Classifying Tales of Terror

Audrey Maldonado - April 2, 2021

About Me

- Upstate New Yorker
- Chicken Farmer
- Librarian

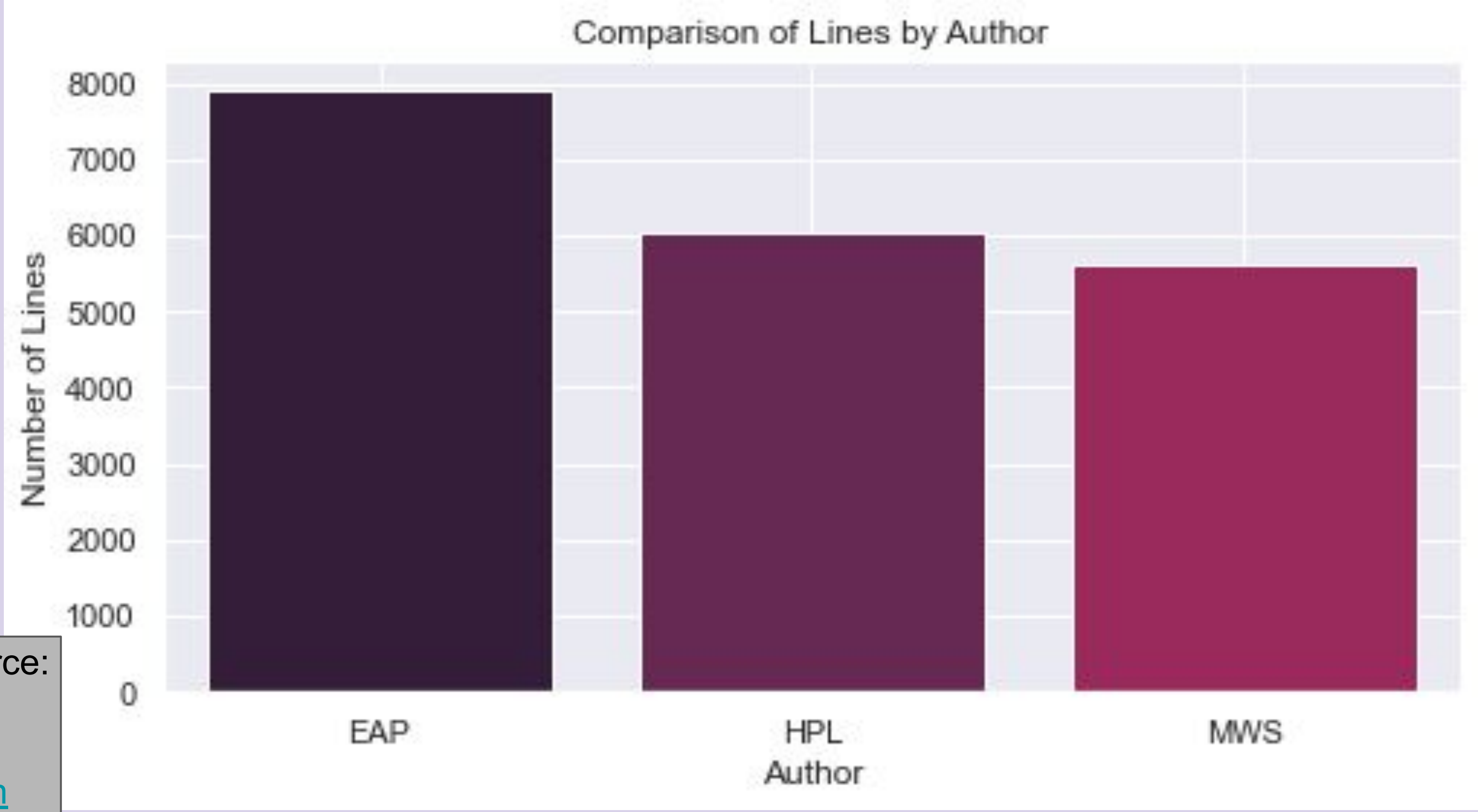
Tech Trivia: In this photo I am on my way to a wig party, held directly across the street from where Samuel F.B. Morse developed the first working telegraphic machine. What town am I in?

Why do we need a taxonomy of terror?

People have strong feelings about what makes a good book. How can book sellers and libraries help individual readers quickly find books they love?

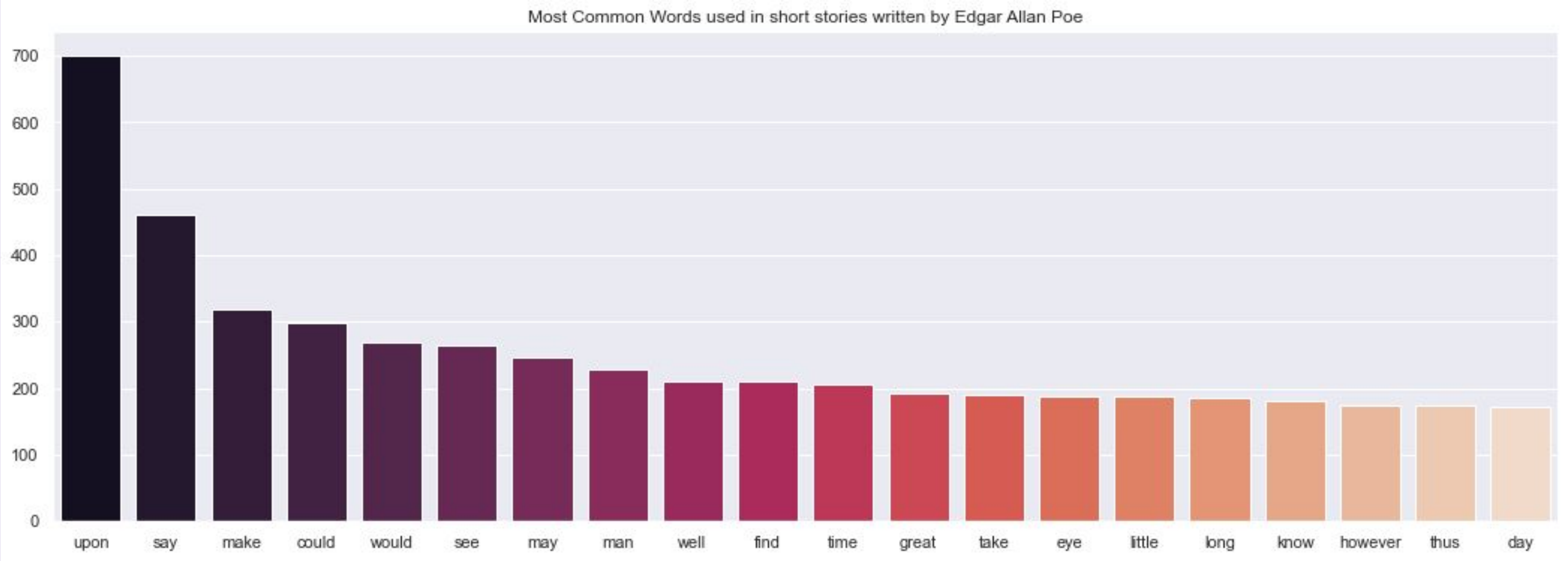
I chose to begin my work with the horror genre because I hypothesized that this highly recognizable genre would be one of the easiest to parse. Categorizing by author is a valuable first step in determining genre and subgenre.

The Spooky Story Dataset : ~20k lines of text



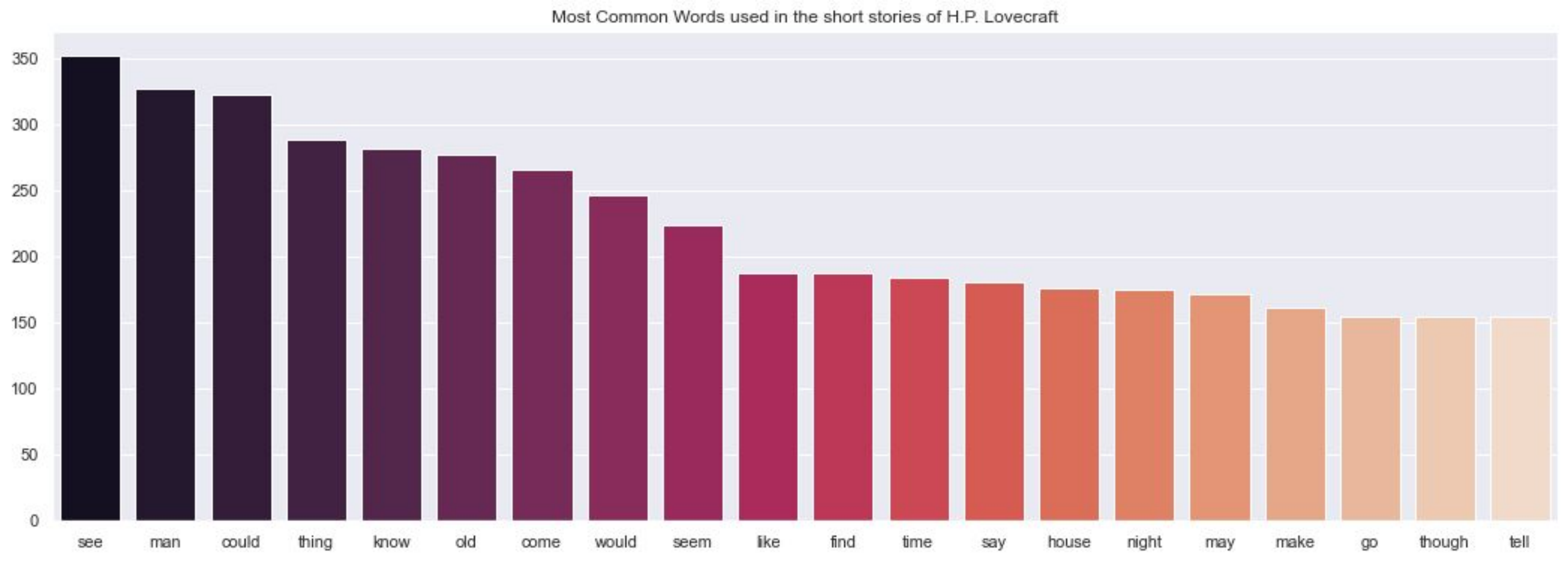
Data Source: [Kaggle 2017 Halloween Competition]

Getting to know Poe



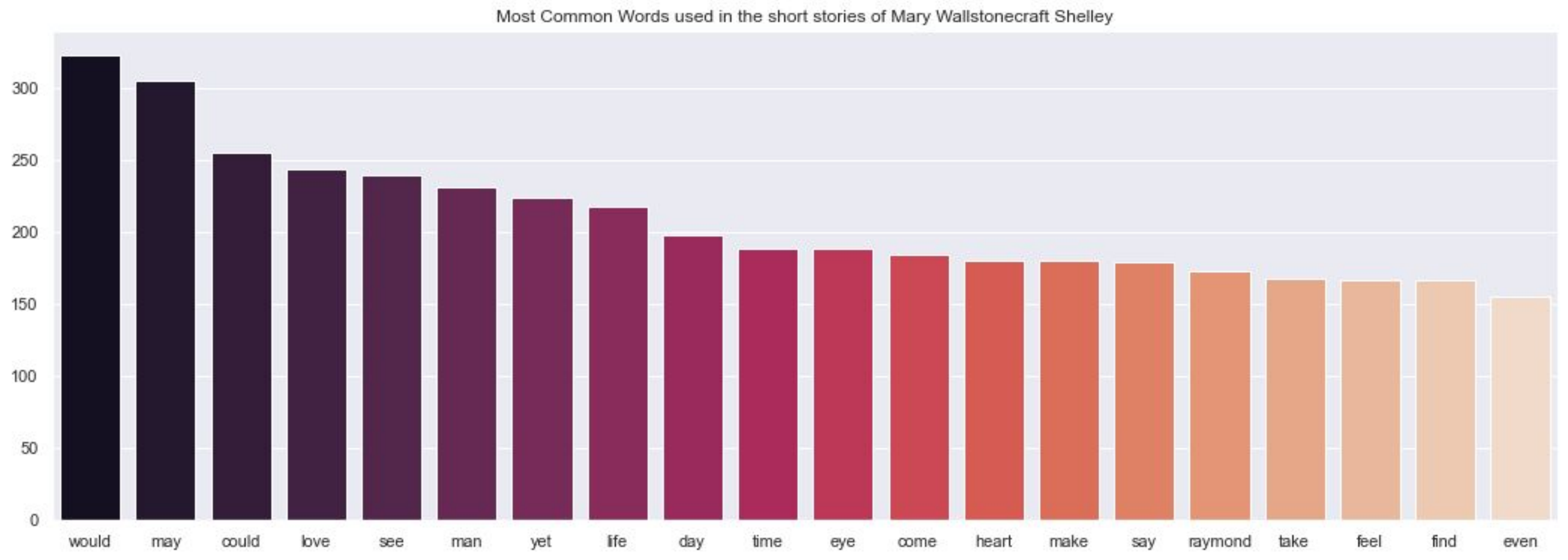
Vocabulary consists of 15,261 words

Getting to know Lovecraft



Vocabulary consists of 14,504 words

Getting to know Shelley



Vocabulary consists of 11,515 words

On Victorian Stopwords:

In his nearly 8000 lines of text, **Poe uses the word “upon” nearly 700 times** and that plays a role in differentiating him.

This leads me to wonder: Would allowing this kind of marker to persist affect the ability to apply this model to other data? **Do the victorian stopwords cause my model to overfit or are they useful in a purpose-built application?**

Word2Vec → Similarity Explorations

Word	Some “Similar” Words
fear	Miserable, innocent, suffer, starvation, hurt, alas
death	Life, conscious, wondering, dreaded, struggle, loss, reality, pain
sadness	Dining, division, brilliancy, exceptional, crypt, floor
bones	Depth, lore, majestic, dollars, vast, yard, monsters, rattling
ghost	Tongue, angel, kingdom, intolerable, transom, daemon
bird	Grave, fiendish, morbid, grotesquely, aged, overhead
madness	Destined, memory, earthly, mood, torture, honor, purpose
evil	Animal, hideous, monstrous, science, bewildered, editor

Pre-processing → CountVectorizer → Naive Bayes

I tested out Linear Support Vector Classification, but found that to be less predictive than the Multinomial Naive Bayes model for this dataset.

Author Classifiers	Accuracy	Predictive Feature Samples
Poe | Lovecraft | Shelley	81%	Aberrant, abhor, little, good, great, time, eye, think, far, word, mean, look
Poe | Lovecraft	88%	Abandonment, abdication, man, old, house, say, night, know, night, dream, day, look
Poe | Shelley	83%	Love, life, man, eye, heart, time, father, hope, little, friend, death, think

Conclusions and Next Steps

This exploration leads me to believe a taxonomy of terror will be useful in creating a metric of spookiness and that author classification is part of that process.

← **Pathways** →

1) Explore AI text generation and chatbots
2) Explore ways models might “reward” words from the taxonomy of terror in classifying within the genre
3) Incorporate more authors and time-periods in the corpus

Time for Questions and Contact Info:

@adorism80 - Mostly makerspace and library stuff

@berlbane4d - Mostly programming and data science stuff

Email: audrey.maldonado@gmail.com