SPOOKY AUTHOR IDENTIFICATION!

By: Tarfah Alabbad & Muneera alshunifi

CONTENTS

INTRODUCTION

"This suspense is terrible.
I hope it will last."

- Oscar Wilde, *The Importance of Being Earnest*

INTRODUCTION

Problem

Plagiarism and author's impersonation and some other crimes has increased nowadays

Solution

Identify the author of given sentence to help reducing those crimes

INTRODUCTION

Goal

- Predict the author of given sentence
- Topic modeling for each text

TOOLS

- Different Python packages using Google colab
- Pandas and NumPy Matplotlib, seaborn for modeling
- NLTK and genism and spacy for text pre-processing and cleaning
- Plotly and wordcloud for visualization

WORK FLOW

Acquire data

preprocessing

Modeling and evaluation

conclusions

01 DATA

Source: Kaggle website

Description: Contains text from works of fiction written by spooky authors public domain Edgar Allan Poe, HP Lovecraft and Mary Shelley

Shape: 19579 x 3

Feature	Description	Data Type
ID	Unique identifier for each sentence	object
Text	Some text written by one of the authors	object
Author	Author of the sentence in a shortcut format (EAP: Edgar Allan Poe, HPL: HP Lovecraft, MWS: Mary Wollstonecraft Shelley)	object

02 PREPROCESSING

EDA
Using plotly

LEMMETIZING
Add it as new feature

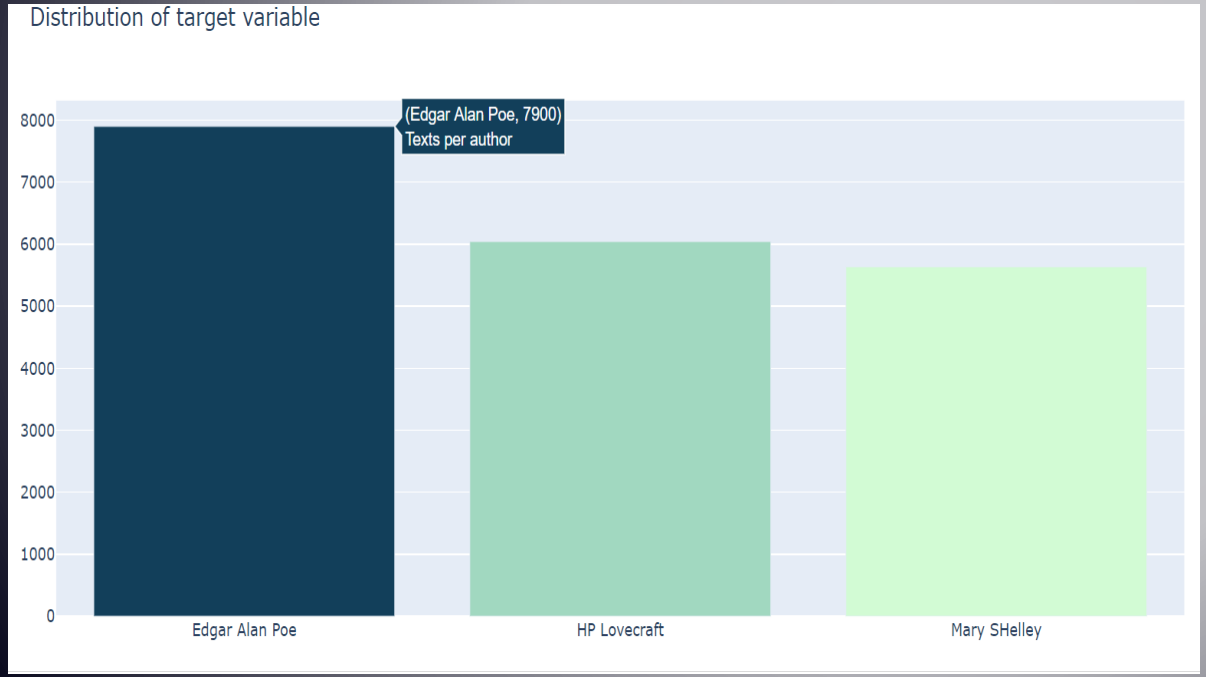
STOP WORDS REMOVAL
Words as “the, it..etc”

LOWERCASING
Convert text to lowercase

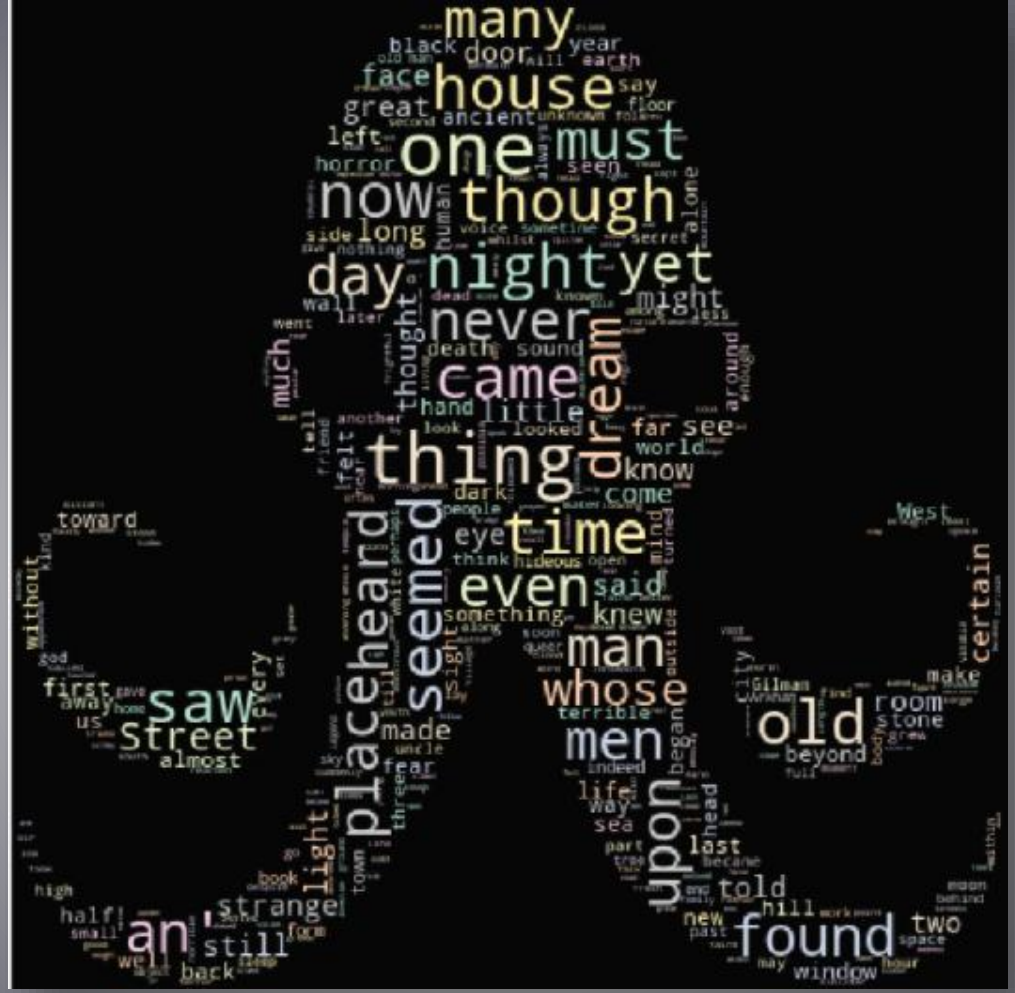
FEATURE ENGINEERING
Count Vectorize
TF-IDF

RAW LENGTH
Add it as new feature
(count words in text)

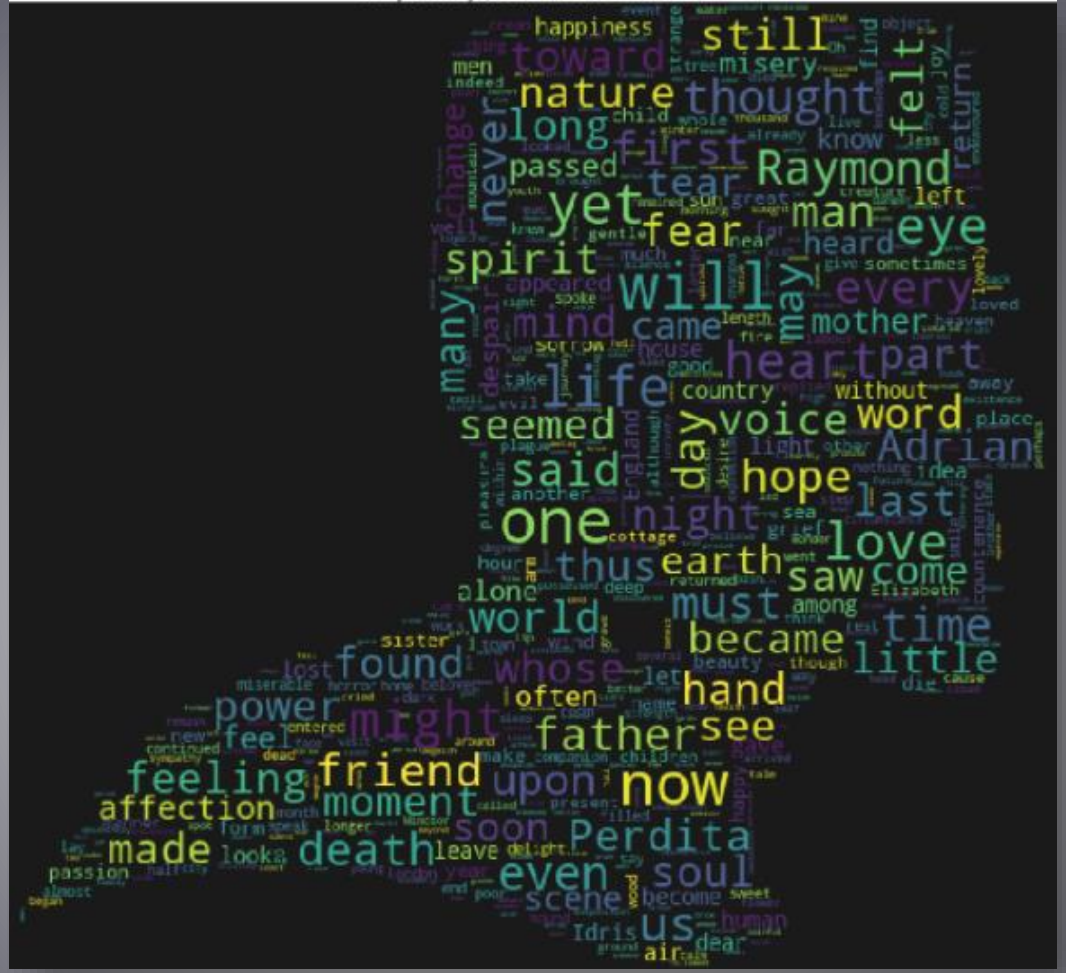
EDA



The plot shows that Edgar Alan Poe has the most written texts.

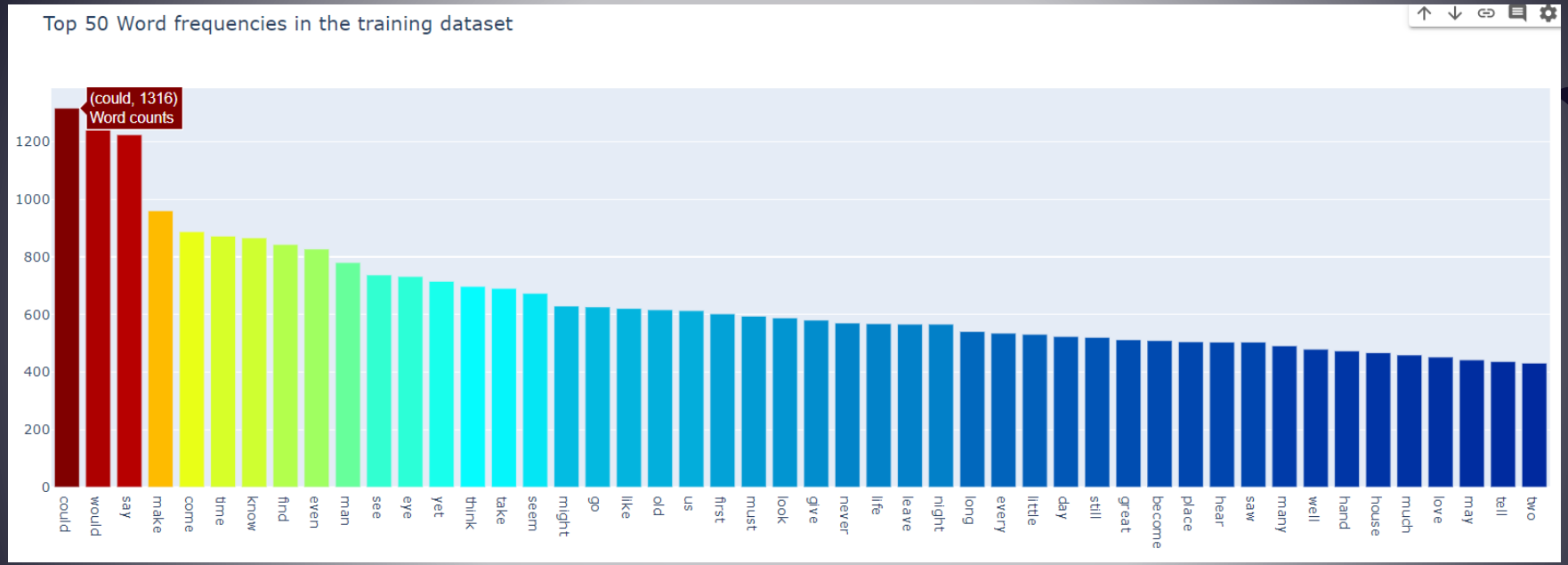


HP love craft cloud word (squid)

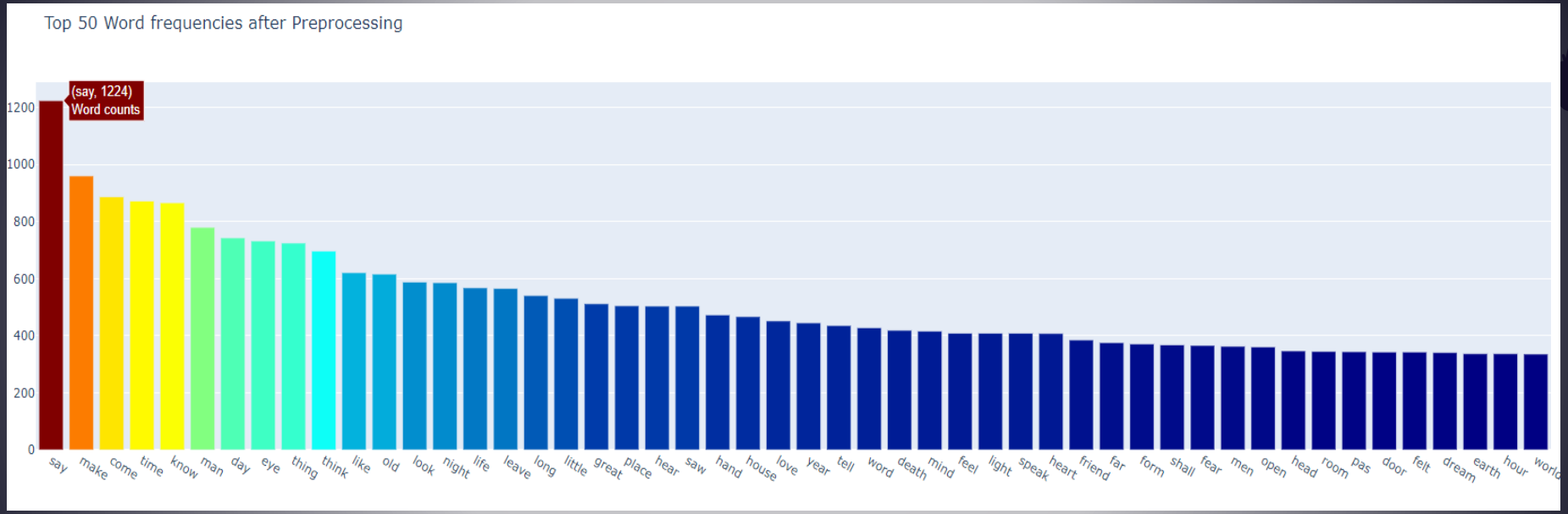


Marry Shelley cloud word (Frankenstein monster)

Edgar Allen Poe cloud word (Raven)

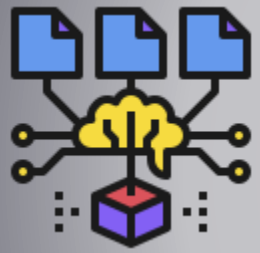


Top 50 word frequencies in the data



Top 50 word frequencies in the data after preprocessing

03 MODELING AND EVALUATION



Multinomial Naïve bayes

SVM

LDA

MODELING AND EVALUATION

- Tried modeling two times, once using Count Vectorize and once using TF-IDF
- Splitting data to 30% test and 70% train
- Classification evaluation metric: F1-score

EVALUATION SCORES (TF-IDF)

	SVM	SVM AFTER SVD AND SCALING	MULTINOMIAL NAïVE BAYES
F1-SCORE	80%	70%	80%

EVALUATION SCORES (COUNT VECTPRIZE)

	SVM	SVM AFTER SVD AND SCALING	MULTINOMIAL NAïVE BAYES
F1-SCORE	76%	65%	81%

EVALUATION SCORES (COUNT VECTPRIZE)

	SVM	SVM AFTER SVD AND SCALING	MULTINOMIAL NAïVE BAYES ★
F1-SCORE	76%	65%	81%

LDA

01 BASE-MODEL

02 LDA TUNING

03 TOPIC-DOCUMENT

04 WORD-TOPIC

TOPIC 1 NATURE

Light - Earth –Tree – Wind – Sun -Air

TOPIC 2 DRAMA

Love- Feel- Dream –Friend –Miserable-Madness

TOPIC 3 GHOSTS

Spirit – Horror-Death –Fear –Kill- Terror

04 CONCLUSION

- **The best classification model:** Multinomial naïve bayes with 81% fi score
- **Topic modeling result:** 3 different topics related to each author we have nature for Edgar Allan Poe, ghost HP Lovecraft and drama Mary Shelley
- **Future work:** recommendation system and deploy it

THANK YOU !

Any questions?