NLP CLASSIFICATION OF JEOPARDY CLUES

BY: JOSHUA BURNS

PANEL

PROBLEM STATEMENT	EDA	DALE-CHALL READABILITY SCORE	MODELING	CONCLUSIONS	WHAT NEXT?
$200	$200	$200	$200	$200	$200
$400	$400	$400	$400	$400	$400
$600	$600	$600	$600	$600	$600
$800	$800	$800	$800	$800	$800
$1000	$1000	$1000	$1000	$1000	$1000

PROBLEM STATEMENT · $200

Background:

Jeopardy, America's favorite quiz show.

- Created in 1964 by Merv Griffin
- Format: 3 contestants, 2 rounds, 6 categories, 5 questions each, followed by Final Jeopardy
- Winner has the mode money at the end of the game
- Modern syndicated format adopted in 1984

⬅ BACK TO PANEL

PROBLEM STATEMENT · $400

Since 1984, Jeopardy has had over 350,000 clues for contestants in tens of thousands of different categories.

⬅ BACK TO PANEL

PROBLEM STATEMENT · $600

What if we could have a machine classify the questions into a category?

Here, we're looking to use machine learning algorithms to correctly classify a Jeopardy clue into it's given category.

⬅ BACK TO PANEL

PROBLEM STATEMENT · $800

The Data:

- A collection of 350k clues collected since daily syndication in 1984 through 2019
- Features include: answer(clue), question(correct response), clue value, round number and air date
- Almost 50k unique categories appear in the dataset.
- Data sourced from Kaggle, where it had been previously cleaned and sorted chronologically

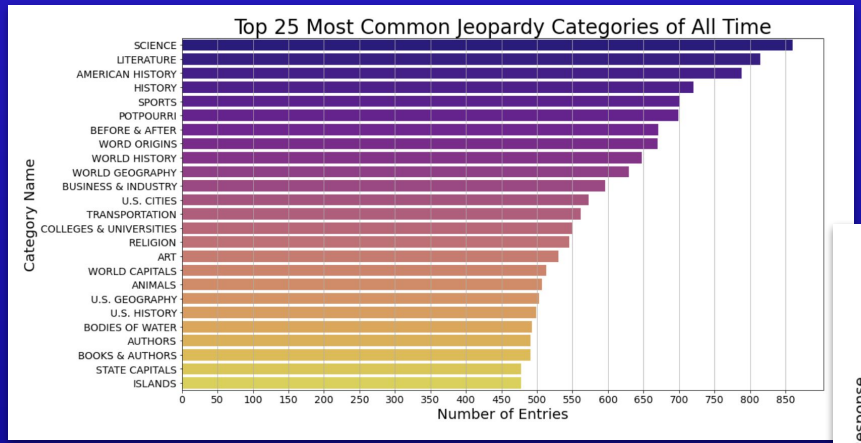
BACK TO PANEL

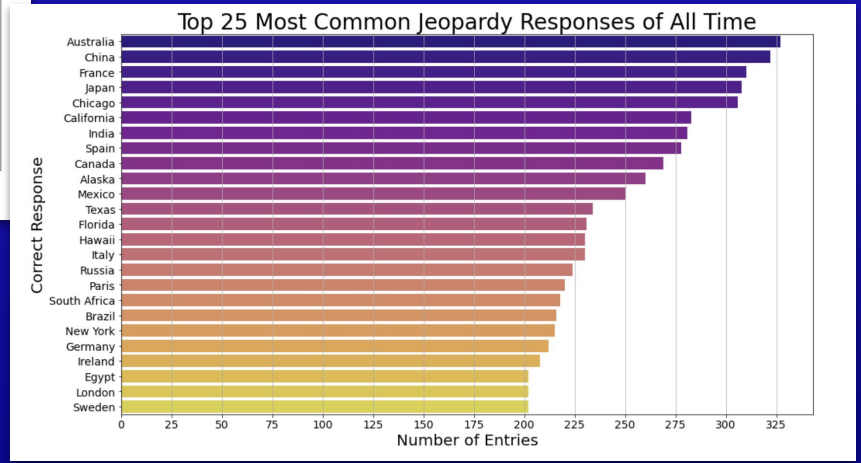
PANEL

PROBLEM STATEMENT	EDA	DALE-CHALL READABILITY SCORE	MODELING	CONCLUSIONS	WHAT NEXT?
	$200	$200	$200	$200	$200
	$400	$400	$400	$400	$400
	$600	$600	$600	$600	$600
	$800	$800	$800	$800	$800
$1000	$1000	$1000	$1000	$1000	$1000


DAILY DOUBLE!

EDA · $200



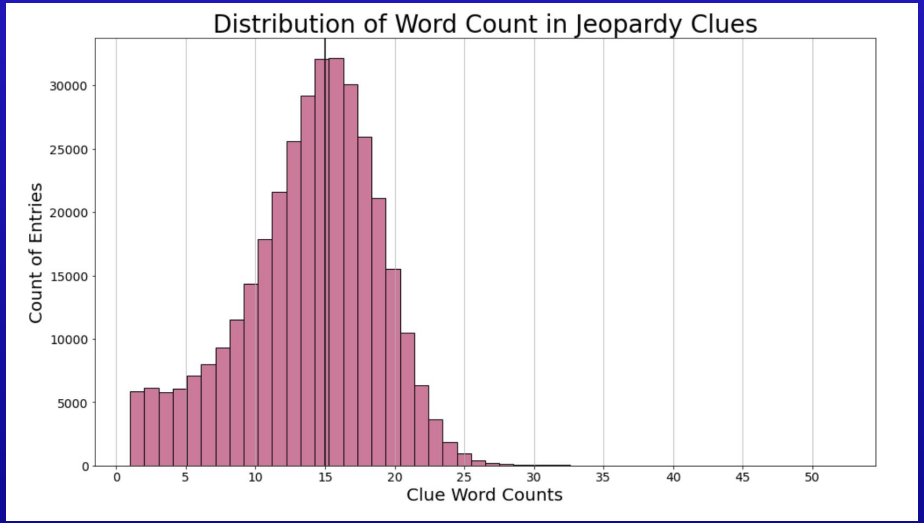


BACK TO PANEL

EDA · $400

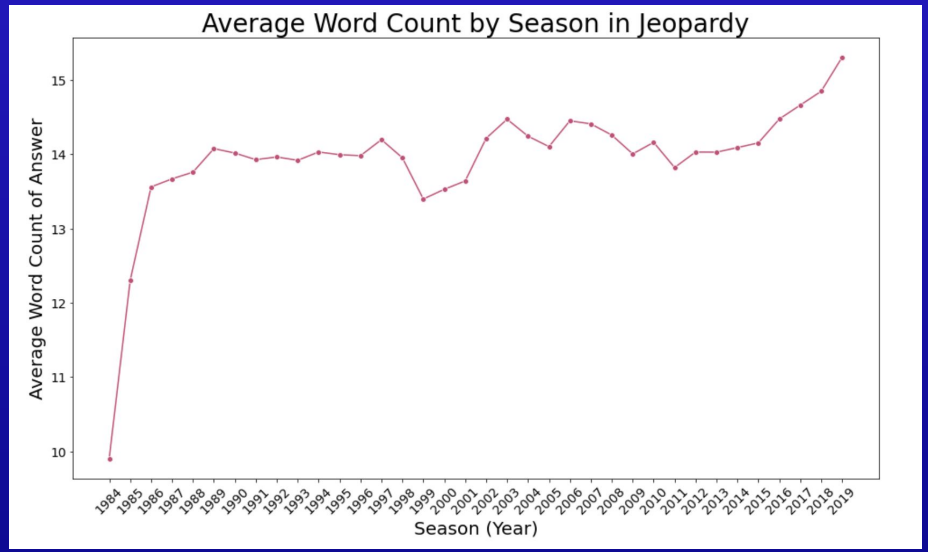
Word counts:

- Median word count: 15
- Minimum word count: 1
- Max word count: 52



⬅ BACK TO PANEL

EDA · $600



Word Count Over Time

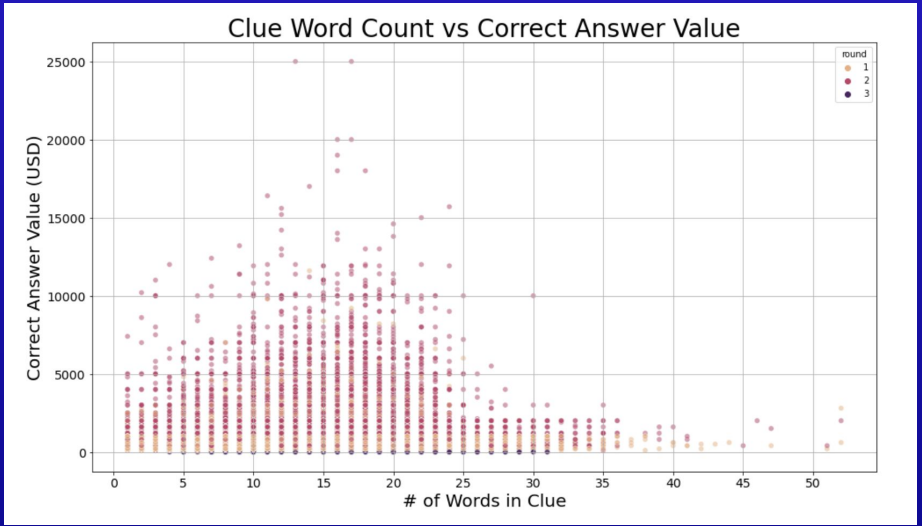
- Broke down clues by year and averaged word count
- Seemingly large increase from 1984 to 1986 (only 3.75 words per year on average.
- Speculation on increase: Popularity of show, technological advances to display capabilities, better clue writers

BACK TO PANEL

EDA · $800

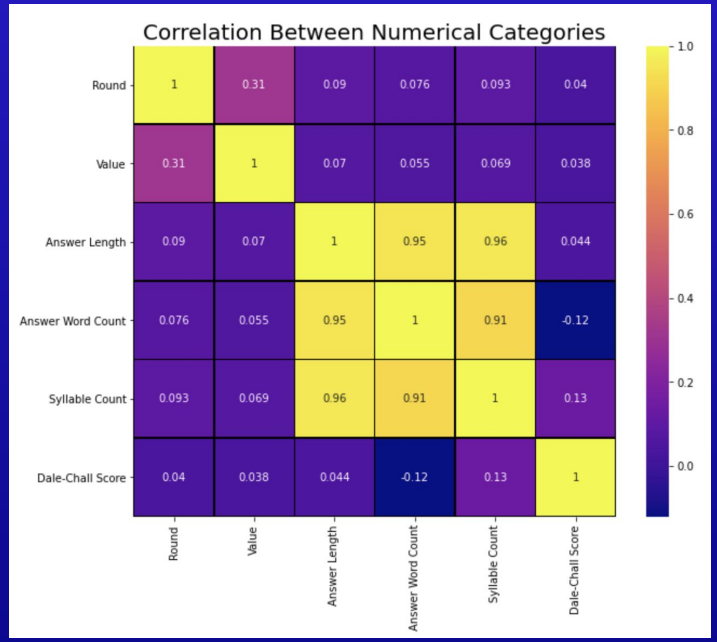
Clue Length vs Correct Answer Value

- Looking for any trend in clue length vs dollars earned.
- Perception = Higher Value:Higher Difficulty
- Outliers = Daily Doubles



BACK TO PANEL

EDA · $1000



Correlations:

- Heatmap for numerical data
- Confirms previous slide:
  - No strong correlation with value & clue length.
- Value/Round correlation makes sense because round 2 values are doubled.
- High correlation block in the center?

BACK TO PANEL

PANEL

PROBLEM STATEMENT		DALE-CHALL READABILITY SCORE	MODELING	CONCLUSIONS	WHAT NEXT?
		$200	$200	$200	$200
		$400	$400	$400	$400
		$600	$600	$600	$600
		$800	$800	$800	$800
$1000		$1000	$1000	$1000	$1000

DALE-CHALL READABILITY SCORE · $200

Feature Creation:

- Textstat library: offers syllable and sentence count, various readability scores.
- Dale-Chall Readability Score: numeric gauge of the comprehension difficulty that readers come upon in a text
- Uses a list of 3,000 words that groups of fourth-grade American students could reliable understand. Any word not on the list is considered to be difficult.

⬅ BACK TO PANEL

DALE-CHALL READABILITY SCORE · $400

- Original formula from 1948:

$$0.1579\left(\frac{\text{difficult words}}{\text{words}} \times 100\right) + 0.0496\left(\frac{\text{words}}{\text{sentences}}\right)$$

- Updated in 1995 to increase the vocabulary to 3,000 words
- If the percentage of difficult words is above 5%, add 3.6365 to the raw score to get the adjusted score, otherwise the raw score is used.

BACK TO PANEL

DALE-CHALL READABILITY SCORE · $600

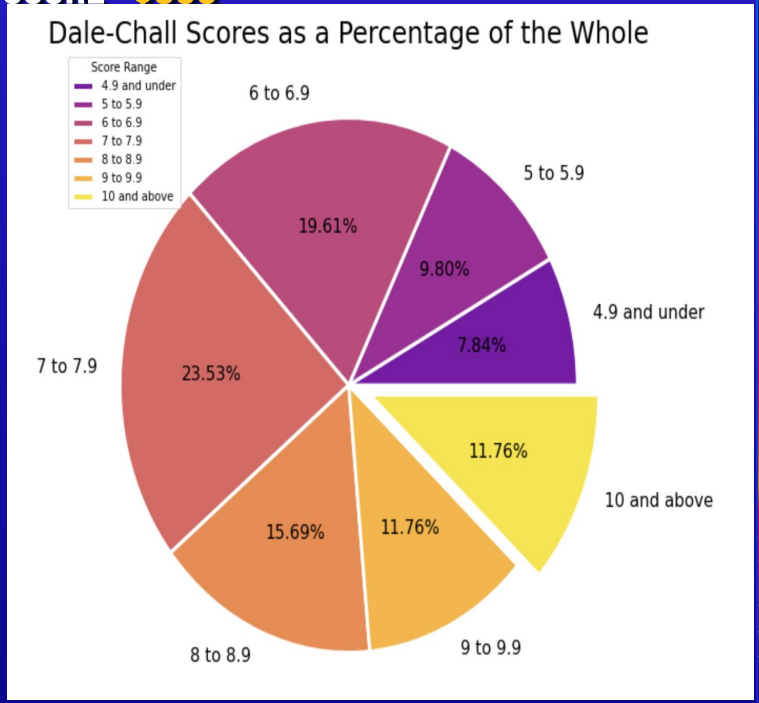
Dale-Chall Results Chart:

Score	Notes
4.9 or lower	easily understood by an average 4th-grade student or lower
5.0–5.9	easily understood by an average 5th or 6th-grade student
6.0–6.9	easily understood by an average 7th or 8th-grade student
7.0–7.9	easily understood by an average 9th or 10th-grade student
8.0–8.9	easily understood by an average 11th or 12th-grade student
9.0–9.9	easily understood by an average 13th to 15th-grade (college) student

BACK TO PANEL

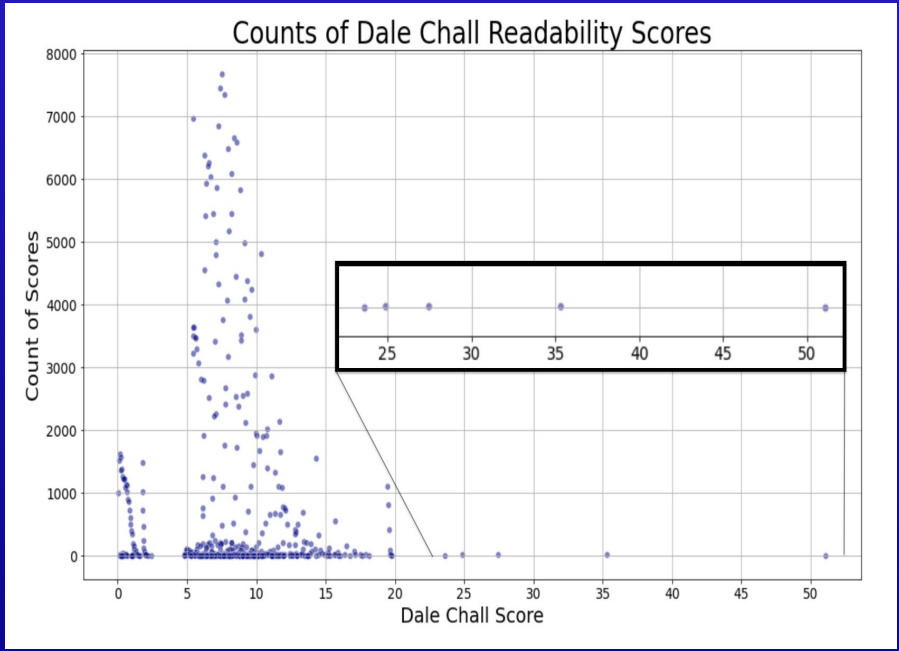
DALE-CHALL READABILITY SCORE - $800

- Dale-Chall scores broken down by range.
- Highest concentration 7 to 7.9: 9th to 10th-grade readability
- ≈12% over 10: post graduate level readability
- Look at those in a moment...



BACK TO PANEL

DALE-CHALL READABILITY SCORE · $1000



- Plot of value counts of Dale-Chall scores.
- Outliers: let's focus on the extreme scores.
- Most of the outliers had no spaces or several special characters.

BACK TO PANEL

PANEL

PROBLEM STATEMENT			MODELING	CONCLUSIONS	WHAT NEXT?
			$200	$200	$200
			$400	$400	$400
			$600	$600	$600
			$800	$800	$800
$1000			$1000	$1000	$1000

MODELING · $200

- After EDA, the dataset was shortened to the Top 25 categories we looked at previously.
- The dataset was split into training and testing data and baseline accuracy was established.
- Each category had between a 3% and 5% chance of being pulled from a random selection.

⬅ BACK TO PANEL

MODELING · $400

- The data was then vectorized:
  - Count Vectorizer
  - TFIDF Vectorizer
- Both vectorizations were fit on default parameter Multinomial Naive Bayes model to see which performed better on training data.
- Average accuracy scores are as follows:
  - Count Vectorizer: 0.7144
  - TFIDF Vectorizer: 0.7929

⬅ BACK TO PANEL

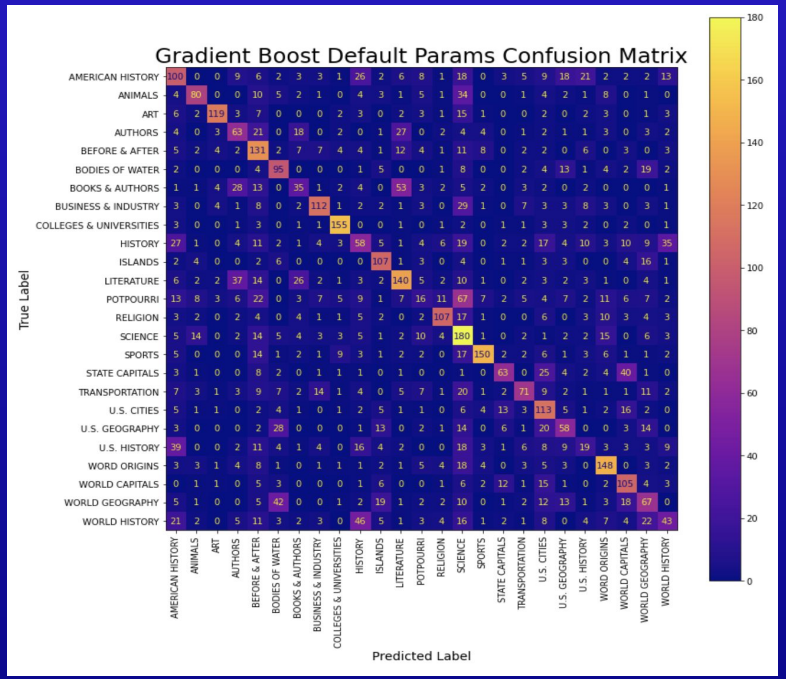
MODELING · $600

- With a vectorizer in hand, several more models were fit with default parameters to look for better performance:
  - Decision Tree
  - Random Forest
  - AdaBoost / Gradient Boost
  - K-Nearest Neighbors

⬅ BACK TO PANEL

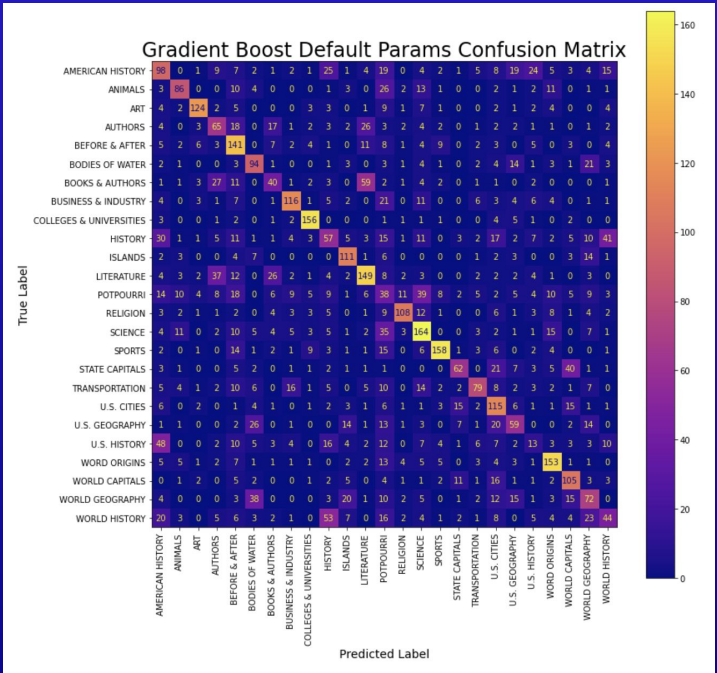
MODELING · $800

- Training average accuracy ranged from 0.243 to 0.999
- Best performance (without overfitting): Gradient Boost - 0.896 accuracy score
- Testing accuracy: 0.471



BACK TO PANEL

MODELING · $1000



Gridsearch Optimization

- 2 iterations or gridsearches
  - Mix of default and updated parameters.
  - notable : n_estimators kept increasing as score improved.
- Final scores:
  - First: 0.483
  - Final: 0.486

BACK TO PANEL

PANEL

PROBLEM STATEMENT				CONCLUSIONS	WHAT NEXT?
				$200	$200
				$400	$400
				$600	$600
				$800	$800
$1000				$1000	$1000

CONCLUSIONS · $200

Initial Reactions:

- I was hopeful for a good model after such great overall training scores.
- Unfortunately, those hopes were dashed with results.

⬅ BACK TO PANEL

CONCLUSIONS · $400

- My suspicion is that the similarity in some of the category names aided in the confusion of the model.
- Categories such as: U.S. History and American History can easily be confused by the machine (evidenced in confusion matrix).

⬅ BACK TO PANEL

CONCLUSIONS · $600

- Time and computational resources were the enemy here.
- Models took several hours to train even while using the maximum processing cores available.

⬅ BACK TO PANEL

PANEL

PROBLEM STATEMENT				CONCLUSIONS	WHAT NEXT?
					$200
					$400
					$600
				$800	$800
$1000				$1000	$1000

WHAT'S NEXT? · $200

With a direction on how to proceed, future iterations of this model can be further optimized to confuse the classifier less.

⬅ BACK TO PANEL

WHAT'S NEXT? · $400

After the initial models performed so poorly, I attempted some clustering on the clues to make for better classification, however, the first round of clustering returned mostly noise

⬅ BACK TO PANEL

WHAT'S NEXT? · $600

It would likely be wiser in future versions to cluster categories instead of questions to see if that produces better results

⬅ BACK TO PANEL

WHAT'S NEXT? · $800

Combining redundant categories under one umbrella, such as "History" or "Literature" as opposed to sub-categories.

BACK TO PANEL

PANEL

PROBLEM STATEMENT					WHAT NEXT?
				$800	
$1000				$1000	$1000

QUESTIONS?


THANK YOU!

CREDITS

Special thanks to all the people who made and released these awesome resources for free:

- Presentation template by SlidesCarnival

PRESENTATION DESIGN

This presentation uses the following typographies:

- Titles: Bebas Neue
- Body copy: Della Respira

Download for free at:

https://www.fontsquirrel.com/fonts/bebas-neue

https://www.1001fonts.com/della-respira-font.html

You don't need to keep this slide in your presentation. It's only here to serve you as a design guide if you need to create new slides or download the fonts to edit the presentation in PowerPoint®