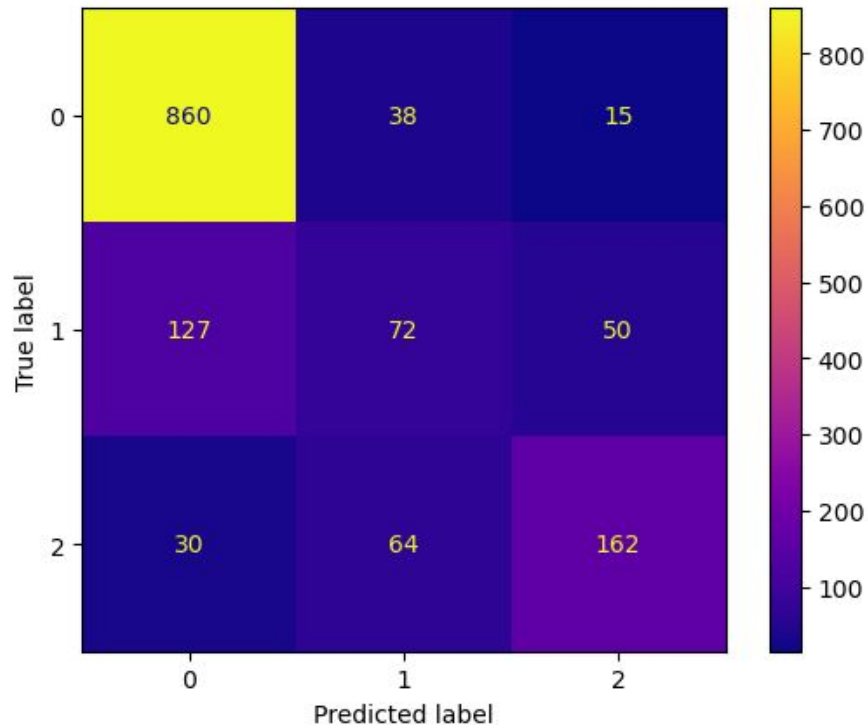
Corpus Text by Difficulty



My Model: XGBOOST Classifier

Test Accuracy Score: 0.77

Train Accuracy Score: 1.0



My Model: Highest TF-IDF for Each Category



“Easy”	“Medium”	“Hard”
<ol style="list-style-type: none">1. Shark2. Marshmallow3. Ia	<ol style="list-style-type: none">1. Color2. Independence3. Spacecraft	<ol style="list-style-type: none">1. Solvent2. Attention3. Carbohydrate

^ Many Science Words

Recommendations:

I would recommend that teachers EVENTUALLY use this model for the following:

1. Classify the readability of novel/new text
2. Get grade appropriate word recommendations for differentiating text



Can YOU tell the difference between these two texts??

“While we learn, the cells of the **brain** (called neurons) connect to each other by reaching out their tiny arms (called axons) or even by growing new arms.”



8th Grade

“Easy”

My model
classified
both as
“Medium”

“The **brain** is constantly trying to find better ways to deliver and deal with information by creating or removing connections between neurons.”

College

“Hard”



Next Steps:

In the next few days, I plan on implementing the following step in order to improve my model:

1. Scrape more diverse texts to my corpus to better train my model especially for “Medium” texts.
2. Feature engineer for authors and diagrams to add more “context” to my model
3. Find a better target for text readability scores
4. Run cosine similarity in order to create book recommendations



Thanks!

Do you have any questions?

ngreen151@gmail.com

<https://github.com/88ngreen88>

CREDITS: This template has been created by **Slidesgo**, and includes icons by **Flaticon** and infographics & images by **Freepik**

Please keep this slide for attribution

