Who's That Author? Authorship attribution of free online texts with computational stylometry.

Abhinil Dutt*
Andrew McQueen*
Murray Ahmed*
University of Illinois Urbana-Champaign
Champaign, IL, USA

1 OVERVIEW

Authorship attribution – inferring the author(s) of written works – is the historical basis for the analysis of linguistic style – formally known as "stylometry" in academic vernacular. Modern applications include revealing ghostwriters, identifying plagiarism, and investigating the origins of ancient literary works.

2 PROJECT SUMMARY

2.1 Description

We seek to implement a text model capable of authorship attribution of free online texts by analyzing their stylistic and lexical features. Our implementation will focus on comparing the differences between distributions of stop words within different text segments of a document. This assumption is motivated from a writer's linguistic style depending greatly on how they structure clauses.

In the search for data resources, we found Project Gutenberg to be a suitable place, since it's online collection is free to access and use.

2.2 Collaboration

Every two weeks, or before milestones, all group members will meet to produce the deliverable and a written summary of the work that each has completed, an updated list of work that still needs to be done, and anything else notable.

(1) **Bi-weekly meetings**: assessing project progress and planning then next steps.
(2) **Asynchronous work**: facilitated through Git for Version Control and Discord for communication.
(3) **Progress Documentation**: written summaries after each meeting.

2.3 Timeline & Responsibilities

The timeline roughly follows a list of task that must be completed chronologically.

Phase 1: *Data Collection & Preparation (Weeks 1 - 3)* – Obtain and clean data (Andrew), preprocess using NLTK (Abhinil), split datasets (Andrew).

Phase 2: *Model Design & Implementation (Weeks 4 - 7)* – Research models (Murray), implement stopword-based model (Abhinil), optimize features (Andrew), initial testing (Murray).

Phase 3: *Evaluation & Testing (Weeks 8 - 9)* – Define metrics (Abhinil), tune model (Andrew), final testing and baseline comparison (Murray).

Phase 4: *Refinement & Results (Weeks 10 - 11)* – Analyze performance (Abhinil), refine model (Andrew), re-test (Murray).

Phase 5: *Documentation & Presentation (Weeks 12 - 13)* – Write report (Abhinil), create poster and slides (Murray), finalize documentation (Andrew).

2.4 Evaluation Strategy

Our (approximate) evaluation strategy will:

(1) *Baseline Comparison:* against word-frequency-based classifiers.
(2) *Cross-Validation:* k-fold cross-validation.
(3) *Performance Metrics:* accuracy, precision, recall, and F1-score.
(4) *Adversarial Testing:* to alter text samples to test robustness.
(5) *Real-world Evaluation:* testing on unseen text documents.

3 RESOURCES

In the preparation of our project we reviewed several sources to learn historical background, keywords and vocabularly, and understand some motivating literature on the subject.

(1) Historical background from the Wikipedia article on Stylometry. (https://en.wikipedia.org/wiki/Stylometry)
(2) Data will be sourced from Project Gutenberg, an online collection of free texts. (https://www.gutenberg.org/)
(3) Burrow's Delta is a well established metric used in Authorship Attribution. (https://aclanthology.org/W15-0709.pdf)
(4) Text Distortion is a novel approach which seeks to better prepare the data. (https://aclanthology.org/E17-1107.pdf)

*All authors contributed equally to this project.