Author Identification

Natural Language Processing Project

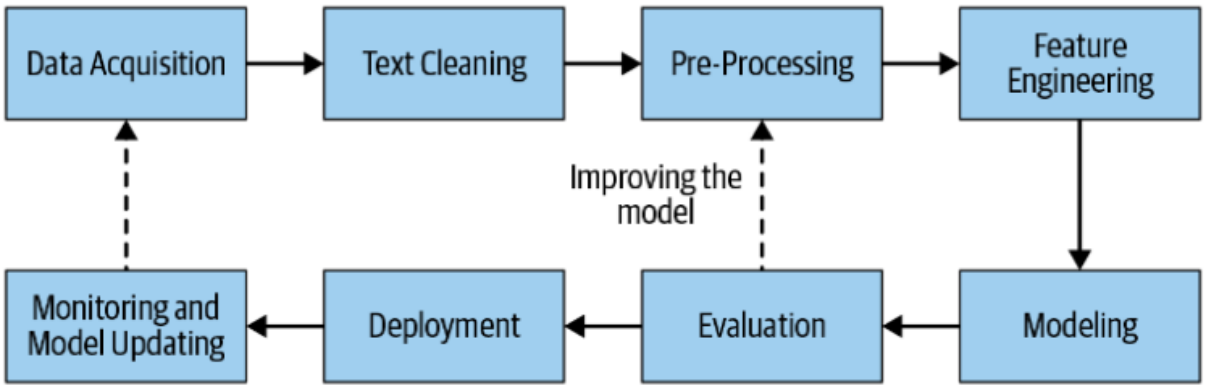
By: Muneera Alshunaifi, Tarfah Alabbad

Overview

In recent years, authorship analysis of anonymous texts in the Internets has received some attention in cyber forensic and data mining communities. Authorship analysis is the study of linguistic and computational characteristics of the documents written by known or unknown authors.

The problem is determining the true author of texts was a task of social interest from the moment it was possible to attribute the authorship of words. With the development of statistical techniques and due to the wide availability of data that can be accessed from the internet, authorship analysis has become a very practical option.

The aim of this project is to predict the author of given sentence, the key idea is to exploit the differences of the writing styles of the authors and use this information to build our models. We will try many experiments in our project using varies machine learning models and a pre-trained language model and we will focus on some performance measure metrics to evaluate the models such as: Precision, Recall, F1-score and we will see the accuracy confusion matrix. To accomplish the project we will follow the following Natural Language Processing pipeline:



Data Description

The dataset obtained from Kaggle website: https://www.kaggle.com/c/spooky-author-identification/data, it contains text from works of fiction written by spooky authors of the public domain: Edgar Allan Poe, HP Lovecraft and Mary Shelley

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

Tools

To explore and analyze the data and do the prediction models in python, we will use Jupyter notebook and Python packages, such as: Pandas and NumPy Matplotlib, seaborn and SKLearn for modeling. NLTK and Textblob for text pre-processing and cleaning the text Also we may use extra tools such as PowerBi for visualization and Flask framework for deployment.

Algorithms

We will try many experiments in our project using varies machine learning models and a pre-trained language model, the pre trained language model consider as feature engineering step and we will be using TF-IDF which is statistical measure used to evaluate how important a word is to a document in a collection of documents or corpus. This importance is directly proportional to the number of times a word appears in the document. However, the next step of feature engineering methods we will use is Count Vectorize which is a feature extraction tool that select the words/features/terms which occur the most frequently. In addition, , we will obtain two different types of machine learning models:

- Supervised: Multinomial Naïve bayes
- Unsupervised: Topic modeling (Latent Dirichlet Allocation LDA)

Conclusion

We aim to know the author who wrote a sentence and discover the difference of written styles for the authors by using machine learning algorithms by following the NLP standards and methods. To sum up, predicting the true author of sentence will help avoiding crimes such as author identity theft.