Spooky Author Identification

Capstone Project By Karthik Pendyala

Mentor – Rohan Rao

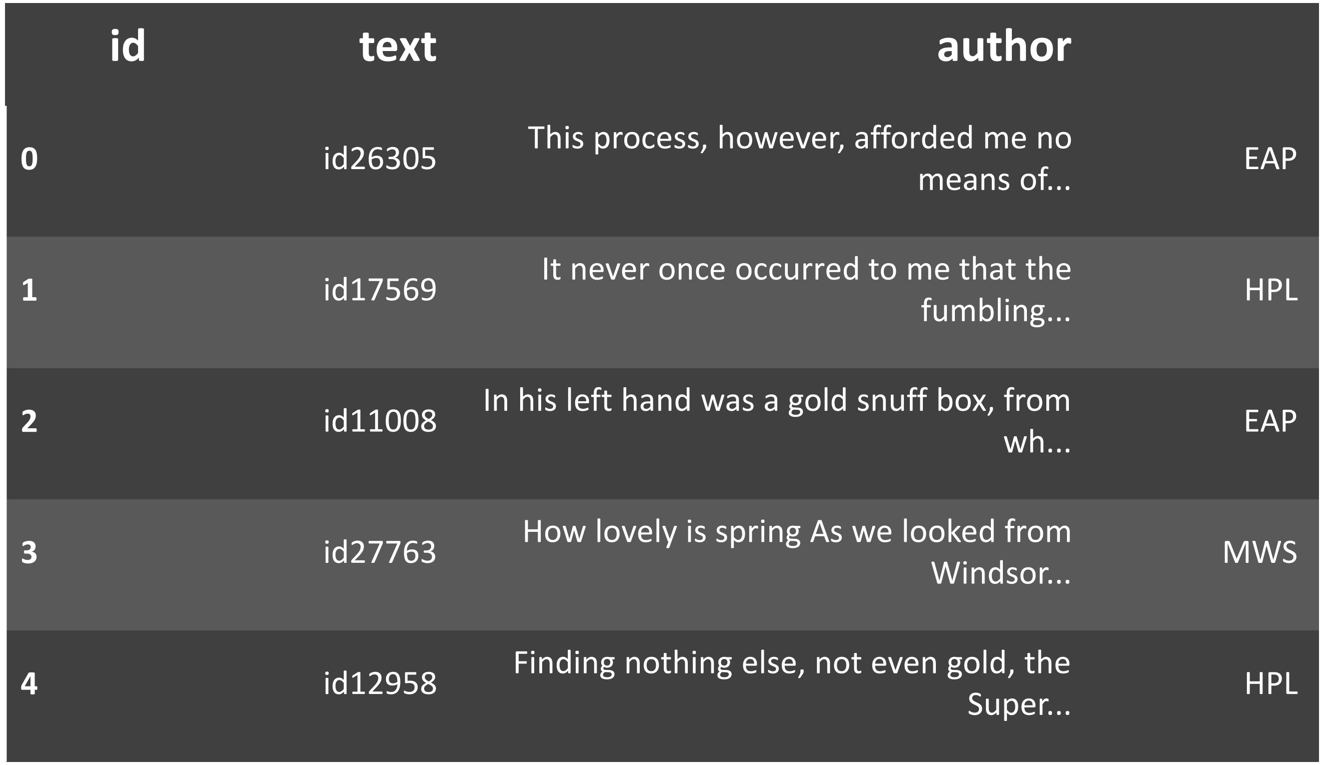
Springboard AI/ML Career Track

Problem Statement

- The problem this capstone project aims at identifying horror authors from their writings. By analyzing the author style and the way of writing, the project aims at providing a model that could accurately detect the name of the author given an input text.

Data

- The dataset contains text from works of fiction written by spooky authors of the public domain: Edgar Allan Poe, HP Lovecraft and Mary Shelley. The data was prepared by chunking larger texts into sentences using CoreNLP's MaxEnt sentence tokenizer.



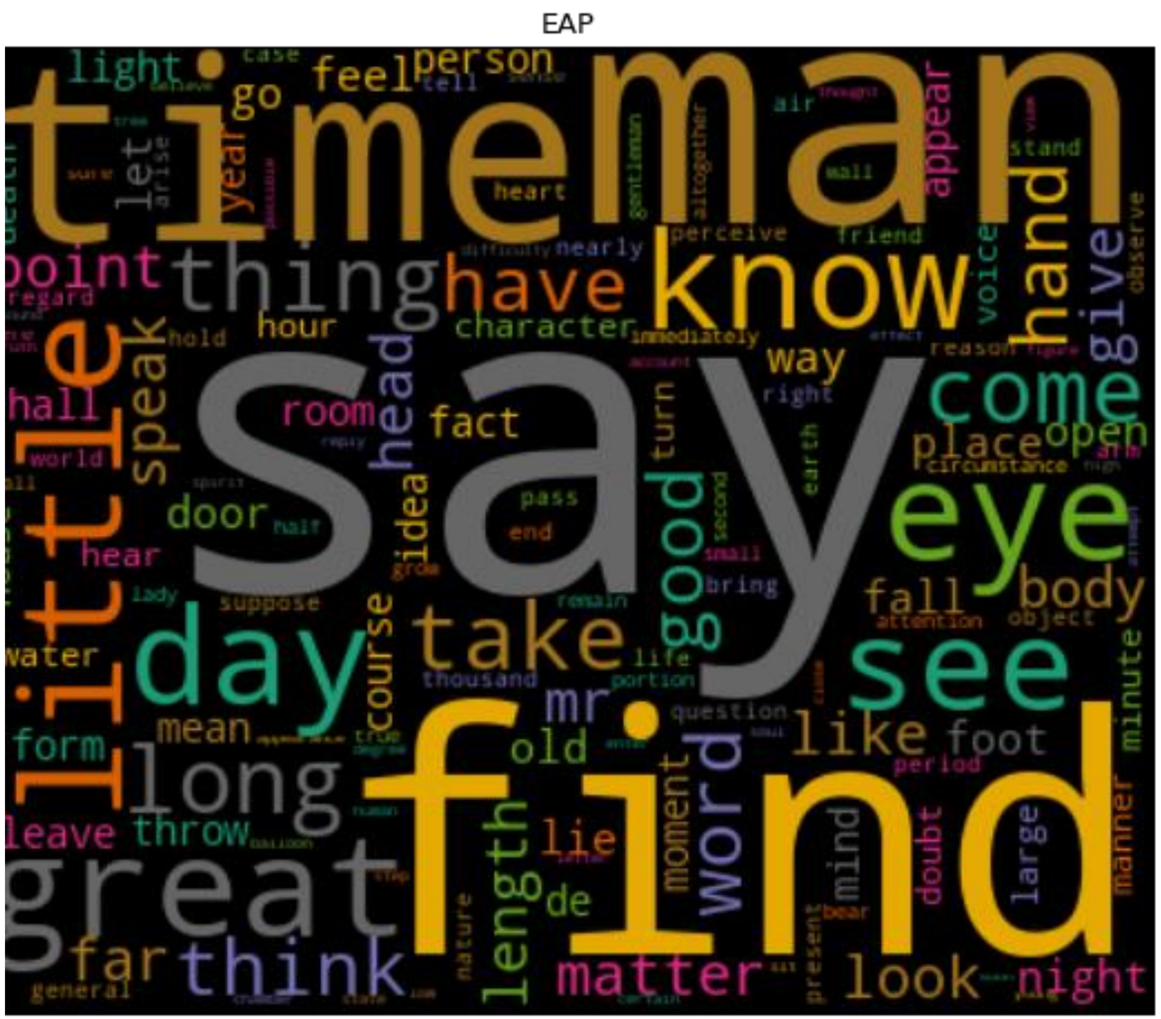
id	text	author	
0	id26305	This process, however, afforded me no means of...	EAP
1	id17569	It never once occurred to me that the fumbling...	HPL
2	id11008	In his left hand was a gold snuff box, from wh...	EAP
3	id27763	How lovely is spring As we looked from Windsor...	MWS
4	id12958	Finding nothing else, not even gold, the Super...	HPL

Preprocessing

- Text data typically requires some cleanup before it can be processed further and fed to a model. The dataset was cleaned as follows:
    - Removal of Punctuation Marks
    - Lemmatisation
    - Removal of Stopwords
    - Label encoding the output label - Convert Author Names into numeric format for training purpose

For Edgar Allan Poe

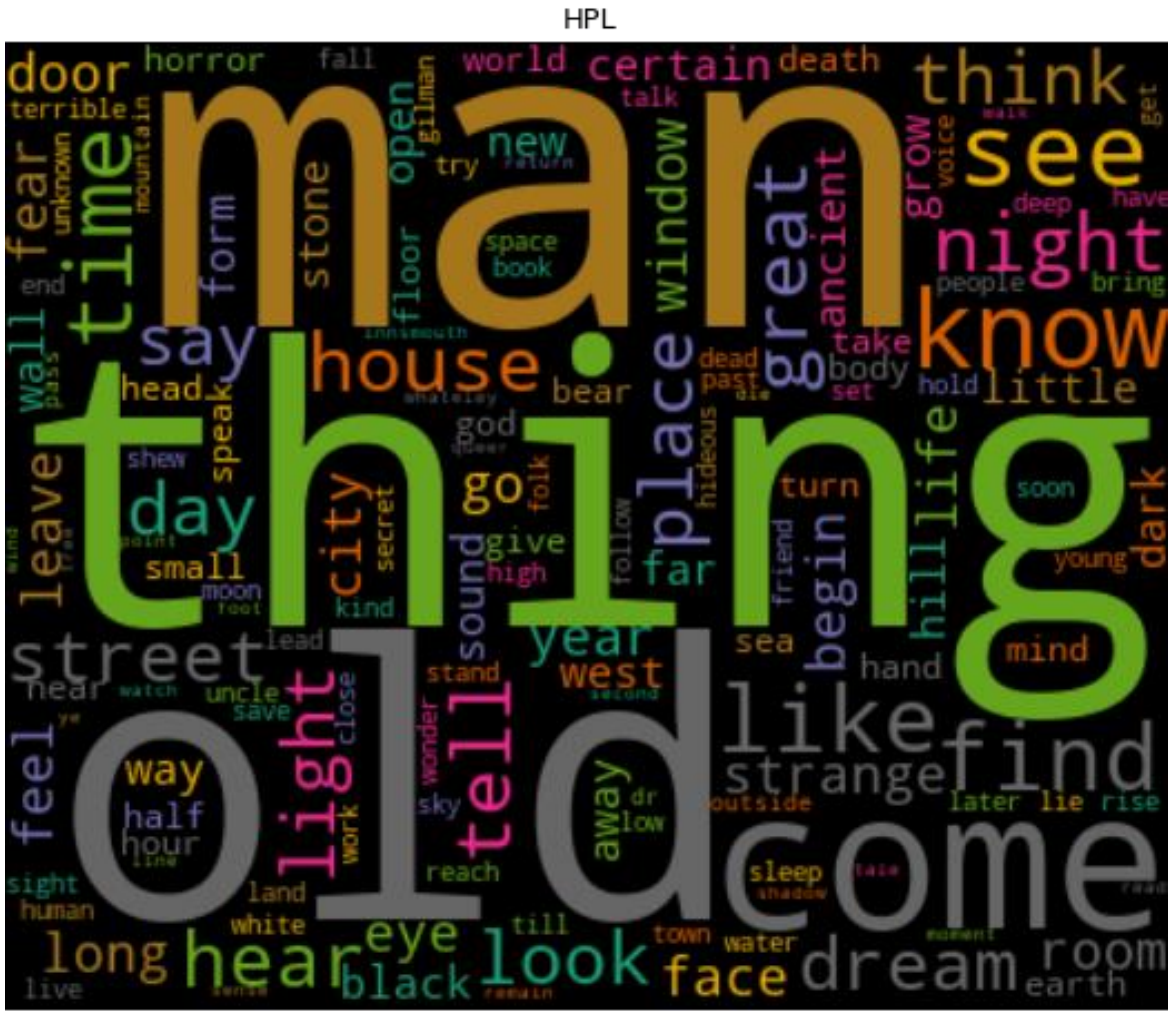
WordClouds



EAP

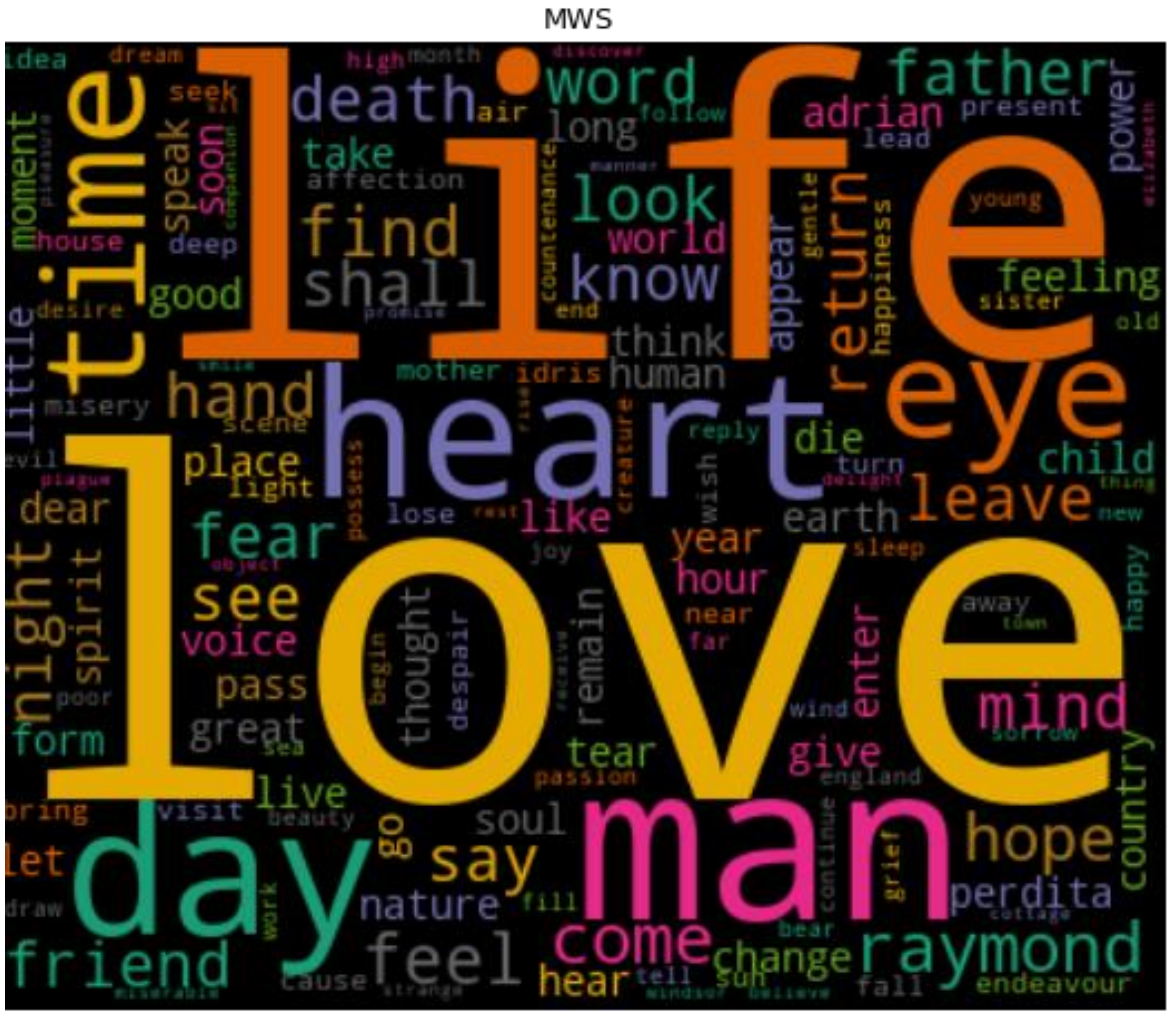
For HP Lovecraft

WordClouds



For Mary Shelley

WordCloud



Modeling

Logistic Regression

Label	Precision	Recall	F1-score	Support
0	0.77	0.86	0.82	1562
1	0.86	0.76	0.80	1149
2	0.83	0.80	0.81	1205

Modeling

Multinomial Naïve Bayes

Label	Precision	Recall	F1-score	Support
0	0.85	0.82	0.84	1562
1	0.87	0.83	0.85	1149
2	0.81	0.87	0.84	1205

Modeling

- To try out some deep learning models I made use of Hugging Face which has several pre-trained models to train on.
- I trained the dataset on different transformers such as:
  - Bert Base Cased
  - Bert Base Uncased
  - Bert Large Cased
  - Bert Large Uncased
  - Distilbert Base Cased
  - Distilbert Base Uncased
  - Roberta Base
  - Roberta Large
  - XLM Roberta Base
  - XLM Roberta Large
- After trying different hyperparameters and considering the accuracy of the models, Bert Large Cased has been used as the final model.
- It was able to achieve an accuracy of 88% on the validation set and better than all the machine learning and deep learning models.

Conclusion & Future Work

- The model evaluation results indicate that deep learning models perform better than Naïve Bayes models and are efficient in predicting the result of the problem statement.
- Future work will be mainly focused on exploring other deep learning architectures and trying out different model tuning methods.

References

- TPU Sherlocked: One-stop for Hugging Face with TF
- Deploy the final Model
- Multinomial NB
- Logistic Regression
- Hugging Face Transformers