✓ Data Description:

To decrease the bias and create a reliable authorship attribution dataset the following criteria have been chosen to filter out authors in Gdelt database: English language writing authors, authors that have enough books available (at least 5), 19th century authors. With these criteria 50 authors have been selected and their books were queried through Big Query Gdelt database. The next task has been cleaning the dataset due to OCR reading problems in the original raw form. To achieve that, firstly all books have been scanned through to get the overall number of unique words and each words frequencies. While scanning the texts, the first 500 words and the last 500 words have been removed to take out specific features such as the name of the author, the name of the book and other word specific features that could make the classification task easier. After this step, we have chosen top 10,000 words that occurred in the whole 50 authors text data corpus. The words that are not in top 10,000 words were removed while keeping the rest of the sentence structure intact. Afterwards, the words are represented with numbers from 1 to 10,000 reverse ordered according to their frequencies. The entire book is split into text fragments with 1000 words each. We separately maintained author and book identification number for each one of them in different arrays. Text segments with less than 1000 words were filled with zeros to keep them in the dataset as well. 1000 words make approximately 2 pages of writing, which is long enough to extract a variety of features from the document. The reason why we have represented top 10,000 words with numbers is to keep the anonymity of texts and allow researchers to run feature extraction techniques faster. Dealing with large amounts of text data can be more challenging than numerical data for some feature extraction techniques.

Variable information in the dataset:

✓ How much data do you have (e.g., 10GB, 500GB, 2TB, etc.)

205 MB

✓ What is contained in the data? Variables, fields

Name	Size	Bytes	Class
WW	50x3500	1400000	double
aid	93600x1	748800	double
bid	93600x1	748800	double
ind	93600x1	748800	double
shortened_vocab	1x10000	1254644	cell
test_ind	93600x1	93600	logical
tfidf	1113x50920	453391680	double
train_ind	93600x1	93600	logical
txt_pieces	93600x1000	748800000	double
vocab	1x50920	6387934	cell

✓ File format(s):

.mat

✓ File index or description of the naming conventions:

Name	Description
WW	Author Word list
aid	Author Id
bid	Book Id
ind	Index numbers
shortened_vocab	10000 Vocabulary list
test_ind	Testing Indexes
tfidf	Tfidf Scores
train_ind	Training Indexes
txt_pieces	All one hot encoded data
vocab	All vocabulary list

✓ The temporal coverage of the data, if relevant

18th and 19th Century English and American Authors Book

✓ How did you obtain the data?

Through Google Big Query https://cloud.google.com/bigquery/public-data/gdelt-books

✓ Process or workflow, including source link or API

https://github.com/agungor2/Authorship_Attribution

✓ How did you transform, edit, or clean the data to prepare it for processing?

Benchmarking Authorship Attribution Over a Thousand Books By Victorian Era Authors, Section 3.1