Author Identification – project proposal

Business Problem

The aim of this project is to predict the author of some given paragraphs based on the author's writing style. In particular, we will be analyzing texts from Edgar Allan Poe, Mary Shelley, and HP Lovecraft with the goal to predict the probability for each excerpt to be written by one of these authors.

This proposed project might be helpful for an online bookstore by helping a prospective user to find the (most likely) author of a preferred piece, as well as other writers with similar literary style. This is useful when the user does not know the author, or is looking for other similar authors. Thus, this author identification project might be an enhanced search tool that will improve user's experience, also resulting in increased sales for the online bookstore.

For this project the challenge will be to find the best classification algorithm that can yield good results. This will imply testing several algorithms and will require high computation effort therefore, a trade-off between the precision of results and the needed computational power must be taken into consideration.

Data

Data is sourced from Kaggle: https://www.kaggle.com/c/spooky-author-identification.
Dataset contains text from works of fiction written by: Edgar Allan Poe (EAP), HP Lovecraft (HPL) and Mary Wollstonecraft Shelley (MWS). There are separate files for training and test: train.csv and test.csv and will be loaded into the project using pandas read_csv command. The data fields for both files are:
Data Fields - train.csv, 19,579 samples

- id - a unique identifier for each sentence
- text - some text written by one of the authors
- Author(label) - the author of the sentence

Data Fields - test.csv, 8,392 samples

- id - a unique identifier for each sentence
- text - some text written by one of the authors

Techniques

The project will be approached with the typical steps involved in supervised machine learning: exploratory data analysis (EDA), visualization, feature engineering, classification modeling and evaluation.

In addition, I will apply the steps specific to the natural language processing (NLP) such as text preprocessing, text vectorization and text classification. After cleaning the text (removing stop words and punctuation), for text preprocessing I will apply sentence/word tokenizing, part of speech tagging and to normalize tokens I'll be using stemming and lemmatization. For feature engineering and text vectorization I will use BoW (bag of words) and TF-IDF (Term Frequency and Inverse Document Frequency) vectorization.