Pull out all the stops: Textual analysis via punctuation sequences

ALEXANDRA N. M. DARMON[1], MARYA BAZZI[1,2,3], SAM D. HOWISON[1], and MASON A. PORTER[1,4]

[1] *Oxford Centre for Industrial and Applied Mathematics, Mathematical Institute, University of Oxford, Oxford OX2 6GG, United Kingdom*
[2] *The Alan Turing Institute, London NW1 2DB, United Kingdom*
[3] *Warwick Mathematics Institute, University of Warwick, Coventry CV4 7AL, United Kingdom*
[4] *Department of Mathematics, University of California, Los Angeles, Los Angeles, California 90095, USA*

(*Received 17 January 2020*)

"I'm tired of wasting letters when punctuation will do, period."
— Steve Martin, *Twitter*, 2011

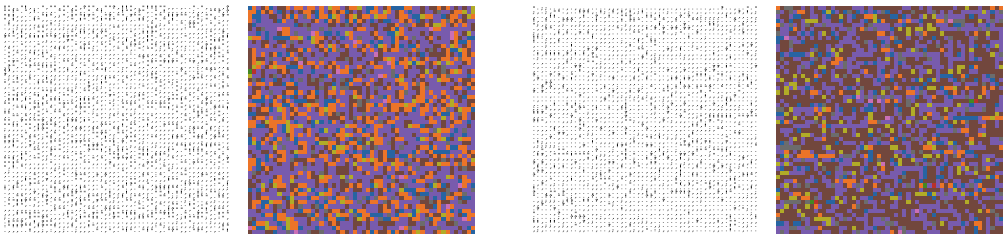
Whether enjoying the lucid prose of a favorite author or slogging through some other writer's cumbersome, heavy-set prattle (full of parentheses, em dashes, compound adjectives, and Oxford commas), readers will notice stylistic signatures not only in word choice and grammar, but also in punctuation itself. Indeed, visual sequences of punctuation from different authors produce marvelously different (and visually striking) sequences. Punctuation is a largely overlooked stylistic feature in "stylometry", the quantitative analysis of written text. In this paper, we examine punctuation sequences in a corpus of literary documents and ask the following questions: Are the properties of such sequences a distinctive feature of different authors? Is it possible to distinguish literary genres based on their punctuation sequences? Do the punctuation styles of authors evolve over time? Are we on to something interesting in trying to do stylometry without words, or are we full of sound and fury (signifying nothing)?

Key Words: Stylometry, computational linguistics, natural language processing, digital humanities, computational methods, mathematical modeling, Markov processes, categorical time series

1 Introduction

"Yesterday Mr. Hall wrote that the printer's proof-reader was improving my punctuation for me, & I telegraphed orders to have him shot without giving him time to pray."
— Mark Twain, *Letter to W. Howells*, 1889

(, ,) . ; . , " " , : (,) ; , ? ; , ? ?

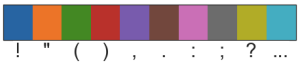


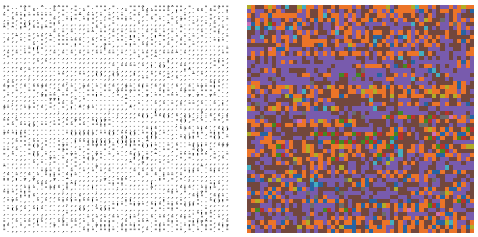
(a) Punctuation sequence: *Sharing Her Crime*, Agnes May Fleming

(b) Heat map: *Sharing Her Crime*, Agnes May Fleming

(c) Punctuation sequence: *King Lear*, William Shakespeare

(d) Heat map: *King Lear*, William Shakespeare





(e) Punctuation sequence: *The History of Mr. Polly*, Herbert George Wells

(f) Heat map: *The History of Mr. Polly*, Herbert George Wells

Figure 1. (a,c,e) Excerpts of ordered punctuation sequences and (b,d,f) corresponding heat maps for books by three different authors: (a,b) Agnes May Fleming; (c,d) William Shakespeare; and (e,f) Herbert George Wells. Each depicted punctuation sequence consists of 3000 successive punctuation marks starting from the midpoint of the full punctuation sequence of the corresponding document. The color bar gives the mapping between punctuation marks and colors.

The sequence of punctuation marks above is what remains of this opening paragraph of our paper (but, to avoid recursion, without the sequence itself) after we remove all of the words. It is perhaps hard to credit that such a minimal sequence encodes any useful information at all; yet it does. In this paper, we investigate the information content of "de-worded" documents, asking questions like the following: Do authors have identifiable punctuation styles (see fig. 1, which was inspired by the visualizations from [3, 4]); if so, can we use them to attribute texts to authors? Do different genres of text differ in their punctuation styles; if so, how? How has punctuation usage evolved over the last few centuries?

In the present paper, we study sequences of punctuation marks (see fig. 1) and the number of words that separate punctuation marks. We use Project Gutenberg [17] to obtain a large literary corpus. We do not attempt to distinguish between an editor's style and an author's style for the documents in our corpus; doing so for a large corpus in

an automated way is a daunting challenge, and we leave it for future efforts. In our work, we investigate whether it is possible to algorithmically assign documents to their authors, irrespective of the documents' edition(s). For ease of writing, we associate documents to authors rather than to both authors and editors throughout our paper, although we recognize that a document's writing and punctuation style can be (and usually is) a product of both.

Our paper contributes to research areas such as computational linguistics and stylometry. *Computational linguistics* is a research area that, broadly speaking, focuses on the development of computational approaches for processing and analyzing natural language. *Stylometry*, a part of computational linguistics — as well as cultural analytics, in the broader context of digital humanities — encompasses quantitative analysis of written text, with the goal of characterizing authorship or other characteristics [19, 35, 45]. Some of the earliest attempts at quantifying the writing style of a document include Mendenhall's work on William Shakespeare's plays in 1887 [33] and Mosteller et al.'s work on *The Federalist Papers* in 1964 [34]. The latter is often regarded as the foundation of computer-assisted stylometry (in contrast with methods based on human expertise) [35, 45]. Uses of stylometry include (1) authorship attribution, recognition, or detection (which aims to determine whether a document was written by a given author); (2) authorship verification (which aims to determine whether a set of documents were written by the same author); (3) plagiarism detection (which aims to determine similarities between two documents); (4) authorship profiling (which aims to determine certain demographics, such as gender, or other characteristics without directly identifying an author);[1] (5) stylochronometry (which is the study and detection of changes in authorial style over time); and (6) adversarial stylometry (which aims to evade authorship attribution via alteration of style).

There has been extensive work on author recognition using a wide variety of stylometric features, including "lexical features" (e.g., number of words and mean sentence length), "syntactic features" (e.g., frequency of different punctuation marks), "semantic features" (e.g., synonyms), and "structural features" (e.g., paragraph length and number of words per paragraph). Two common stylometric features for author recognition are "n-grams" (e.g., in the form of n contiguous words or characters) and "function words" (e.g., pronouns, prepositions, and auxiliary verbs). In this paper, in contrast to prior work, we focus on punctuation, rather than on words or letters. We explore several stylometric tasks through the lens of punctuation, illustrating their distinctive role in text.

According to the definition in [27], *punctuation* refers to the various systems of dots and other marks that accompany letters as part of a writing system. Punctuation is distinct from *diacritic marks*, which are typically modifications of individual letters (e.g., ç, ö, and ő) and *logographs*, which are symbolic representations of lexical items (e.g., # and &). Other common symbols, such as the slash to indicate alternation (e.g., and/or) and the asterisk " * ", do not fall squarely into one of these categories, but they are not considered to be true punctuation marks [27]. Common punctuation marks are the

[1] For an example of "quantitative profiling", see Neidorf et al. [36], who used stylometry to investigate stylistic features (some of which are punctuation-like, as discussed in `https://arstechnica.com/science/2019/04/tolkien-was-right-scholars-conclude-beowulf-likely-the-work-of-single-author/`) of *Beowulf* and concluded that it is likely the work of a single author.

period (i.e., full stop) “ . ”; the comma “ , ”; the colon “ : ”; the semicolon “ ; ”; the left and right parentheses, “ (” and “) ”; the question mark “ ? ”; the exclamation point (which is also called the exclamation mark) “ ! ”; the hyphen “ - ”; the en dash “ – ”; the em dash “ — ”; the opening and closing single quotation marks (i.e., inverted commas), “ ‘ ” and “ ’ ”; the opening and closing double quotation marks (which are also known as inverted commas), “ “ ”, and “ ” ”; the apostrophe “ ’ ”; and the ellipsis “ ... ”.

The aforementioned punctuation set (with minor variations) is used today in a large number of alphabetic writing systems and alphabetic languages [27]. In this sense, for a large number of languages, punctuation is a “supra-linguistic” representational system. However, punctuation varies significantly across individuals, and there is no consensus on how it should be used [13, 29, 37, 39, 47]; authors, editors, and typesetters can sometimes get into emphatic disagreements about it.[2] Accordingly, as a representational system, punctuation is not standardized, and it may never achieve standardization [27].

For our study, we use Project Gutenberg [17] to obtain a large corpus of documents, and we extract a sequence of punctuation marks for each document in the corpus (see section 2). Broadly, our goal is to investigate the following question: Do punctuation marks encode stylistic information about an author, a genre, or a time period? (Recall that we do not distinguish between the roles of authors and editors in a document, so our use of the word “author” is an expository shortcut.) Different writers have different writing styles (e.g., long versus short sentences, frequent versus sparse dialogue, and so on), and a writer’s style can also evolve over time or differ across different types of works. It is plausible that an author’s use of punctuation is — consciously or unconsciously — at least partly indicative of an idiosyncratic style, and we seek to explore the extent to which this is the case. Although there is a wealth of work that focuses on quantitative analysis of writing styles, punctuation marks and their (conscious or unconscious) stylistic footprints have largely been overlooked. Analysis of punctuation is also pertinent to “prosody”, the study of the tune and rhythm of speech[3] and how these features contribute to meaning [18].

To the best of our knowledge, very few researchers have explored author recognition using only stylometric features that are punctuation-focused [7, 16]. Additionally, the few existing works that include a punctuation-focused analysis used a very small author corpus (40 authors in [16] and 5 authors in [7]) and focused on the frequency with which different punctuation marks occur (ignoring, e.g., the order in which they occur). In the present paper, we investigate author recognition using features that account for both the frequency and the order of punctuation marks in a corpus of 651 authors and 14947 documents that we draw from the Project Gutenberg database (see section 3). Although Project Gutenberg is a popular database for the statistical analysis of language, most previous studies that have used it have considered only a small number of manually selected documents [14]. We also use Project Gutenberg to explore genre recognition [9, 23, 41, 42] from a punctuation perspective and stylochronometry [6, 12, 21, 22, 38, 46, 50]

[2] Not that any of us would ever descend to this.

[3] An amusing illustration is the contrast between the Oxford comma, the Walken comma, and the Shatner comma. For one example, see `https://www.scoopnest.com/user/JournalistsLike/529351917986934784`.

in section 4 and section 5, respectively. There are not many studies of stylochronometry, and existing ones tend to be rather specific in nature (e.g., focused on particular authors, such as Shakespeare [50] and band members from the Beatles [22], or on particular time frames) [35,46]. Literary genre recognition (e.g., fiction, philosophy, etc.) has also received limited attention, and we are not aware of even a single study that has attempted genre recognition solely using punctuation. We wish to examine (1) whether punctuation is at all indicative of the style of an author, genre, or time period; and, if so, (2) the strength of stylistic signatures when one ignores words. In short, how much can one learn from punctuation alone?

Importantly, we do not seek to try to identify the best set of features for a given stylometric task, nor do we seek to conduct a thorough comparison of different methods for a given stylometric task. Instead, our goal is to give punctuation, an unsung hero of style, some overdue credit through an initial quantitative study of punctuation-focused stylometry. To do this, we focus on a small number of punctuation-related stylometric features and use this set of features to investigate questions in author recognition, genre recognition, and stylochronometry. To reiterate an important point, we do not account for an editor's effect on an author's style in our analysis, and it is important to interpret all of our findings with that caveat in mind. Given the supra-linguistic nature of punctuation and our reliance on punctuation-based features, one can perform an analysis like ours across different languages that use the same set of punctuation (e.g., across different translations). We offer a novel perspective on stylometry that we hope others will carry forward in their own punctuational pursuits, which include many exciting future directions.

Our paper proceeds as follows. We describe our data set (as well as our filtering and cleaning of it), punctuation-based features, and classification techniques in section 2. We compare the use of punctuation across authors in section 3, across genres in section 4, and over time in section 5. We conclude and offer directions for future work in section 6. The data set of punctuation sequences that we use in this paper is available at `https://dx.doi.org/10.5281/zenodo.3605100`, and the code that we use to analyze punctuation sequences is available at `https://github.com/alex-darmon/punctuation-stylometry`.

2 Data and methodology

"This sentence has five words. Here are five more words. Five-word sentences are fine. But several together become monotonous. Listen to what is happening. The writing is getting boring. The sound of it drones. It's like a stuck record. The ear demands some variety. Now listen. I vary the sentence length, and I create music. Music. The writing sings. It has a pleasant rhythm, a lilt, a harmony. I use short sentences. And I use sentences of medium length. And sometimes, when I am certain the reader is rested, I will engage him with a sentence of considerable length, a sentence that burns with energy and builds with all the impetus of a crescendo, the roll of the drums, the crash of the cymbals — sounds that say listen to this, it is important."

— Gary Provost, *100 Ways to Improve Your Writing*, 1985.

2.1 Data set

We use the API functionality of Project Gutenberg [17] to obtain our document corpus and the natural-language-processing (NLP) library spaCy [20] to extract a punctuation sequence from each document.[4] Using data from Project Gutenberg requires several filtering and cleaning steps before it is meaningful to perform statistical analysis [14]. We describe our steps below.

We retain only documents that are written in English (a document's language is specified in metadata). We remove the author labels "Various", "Anonymous", and "Unknown". To try and mitigate, in an automated way, the issue of a document appearing more than once in our corpus (e.g., "Tales and Novels of J. de La Fontaine – Complete", "The Third Part of King Henry the Sixth", "Henry VI, Part 3", "The Complete Works of William Shakespeare", and "The History of Don Quixote, Volume 1, Complete"), we ensure that any given title appears only once, and we remove all documents with the word "complete" in the title.[5] (Note that the word "anthology" does not appear in any titles in our final corpus.) We also adjust some instances where a punctuation mark or a space appears incorrectly in the Project Gutenberg raw data (specifically, instances in which a double quotation appears as unicode or the spacing between words and punctuation marks is missing), and we remove any documents in which double quotations do not appear.[6] Among the remaining documents, we retain only authors who have written at least 10 documents in our corpus. For each of these documents, we remove headers using the python function "strip_headers", which is available in Gutenberg's Python package. This yields a data set with 651 authors and 14947 documents. We show this final list of authors in appendix A. We show the distribution of documents per author in fig. 2. The documents in our corpus have various metadata, such as author birth year, author death year, document "bookshelf" (with at most one unique bookshelf per document), document subject (with multiple subjects possible per document), document language, and document rights. In some of our computational experiments, we use the following metadata: author birth year, author death year, and document "bookshelf" (which we term document "genre", as that is what it appears to represent). Gerlach and Font-Clos [14] pointed out recently that "bookshelf" may be better suited than "subject" for practical purposes such as text classification, because the former constitute broader categories and provide a unique assignment of labels to documents.

For each document, we extract a sequence of the following 10 punctuation marks: the period " . "; the comma " , "; the colon " : "; the semicolon " ; "; the left parenthesis " ("; the right parenthesis ") "; the question mark " ? "; the exclamation mark " ! "; double quotation marks, " " " and " " " (which are not differentiated consistently in Project Gutenberg's raw data); single quotation marks, " ‘ " and " ’ " (which are also not differentiated consistently in Project Gutenberg's raw data), which we amalgamate with

[4] Many abbreviations, such as "Dr." and "Mr.", are treated as words in spaCy. Therefore, spaCy does not count the periods in them as punctuation marks.

[5] It is still possible for a document to appear more than once in our corpus (e.g., "The Third Part of King Henry the Sixth" and "Henry VI, Part 3"). We manually remove such duplicates when investigating specific authors over time (see section 5).

[6] The latter may be legitimate documents, but we remove them to err on the side of caution.

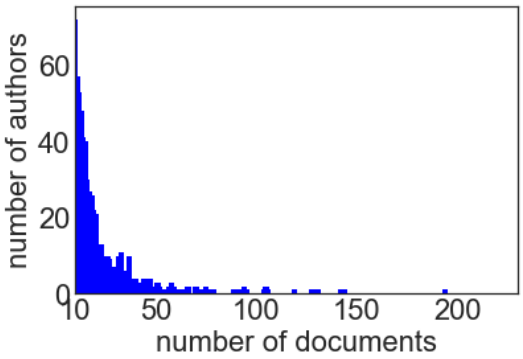


Figure 2. Histogram of the number of documents per author in our corpus.

double quotation marks; and the ellipsis " ... ". To promote a language-independent approach to punctuation (e.g., apostrophes in French can arise as required parts of words), we do not include apostrophes in our analysis. We also do not include hyphens, en dashes, or em dashes, as these are not differentiated consistently in Project Gutenberg's raw data and we find the choices among these marks in different documents — standard rules of language be damned — to be unreliable upon a visual inspection of some documents in our corpus.

2.2 Features

Using standard terminology from the machine-learning literature, we use the word "feature" to refer to any quantitative characteristic of a document or set of documents. We compute six feature vectors for each document k in our corpus to quantify the frequency with which punctuation marks occur, the order in which they occur, and the number of words that tend to occur between them. Specifically, we compute the following:

(1) $\boldsymbol{f}^{1,k}$, the frequency vector for punctuation marks in a given document k;
(2) $\boldsymbol{f}^{2,k}$, an empirical approximation of the conditional probability of the successive occurrence of elements in an ordered pair of punctuation marks in document k;
(3) $\boldsymbol{f}^{3,k}$, an empirical approximation of the joint probability of the successive occurrence of elements in an ordered pair of punctuation marks in document k;
(4) $\boldsymbol{f}^{4,k}$, the frequency vector for sentence lengths in a given document k, where we consider the end of a sentence to be marked by a period, exclamation mark, question mark, or ellipsis;
(5) $\boldsymbol{f}^{5,k}$, the frequency vector for the number of words between successive punctuation marks in a given document k; and
(6) $\boldsymbol{f}^{6,k}$, the mean number of words between successive occurrences of the elements in an ordered pair of punctuation marks in document k.

We summarize these features in table 1 and define each of these six features below. When appropriate, we suppress the superscript k (which indexes the document for which we compute a feature) from $\boldsymbol{f}^{i,k}$ for ease of writing.

Let $\Theta = \{\theta_1, \ldots, \theta_{10}\}$ denote the (unordered) set of 10 punctuation marks (see sec-

Table 1. Summary of the punctuation-sequence features that we study. See the text for details and mathematical formulas.

Feature	Description	Formula
$\boldsymbol{f}^1$	Punctuation-mark frequency	(2.1)
$\boldsymbol{f}^2$	Conditional frequency of successive punctuation marks	(2.2)
$\boldsymbol{f}^3$	Frequency of successive punctuation marks	(2.3)
$\boldsymbol{f}^4$	Sentence-length frequency	(2.4)
$\boldsymbol{f}^5$	Frequency of number of words between successive punctuation marks	(2.5)
$\boldsymbol{f}^6$	Mean number of words between successive occurrences of the elements in ordered pairs of punctuation marks	(2.7)

tion 2.1). Let n denote the total number of documents in our corpus; and let $D_k = \{\theta_1^k, \ldots, \theta_{n_k}^k\}$, with $k \in \{1, \ldots, n\}$, denote the sequence of n_k punctuation marks in document k. As an example, consider the following quote by Ursula K. Le Guin (from an essay in her 2004 collection, *The Wave in the Mind*):

`I don't have a gun and I don't have even one wife and my sentences tend to go on and on and on, with all this syntax in them. Ernest Hemingway would have died rather than have syntax. Or semicolons. I use a whole lot of half-assed semicolons; there was one of them just now; that was a semicolon after "semicolons," and another one after "now."`

The sequence D_k for this quote is {, | . | . | ; | ; | “ | , | ” | “ | . | ”},[7] and there are $n_k = 12$ punctuation marks. From D_k, we can calculate $\boldsymbol{f}^{1,k}$, $\boldsymbol{f}^{2,k}$, and $\boldsymbol{f}^{3,k}$.

We determine each entry of $\boldsymbol{f}^{1,k}$ from the number of times that the associated punctuation mark appears in a document, relative to the total number of punctuation marks in a document:

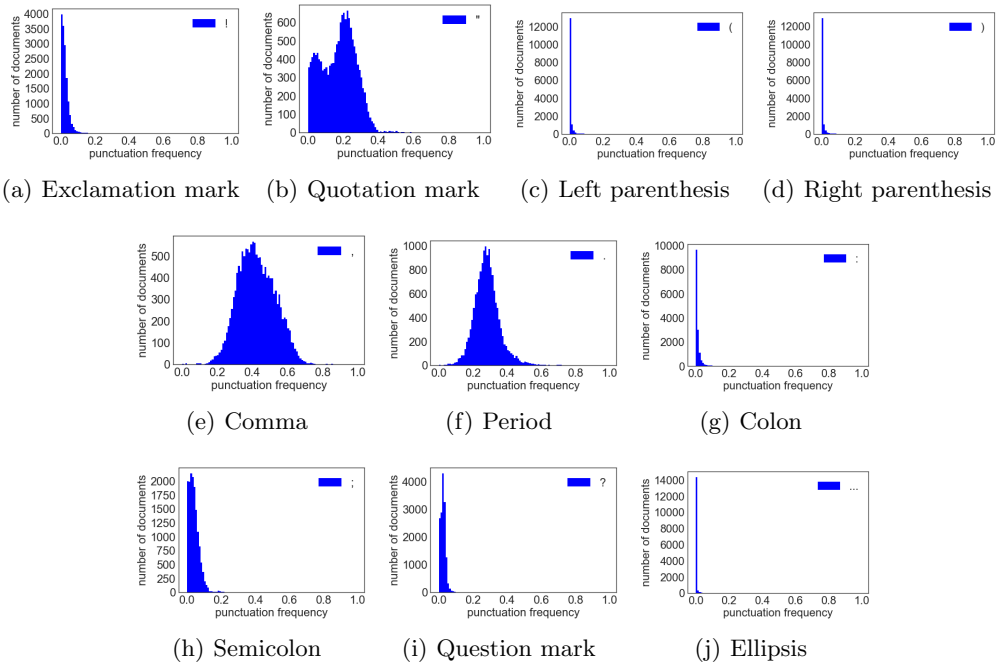
$$f_i^{1,k} = \frac{|\{\theta_l^k \in D_k \,|\, \theta_l^k = \theta_i\}|}{n_k} \,. \tag{2.1}$$

The feature $\boldsymbol{f}^{1,k}$ induces a discrete probability distribution on the set of punctuation marks for each document in our corpus (i.e., $\sum_{i=1}^{|\Theta|} f_i^{1,k} = 1$ for all k) and is independent of the order of the punctuation marks. For the Le Guin quote,

$$\boldsymbol{f}^1 = \begin{bmatrix} \mathbf{!} & \mathbf{"} & \mathbf{(} & \mathbf{)} & \mathbf{,} & \mathbf{.} & \mathbf{:} & \mathbf{;} & \mathbf{?} & \mathbf{...} \\ 0 & \frac{1}{3} & 0 & 0 & \frac{1}{6} & \frac{1}{3} & 0 & \frac{1}{6} & 0 & 0 \end{bmatrix},$$

where the second row indicates the elements of the vector and the first row indicates the

[7] Because there can be commas in the elements of some of the sets and sequences that we consider (e.g., the sequence D_k), we use vertical lines instead of commas to separate elements in sets and sequences with punctuation marks to avoid confusion.



(a) Exclamation mark (b) Quotation mark (c) Left parenthesis (d) Right parenthesis

(e) Comma (f) Period (g) Colon

(h) Semicolon (i) Question mark (j) Ellipsis

Figure 3. Histogram of punctuation-mark frequencies of the documents in our corpus. The horizontal axis of each panel gives the frequency of a punctuation mark binned by 0.01, and the vertical axis of each panel gives the total number of documents in our corpus with a punctuation-mark frequency in the bin. That is, the first bar of a panel for punctuation mark θ_i indicates the number of documents in our corpus for which $0 \leqslant f_i^{1,k} < 0.01$, the second bar indicates the number of documents in our corpus for which $0.01 \leqslant f_i^{1,k} < 0.02$, and so on. In descending order, the means (rounded to the third decimal) of each set $\{f_i^{1,k}, k = 1, \ldots, n\}$ (which we use to construct our plot for θ_i) are 0.024 (exclamation mark), 0.175 (apostrophe), 0.006 (left parenthesis), 0.006 (right parenthesis), 0.425 (comma), 0.283 (period), 0.013 (colon), 0.041 (semicolon), 0.025 (question mark), and 0.002 (ellipsis). These numbers imply that, on average, 42.5% of the punctuation of a document in our corpus consists of commas, 28.3% of it consists of periods, 4.1% of it consists of semicolons, and so on.

corresponding punctuation marks. (Recall from section 2.1 that we amalgamate opening and closing double and single quotation marks into a single punctuation mark, so that entry refers to the appearance of either of those two marks.) An alternative is to consider the frequency of punctuation marks relative to the number of characters or words in a document [16]. In fig. 3, we show the histograms of punctuation-mark frequencies (which are given by $\boldsymbol{f}^1$) across all documents in our corpus. These plots give an idea of the overall usage of each punctuation mark in our corpus. For instance, we see that commas and periods are (unsurprisingly) the most common punctuation marks in the corpus documents. We also observe that the comma frequency varies more across documents than the period frequency. Another observation is that there appear to be two peaks in quotation-mark frequency: a lower peak at about 0.1 (with a height of approximately

450 documents) and a higher peak at about 0.25 (with a height of approximately 650 documents). No other punctuation mark has more than one noticeable peak; this may suggest that one can cluster documents in our corpus into two sets whose characteristic feature is how often they use quotation marks.

To compute $\boldsymbol{f}^{2,k}$ and $\boldsymbol{f}^{3,k}$, we consider a categorical Markov chain on the sequence of punctuation marks and associate each punctuation mark with a state of the Markov chain. We first need two types of transition matrices. We calculate the matrix $\boldsymbol{P}^k \in [0,1]^{|\Theta|\times|\Theta|}$ from the number of times that elements in an ordered pair of punctuation marks occur successively in a document, relative to the number of times that the first punctuation mark in this pair occurs in the document:

$$P_{ij}^k = \frac{|\{\theta_l^k \in D_k | \theta_l^k = \theta_i \text{ and } \theta_{l+1}^k = \theta_j\}|}{|\{\theta_l^k \in D_k \,|\, \theta_l^k = \theta_i\}|}\,, \quad \text{with} \quad \sum_j P_{ij}^k = 1\,. \tag{2.2}$$

When a punctuation mark θ_i does not appear in a document, we set all entries in the corresponding row to 0. We calculate the matrix $\widetilde{\boldsymbol{P}}^k \in [0,1]^{|\Theta|\times|\Theta|}$ from the number of times that elements in an ordered pair of successive punctuation marks occur in a document, relative to the total number of punctuation marks in the document:

$$\widetilde{P}_{ij}^k = \frac{|\{\theta_l^k \in D_k | \theta_l^k = \theta_i \text{ and } \theta_{l+1}^k = \theta_j\}|}{n_k}\,, \quad \text{with} \quad \sum_{i,j} \widetilde{P}_{ij}^k = 1\,. \tag{2.3}$$

Note that $\widetilde{P}_{ij}^k = P_{ij}^k f_i^{1,k}$.

The transition matrix $\boldsymbol{P}^k$ is an estimate of the conditional probability of observing punctuation mark θ_j after punctuation mark θ_i in document k, and the transition matrix $\widetilde{\boldsymbol{P}}^k$ is an estimate of the joint probability of observing the punctuation marks θ_i and θ_j in succession in document k. The relationship $\widetilde{P}_{ij}^k = P_{ij}^k f_i^{1,k}$ ensures that rare (respectively, frequent) events are given less (respectively, more) weight in $\widetilde{\boldsymbol{P}}$ than in $\boldsymbol{P}$. For example, if an author seldom uses the ellipsis "..." in a document, the few ways in which it was used (which, arguably, are not representative of authorial style) are assigned high probabilities in $\boldsymbol{P}$ but low probabilities in $\widetilde{\boldsymbol{P}}$. For the Le Guin quote, $\boldsymbol{P}$ and $\widetilde{\boldsymbol{P}}$ are

$$\boldsymbol{P} = \begin{bmatrix} \texttt{!} & \texttt{"} & \texttt{(} & \texttt{)} & \texttt{,} & \texttt{.} & \texttt{:} & \texttt{;} & \texttt{?} & \texttt{...} \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{4} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad \widetilde{\boldsymbol{P}} = \begin{bmatrix} \texttt{!} & \texttt{"} & \texttt{(} & \texttt{)} & \texttt{,} & \texttt{.} & \texttt{:} & \texttt{;} & \texttt{?} & \texttt{...} \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{9} & 0 & 0 & \frac{1}{9} & \frac{1}{9} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{12} & 0 & 0 & 0 & \frac{1}{12} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{12} & 0 & 0 & 0 & \frac{1}{6} & 0 & \frac{1}{12} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{12} & 0 & 0 & 0 & 0 & 0 & \frac{1}{12} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

where the first row of each matrix indicates the corresponding punctuation mark. Observe

that $\widetilde{P}_{56} < \widetilde{P}_{66}$, even though these entries are equal in $\boldsymbol{P}$, because two successive periods occur more frequently than a period followed by a comma in Le Guin's quote.

We obtain $\boldsymbol{f}^{2,k}$ and $\boldsymbol{f}^{3,k}$ by "flattening" (i.e., concatenating the rows of) the matrices $\boldsymbol{P}^k$ and $\widetilde{\boldsymbol{P}}^k$, respectively. For example, we obtain $\boldsymbol{f}^2$ for the Le Guin quote by appending the rows of $\boldsymbol{P}$ in order and one after the other. The feature $\boldsymbol{f}^{3,k}$ induces a joint probability distribution on the space of ordered punctuation pairs. In contrast to $\boldsymbol{f}^{1,k}$, the features $\boldsymbol{f}^{2,k}$ and $\boldsymbol{f}^{3,k}$ depend on the order in which punctuation marks occur in a document. As we will see in section 3, the feature $\boldsymbol{f}^{3,k}$ is very effective at distinguishing different authors. We account for order with a one-step lag in $\boldsymbol{f}^{2,k}$ and $\boldsymbol{f}^{3,k}$ (i.e., each state depends only on the previous state). One can generalize these features to account for memory or "long-range correlations" [30]. For example, the probability of closing a parenthesis increases after it has been opened.

The features $\boldsymbol{f}^{4,k}$, $\boldsymbol{f}^{5,k}$, and $\boldsymbol{f}^{6,k}$ account for the number of words that occur between punctuation marks. Let $D_k^w = \{w_0^k, w_1^k, \ldots, w_{n_k-1}^k\}$ denote the number of words that occur between successive punctuation marks in D_k, with w_0^k equal to the number of words before the first punctuation mark. Therefore, w_1^k is the number of words between punctuation marks θ_1^k and θ_2^k, and so on. The sequence D_k^w for Le Guin's comment is $\{25, 6, 9, 2, 9, 7, 5, 1, 0, 4, 1, 0\}$, where we count "don't" as two words and we also count "half-assed" as two words. The minimum number of words that can occur between successive punctuation marks is 0, and we cap the maximum number of words that can occur between successive punctuation marks at $n_s = 40$ and the number of words in a sentence at $n_S = 200$. Fewer than 0.05 % of the sentences in our corpus exceed $n_S = 200$ words; similarly, the $n_s = 40$ cap is exceeded by fewer than 0.05 % of the strings between successive punctuation marks.

The entries of the feature $\boldsymbol{f}^{4,k} \in [0,1]^{n_S \times 1}$, which quantifies the frequency of sentence lengths, are

$$f_i^{4,k} = \frac{|\{w_l^k \in D_k^w \,|\, w_l^k = i \text{ and } \theta_l, \theta_{l+1} \in \{. \,|\, \ldots \,|\, ! \,|\, ?\}\}|}{n_k} \,. \tag{2.4}$$

In the Le Guin quote, there are four sentences, with lengths 31, 9, 2, and 27 (in sequential order). The feature $\boldsymbol{f}^{4,k}$, an $n_S \times 1$ vector with $n_S = 200$, thus has the value $1/4$ in the 9^{th}, 2^{nd}, 27^{th}, and 31^{st} positions and the value 0 in all other entries. One can also consider other measures of sentence length (e.g., the number of characters, instead of the number of words) [48].

The entries of the feature $\boldsymbol{f}^{5,k} \in [0,1]^{n_s \times 1}$, which quantifies the frequency of the number of words between successive punctuation marks, are

$$f_i^{5,k} = \frac{|\{w_l^k \in D_k^w \,|\, w_l^k = i\}|}{n_k} \,. \tag{2.5}$$

In the Le Guin quote, recall that $D_k^w = \{25, 6, 9, 2, 9, 7, 5, 1, 0, 4, 1, 0\}$ (which includes 9 unique integers), so the $n_s \times 1$ vector (with $n_s = 40$, as mentioned above) $\boldsymbol{f}^5$ has 9 nonzero entries. For example, $\boldsymbol{f}_1^5 = 2/12$ (because 0 occurs twice out of $n_k = 12$ total punctuation marks) and $\boldsymbol{f}_4^5 = 0$ (because 3 never occurs out of $n_k = 12$ possible times).

The features $\boldsymbol{f}^{4,k}$ and $\boldsymbol{f}^{5,k}$ induce discrete probability distributions on the number of words in sentences and the number of words between successive punctuation marks,

respectively. The expectation of the feature $\boldsymbol{f}^{5,k}$ quantifies the "rate of punctuation" and is equal to the total number of words, relative to the total number of punctuation marks:

$$\mathbb{E}\left[f^{5,k}\right] = \sum_{i=0}^{n_s} i \times f_i^{5,k} = \frac{1}{n_k} \times \sum_{i=0}^{n_s} i \times \left|\{w_l^k \in D_k^w \,|\, w_l^k = i\}\right| = \frac{|D_k^w|}{n_k}\,. \tag{2.6}$$

The feature $\boldsymbol{f}^{5,k}$ tracks word-count frequency between successive punctuation marks, without distinguishing between different punctuation marks.

With $\boldsymbol{f}^{6,k}$, we compute the mean number of words between successive occurrences of the elements in ordered pairs of punctuation marks using a matrix $\boldsymbol{W}^k \in [0, n_s]^{|\Theta|\times|\Theta|}$ with entries

$$W_{ij}^k = \langle \{w_l^k \in D_k^w \,|\, \theta_l = \theta_i \text{ and } \theta_{l+1} = \theta_j\} \rangle\,, \tag{2.7}$$

where $\langle\,\cdot\,\rangle$ denotes the sample mean of a set. The matrix for the Le Guin excerpt is

$$\boldsymbol{W} = \left[\begin{array}{cccccccccc} \texttt{!} & \texttt{"} & \texttt{(} & \texttt{)} & \texttt{,} & \texttt{.} & \texttt{:} & \texttt{;} & \texttt{?} & \texttt{...} \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 5.5 & 0 & 9 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 & 0 & 0 & 0 & 7 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right]\,.$$

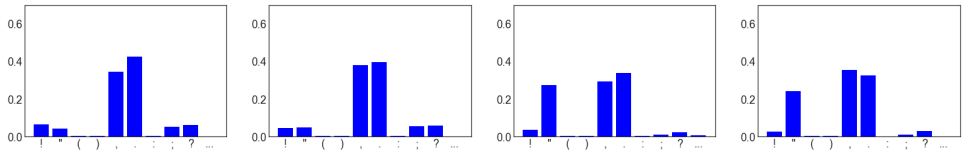
We obtain $\boldsymbol{f}^{6,k}$ by flattening the matrix $\boldsymbol{W}^k$ by concatenating its rows. As variants of this feature, one need not require that punctuation-mark occurrences are successive, and one can subsequently compute the number of words or even the number of (other) punctuation marks between the elements of an ordered pair of punctuation marks.

In the rest of our paper, we focus on the six features $\boldsymbol{f}^1, \ldots, \boldsymbol{f}^6$. We show example histograms of $\boldsymbol{f}^1$ (punctuation frequency) and $\boldsymbol{f}^5$ (mean number of words between successive punctuation marks) for some documents by the same authors in fig. 4.

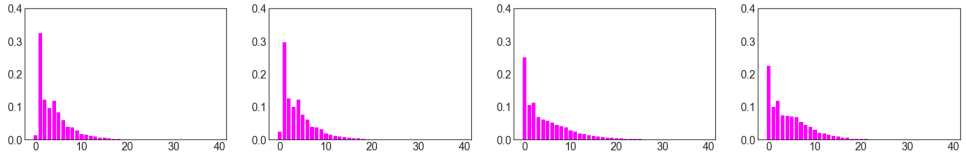
2.3 Kullback–Leibler divergence

To quantify the similarity between two discrete distributions (e.g., between the features $\boldsymbol{f}^1, \boldsymbol{f}^3, \boldsymbol{f}^4$, and $\boldsymbol{f}^5$ from different documents), we use Kullback–Leibler (KL) divergence [25], an information-theoretic measure that is related to Shannon entropy and ideas from maximum-likelihood theory. KL divergence and variants of it have been used in prior research on author recognition [2, 35, 52]. One can also consider other similarity measures, such as chi-square distance [35] and Jensen–Shannon divergence [1, 15, 31].

Consider a random variable X with a discrete, finite support $x \in \mathcal{X}$; and let $\boldsymbol{p} \in [0,1]^{|\mathcal{X}|\times 1}$ and $\boldsymbol{q} \in [0,1]^{|\mathcal{X}|\times 1}$ be two probability distributions for X that we assume are absolutely continuous with respect to each other. Broadly speaking, KL divergence



(a) $\boldsymbol{f}^1$ *King Lear*, W. Shakespeare (b) $\boldsymbol{f}^1$ *Hamlet*, W. Shakespeare (c) $\boldsymbol{f}^1$ *The History of Mr. Polly*, H. G. Wells (d) $\boldsymbol{f}^1$ *The Wheels of Chance*, H. G. Wells



(e) $\boldsymbol{f}^5$ *Hamlet*, W. Shakespeare (f) $\boldsymbol{f}^5$ *Hamlet*, W. Shakespeare (g) $\boldsymbol{f}^5$ *The History of Mr. Polly*, H. G. Wells (h) $\boldsymbol{f}^5$ *The Wheels of Chance*, H. G. Wells

Figure 4. (a,b,c,d) Histograms of punctuation-mark frequency ($\boldsymbol{f}^1$) and (e,f,g,h) the number of words that occur between successive punctuation marks ($\boldsymbol{f}^5$) for two documents by William Shakespeare and two documents by Herbert George Wells.

quantifies how close a probability distribution $\boldsymbol{p} = \{p_i\}$ is to a candidate distribution $\boldsymbol{q} = \{q_i\}$, where p_i (respectively, q_i) denotes the probability that X takes the value i when it is distributed according to $\boldsymbol{p}$ (respectively, $\boldsymbol{q}$) [10]. The KL divergence between the probability distributions $\boldsymbol{p}$ and $\boldsymbol{q}$ is defined as

$$d_{\mathrm{KL}}(p, q) = \sum_{i=1}^{|\mathcal{X}|} p_i \log\left(\frac{p_i}{q_i}\right) \tag{2.8}$$

and satisfies four important properties:

(1) $d_{\mathrm{KL}}(\boldsymbol{p}, \boldsymbol{q}) \geqslant 0$;
(2) $d_{\mathrm{KL}}(\boldsymbol{p}, \boldsymbol{q}) = 0$ if and only if $p_i = q_i$ for all i;
(3) $d_{\mathrm{KL}}(.\,, .)$ is asymmetric in its arguments; and
(4) $d_{\mathrm{KL}}(\boldsymbol{p}, \boldsymbol{q}) = H(\boldsymbol{p}, \boldsymbol{q}) - H(\boldsymbol{p})$,

where $H(\boldsymbol{p}) = \sum_i p_i \log p_i$ denotes the Shannon entropy of $\boldsymbol{p}$ and $H(\boldsymbol{p}, \boldsymbol{q})$ denotes the Shannon entropy of the joint distribution of $\boldsymbol{p}$ and $\boldsymbol{q}$ [28, 43]. Entropy quantifies the "unevenness" of a probability distribution. It represents the mean information that is required to specify an outcome of a random variable, given its probability distribution. It achieves its minimum value 0 for a constant random variable (e.g., $p_1 = 1$ and $p_i = 0$ for $i \neq 1$) and its maximum value $\log(|\mathcal{X}|)$ for a uniform distribution. In some sense, $d_{\mathrm{KL}}(\boldsymbol{p}, \boldsymbol{q})$ measures the "unevenness" of the joint distribution of $\boldsymbol{p}$ and $\boldsymbol{q}$ relative to the distribution of $\boldsymbol{p}$. One can also derive KL divergence from likelihood theory. In particular, one can show that, as the number of samples from the discrete random variable $\mathcal{X}$ tends to infinity,

KL divergence measures the mean likelihood of observing data with the distribution $\boldsymbol{p}$ if the distribution $\boldsymbol{q}$ actually generated the data [11, 44].

To adjust for cases in which $\boldsymbol{p}$ and $\boldsymbol{q}$ are not absolutely continuous with respect to each other (e.g., one document has one or more ellipses, but another does not, resulting in unequal supports), we remove any frequency component that corresponds to a punctuation mark that is not in the common support and then distribute the weight of the removed frequency uniformly across the other frequencies. For example, suppose that $p_1 \neq 0$ but $q_1 = 0$. We then define $\widetilde{\boldsymbol{p}}$ such that $\widetilde{\boldsymbol{p}} = \{p_i/(1-p_1)\,,\ \ i \neq 1\}$ and compute $d_{\mathrm{KL}}(\widetilde{\boldsymbol{p}}, \boldsymbol{q})$.

2.4 Classification models

We describe the two classification approaches that we use for author recognition (see section 3.2) and genre recognition (see section 4.2). Much of the existing classification work on author recognition uses machine-learning classifiers (e.g., support vector machines or neural networks) or similarity-based classification techniques (e.g., using KL divergence) [35, 45]. We use neural networks and similarity-based classification with KL divergence for both author and genre classification. Following standard practice, we split the n documents in our data set into into a training set and a testing set. Broadly speaking, a training set calibrates a classification model (e.g., to "feed" a neural network and adjust its parameters), and one then uses a testing set to evaluate the accuracy of a calibrated model. We ensure that all authors or genres (i.e., all "classes") that appear in the testing set also appear in the training corpus; this is known as "closed-set attribution" and is common practice in author recognition [35, 45]. For a given data set, we place 80% of the documents in the training set and the remaining 20% of documents in the testing set. (A training:testing ratio of 80:20 is a common choice.) A given data set is sometimes the entire corpus (i.e., 14947 documents and 651 authors), and it is sometimes a subset of it. In our summary tables (see section 3.2 and section 4.2), we explicitly specify the sizes of the training and testing sets of our experiments.

2.4.1 *Similarity-based classification*

We label our p classes by $c_1, c_2, \ldots, c_p$ (recall that these can correspond to authors or genres), and we denote the set of training documents for class c_j by $\mathcal{D}_j$. For each class c_j, we define a class-level feature $\boldsymbol{f}^{l,c_j}$, with $l \in \{1, \ldots, 6\}$ and $j \in \{1, \ldots, p\}$, by averaging the features across the training documents in that class. That is, the i^{th} entry of $\boldsymbol{f}^{l,c_j}$ is

$$f_i^{l,c_j} = \frac{1}{|\mathcal{D}_j|} \sum_{k \in \mathcal{D}_j} f_i^{l,k}\,, \tag{2.9}$$

where $l \in \{1, \ldots, 6\}$ and we use the features $\boldsymbol{f}^{1,k}, \boldsymbol{f}^{2,k}, \ldots, \boldsymbol{f}^{6,k}$ from section 2.2. This yields a set $\phi^k = \{\boldsymbol{f}^{1,k}, \ldots, \boldsymbol{f}^{6,k}\}$ of features for each document and a set $\phi^{c_j} = \{\boldsymbol{f}^{1,c_j}, \ldots, \boldsymbol{f}^{6,c_j}\}$ of features for each class.

To determine which class is "most similar" to a document k in our testing set, we solve the following minimization problem:

$$\mathrm{argmin}_{j \in \{1,\ldots,p\}} d(\phi^k, \phi^{c_j})\,, \tag{2.10}$$

for some choice of similarity measure $d(\,.\,,\,.\,)$. In our numerical experiments of section 3, we use the KL-divergence similarity measure d_{KL} to define $d(\,.\,,\,.\,)$ as

$$d(\phi^k, \phi^{c_j}) = \mathrm{argmin}_{l \in \mathcal{L}} d_{\mathrm{KL}}(\boldsymbol{f}^{l,c_j}, \boldsymbol{f}^{l,k})\,, \tag{2.11}$$

where we restrict the set of features to those that induce discrete probability distributions and consider each feature individually (i.e., $\mathcal{L} = \{1\}$, $\mathcal{L} = \{3\}$, $\mathcal{L} = \{4\}$, or $\mathcal{L} = \{5\}$).

2.4.2 *Neural networks*

We use feedforward neural networks with the standard backpropagation algorithm as a machine-learning classifier [24]. A neural network uses the features of a training set to automatically infer rules for recognizing the classes of a testing set by adjusting the weights of each "neuron" using a stochastic gradient-descent-based learning algorithm. In contrast with neural networks for classical NLP classification, where it is standard to use word embeddings and employ convolutional or recurrent neural networks [26] to ensure that input vectors have equal lengths, we have already defined our features such that they have equal length. It thus suffices for us to use feedforward neural networks. The input vector that corresponds to each document is a concatenation of the six features (or a subset thereof) in section 2.2, and the output is a probability vector, which one can interpret as the likelihood that a given document belongs to a given class. We assign each document in our testing set to the class with highest probability.

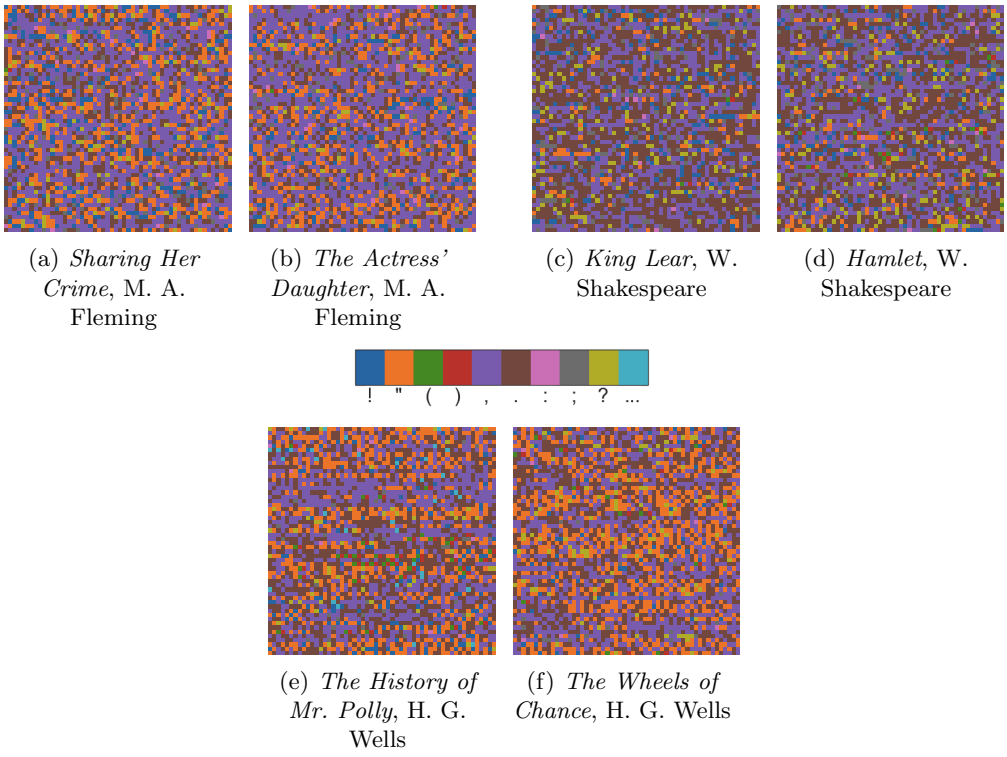
2.5 Model evaluation

For each test of a classification model, we consider a data set with a fixed number of classes (e.g., 651 classes if we perform author recognition on all authors in our corpus), a uniformly-randomly sampled training set (80% of the data set), and a testing set (the remaining 20% of the data set). We measure "accuracy" as the ratio of correctly assigned documents relative to the total number of documents in a testing set. For each test of a classification model, we report two quantities: (1) the accuracy of the classification model on the testing set; and (2) the accuracy of a baseline classifier on the testing set, which we obtain by assigning each document in the testing set to each class with a probability that is proportional to the class's size in the training set.

3 Case study: Author analysis

"It is almost always a greater pleasure to come across a semicolon than a period. The period tells you that that is that; if you didn't get all the meaning you wanted or expected, anyway you got all the writer intended to parcel out and now you have to move along. But with a semicolon there you get a pleasant little feeling of expectancy; there is more to come; to read on; it will get clearer."

— Thomas Lewis, *Notes on Punctuation*, 1979



(a) *Sharing Her Crime*, M. A. Fleming

(b) *The Actress' Daughter*, M. A. Fleming

(c) *King Lear*, W. Shakespeare

(d) *Hamlet*, W. Shakespeare

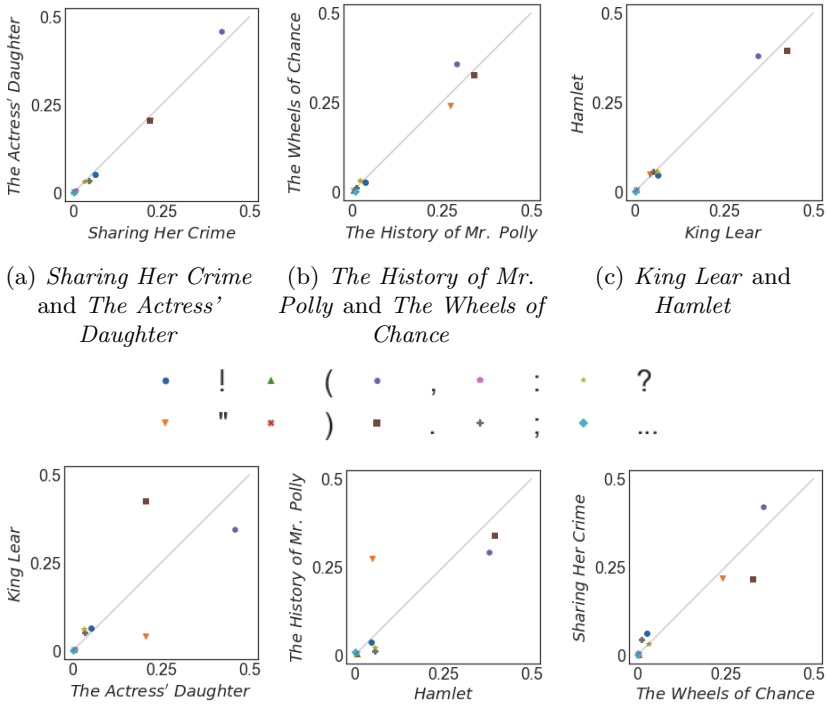
(e) *The History of Mr. Polly*, H. G. Wells

(f) *The Wheels of Chance*, H. G. Wells

Figure 5. Sequences of successive punctuation marks that we extract from documents by (a,b) May Agnes Fleming, (c,d) William Shakespeare, and (e,f) Herbert George Wells. We map each punctuation mark to a distinct color. We cap the length of the punctuation sequence at 3000 entries, which start at the midpoint of the punctuation sequence of the corresponding document.

3.1 Consistency

We explore punctuation sequences of a few authors to gain some insight into whether certain authors have more distinguishable punctuation styles than others. (Once again, recall our cautionary note that we do not distinguish between the roles of authors and editors for the documents in our corpus.) In fig. 5, we show (augmenting fig. 1) raw sequences of punctuation marks for two books by each of the following three authors: May Agnes Fleming, William Shakespeare, and Herbert George (H. G.) Wells. We observe for this document sample that, visually, one can correctly guess which documents were written by the same author based only on the sequences of punctuation marks. This striking possibility was illustrated previously in A. J. Calhoun's blog entry [4], which motivated our research. From fig. 5, we see that Wells appears to use noticeably more quotation marks than the other two authors. We also observe that Shakespeare appears to use more periods than Wells. These observations are consistent with the histograms in fig. 4 (where we also observe that Shakespeare appears to use more exclamation marks and question marks than Wells), which we compute from the entire documents, so our observations from the samples in fig. 5 appear to hold throughout those documents.



(a) *Sharing Her Crime* and *The Actress' Daughter*

(b) *The History of Mr. Polly* and *The Wheels of Chance*

(c) *King Lear* and *Hamlet*

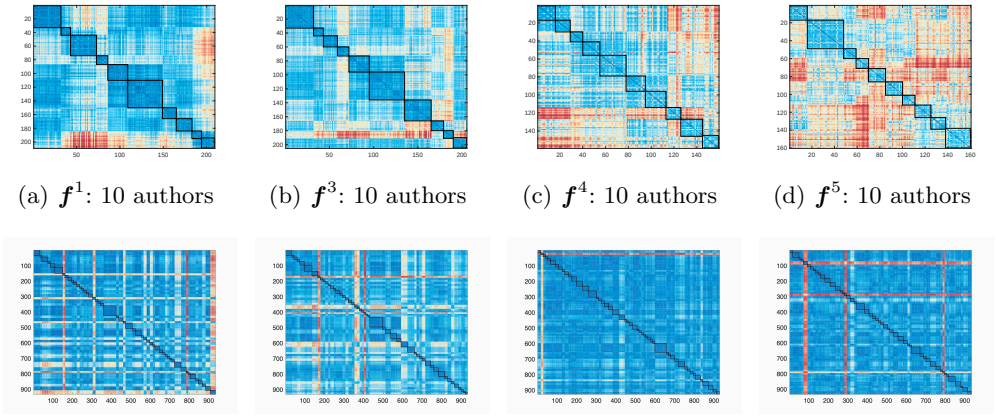
(d) *The Actress' Daughter* and *King Lear*

(e) *Hamlet* and *The History of Mr.Polly*

(f) *The Wheels of Chance* and *Sharing Her Crime*

Figure 6. Scatter plots of frequency vectors (i.e., $\boldsymbol{f^1}$) of punctuation marks to compare books from the same author: (a) *Sharing Her Crime* and *The Actress' Daughter* by May Agnes Fleming, (c) *King Lear* and *Hamlet* by William Shakespeare, and (e) *The History of Mr. Polly* and *The Wheels of Chance* by H. G. Wells. Scatter plots of frequency vectors of punctuation marks to compare books from different authors: (b) *The Actress' Daughter* and *King Lear*, (d) *Hamlet* and *The History of Mr. Polly*, and (f) *The Wheels of Chance* and *Sharing Her Crime*. We represent each punctuation mark by a colored marker, with coordinates given by the punctuation frequencies in a vector that is associated to each document. The gray line represents the identity function. More similar frequency vectors correspond to dots that are closer to the gray line.

In fig. 6, we plot examples of the punctuation frequency (i.e., $\boldsymbol{f^1}$) of one document versus that of another document by the same author (top row) and a document by a different author (bottom row). We base these plots on the "rank order" plots in [51], who used such plots to illustrate the top-ranking words in various texts. In our plots, any punctuation mark (which we represent by a colored marker) that has the same frequency in both documents lies on the gray diagonal line. Any marker above (respectively, below) the gray line signifies that it is used more (respectively, less) frequently by the author on the vertical axis (respectively, horizontal axis). In these examples, we see for documents by the same author that the markers tend to be closer to the gray line than for documents by



(a) $\boldsymbol{f}^1$: 10 authors (b) $\boldsymbol{f}^3$: 10 authors (c) $\boldsymbol{f}^4$: 10 authors (d) $\boldsymbol{f}^5$: 10 authors

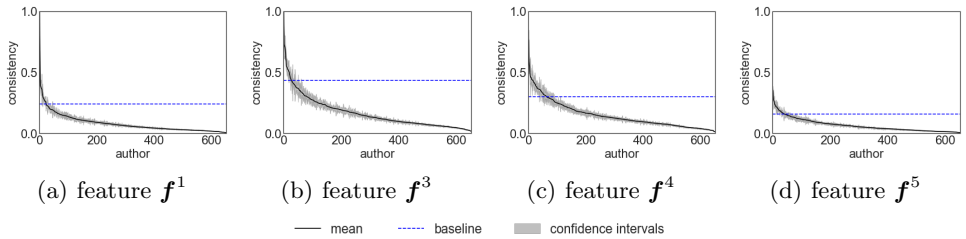
(e) $\boldsymbol{f}^1$: 50 authors (f) $\boldsymbol{f}^3$: 50 authors (g) $\boldsymbol{f}^4$: 50 authors (h) $\boldsymbol{f}^5$: 50 authors

Figure 7. Heat maps showing KL divergence between the features (a,e) $\boldsymbol{f}^1$, (b,f) $\boldsymbol{f}^3$, (c,g) $\boldsymbol{f}^4$, and (d,h) $\boldsymbol{f}^5$ for different sets of authors. We show the 10 most-consistent (see the main text for our notion of "consistency") authors for each feature in the top row and the fifty most-consistent authors for each feature in the bottom row. The diagonal blocks that we enclose in black indicate documents by the same author. Authors can differ across panels, because author consistency can differ across features. The colors scale (nonlinearly) from dark blue (corresponding to a KL divergence of 0) to dark red (corresponding to the maximum value of KL divergence in the underlying matrix). For ease of exposition, we suppress color bars (they span the interval $[0, 3.35]$), given that the purpose of this figure is to illustrate the presence and/or absence of high-level structure. When determining the 10 most-consistent authors, we exclude the author "United States. Warren Commission" (see row 7 in table A 1) in panels (a–c) and we exclude the author "United States. Central Intelligence Agency" (see row 116 in table A 1) in panel (d). In each case, the we replace them with the next most-consistent author and proceed from there. Works by these two authors consist primarily of court testimonies or lists of definitions and facts (with rigid styles); they manifested as pronounced dark-red stripes that masked salient block structures.

different authors. In fig. 6(d), for example, we observe that Fleming used more quotation marks and commas in *The Actress' Daughter* than Shakespeare did in *King Lear*, whereas Shakespeare used more periods in *King Lear* than Fleming did in *The Actress' Daughter*. One can make similar observations about panels (e) and (f) of fig. 6. These observations are consistent with those of fig. 4 and fig. 5.

Our illustrations in fig. 5 and fig. 6 use a very small number of documents by only a few authors. To quantify the "consistency" of an author across all documents by that author in our corpus, we use KL divergence.

In fig. 7, we show heat maps of KL divergence between discrete probability distributions induced by the feature vectors $\boldsymbol{f}^1$, $\boldsymbol{f}^3$, $\boldsymbol{f}^4$, and $\boldsymbol{f}^5$. We define the "consistency" of an author relative to a feature as the mean KL divergence for that feature computed across



(a) feature $\boldsymbol{f}^1$ (b) feature $\boldsymbol{f}^3$ (c) feature $\boldsymbol{f}^4$ (d) feature $\boldsymbol{f}^5$

Figure 8. Evaluations of author consistency. In each panel, we show author consistency (3.1) for the features (a) $\boldsymbol{f}^1$, (b) $\boldsymbol{f}^3$, (c) $\boldsymbol{f}^4$, and (d) $\boldsymbol{f}^5$ using a solid black curve. In gray, we plot confidence intervals of KL divergence across pairs of documents for each author. To compute the confidence intervals, we assume that the KL divergence values across pairs of distinct documents for each author are normally distributed. There are at least 10 documents by each author in our corpus (see section 2.1), so the number of KL values across pairs of distinct documents by a given author is at least 90. The dotted blue line indicates a consistency baseline, which we obtain by choosing, uniformly at random, 1000 ordered pairs of documents by distinct authors and computing the mean KL divergence between the features of these document pairs.

all pairs of documents by that author. That is,

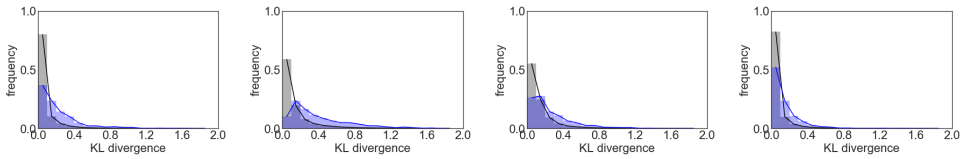
$$\mathcal{C}_{\boldsymbol{f}^i}(a) = \frac{2}{|\mathcal{D}_a - 1||\mathcal{D}_a|} \sum_{k,k' \in \mathcal{D}_a} d_{\mathrm{KL}}(\boldsymbol{f}^{i,k}, \boldsymbol{f}^{i,k'}) \,, \tag{3.1}$$

where a denotes an author in our corpus and $\mathcal{D}_a$ is the set of documents by author a. For each feature in fig. 7, we show the 10 (respectively, 50) most-consistent authors in the top row (respectively, bottom row). Diagonal blocks with black outlines correspond to documents by the same author. Although there appears to be greater similarity within diagonal blocks than between them for several of the authors, it is difficult to interpret the heat maps when there are many authors (and it becomes increasingly difficult as one considers progressively more authors).

In fig. 8, we show author consistency in our entire corpus for the feature vectors $\boldsymbol{f}^1$, $\boldsymbol{f}^3$, $\boldsymbol{f}^4$, and $\boldsymbol{f}^5$. In each panel, we show a baseline (in blue), which we obtain by choosing, uniformly at random, 1000 ordered pairs of documents by distinct authors and computing the mean KL divergence between the features of these document pairs. One pair is a single element of an off-diagonal block of a matrix like those in fig. 7.

We order each panel from the least-consistent author to the most-consistent author. Authors can differ across panels, because the consistency measure (3.1) is a feature-dependent quantity. We observe in all panels of fig. 8 that most authors are more consistent on average than the baseline. (The black curve lies below the blue horizontal line for most authors.) The differences between authors relative to the baseline are most pronounced for the feature $\boldsymbol{f}^3$ (see table 1). This suggests that $\boldsymbol{f}^3$ may carry more information than our other five features about an author's idiosyncratic style. We come back to this observation in section 3.2.

In fig. 9, we show the distribution of KL divergence values between documents by the same authors (in black) and between documents by distinct authors (in blue). For fig. 8,



(a) feature $\boldsymbol{f}^1$ (b) feature $\boldsymbol{f}^3$ (c) feature $\boldsymbol{f}^4$ (d) feature $\boldsymbol{f}^5$

Figure 9. Distributions of KL divergence for authors. In each panel, we show the distributions of KL divergence between all pairs of documents in the corpus by the same author (in black) and between 1000 ordered pairs of documents by distinct authors (in blue). We choose the ordered pairs uniformly at random from the set of all ordered pairs of documents by distinct authors. Each panel corresponds to a distinct feature. The means of the distributions of each panel are (a) 0.0828 (black) and 0.240 (blue), (b) 0.167 (black) and 0.433 (blue), (c) 0.149 (black) and 0.275 (blue), and (d) 0.0682 (black) and 0.154 (blue).

we use the former to compute author consistency (by taking the mean of the values for each author) and the latter to compute the consistency baseline (by taking the mean of all values). For all features, we see from a Kolmogorov–Smirnov (KS) test that the difference between the empirical distributions is statistically significant. (In all cases, the p-value is less than or equal to 1.218×10^{-79}.)

3.2 Author recognition

We use the classification techniques from section 2.4 to perform author recognition. We show our results using KL divergence (see section 2.4.1) in table 2 and using neural networks (see section 2.4.2) in table 3. In each table, we specify the number of authors ("No. authors"), the number of documents in the training set ("Training size"), the number of documents in the testing set ("Testing size"), the accuracy of the test using various sets of features, and the baseline accuracy (as defined in section 2.5). Each row in a table corresponds to an experiment on a set of distinct authors, which we choose uniformly at random. (The set consists of the entire corpus when the number of authors is 651.) For a given number of authors, we use the same sample across both tables to allow a fair comparison.

We show classification results using KL divergence in table 2 using each individual frequency feature vector as input. As we consider more authors, the accuracy on the testing set tends to decrease significantly. The issue of developing a method that scales well as one increases the number of authors is an open problem in author recognition even when using words from text [35], and we are exploring stylistic signatures from punctuation only, a much smaller set of information. Remarkably, we are able to achieve an accuracy of 66% on a sample of 50 authors using only the feature $\boldsymbol{f}^3$. This is consistent with the plots in fig. 8, where $\boldsymbol{f}^3$ gave the best improvement from the baseline.

We show classification results using a one-layer neural network with 2000 neurons in table 3 using various sets of input vectors (which, contrary to when one uses KL divergence, need not be feature vectors that induce probability distributions). We also

Table 2. Results of our author-recognition experiments using a classification based on KL divergence (see section 2.4.1) for author samples of various sizes and using the individual features $\boldsymbol{f}^1$, $\boldsymbol{f}^3$, $\boldsymbol{f}^4$, and $\boldsymbol{f}^5$ as input. We measure accuracy as the ratio of correctly assigned documents relative to the total number of documents in the testing set. (See section 2.5 for a description of the baseline.)

No. authors	Training size	Testing size	Accuracy on the testing set				
			$\boldsymbol{f}^1$	$\boldsymbol{f}^3$	$\boldsymbol{f}^4$	$\boldsymbol{f}^5$	baseline
10	216	55	0.69	0.74	0.52	0.63	0.21
50	834	209	0.54	0.66	0.30	0.31	0.029
100	2006	502	0.37	0.49	0.25	0.23	0.019
200	3549	888	0.30	0.47	0.16	0.20	0.0079
400	7439	1860	0.27	0.41	0.15	0.16	0.0047

observe in table 3 that accuracy on the testing set tends to decrease significantly as one increases the number of authors. Overall, however, the neural network outperforms our KL divergence-based classification. We achieve an accuracy of 62% when using only $\boldsymbol{f}^3$ and an accuracy of 72% when using all feature vectors on a sample of 651 authors (i.e., on the entire corpus). Interestingly, in some of our experiments, using the features $\{\boldsymbol{f}^1, \boldsymbol{f}^3, \boldsymbol{f}^4, \boldsymbol{f}^5\}$ gives slightly better accuracy than using all features.

Based on preliminary experiments, our accuracy results in table 2 and table 3 seem to be robust to (1) different author samples of the same size and (2) different training and testing samples for a given author sample. However, the heterogeneity in accuracy across different author samples of the same size is more pronounced than the heterogeneity that we observe from different training and testing samples for a given author sample, as different author samples can sometimes yield significantly different training and testing set sizes (see fig. 2). Such heterogeneity across different author samples decreases as one increases the number of authors.

To the best of our knowledge, most attempts thus far at author recognition of literary documents have used data sets that are of significantly smaller scale than our corpus [14,35]. One recent example of author analysis from a corpus extracted from Project Gutenberg is the one in Qian *et al.* [40]. Their corpus consists of 50 authors (with their choices of authors based on a popularity criterion) and 900 single-paragraph excerpts for each author. (For a given author, they extracted their excerpts from several books.) Using word-based features and machine-learning classifiers, they achieved an accuracy of 89.2% using 90% of their data for training and 10% of it for testing.

Table 3. Results of our author-recognition experiments using a one-layer, 2000-neuron neural network (see section 2.4.2) for author samples of various sizes and using different features or sets of features as input: $\boldsymbol{f}^1$, $\boldsymbol{f}^3$, $\boldsymbol{f}^4$, $\boldsymbol{f}^5$, $\{\boldsymbol{f}^1, \boldsymbol{f}^3, \boldsymbol{f}^4, \boldsymbol{f}^5\}$, and $\{\boldsymbol{f}^1, \boldsymbol{f}^2, \boldsymbol{f}^3, \boldsymbol{f}^4, \boldsymbol{f}^5, \boldsymbol{f}^6\}$ (which we label as "all"). We measure accuracy as the ratio of correctly assigned documents relative to the total number of documents in the testing set. (See section 2.5 for a description of the baseline.)

No. authors	Training size	Testing size	Accuracy on testing set						
			$\boldsymbol{f}^1$	$\boldsymbol{f}^3$	$\boldsymbol{f}^4$	$\boldsymbol{f}^5$	$\{\boldsymbol{f}^1, \boldsymbol{f}^3, \boldsymbol{f}^4, \boldsymbol{f}^5\}$	all	baseline
10	216	55	0.89	0.93	0.64	0.80	0.89	0.87	0.21
50	834	209	0.65	0.81	0.44	0.49	0.81	0.82	0.029
100	2006	502	0.55	0.79	0.37	0.39	0.79	0.80	0.019
200	3549	888	0.46	0.71	0.23	0.32	0.71	0.75	0.0079
400	7439	1860	0.39	0.70	0.23	0.27	0.71	0.73	0.0047
600	11102	2776	0.37	0.70	0.21	0.25	0.61	0.74	0.0029
651	11957	2990	0.36	0.62	0.20	0.23	0.67	0.72	0.0024

4 Case study: Genre analysis

"Cut out all those exclamation marks. An exclamation mark is like laughing at your own jokes."

— Attributed to F. Scott Fitzgerald, as conveyed by Sheilah Graham and Gerold Frank in *Beloved Infidel: The Education of a Woman*, 1958

"'Multiple exclamation marks,' he went on, shaking his head, 'are a sure sign of a diseased mind.'"

— Terry Pratchett, *Eric*, 1990

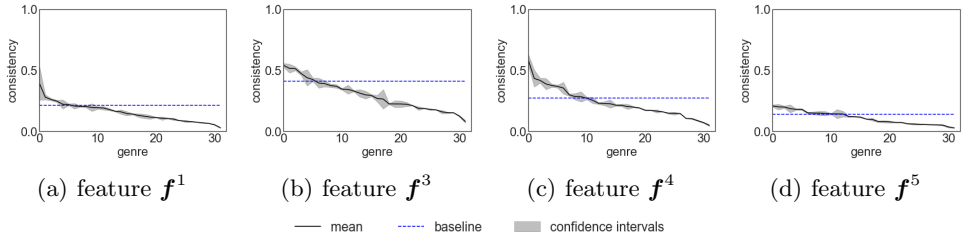
We now use genres as our classes. Among the 121 genre ("bookshelf") labels that are available in Gutenberg[8], we keep those that include at least 10 documents. Among the remaining genres, we select 32 relatively unspecialized genre labels. We show this final list of genres in appendix A. This yields a data set with 2413 documents.

4.1 Consistency

In fig. 10, we show consistency plots (of the same type as in fig. 8), but now we use genres (instead of authors) as our classes. We observe that the KL-divergence consistency relative to the baseline is less pronounced for genres than it was for authors. Nevertheless, most genres are more consistent than the baseline, and the frequency feature vector $\boldsymbol{f}^3$ appears to be the most helpful of our features for evaluating a genre's punctuation style.

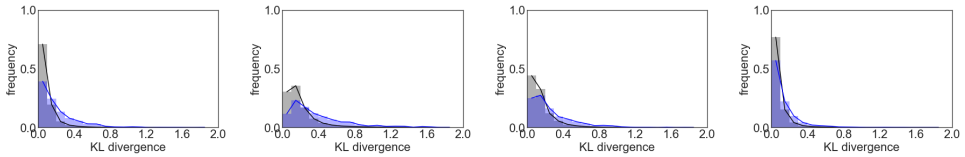
In fig. 11, we show the distributions of KL divergence between documents from the same genre (in black) and between documents from different genres (in blue). One can

[8] Every document in our corpus has at most one genre, but most documents are not assigned a genre.



(a) feature $\boldsymbol{f}^1$ (b) feature $\boldsymbol{f}^3$ (c) feature $\boldsymbol{f}^4$ (d) feature $\boldsymbol{f}^5$

Figure 10. Evaluation of genre consistency. In each panel, we show the genre consistency (specifically, we use equation (3.1), but with genres, instead of authors) for (a) $\boldsymbol{f}^1$, (b) $\boldsymbol{f}^3$, (c) $\boldsymbol{f}^4$, and (d) $\boldsymbol{f}^5$ as a solid black curve. In gray, we show confidence intervals of KL divergence across pairs of documents for each genre. To compute the confidence intervals, we assume that the KL divergence across pairs of distinct documents for each genre are normally distributed. There are at least 10 documents for each genre in our corpus (see the introduction of section 4), so the number of KL values across pairs of distinct documents for each genre is at least 90. The dotted blue line indicates a consistency baseline, which we obtain by choosing, uniformly at random, 1000 ordered pairs of documents from distinct genres and computing the mean KL divergence between the features of these document pairs.



(a) feature $\boldsymbol{f}^1$ (b) feature $\boldsymbol{f}^3$ (c) feature $\boldsymbol{f}^4$ (d) feature $\boldsymbol{f}^5$

Figure 11. Distributions of KL divergence for genre. In each panel, we show the distributions of KL divergence between all pairs of documents in the corpus from the same genre (in black) and between 1000 ordered pairs of documents from distinct genres (in blue). We choose the ordered pairs uniformly at random from the set of all ordered pairs of documents from distinct genres. Each panel corresponds to a distinct feature. The means of the distributions of each panel are (a) 0.102 (black) and 0.215 (blue), (b) 0.206 (black) and 0.412 (blue), (c) 0.154 (black) and 0.272 (blue), and (d) 0.0821 (black) and 0.138 (blue).

use the former to compute genre consistency in fig. 10 (by taking the mean of the values for each genre) and the latter to compute the consistency baseline in fig. 10 (by taking the mean of all values). For all features, we see from a KS test that the difference between the empirical distributions is statistically significant. (In all cases, the p-value is less than or equal to 2.247×10^{-36}.)

4.2 Genre recognition

We perform genre recognition using neural networks and show our results in table 4. We are less successful at genre detection than we were at author detection. This is consistent

Table 4. Results of our genre-recognition experiments using a one-layer, 2000-neuron neural network (see section 2.4.2) using different features or sets of features as input: $\boldsymbol{f}^1$, $\boldsymbol{f}^3$, $\boldsymbol{f}^4$, $\boldsymbol{f}^5$, $\{\boldsymbol{f}^1, \boldsymbol{f}^3, \boldsymbol{f}^4, \boldsymbol{f}^5\}$, and $\{\boldsymbol{f}^1, \boldsymbol{f}^2, \boldsymbol{f}^3, \boldsymbol{f}^4, \boldsymbol{f}^5, \boldsymbol{f}^6\}$ (which we label as "all"). We measure accuracy as the ratio of correctly assigned documents relative to the total number of documents in the testing set. (See section 2.5 for a description of the baseline.)

No. genres	Training size	Testing size	Accuracy on testing set						
			$\boldsymbol{f}^1$	$\boldsymbol{f}^3$	$\boldsymbol{f}^4$	$\boldsymbol{f}^5$	$\{\boldsymbol{f}^1, \boldsymbol{f}^3, \boldsymbol{f}^4, \boldsymbol{f}^5\}$	all	baseline
32	1930	483	0.56	0.65	0.37	0.40	0.61	0.64	0.094

with our genre consistency plots (see fig. 10), which indicated a smaller differentiation from the baseline than in our author consistency plots (see fig. 8). Our highest accuracy for genre recognition is 65%; we achieve it when using only the feature $\boldsymbol{f}^3$ as input. These observations are robust to different samples of the training and testing sets.

5 Case study: Temporal analysis

"Whatever it is that you know, or that you don't know, tell me about it. We can exchange tirades. The comma is my favorite piece of punctuation and I've got all night."

— Rasmenia Massoud, *Human Detritus*, 2011

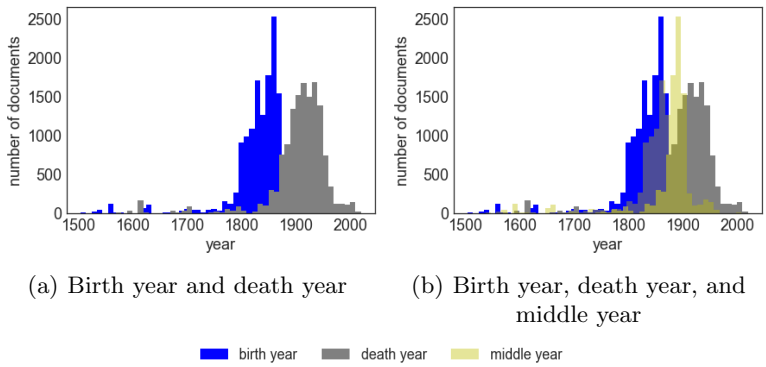
"Who gives a @!#?@! about an Oxford comma?
I've seen those English dramas too
They're cruel"

— Vampire Weekend, *Oxford Comma*, 2008

We perform experiments to obtain preliminary insight into how punctuation has changed over time. In our corpus, we have access to the birth year and death year of 614 and 615 authors, respectively, of the 651 total authors. We have both the birth and death years for 607 authors. In fig. 12, we show the distribution of the number of documents by author birth year, death year, and "middle year".[9] (See the caption of fig. 12 for the definition of middle year.) We restrict our analysis to authors with a middle year between 1500 and 2012. Of the authors for whom we possess either a birth year or a death year, 616 of them have a middle year between 1500 and 2012. We show the evolution of punctuation marks over time for these 616 authors in fig. 13 and fig. 14, and we examine the punctuation usage of specific authors over time in fig. 15. Based on our experiments, it appears from fig. 13 that the use of quotation marks and periods has increased over time (at least in our corpus), but that the use of commas has decreased over time. Less noticeably, the use of semicolons has also decreased over time.[10] In fig. 14, we observe

[9] We use "middle year" as a proxy for "publication year", which is unavailable in the metadata of Project Gutenberg. Our results are qualitatively similar when we use birth year or death year instead of middle year.

[10] See [49] for a "biography" of the semicolon, which reportedly was invented in 1494.

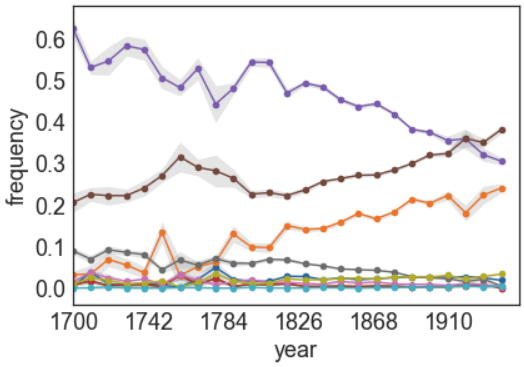


(a) Birth year and death year

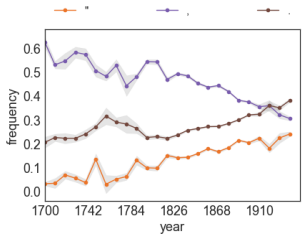
(b) Birth year, death year, and middle year

Figure 12. Distribution of author dates over time in our corpus. The bars represent the number of documents by author birth year (blue) and death year (gray) split into bins, where each bin represents a 10-year period. (We start at 1500.) For ease of visualization, we only show documents for authors who were born in 1500 or later. (Only six of our authors for whom we have birth years were born before 1500.) We determine the "middle year" of an author by taking the mean of the birth year and the death year if they are both available. If we know only the birth year, we assume that the middle year of an author is 30 years after the birth year; if we know only the death year, we assume that the middle year is 30 years prior to the death year.

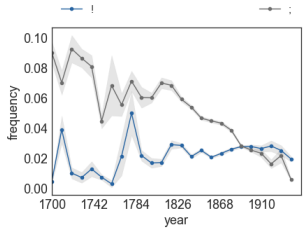
that the punctuation rate (given by the formula (2.6)) tends to decrease over time in our corpus. However, this observation requires further statistical testing, especially given the large variance in fig. 14. Because of our relatively small number of documents per author and the uneven distribution of documents in time, our experiments in fig. 15 give only preliminary insights into the temporal evolution of punctuation, which merits a thorough analysis with a much larger (and more appropriately sampled) corpus. Nevertheless, this case study illustrates the potential for studying the temporal evolution of punctuation styles of authors, genres, and literature (and other text) more generally.



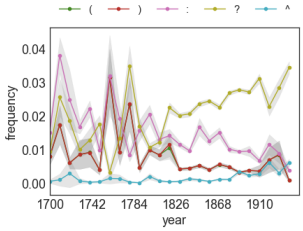
(a) Punctuation marks over time



(b) Quotation mark, period, and comma



(c) Exclamation mark and semicolon



(d) Left parenthesis, right parenthesis, colon, question mark, and ellipsis

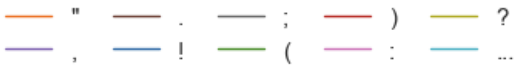


Figure 13. Mean frequency of punctuation marks versus the middle years of authors. Recall that $\boldsymbol{f}^{1,k}$ is the frequency of punctuation marks for document k. We bin middle years into 10-year periods that start at 1700. In (a), we show the temporal evolution of all punctuation marks. For clarity, we also separately plot (b) the three punctuation marks with the largest frequencies in the final year of our data set, (c) the next two most-frequent punctuation marks, and (d) the remaining punctuation marks. The gray shaded area indicates confidence intervals. To compute the confidence intervals, we assume that the values of $\boldsymbol{f}^{1,k}$ are normally distributed for each year.

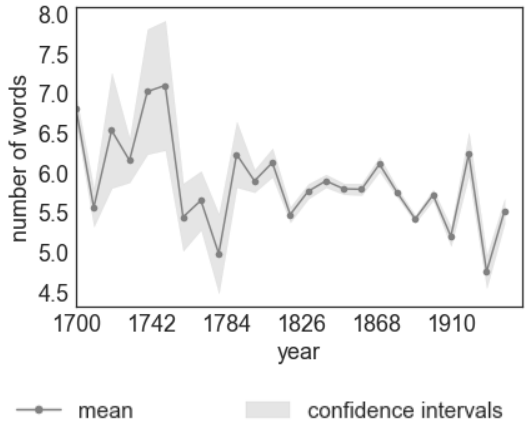
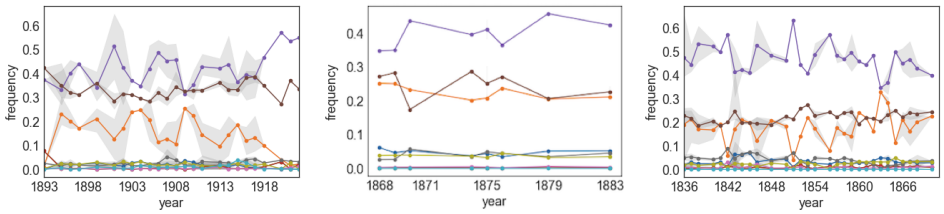


Figure 14. Temporal evolution of the mean number of words between two consecutive punctuation marks (i.e., $\mathbb{E}\left[\boldsymbol{f}^{5,k}\right]$ from formula (2.6)) versus author middle years, which we bin into 10-year periods that start at 1700. The gray shaded area indicates confidence intervals. To compute the confidence intervals, we assume that the values of $\mathbb{E}\left[\boldsymbol{f}^{5,k}\right]$ are normally distributed for each year. This reflects how the punctuation rate in our corpus has changed over time.



(a) H. G. Wells over time (b) A. M. Fleming over time (c) C. Dickens over time

Figure 15. Mean frequency of punctuation marks versus publication date for works by (a) Herbert George Wells, (b) Agnes May Fleming, and (c) Charles Dickens. Recall that $\boldsymbol{f}^{1,k}$ is the frequency of punctuation marks for document k. The gray shaded area indicates the minimum and maximum value of $\boldsymbol{f}^{1,k}$ for each year. (Because of the small sample sizes, we do not show confidence intervals.)

6 Conclusions and Discussion

"La punteggiatura è come l'elettroencefalogramma di un cervello che sogna — non dà le immagini ma rivela il ritmo del flusso sottostante."
— Andrea Moro, *Il Segreto di Pietramala*, 2018

We have explored whether punctuation is a sufficiently rich stylistic feature to distinguish between different authors and between different genres, and we have also examined how it has evolved over time. Using a large corpus of documents from Project Gutenberg, we observed that simple punctuation-based quantitative features (which account for both frequency and order) can distinguish accurately between the styles of different authors. These features can also help distinguish between genres, although less successfully than for authors. One feature, which we denote by $\boldsymbol{f}^3$, measures the frequency of successive punctuation marks (and thereby accounts for the order in which punctuation marks appear). Among the features that we studied, it revealed the most information about punctuation style across all of our experiments. It is worth noting that, unlike $\boldsymbol{f}^2$, which also accounts for the order of punctuation marks, $\boldsymbol{f}^3$ gives less weight to rare events and more weight to frequent events (see eq. (2.3)). This characteristic of $\boldsymbol{f}^3$, coupled with the fact that it accounts for the order of punctuation marks, may explain some of its success in our experiments. It would be interesting to investigate whether particular entries of $\boldsymbol{f}^3$ have more predictive power than others, and it is also worth exploring accuracy as a function of the length of the punctuation sequences that one extracts from a document. The latter may shed light on how much of a "punctuation signal" is necessary to determine an author's stylistic footprint. In preliminary explorations, we also observed changes in punctuation style across time, but it is necessary to conduct more thorough investigations of temporal usage patterns.

To assess whether our observations extend beyond our Project Gutenberg corpus, it is necessary to conduct further experiments (e.g., on a larger corpus, across different e-book sources, and so on). For example, it is desirable to repeat our analysis using the "Text data" level of granularity in the recently introduced Standardized Project Gutenberg Corpus [14]. We also reiterate that although we associate documents to authors throughout our paper as an expository shortcut, authors and editors both influence a document's writing and punctuation style, and we do not distinguish between the two in our analysis. It would be interesting (although daunting and computationally challenging for Project Gutenberg) to try to gauge whether and how much different editors affect authorial style.[11] It is also worth reiterating that Project Gutenberg has limitations with the cleanliness of its data. (See our discussion in section 2.1 for examples of such issues.) These issues may be inherited from the e-books themselves, they may be related to how the documents were entered into Project Gutenberg, or both issues may be present. Although we extensively clean the Project Gutenberg data to ameliorate some of its limitations, important future work is comparing documents that one extracts from Project Gutenberg with the same documents from other data sources.

[11] Such an analysis may be easier with academic papers, as one can compare papers on arXiv to their published versions.

Our framework allows the exploration of numerous other fascinating ideas. For example, we expect it to be fruitful to examine higher-order categorical Markov chains when accounting for punctuation order. Additionally, we look forward to extensions of our work that explore other features, such as the number of words between elements in ordered pairs of punctuation marks (even when they are not successive) and different ways of measuring punctuation frequency [16] and sentence length [48]), and that try to quantify how large a sample of a document is necessary to correctly identify its features of punctuation style. If this size is sufficiently small, it may even be possible to identify punctuation style from collections of short text (such as tweets from politicians with limited coherence). It is also likely to be useful to exploit more sophisticated machine-learning classifiers that can take raw punctuation sequences (rather than features that one produces from them) as input and exploit "long-range correlations" [30] between punctuation marks.

Building on our analysis, it will be interesting to investigate other aspects of stylometry — such as author pacing or the influence on an author of gender, culture, other demographics, local history, or other aspects of humanity — and to compare the results of punctuation-based stylometry with existing (word-based) approaches in NLP on the same tasks. One can also explore how successful punctuation-based features are at plagiarism detection and investigate whether the punctuation in a part of a document (e.g., one chapter) is representative of the punctuation in a whole document. Further investigations of a punctuation-based approach to stylometry also provide an opportunity to apply other methods for analyzing categorical time series (e.g., an extension of rough-path signatures [8, 32] to categorical time series).

On a more general front, relevant stylometric applications include analysis of stylistic differences in punctuation between politicians from different political parties [5] and comparisons between different editions of the same book. It would also be interesting to explore the effects of an editor's or journal's style on documents by a given author (an especially relevant study, in light of the potential to confound such contributions in corpuses like Project Gutenberg), as well as the effects of a translator's style on documents. We envisage that the latter application is particularly well-suited to punctuation-based stylometry, as punctuation marks depend far less than words on the specific choice of language. We also expect there to be commercial applications (e.g., using online data sources) of time-series analysis of symbols without the use of words.

Acknowledgements

The original inspiration for this project was Adam Calhoun's blog entry [4] and its striking visualizations of punctuation sequences. We thank Mariano Beguerisse Díaz, Arthur Benjamin, Bryan Bischof, Chris Brew, Cynthia Gong, Joanna Innes, Jalil Kazemitabar, Aisling Kelliher, Terry Lyons, Ursula Martin, Stephen Pulman, Massimo Stella, Adam Tsakalidis, Dmitri Vainchtein, Bo Wang, and two anonymous referees for helpful comments. Other attendees at SDH's 60th birthday workshop (see `https://www.maths.ox.ac.uk/groups/mathematical-finance/sam-howisons-60th-birthday-workshop-2018`) also made helpful comments. For part of this project, MB was supported by The Alan Turing Institute under EPSRC grant EP/N510129/1. MAP and SDH thank their stu-

dents and postdocs for putting up with many long discussions about punctuation when they perhaps should have been discussing other elements of their scholarship. (It was inevitable that we would eventually write an article like this.) MAP thanks SDH for his collaboration and friendship, and he wishes him a very happy birthday filled with British spelling, the word "which" (and occasionally "that"), and minimal commas (and parenthetical remarks).

Appendix A Author and genre lists

`"Mr Speaker, I said the honourable Member was a liar it is true and I am sorry for it. The honourable Member may place the punctuation where he pleases."`

— Attributed to Richard Brinsley Sheridan (1751–1816), responding to a rebuke from the Chair for calling a fellow Member of Parliament a liar.

In table A 1, we list the authors that we use in our study. We order them based on their $\boldsymbol{f}^3$ consistency, where smaller numbers indicate greater consistency. (See (3.1) for the definition of "consistency".) The author order proceeds down the first column and then down the second column. We structure each row as follows: "Author name (number of documents by that author in our corpus, test-set size for our experiments on the full corpus with the full set of features, author $\boldsymbol{f}^3$ consistency in our corpus, author accuracy on test set)". Accuracy values that are closer to 1 indicate that we correctly assign a larger fraction of books by that author. (See (2.5) for the definition of "accuracy".) The designation "NA" indicates that an author is not in the test set. We number each row in table A 1 to facilitate the referencing of specific authors. One number references two distinct authors (with one in each column), and we increment the row number from page to page in a way that accounts for the number of authors in the second column.

In table A 2, we list the genres that we use in our study. We order them based on their $\boldsymbol{f}^3$ consistency. The genre order proceeds down the first column and then down the second column. We structure each row as follows: "Genre (number of documents in the genre, test-set size for our experiments on the full corpus with the full set of features, author $\boldsymbol{f}^3$ consistency in our corpus, genre accuracy on test set)". Consistency values that are closer to 0 correspond to genres that are more consistent, and accuracy values that are closer to 1 indicate that we correctly assign a larger fraction of books of that author. We number each row in table A 2 to facilitate the referencing of specific genres. One number references two distinct genres (with one in each column).

Table A 1. The authors that we use in our study.

	author (No. documents - test size - consistency - accuracy)	author (No. documents - test size - consistency - accuracy)
0	Matthews, Stanley R. (32 - 5.0 - 0.018 - 1.0)	Werner, E. (18 - 2.0 - 0.053 - 1.0)
1	Hill, Grace Brooks (11 - 2.0 - 0.02 - 0.0)	Kyne, Peter B. (Peter Bernard) (10 - 2.0 - 0.054 - 0.0)
2	Dell, Ethel M. (Ethel May) (16 - 5.0 - 0.02 - 0.8)	Wood, Henry, Mrs. (24 - 6.0 - 0.055 - 1.0)
3	Goodwin, Harold L. (Harold Leland) (13 - 2.0 - 0.021 - 1.0)	King, Charles (27 - 6.0 - 0.055 - 1.0)
4	Young, Clarence (23 - 7.0 - 0.023 - 0.857)	Bassett, Sara Ware (16 - 3.0 - 0.055 - 0.333)
5	Hancock, H. Irving (Harrie Irving) (40 - 9.0 - 0.024 - 1.0)	Abbott, John S. C. (John Stevens Cabot) (23 - 7.0 - 0.056 - 1.0)
6	Wirt, Mildred A. (Mildred Augustine) (30 - 5.0 - 0.024 - 1.0)	Gregory, Jackson (10 - NA - 0.056 - NA)
7	United States. Warren Commission (12 - 1.0 - 0.025 - 1.0)	Maclaren, Alexander (20 - 7.0 - 0.056 - 1.0)
8	Merriman, Henry Seton (14 - 3.0 - 0.026 - 1.0)	De Quincey, Thomas (20 - 3.0 - 0.056 - 1.0)
9	Brame, Charlotte M. (11 - 1.0 - 0.026 - 1.0)	Aimard, Gustave (29 - 5.0 - 0.056 - 0.8)
10	Patchin, Frank Gee (15 - NA - 0.026 - NA)	Mundy, Talbot (13 - 3.0 - 0.056 - 1.0)
11	Norris, Kathleen Thompson (11 - 3.0 - 0.027 - 1.0)	Carey, Rosa Nouchette (11 - 2.0 - 0.056 - 0.5)
12	Hayes, Clair W. (Clair Wallace) (18 - 2.0 - 0.028 - 1.0)	Barbour, Ralph Henry (32 - 6.0 - 0.056 - 1.0)
13	Hocking, Joseph (11 - 3.0 - 0.028 - 0.333)	Goldfrap, John Henry (37 - 7.0 - 0.056 - 0.857)
14	Locke, William John (21 - 4.0 - 0.028 - 0.75)	Nicholson, Meredith (13 - 3.0 - 0.056 - 0.667)
15	Henry, O. (13 - 2.0 - 0.028 - 1.0)	Tarkington, Booth (19 - 4.0 - 0.056 - 0.75)
16	Parrish, Randall (15 - 4.0 - 0.03 - 1.0)	Packard, Frank L. (Frank Lucius) (11 - 2.0 - 0.057 - 1.0)
17	Bowen, Robert Sidney (15 - 3.0 - 0.031 - 1.0)	Dowling, Richard (16 - 1.0 - 0.058 - 0.0)
18	Lynde, Francis (17 - 2.0 - 0.031 - 1.0)	Ainsworth, William Harrison (20 - 1.0 - 0.058 - 1.0)
19	Bloundelle-Burton, John (14 - 3.0 - 0.031 - 0.667)	Everett-Green, Evelyn (19 - 6.0 - 0.058 - 0.833)
20	Suetonius (14 - 3.0 - 0.031 - 1.0)	Saint-Simon, Louis de Rouvroy, duc de (15 - 2.0 - 0.058 - 0.5)
21	Wairy, Louis Constant (12 - 3.0 - 0.033 - 1.0)	Thorne, Guy (15 - 3.0 - 0.059 - 0.667)
22	Blanchard, Amy Ella (12 - 3.0 - 0.033 - 0.0)	Seltzer, Charles Alden (10 - 2.0 - 0.059 - 0.5)
23	Cholmondeley, Mary (11 - 4.0 - 0.036 - 0.25)	Meade, L. T. (52 - 12.0 - 0.059 - 0.667)
24	Buffon, Georges Louis Leclerc, comte de (10 - 1.0 - 0.036 - 1.0)	Douglas, Amanda M. (19 - 2.0 - 0.059 - 0.5)
25	Walton, Amy (10 - 2.0 - 0.036 - 1.0)	Fitzhugh, Percy Keese (22 - 4.0 - 0.06 - 1.0)
26	Ferber, Edna (10 - 2.0 - 0.036 - 1.0)	Oppenheim, E. Phillips (Edward Phillips) (58 - 14.0 - 0.06 - 1.0)
27	Hope, Laura Lee (64 - 11.0 - 0.037 - 0.818)	Stephens, Ann S. (Ann Sophia) (13 - 5.0 - 0.06 - 0.6)
28	Chadwick, Lester (16 - 4.0 - 0.037 - 0.75)	Fyfe, H. B. (Horace Bowne) (16 - 2.0 - 0.061 - 1.0)
29	Mitford, Bertram (27 - 2.0 - 0.038 - 1.0)	Wodehouse, P. G. (Pelham Grenville) (37 - 4.0 - 0.061 - 1.0)
30	Appleton, Victor (31 - 5.0 - 0.038 - 0.4)	Deland, Margaret Wade Campbell (11 - 3.0 - 0.061 - 0.667)
31	Penrose, Margaret (22 - 2.0 - 0.039 - 0.5)	Holt, Emily Sarah (22 - 5.0 - 0.061 - 0.8)
32	Collingwood, Harry (33 - 6.0 - 0.039 - 1.0)	Carter, Herbert, active 1909-1917 (12 - NA - 0.061 - NA)
33	Finley, Martha (35 - 8.0 - 0.04 - 0.625)	Porter, Eleanor H. (Eleanor Hodgman) (13 - 4.0 - 0.062 - 0.75)
34	Mackintosh, Charles Henry (11 - 2.0 - 0.04 - 1.0)	Moore, Frank Frankfort (19 - 6.0 - 0.062 - 1.0)
35	Phillips, David Graham (14 - 2.0 - 0.04 - 0.5)	Farjeon, B. L. (Benjamin Leopold) (29 - 4.0 - 0.062 - 1.0)
36	Boldrewood, Rolf (15 - NA - 0.04 - NA)	Snell, Roy J. (Roy Judson) (40 - 10.0 - 0.062 - 1.0)
37	Harper, Charles G. (Charles George) (16 - NA - 0.04 - NA)	Kock, Paul de (18 - 6.0 - 0.062 - 0.833)
38	Weyman, Stanley John (28 - 11.0 - 0.041 - 1.0)	Johnson, Owen (11 - 3.0 - 0.062 - 0.667)
39	Roy, Lillian Elizabeth (16 - 3.0 - 0.041 - 1.0)	Walsh, James J. (James Joseph) (12 - 3.0 - 0.063 - 1.0)
40	Emerson, Alice B. (23 - 2.0 - 0.042 - 0.0)	Blackwood, Algernon (22 - 6.0 - 0.063 - 0.667)
41	McCutcheon, George Barr (33 - 6.0 - 0.043 - 0.667)	Craik, Dinah Maria Mulock (15 - 3.0 - 0.063 - 0.333)
42	Reeve, Arthur B. (Arthur Benjamin) (14 - NA - 0.044 - NA)	Marlowe, Stephen (16 - 4.0 - 0.063 - 0.5)
43	Shaler, Robert (18 - 5.0 - 0.044 - 0.4)	Harben, Will N. (Will Nathaniel) (13 - NA - 0.063 - NA)
44	Bourrienne, Louis Antoine Fauvelet de (16 - 3.0 - 0.044 - 0.333)	Robertson, Margaret M. (Margaret Murray) (11 - 1.0 - 0.064 - 1.0)
45	Vaizey, George de Horne, Mrs. (22 - NA - 0.044 - NA)	De Mille, James (17 - 3.0 - 0.064 - 0.667)
46	Mathews, Joanna H. (Joanna Hooe) (13 - 2.0 - 0.044 - 1.0)	Rockwood, Roy (16 - 1.0 - 0.064 - 0.0)
47	Vance, Louis Joseph (12 - NA - 0.045 - NA)	Holmes, Mary Jane (21 - 3.0 - 0.064 - 1.0)
48	Duncan, Sara Jeannette (10 - 1.0 - 0.045 - 0.0)	Mühlbach L. (Luise) (20 - 2.0 - 0.064 - 0.5)
49	Pansy (11 - 1.0 - 0.045 - 0.0)	Leslie, Madeline (20 - 4.0 - 0.065 - 1.0)
50	Raine, William MacLeod (22 - 4.0 - 0.046 - 1.0)	Oliphant, Mrs. (Margaret) (70 - 14.0 - 0.066 - 0.929)
51	Douglas, Alan, Captain (10 - 4.0 - 0.046 - 0.75)	Boothby, Guy (16 - 4.0 - 0.066 - 0.25)
52	MacGrath, Harold (21 - NA - 0.048 - NA)	Green, Anna Katharine (35 - 5.0 - 0.066 - 0.8)
53	Cannon, Richard (26 - 5.0 - 0.048 - 1.0)	Williamson, C. N. (Charles Norris) (19 - 5.0 - 0.066 - 0.4)
54	Warner, Susan (25 - 2.0 - 0.048 - 1.0)	Hale, Edward Everett (10 - 5.0 - 0.066 - 0.2)
55	Cody, H. A. (Hiram Alfred) (12 - 3.0 - 0.048 - 1.0)	Aycock, Roger D. (12 - 4.0 - 0.066 - 0.75)
56	Brazil, Angela (27 - 4.0 - 0.048 - 1.0)	Daviess, Maria Thompson (11 - 2.0 - 0.067 - 1.0)
57	Barr, Robert (20 - 4.0 - 0.048 - 0.75)	Day, Holman (11 - NA - 0.067 - NA)
58	Rice, Alice Caldwell Hegan (10 - 4.0 - 0.049 - 0.0)	Chambers, Robert W. (Robert William) (43 - 9.0 - 0.067 - 0.889)
59	Frey, Hildegard G. (10 - NA - 0.049 - NA)	Munroe, Kirk (15 - 3.0 - 0.067 - 0.667)
60	Southworth, Emma Dorothy Eliza Nevitte (13 - 2.0 - 0.049 - 1.0)	Blackmore, R. D. (Richard Doddridge) (23 - 5.0 - 0.068 - 1.0)
61	Standish, Burt L. (25 - 2.0 - 0.049 - 1.0)	Mansfield, M. F. (Milburg Francisco) (16 - 4.0 - 0.068 - 0.75)
62	Tracy, Louis (27 - 5.0 - 0.049 - 0.6)	Crockett, S. R. (Samuel Rutherford) (19 - 4.0 - 0.068 - 0.75)
63	Altsheler, Joseph A. (Joseph Alexander) (33 - 8.0 - 0.049 - 1.0)	Chase, Josephine (32 - 4.0 - 0.068 - 0.75)
64	Skinner, Charles M. (Charles Montgomery) (10 - 2.0 - 0.05 - 1.0)	Heyse, Paul (10 - 4.0 - 0.068 - 0.25)
65	Hutcheson, John C. (John Conroy) (17 - 1.0 - 0.05 - 1.0)	Buck, Charles Neville (11 - 1.0 - 0.068 - 1.0)
66	Braddon, M. E. (Mary Elizabeth) (30 - 5.0 - 0.05 - 1.0)	Mangasarian, M. M. (Mangasar Mugurditch) (12 - NA - 0.069 - NA)
67	Comstock, Harriet T. (Harriet Theresa) (10 - 4.0 - 0.051 - 0.5)	Shakespeare (spurious and doubtful works) (10 - 1.0 - 0.069 - 0.0)
68	Glasgow, Ellen Anderson Gholson (12 - 3.0 - 0.051 - 0.667)	Riis, Jacob A. (Jacob August) (11 - 2.0 - 0.069 - 0.0)
69	Beach, Rex (16 - 4.0 - 0.052 - 0.75)	Miller, Alex. McVeigh, Mrs. (17 - 2.0 - 0.069 - 1.0)
70	Cullum, Ridgwell (17 - 2.0 - 0.052 - 1.0)	Westerman, Percy F. (Percy Francis) (34 - 10.0 - 0.07 - 0.9)
71	Stratemeyer, Edward (75 - 13.0 - 0.052 - 0.923)	Ewing, Juliana Horatia Gatty (20 - 3.0 - 0.07 - 0.667)
72	May, Sophie (25 - 2.0 - 0.052 - 1.0)	Schubin, Ossip (10 - 2.0 - 0.07 - 0.0)
73	Bower, B. M. (29 - 6.0 - 0.052 - 1.0)	Lavell, Edith (11 - 1.0 - 0.071 - 1.0)
74	Fleming, May Agnes (11 - 2.0 - 0.052 - 0.5)	James, G. P. R. (George Payne Rainsford) (49 - 7.0 - 0.071 - 1.0)

	author (No. documents - test size - consistency - accuracy)	author (No. documents - test size - consistency - accuracy)
150	Sheckley, Robert (18 - 6.0 - 0.03 - 0.667)	Cawein, Madison Julius (19 - 1.0 - 0.039 - 1.0)
151	Williamson, C. N. (Charles Norris) (19 - 5.0 - 0.03 - 0.4)	King, Basil (10 - 4.0 - 0.039 - 1.0)
152	Vandercook, Margaret (24 - 4.0 - 0.03 - 1.0)	Jókai, Mór (28 - 9.0 - 0.039 - 0.444)
153	Schubin, Ossip (10 - 2.0 - 0.03 - 0.0)	Schmitz, James H. (10 - NA - 0.039 - NA)
154	Mangasarian, M. M. (Mangasar Mugurditch) (12 - NA - 0.03 - NA)	Rohmer, Sax (17 - 1.0 - 0.039 - 0.0)
155	Chambers, Robert W. (Robert William) (43 - 9.0 - 0.03 - 0.889)	Sue, Eugène (44 - 11.0 - 0.04 - 0.818)
156	Heyse, Paul (10 - 4.0 - 0.03 - 0.25)	Reynolds, Mack (24 - 5.0 - 0.04 - 1.0)
157	Holinshed, Raphael (27 - 3.0 - 0.03 - 1.0)	Maspero, G. (Gaston) (10 - 3.0 - 0.04 - 1.0)
158	Moore, Frank Frankfort (19 - 6.0 - 0.03 - 1.0)	Stoddard, William Osborn (12 - 2.0 - 0.04 - 1.0)
159	Steel, Flora Annie Webster (20 - 6.0 - 0.03 - 1.0)	Haggard, H. Rider (Henry Rider) (51 - 9.0 - 0.04 - 0.778)
160	Bindloss, Harold (43 - 11.0 - 0.03 - 1.0)	Ward, Humphry, Mrs. (33 - 8.0 - 0.04 - 0.625)
161	Smith, E. E. (Edward Elmer) (10 - 3.0 - 0.031 - 0.667)	Strang, Herbert (32 - 7.0 - 0.04 - 1.0)
162	Lavell, Edith (11 - 1.0 - 0.031 - 1.0)	Black, William (20 - 5.0 - 0.04 - 0.8)
163	Ellis, Havelock (12 - 2.0 - 0.031 - 1.0)	Lincoln, Joseph Crosby (18 - 2.0 - 0.041 - 1.0)
164	Munroe, Kirk (15 - 3.0 - 0.031 - 0.667)	Mitford, Mary Russell (13 - 1.0 - 0.041 - 1.0)
165	Jefferson, Thomas (17 - 6.0 - 0.031 - 1.0)	Maupassant, Guy de (33 - 8.0 - 0.041 - 0.75)
166	Mulford, Clarence Edward (10 - 3.0 - 0.031 - 1.0)	Crane, Stephen (13 - 3.0 - 0.041 - 0.333)
167	Brand, Max (14 - 1.0 - 0.031 - 1.0)	Norton, Andre (14 - 5.0 - 0.041 - 0.6)
168	Oxley, J. Macdonald (James Macdonald) (10 - 1.0 - 0.031 - 1.0)	Hay, Ian (13 - 1.0 - 0.042 - 1.0)
169	James, William (11 - 1.0 - 0.031 - 1.0)	Pater, Walter (13 - 1.0 - 0.042 - 0.0)
170	Hope, Anthony (33 - 5.0 - 0.031 - 0.6)	Sharp, Dallas Lore (10 - 3.0 - 0.042 - 1.0)
171	Smiles, Samuel (14 - 1.0 - 0.032 - 0.0)	Macaulay, Thomas Babington, Baron (19 - 1.0 - 0.042 - 0.0)
172	Nye, Bill (11 - NA - 0.032 - NA)	Ford, Sewell (12 - 3.0 - 0.042 - 1.0)
173	James, G. P. R. (George Payne Rainsford) (49 - 7.0 - 0.032 - 1.0)	Spyri, Johanna (15 - 4.0 - 0.042 - 1.0)
174	Hume, Fergus (63 - 17.0 - 0.032 - 0.941)	Symonds, John Addington (15 - 2.0 - 0.042 - 0.5)
175	Speed, Nell (16 - 5.0 - 0.032 - 0.8)	Burroughs, John (23 - 5.0 - 0.042 - 1.0)
176	United States. Central Intelligence Agency (21 - 4.0 - 0.032 - 1.0)	Molesworth, Mrs. (55 - 8.0 - 0.043 - 1.0)
177	Bailey, Arthur Scott (40 - 10.0 - 0.032 - 1.0)	Orczy, Emmuska Orczy, Baroness (18 - 3.0 - 0.043 - 1.0)
178	Harris, Frank (10 - NA - 0.032 - NA)	Murfree, Mary Noailles (26 - 2.0 - 0.043 - 0.5)
179	Loti, Pierre (11 - 3.0 - 0.032 - 0.667)	Buchanan, Robert Williams (10 - 1.0 - 0.044 - 1.0)
180	Stephens, Robert Neilson (10 - 4.0 - 0.032 - 0.25)	Wood, William Charles Henry (12 - NA - 0.044 - NA)
181	Hendryx, James B. (James Beardsley) (10 - 1.0 - 0.033 - 1.0)	Walpole, Hugh (12 - 2.0 - 0.044 - 0.5)
182	Gale, Zona (10 - 3.0 - 0.033 - 0.667)	Fletcher, J. S. (Joseph Smith) (17 - 2.0 - 0.044 - 1.0)
183	Castlemon, Harry (38 - 8.0 - 0.033 - 0.875)	Russell, William Clark (18 - 10.0 - 0.044 - 0.4)
184	Arthur, T. S. (Timothy Shay) (32 - 10.0 - 0.033 - 0.6)	Marsh, Richard (19 - 5.0 - 0.044 - 0.4)
185	Jameson, Mrs. (Anna) (10 - NA - 0.033 - NA)	Ouida (22 - 2.0 - 0.045 - 1.0)
186	Habberton, John (11 - 2.0 - 0.033 - 0.5)	Bensusan, S. L. (Samuel Levy) (11 - 3.0 - 0.045 - 0.333)
187	Wallace, F. L. (Floyd L.) (13 - 1.0 - 0.033 - 1.0)	Pepys, Samuel (76 - 18.0 - 0.045 - 1.0)
188	Onions, Oliver (11 - 3.0 - 0.033 - 0.667)	Johnston, Annie F. (Annie Fellows) (37 - 7.0 - 0.045 - 0.571)
189	Bacon, Josephine Daskam (13 - 1.0 - 0.033 - 1.0)	Smith, Francis Hopkinson (26 - 3.0 - 0.045 - 0.667)
190	Shakespeare (spurious and doubtful works) (10 - 1.0 - 0.034 - 0.0)	Smith, Evelyn E. (15 - 3.0 - 0.046 - 1.0)
191	Burgess, Thornton W. (Thornton Waldo) (37 - 8.0 - 0.034 - 1.0)	Norris, Frank (10 - 4.0 - 0.046 - 0.25)
192	Barr, Amelia E. (26 - 4.0 - 0.034 - 1.0)	Smith, George O. (George Oliver) (10 - 1.0 - 0.046 - 1.0)
193	Brereton, F. S. (Frederick Sadleir) (18 - 5.0 - 0.034 - 0.6)	Stacpoole, H. De Vere (Henry De Vere) (20 - 6.0 - 0.046 - 0.167)
194	Hill, Grace Livingston (15 - 4.0 - 0.034 - 0.5)	Pemberton, Max (11 - 3.0 - 0.046 - 0.333)
195	Mill, John Stuart (14 - 2.0 - 0.034 - 1.0)	Lord, John (18 - 5.0 - 0.046 - 0.6)
196	Alcott, Louisa May (37 - 5.0 - 0.034 - 0.8)	Hume, David (13 - 2.0 - 0.046 - 1.0)
197	Moody, Dwight Lyman (14 - 2.0 - 0.035 - 1.0)	Irving, Washington (20 - 3.0 - 0.047 - 0.667)
198	Hale, Edward Everett (10 - 5.0 - 0.035 - 0.2)	MacGregor, Mary Esther Miller (10 - 3.0 - 0.047 - 0.0)
199	Machen, Arthur (10 - 3.0 - 0.035 - 0.333)	Ballantyne, R. M. (Robert Michael) (91 - 21.0 - 0.047 - 0.952)
200	Perkins, Lucy Fitch (13 - 4.0 - 0.035 - 0.5)	Lowndes, Marie Belloc (15 - NA - 0.048 - NA)
201	Chapman, Allen (25 - 2.0 - 0.035 - 0.5)	Alger, Horatio, Jr. (95 - 21.0 - 0.048 - 0.857)
202	Fox, John (13 - 3.0 - 0.035 - 0.667)	Webster, Frank V. (19 - 3.0 - 0.048 - 0.333)
203	James, George Wharton (11 - 2.0 - 0.035 - 0.0)	Richards, Laura Elizabeth Howe (42 - 6.0 - 0.048 - 0.833)
204	Connor, Ralph (14 - 4.0 - 0.035 - 1.0)	Cervantes Saavedra, Miguel de (47 - 13.0 - 0.049 - 0.462)
205	Whyte-Melville, G. J. (George John) (10 - 5.0 - 0.036 - 0.0)	Church, Alfred John (12 - 3.0 - 0.049 - 0.0)
206	Marryat, Frederick (36 - 6.0 - 0.036 - 1.0)	Garrett, Randall (43 - 10.0 - 0.049 - 0.6)
207	Williamson, A. M. (Alice Muriel) (15 - 3.0 - 0.036 - 0.333)	Farnol, Jeffery (14 - 2.0 - 0.049 - 1.0)
208	Von Arnim, Elizabeth (12 - 2.0 - 0.036 - 1.0)	Le Queux, William (66 - 8.0 - 0.049 - 0.875)
209	Harland, Henry (12 - 5.0 - 0.036 - 0.8)	Romanes, George John (11 - 3.0 - 0.049 - 1.0)
210	Grey, Zane (26 - 4.0 - 0.037 - 1.0)	Parkman, Francis (15 - 2.0 - 0.05 - 0.5)
211	Saunders, Marshall (13 - 2.0 - 0.037 - 0.0)	Saintsbury, George (12 - 3.0 - 0.05 - 0.667)
212	Sedgwick, Anne Douglas (14 - 2.0 - 0.038 - 0.5)	Fiske, John (18 - 4.0 - 0.05 - 0.5)
213	Hornung, E. W. (Ernest William) (26 - 2.0 - 0.038 - 1.0)	Turgenev, Ivan Sergeevich (22 - 5.0 - 0.051 - 0.6)
214	Del Rey, Lester (12 - NA - 0.038 - NA)	Duellman, William Edward (12 - 2.0 - 0.051 - 0.5)
215	Fenn, George Manville (128 - 28.0 - 0.038 - 0.964)	Santayana, George (10 - 2.0 - 0.051 - 0.0)
216	Richmond, Grace S. (Grace Smith) (15 - 4.0 - 0.038 - 0.5)	Garland, Hamlin (23 - 2.0 - 0.051 - 1.0)
217	Hulbert, Archer Butler (17 - 1.0 - 0.038 - 1.0)	Marks, Winston K. (12 - 3.0 - 0.051 - 0.333)
218	Catherwood, Mary Hartwell (20 - 8.0 - 0.038 - 0.625)	Bellamy, Edward (20 - 4.0 - 0.051 - 0.25)
219	Kingston, William Henry Giles (131 - 32.0 - 0.038 - 0.938)	Gissing, George (24 - 9.0 - 0.051 - 0.667)
220	Auerbach, Berthold (10 - 2.0 - 0.038 - 0.5)	Doctorow, Cory (13 - 4.0 - 0.051 - 1.0)
221	Burke, Edmund (15 - 3.0 - 0.038 - 1.0)	Dante Alighieri (32 - 5.0 - 0.051 - 0.6)
222	Vasari, Giorgio (11 - 1.0 - 0.039 - 1.0)	Hoare, Edward (32 - 8.0 - 0.051 - 0.5)
223	Frederic, Harold (14 - 2.0 - 0.039 - 0.5)	Rathborne, St. George (14 - 2.0 - 0.052 - 0.5)
224	Spencer, Herbert (10 - 1.0 - 0.039 - 1.0)	Motley, John Lothrop (89 - 17.0 - 0.052 - 0.882)

	author (No. documents - test size - consistency - accuracy)	author (No. documents - test size - consistency - accuracy)
300	Wiggin, Kate Douglas Smith (33 - 6.0 - 0.052 - 0.667)	Chesterfield, Philip Dormer Stanhope, Earl of (12 - 2.0 - 0.069 - 1.0)
301	Samachson, Joseph (12 - 1.0 - 0.052 - 0.0)	Collins, Wilkie (35 - 5.0 - 0.069 - 0.8)
302	Mitton, G. E. (Geraldine Edith) (12 - 2.0 - 0.052 - 0.0)	Atherton, Gertrude Franklin Horn (25 - 5.0 - 0.07 - 0.8)
303	Curwood, James Oliver (27 - NA - 0.052 - NA)	Huneker, James (11 - 1.0 - 0.07 - 0.0)
304	Kjelgaard, Jim (11 - 4.0 - 0.052 - 0.75)	Swift, Jonathan (16 - 1.0 - 0.07 - 0.0)
305	Crawford, F. Marion (Francis Marion) (47 - 11.0 - 0.052 - 0.818)	Huxley, Thomas Henry (48 - 13.0 - 0.07 - 0.692)
306	Nourse, Alan Edward (23 - 3.0 - 0.053 - 0.667)	Rinehart, Mary Roberts (29 - 6.0 - 0.07 - 0.333)
307	Thoreau, Henry David (11 - 2.0 - 0.053 - 0.0)	Trollope, Anthony (78 - 24.0 - 0.071 - 0.75)
308	Cable, George Washington (14 - 1.0 - 0.053 - 1.0)	Abbott, Eleanor Hallowell (10 - 4.0 - 0.071 - 0.75)
309	Leblanc, Maurice (16 - 6.0 - 0.053 - 1.0)	Howard, Robert E. (Robert Ervin) (12 - 4.0 - 0.071 - 1.0)
310	Parker, Gilbert (106 - 18.0 - 0.053 - 0.778)	Müller, F. Max (Friedrich Max) (10 - 2.0 - 0.071 - 0.5)
311	Mahan, A. T. (Alfred Thayer) (15 - 2.0 - 0.053 - 0.5)	Gautier, Théophile (11 - 1.0 - 0.071 - 0.0)
312	Foote, G. W. (George William) (10 - 1.0 - 0.054 - 1.0)	Lever, Charles James (53 - 12.0 - 0.072 - 0.75)
313	Duncan, Norman (10 - 2.0 - 0.054 - 0.5)	Hegel, Georg Wilhelm Friedrich (10 - 2.0 - 0.072 - 1.0)
314	Couperus, Louis (13 - 1.0 - 0.054 - 1.0)	Hichens, Robert (27 - 5.0 - 0.072 - 1.0)
315	Fanny, Aunt (13 - 5.0 - 0.055 - 0.6)	Emerson, Ralph Waldo (12 - 3.0 - 0.072 - 0.667)
316	Laumer, Keith (12 - 3.0 - 0.055 - 0.667)	Coolidge, Susan (14 - 4.0 - 0.073 - 0.75)
317	Lamb, Charles (10 - 3.0 - 0.055 - 0.667)	Corelli, Marie (14 - 3.0 - 0.073 - 0.333)
318	Harrison, Harry (10 - 3.0 - 0.055 - 0.667)	Wright, Harold Bell (10 - 1.0 - 0.073 - 0.0)
319	Grant, James, archaeologist (12 - 2.0 - 0.055 - 1.0)	Benson, Robert Hugh (11 - 3.0 - 0.073 - 0.667)
320	Beerbohm, Max, Sir (10 - 3.0 - 0.055 - 0.333)	Woolson, Constance Fenimore (14 - 3.0 - 0.074 - 0.667)
321	Gaskell, Elizabeth Cleghorn (23 - 1.0 - 0.055 - 1.0)	Glyn, Elinor (17 - 5.0 - 0.075 - 0.6)
322	Ingersoll, Robert Green (30 - 6.0 - 0.056 - 1.0)	Wade, Mary Hazelton Blanchard (21 - 4.0 - 0.075 - 0.75)
323	Butler, Samuel (18 - 1.0 - 0.056 - 0.0)	Bates, Arlo (14 - 5.0 - 0.075 - 0.6)
324	Merwin, Samuel (13 - 3.0 - 0.056 - 0.333)	Griffiths, Arthur (18 - 2.0 - 0.076 - 0.5)
325	Dewey, John (15 - 1.0 - 0.056 - 1.0)	Sienkiewicz, Henryk (18 - 3.0 - 0.076 - 0.0)
326	Hardy, Thomas (26 - 4.0 - 0.057 - 0.75)	Scott, Walter (56 - 9.0 - 0.076 - 0.778)
327	Bacheller, Irving (18 - 5.0 - 0.057 - 0.6)	Benson, E. F. (Edward Frederic) (28 - 8.0 - 0.077 - 0.875)
328	Follen, Eliza Lee Cabot (10 - 4.0 - 0.057 - 0.5)	Eliot, George (15 - 4.0 - 0.077 - 0.75)
329	Morley, John (30 - 4.0 - 0.057 - 1.0)	Guiney, Louise Imogen (13 - NA - 0.078 - NA)
330	Peattie, Elia Wilkinson (10 - 2.0 - 0.058 - 0.0)	Le Fanu, Joseph Sheridan (31 - 9.0 - 0.078 - 0.667)
331	Ritchie, J. Ewing (James Ewing) (20 - 2.0 - 0.058 - 0.0)	Blasco Ibáñez, Vicente (14 - 2.0 - 0.078 - 1.0)
332	Holley, Marietta (16 - 4.0 - 0.058 - 0.75)	Benson, Arthur Christopher (16 - 3.0 - 0.078 - 0.667)
333	Stables, Gordon (26 - 6.0 - 0.059 - 0.5)	Hurll, Estelle M. (Estelle May) (13 - 4.0 - 0.079 - 1.0)
334	Birmingham, George A. (15 - 3.0 - 0.059 - 0.667)	Stowe, Harriet Beecher (31 - 4.0 - 0.079 - 0.25)
335	Edgeworth, Maria (18 - 6.0 - 0.059 - 0.5)	Whittier, John Greenleaf (37 - 5.0 - 0.079 - 1.0)
336	Stephen, Leslie (11 - 2.0 - 0.059 - 1.0)	Burney, Fanny (14 - 2.0 - 0.079 - 1.0)
337	Ruskin, John (47 - 13.0 - 0.06 - 0.538)	Dostoyevsky, Fyodor (11 - 2.0 - 0.079 - 1.0)
338	Piper, H. Beam (33 - 6.0 - 0.06 - 1.0)	Reed, Helen Leah (10 - NA - 0.079 - NA)
339	De la Mare, Walter (10 - 3.0 - 0.06 - 0.333)	Hough, Emerson (25 - 4.0 - 0.08 - 0.25)
340	Reed, Talbot Baines (16 - 4.0 - 0.06 - 0.75)	Beaumont, Francis (10 - 2.0 - 0.08 - 0.5)
341	Pyle, Howard (16 - 1.0 - 0.06 - 0.0)	Roosevelt, Theodore (17 - 1.0 - 0.08 - 0.0)
342	Ebers, Georg (144 - 28.0 - 0.06 - 1.0)	Hawthorne, Julian (12 - 3.0 - 0.08 - 0.0)
343	Roberts, B. H. (Brigham Henry) (14 - 2.0 - 0.06 - 0.5)	Moodie, Susanna (14 - 2.0 - 0.081 - 0.5)
344	Thackeray, William Makepeace (35 - 7.0 - 0.061 - 0.571)	Pyle, Katharine (11 - 3.0 - 0.081 - 1.0)
345	Roe, Edward Payson (19 - 5.0 - 0.061 - 1.0)	Doyle, Arthur Conan (61 - 13.0 - 0.081 - 1.0)
346	Spence, Lewis (10 - 1.0 - 0.061 - 1.0)	Lee, Vernon (15 - 4.0 - 0.081 - 0.5)
347	Morris, Charles (18 - 4.0 - 0.062 - 1.0)	Willis, Nathaniel Parker (10 - 1.0 - 0.082 - 1.0)
348	Russell, Bertrand (11 - 3.0 - 0.062 - 0.667)	Wordsworth, William (14 - NA - 0.082 - NA)
349	Quiller-Couch, Arthur (40 - 7.0 - 0.062 - 0.571)	Adams, Samuel Hopkins (13 - 2.0 - 0.082 - 0.5)
350	Euripides (10 - 2.0 - 0.063 - 1.0)	Murray, David Christie (14 - 2.0 - 0.083 - 1.0)
351	Andersen, H. C. (Hans Christian) (14 - 3.0 - 0.063 - 0.667)	Franklin, Benjamin (10 - 3.0 - 0.083 - 0.333)
352	Schopenhauer, Arthur (12 - 1.0 - 0.063 - 1.0)	Burton, Richard Francis, Sir (20 - 6.0 - 0.083 - 0.833)
353	Froude, James Anthony (12 - 4.0 - 0.063 - 0.75)	Montaigne, Michel de (21 - 3.0 - 0.084 - 1.0)
354	Ralphson, G. Harvey (George Harvey) (14 - 5.0 - 0.064 - 0.8)	Dryden, John (20 - 5.0 - 0.084 - 0.8)
355	Jonson, Ben (12 - 1.0 - 0.064 - 0.0)	Moore, George Augustus (16 - 2.0 - 0.085 - 0.5)
356	Allen, James Lane (13 - 6.0 - 0.064 - 0.333)	Hewlett, Maurice (15 - 3.0 - 0.085 - 0.0)
357	Sabatini, Rafael (18 - 4.0 - 0.064 - 1.0)	Lincoln, Abraham (19 - 5.0 - 0.085 - 0.2)
358	Harte, Bret (57 - 12.0 - 0.065 - 0.75)	Zangwill, Israel (15 - 4.0 - 0.086 - 0.5)
359	Reid, Mayne (50 - 11.0 - 0.066 - 0.727)	Brinton, Daniel G. (Daniel Garrison) (18 - 3.0 - 0.086 - 0.667)
360	Jacobs, W. W. (William Wymark) (105 - 30.0 - 0.066 - 0.967)	Becke, Louis (39 - 8.0 - 0.086 - 0.75)
361	Le Gallienne, Richard (17 - 4.0 - 0.066 - 0.5)	Hall, E. Raymond (Eugene Raymond) (15 - 3.0 - 0.087 - 0.667)
362	Erckmann-Chatrian (10 - 5.0 - 0.066 - 0.4)	Beers, Henry A. (Henry Augustin) (10 - 2.0 - 0.087 - 0.0)
363	Cooper, James Fenimore (38 - 4.0 - 0.066 - 0.5)	Meynell, Alice (11 - 3.0 - 0.087 - 0.333)
364	Carlyle, Thomas (35 - 10.0 - 0.066 - 0.9)	Hakluyt, Richard (15 - 3.0 - 0.087 - 1.0)
365	Atkinson, William Walker (19 - 2.0 - 0.066 - 1.0)	White, Stewart Edward (23 - 8.0 - 0.088 - 0.875)
366	Frazer, James George (17 - 6.0 - 0.066 - 1.0)	MacDonald, George (60 - 12.0 - 0.088 - 0.75)
367	Melville, Herman (16 - 4.0 - 0.067 - 0.5)	Churchill, Winston (62 - 11.0 - 0.088 - 0.818)
368	Grinnell, George Bird (13 - 2.0 - 0.067 - 1.0)	Baum, L. Frank (Lyman Frank) (54 - 8.0 - 0.089 - 0.875)
369	Singmaster, Elsie (11 - 2.0 - 0.067 - 1.0)	Kingsley, Charles (45 - 4.0 - 0.089 - 1.0)
370	Richardson, Samuel (14 - 4.0 - 0.068 - 0.75)	Ellis, Edward Sylvester (52 - 10.0 - 0.089 - 0.9)
371	Sinclair, May (21 - 6.0 - 0.068 - 1.0)	Cobb, Irvin S. (Irvin Shrewsbury) (24 - 8.0 - 0.089 - 1.0)
372	Lytton, Edward Bulwer Lytton, Baron (194 - 43.0 - 0.068 - 0.93)	Caine, Hall, Sir (17 - 2.0 - 0.09 - 0.5)
373	Buchan, John (11 - 3.0 - 0.068 - 0.0)	Flaubert, Gustave (14 - 5.0 - 0.09 - 0.6)
374	Wallace, Alfred Russel (13 - 2.0 - 0.069 - 0.5)	Dawson, Coningsby (15 - 2.0 - 0.09 - 1.0)

	author (No. documents - test size - consistency - accuracy)	author (No. documents - test size - consistency - accuracy)
450	Hegel, Georg Wilhelm Friedrich (10 - 2.0 - 0.192 - 1.0)	Molière (20 - 4.0 - 0.244 - 0.75)
451	Daudet, Alphonse (17 - 3.0 - 0.193 - 0.333)	Fletcher, John (15 - 4.0 - 0.244 - 0.0)
452	Brinton, Daniel G. (Daniel Garrison) (18 - 3.0 - 0.193 - 0.667)	Lebert, Marie (15 - 4.0 - 0.247 - 0.75)
453	France, Anatole (31 - 3.0 - 0.194 - 0.667)	Schoolcraft, Henry Rowe (13 - 3.0 - 0.247 - 0.333)
454	Hakluyt, Richard (15 - 3.0 - 0.195 - 1.0)	Saltus, Edgar (13 - 4.0 - 0.249 - 0.25)
455	Duellman, William Edward (12 - 2.0 - 0.195 - 0.5)	Ballou, Maturin Murray (19 - 5.0 - 0.249 - 0.4)
456	Janifer, Laurence M. (12 - 2.0 - 0.196 - 1.0)	Page, Thomas Nelson (24 - 6.0 - 0.25 - 0.5)
457	Lincoln, Abraham (19 - 5.0 - 0.196 - 0.2)	Hall, E. Raymond (Eugene Raymond) (15 - 3.0 - 0.252 - 0.667)
458	Franklin, Benjamin (10 - 3.0 - 0.198 - 0.333)	Meredith, George (94 - 27.0 - 0.252 - 0.889)
459	Leacock, Stephen (14 - 2.0 - 0.199 - 0.0)	Moore, Thomas (12 - 2.0 - 0.255 - 0.0)
460	Guiney, Louise Imogen (13 - NA - 0.199 - NA)	Janvier, Thomas A. (Thomas Allibone) (13 - 3.0 - 0.257 - 0.333)
461	Jonson, Ben (12 - 1.0 - 0.199 - 0.0)	Potter, Beatrix (21 - 5.0 - 0.257 - 1.0)
462	Zola, Émile (37 - 11.0 - 0.199 - 0.818)	Wallace, Edgar (16 - 5.0 - 0.257 - 0.6)
463	Warner, Charles Dudley (41 - 10.0 - 0.2 - 0.3)	Boswell, James (12 - 3.0 - 0.258 - 0.667)
464	Cabell, James Branch (13 - 3.0 - 0.2 - 0.667)	Harris, Joel Chandler (14 - 1.0 - 0.258 - 0.0)
465	Burton, Richard Francis, Sir (20 - 6.0 - 0.2 - 0.833)	Young, Filson (11 - 3.0 - 0.26 - 0.333)
466	Dawson, Coningsby (15 - 2.0 - 0.201 - 1.0)	Grote, George (13 - 3.0 - 0.26 - 1.0)
467	Seton, Ernest Thompson (15 - 2.0 - 0.201 - 0.5)	Allen, Grant (29 - 4.0 - 0.263 - 0.25)
468	Reade, Charles (15 - 2.0 - 0.201 - 1.0)	Bone, Jesse F. (Jesse Franklin) (12 - 1.0 - 0.264 - 1.0)
469	Beaumont, Francis (10 - 2.0 - 0.202 - 0.5)	Harland, Marion (13 - 3.0 - 0.266 - 0.667)
470	Bierce, Ambrose (17 - 4.0 - 0.202 - 0.25)	Phillpotts, Eden (19 - 3.0 - 0.268 - 0.333)
471	Aldrich, Thomas Bailey (19 - 5.0 - 0.203 - 0.0)	James, Henry (75 - 10.0 - 0.269 - 1.0)
472	Rousseau, Jean-Jacques (18 - 5.0 - 0.203 - 0.4)	Wallace, Dillon (11 - 3.0 - 0.272 - 0.333)
473	Wharton, Edith (33 - 10.0 - 0.204 - 0.8)	Borrow, George (39 - 2.0 - 0.273 - 1.0)
474	Yonge, Charlotte M. (Charlotte Mary) (59 - 8.0 - 0.204 - 1.0)	Byron, George Gordon Byron, Baron (12 - 3.0 - 0.273 - 0.667)
475	Bunyan, John (14 - 2.0 - 0.205 - 0.0)	Mitchell, S. Weir (Silas Weir) (12 - 2.0 - 0.274 - 0.0)
476	Browning, Robert (10 - 1.0 - 0.205 - 1.0)	Mencken, H. L. (Henry Louis) (10 - 2.0 - 0.275 - 0.5)
477	Dryden, John (20 - 5.0 - 0.206 - 0.8)	Plato (27 - 3.0 - 0.275 - 1.0)
478	Hubbard, Elbert (20 - 3.0 - 0.206 - 1.0)	Weymouth, Richard Francis (25 - 4.0 - 0.275 - 1.0)
479	Hearn, Lafcadio (22 - 7.0 - 0.207 - 0.429)	Lewis, Alfred Henry (15 - 4.0 - 0.278 - 0.75)
480	Paine, Albert Bigelow (29 - 6.0 - 0.208 - 0.667)	Disraeli, Benjamin, Earl of Beaconsfield (17 - 1.0 - 0.279 - 0.0)
481	Roberts, Charles G. D., Sir (26 - 4.0 - 0.208 - 1.0)	Eggleston, Edward (12 - 1.0 - 0.28 - 0.0)
482	Baring-Gould, S. (Sabine) (57 - 9.0 - 0.208 - 0.667)	Baldwin, James (11 - 1.0 - 0.283 - 0.0)
483	Freeman, Mary Eleanor Wilkins (23 - 5.0 - 0.21 - 0.4)	Besant, Walter (19 - 3.0 - 0.283 - 0.333)
484	Twain, Mark (142 - 32.0 - 0.21 - 0.812)	Walpole, Horace (12 - 4.0 - 0.287 - 0.5)
485	Davis, Richard Harding (49 - 9.0 - 0.21 - 0.667)	Laut, Agnes C. (12 - 1.0 - 0.288 - 0.0)
486	Verne, Jules (46 - 13.0 - 0.212 - 0.923)	Stevenson, Robert Louis (70 - 14.0 - 0.29 - 0.857)
487	Leland, Charles Godfrey (10 - 2.0 - 0.212 - 0.0)	Carleton, William (21 - 4.0 - 0.29 - 1.0)
488	Dixon, Thomas (13 - 4.0 - 0.213 - 1.0)	Wells, H. G. (Herbert George) (51 - 12.0 - 0.293 - 0.75)
489	Besant, Annie (17 - 1.0 - 0.214 - 1.0)	Conrad, Joseph (31 - 3.0 - 0.295 - 1.0)
490	Hawthorne, Nathaniel (92 - 22.0 - 0.215 - 0.864)	Van Dyke, Henry (29 - 7.0 - 0.296 - 0.571)
491	Bangs, John Kendrick (37 - 8.0 - 0.216 - 0.875)	Herford, Oliver (13 - 3.0 - 0.296 - 0.0)
492	Coleridge, Samuel Taylor (18 - 2.0 - 0.217 - 1.0)	Herrick, Robert (11 - 3.0 - 0.297 - 0.0)
493	Voltaire (19 - 6.0 - 0.218 - 0.833)	Morris, William (28 - 12.0 - 0.298 - 0.25)
494	Maclaren, Ian (13 - 5.0 - 0.218 - 0.8)	Adams, Andy (10 - 2.0 - 0.301 - 1.0)
495	Dickens, Charles (79 - 15.0 - 0.218 - 0.667)	Marlowe, Christopher (10 - 3.0 - 0.305 - 0.667)
496	Wells, Carolyn (58 - 16.0 - 0.219 - 0.688)	Chesterton, G. K. (Gilbert Keith) (37 - 9.0 - 0.305 - 0.667)
497	Eggleston, George Cary (17 - 3.0 - 0.221 - 0.0)	Chekhov, Anton Pavlovich (23 - 5.0 - 0.306 - 0.8)
498	Hughes, Rupert (12 - 1.0 - 0.221 - 1.0)	London, Jack (50 - 11.0 - 0.306 - 0.818)
499	Nesbit, E. (Edith) (30 - 6.0 - 0.224 - 0.5)	Wilson, Harry Leon (13 - 1.0 - 0.306 - 0.0)
500	Lucas, E. V. (Edward Verrall) (11 - 4.0 - 0.224 - 0.25)	Wilcox, Ella Wheeler (23 - 8.0 - 0.306 - 0.5)
501	Hugo, Victor (15 - 6.0 - 0.224 - 0.833)	Shakespeare, William (105 - 19.0 - 0.307 - 0.842)
502	Field, Eugene (14 - 2.0 - 0.227 - 1.0)	Fielding, Henry (14 - 5.0 - 0.308 - 0.2)
503	Defoe, Daniel (44 - 11.0 - 0.23 - 0.545)	Phelps, Elizabeth Stuart (14 - 4.0 - 0.31 - 0.0)
504	Belloc, Hilaire (27 - 5.0 - 0.231 - 0.8)	‘Abdu’l-Bahá (15 - NA - 0.314 - NA)
505	Darwin, Charles (30 - 2.0 - 0.235 - 0.5)	Graham, Harry (10 - 2.0 - 0.319 - 0.5)
506	Drake, Samuel Adams (10 - 1.0 - 0.235 - 0.0)	Tagore, Rabindranath (19 - 4.0 - 0.322 - 0.0)
507	Cicero, Marcus Tullius (14 - 3.0 - 0.235 - 0.0)	Webster, Jean (10 - 1.0 - 0.325 - 1.0)
508	Newman, John Henry (14 - 1.0 - 0.236 - 1.0)	Masefield, John (17 - 2.0 - 0.327 - 1.0)
509	Balzac, Honoré de (119 - 18.0 - 0.236 - 0.833)	Longfellow, Henry Wadsworth (14 - 4.0 - 0.327 - 0.25)
510	Butler, Ellis Parker (22 - 3.0 - 0.236 - 0.333)	Otis, James (45 - 9.0 - 0.33 - 0.889)
511	Johnston, Mary (18 - 2.0 - 0.236 - 0.5)	Burroughs, Edgar Rice (19 - 1.0 - 0.335 - 1.0)
512	Leinster, Murray (37 - 9.0 - 0.236 - 0.778)	Haeckel, Ernst (13 - 6.0 - 0.337 - 0.333)
513	O’Donnell, Elliott (10 - 2.0 - 0.237 - 0.0)	Johnson, Samuel (23 - 6.0 - 0.337 - 0.5)
514	Wister, Owen (13 - 4.0 - 0.237 - 0.0)	Jewett, Sarah Orne (12 - 2.0 - 0.339 - 0.5)
515	McElroy, John (15 - NA - 0.238 - NA)	Luther, Martin (18 - 4.0 - 0.34 - 0.0)
516	United States. Work Projects Administration (34 - 6.0 - 0.239 - 1.0)	Homer (12 - 5.0 - 0.341 - 0.0)
517	La Fontaine, Jean de (31 - 6.0 - 0.242 - 0.5)	Warner, Anne (10 - 2.0 - 0.35 - 0.0)
518	Lang, Andrew (72 - 17.0 - 0.242 - 0.882)	Bennett, Arnold (44 - 16.0 - 0.35 - 0.875)
519	Brady, Cyrus Townsend (13 - 4.0 - 0.242 - 0.0)	Home, Gordon (15 - 5.0 - 0.351 - 0.4)
520	Burnett, Frances Hodgson (41 - 6.0 - 0.242 - 0.667)	Nietzsche, Friedrich Wilhelm (17 - 1.0 - 0.351 - 1.0)
521	Dumas, Alexandre (58 - 10.0 - 0.243 - 0.8)	Abbott, Jacob (51 - 11.0 - 0.359 - 0.727)
522	Gibbon, Edward (11 - 1.0 - 0.243 - 1.0)	Gibbs, George (15 - 3.0 - 0.364 - 1.0)
523	Duchess (16 - 1.0 - 0.244 - 1.0)	Baker, George M. (George Melville) (19 - 4.0 - 0.365 - 1.0)
524	Eddy, Mary Baker (10 - 2.0 - 0.244 - 0.5)	Rolland, Romain (12 - 3.0 - 0.365 - 0.333)

author (No. documents - test size - consistency - accuracy)

600 Jackson, Helen Hunt (13 - 2.0 - 0.369 - 0.0)
601 Crane, Walter (17 - 4.0 - 0.37 - 0.5)
602 Goethe, Johann Wolfgang von (15 - 2.0 - 0.371 - 0.5)
603 Bjørnson, Bjørnstjerne (16 - 3.0 - 0.372 - 1.0)
604 Carroll, Lewis (19 - 4.0 - 0.373 - 1.0)
605 Kipling, Rudyard (44 - 10.0 - 0.373 - 0.7)
606 Riley, James Whitcomb (17 - 2.0 - 0.377 - 0.5)
607 Jerome, Jerome K. (Jerome Klapka) (32 - 6.0 - 0.379 - 0.333)
608 Stevenson, Burton Egbert (17 - 4.0 - 0.382 - 0.25)
609 Webster, Noah (11 - 1.0 - 0.388 - 1.0)
610 Gorky, Maksim (10 - 1.0 - 0.393 - 0.0)
611 Peck, George W. (George Wilbur) (10 - 2.0 - 0.395 - 1.0)
612 Howells, William Dean (94 - 23.0 - 0.4 - 0.783)
613 Stringer, Arthur (10 - nan - 0.402 - nan)
614 Andreyev, Leonid (11 - nan - 0.403 - nan)
615 Xenophon (16 - 3.0 - 0.405 - 0.667)
616 Swinburne, Algernon Charles (25 - 2.0 - 0.406 - 0.5)
617 Yeats, W. B. (William Butler) (35 - 5.0 - 0.414 - 0.4)
618 Ibsen, Henrik (18 - 4.0 - 0.415 - 0.25)
619 Montgomery, L. M. (Lucy Maud) (12 - 2.0 - 0.416 - 1.0)
620 Library of Congress. Copyright Office (66 - 13.0 - 0.425 - 0.923)
621 Wilson, Ann (12 - 2.0 - 0.426 - 1.0)
622 Morley, Christopher (12 - 2.0 - 0.428 - 1.0)
623 Galsworthy, John (47 - 9.0 - 0.437 - 1.0)
624 Tennyson, Alfred Tennyson, Baron (12 - 2.0 - 0.44 - 0.5)
625 Shoghi, Effendi (17 - 5.0 - 0.445 - 1.0)
626 Reed, Myrtle (13 - 3.0 - 0.451 - 0.333)
627 Holmes, Oliver Wendell (33 - 10.0 - 0.462 - 0.6)
628 Lawrence, D. H. (David Herbert) (20 - 6.0 - 0.473 - 0.667)
629 Shaw, Bernard (42 - 8.0 - 0.481 - 0.75)
630 Anstey, F. (18 - 2.0 - 0.493 - 1.0)
631 Strindberg, August (22 - 4.0 - 0.505 - 0.75)
632 Bahá'u'lláh (11 - 5.0 - 0.508 - 0.2)
633 Burgess, Gelett (11 - 2.0 - 0.515 - 0.0)
634 Tolstoy, Leo, graf (38 - 8.0 - 0.521 - 0.75)
635 Bridges, Robert (11 - 2.0 - 0.523 - 0.5)
636 Spinoza, Benedictus de (12 - 3.0 - 0.525 - 0.333)
637 Wilde, Oscar (25 - 2.0 - 0.536 - 0.5)
638 Poe, Edgar Allan (16 - 2.0 - 0.541 - 0.5)
639 Rice, Cale Young (11 - 3.0 - 0.544 - 0.333)
640 Barrie, J. M. (James Matthew) (25 - 1.0 - 0.55 - 1.0)
641 Dunsany, Lord (16 - 5.0 - 0.575 - 0.2)
642 Maeterlinck, Maurice (18 - 2.0 - 0.61 - 1.0)
643 Sinclair, Upton (24 - 10.0 - 0.653 - 0.5)
644 Schiller, Friedrich (32 - 7.0 - 0.683 - 0.429)
645 Wagner, Richard (11 - 1.0 - 0.702 - 0.0)
646 Aesop (22 - 4.0 - 0.714 - 0.5)
647 Sudermann, Hermann (14 - 1.0 - 0.715 - 0.0)
648 Milne, A. A. (Alan Alexander) (11 - 2.0 - 0.733 - 0.5)
649 Maugham, W. Somerset (William Somerset) (26 - 5.0 - 0.853 - 0.6)
650 Honig, Winfried (11 - 2.0 - 1.081 - 0.0)

Table A 2. The genres that we use in our study.

	genre (No. documents - test size - consistency - accuracy)	genre (No. documents - test size - consistency - accuracy)
0	World War II (11 - 6 - 0.08 - 0.667)	World War I (57 - 17 - 0.294 - 0.235)
1	Crime Fiction (27 - 7 - 0.125 - 0.857)	Art (14 - 1 - 0.299 - 1.0)
2	Historical Fiction (263 - 48 - 0.152 - 0.833)	Animal (16 - 2 - 0.313 - 1.0)
3	Western (76 - 18 - 0.153 - 0.611)	Children's Literature (158 - 26 - 0.33 - 0.462)
4	Horror (16 - 2 - 0.159 - 0.0)	Classical Antiquity (13 - 3 - 0.344 - 0.0)
5	Children's Book Series (354 - 75 - 0.176 - 0.853)	US Civil War (78 - 13 - 0.345 - 0.462)
6	Adventure (37 - 9 - 0.179 - 0.333)	Christmas (44 - 8 - 0.37 - 0.125)
7	Children's Fiction (269 - 58 - 0.188 - 0.879)	Fantasy (48 - 8 - 0.375 - 0.375)
8	Crime Nonfiction (20 - 5 - 0.188 - 0.0)	Poetry (22 - 4 - 0.391 - 0.0)
9	Science Fiction (447 - 87 - 0.211 - 0.851)	Travel (16 - 2 - 0.396 - 0.0)
10	Movie Books (37 - 6 - 0.221 - 0.0)	Children's Picture Books (35 - 8 - 0.424 - 0.5)
11	Biology (15 - 3 - 0.224 - 0.667)	Children's Instructional Books (12 - 2 - 0.44 - 0.0)
12	Children's History (23 - 4 - 0.224 - 0.25)	Harvard Classics (40 - 10 - 0.474 - 0.0)
13	Humor (82 - 19 - 0.225 - 0.474)	Best Books Ever Listings (54 - 10 - 0.514 - 0.3)
14	Precursors of Science Fiction (12 - 1 - 0.264 - 0.0)	One Act Plays (28 - 7 - 0.516 - 0.571)
15	School Stories (33 - 5 - 0.269 - 0.2)	Philosophy (56 - 9 - 0.541 - 0.778)

References

[1] E. G. Altmann, L. Dias, and M. Gerlach, *Generalized entropies and the similarity of texts*, Journal of Statistical Mechanics: Theory and Experiment, 1 (2017), 014002.

[2] R. Arun, V. Suresh, and C. E. V. Madhavan, *Stopword graphs and authorship attribution in text corpora*, in Proceedings of the IEEE International Conference on Semantic Computing, 2009, pp. 192–196.

[3] A. J. Calhoun, *Punctuation code*, 2016. Available at `https://github.com/adamjcalhoun/punctuation`.

[4] ———, *Punctuation in novels*, 2016. Available at `https://medium.com/~@neuroecology/punctuation-in-novels-8f316d542ec4#.brev0b3w1`.

[5] ———, *What does punctuation tell us about Republicans and Democrats?*, 2016. Available at `https://medium.com/@neuroecology/what-does-punctuation-tell-us-about-republicans-and-democrats-bd46b9f98220`.

[6] F. Can and J. M. Patton, *Change of writing style with time*, Computers and the Humanites, 38 (2004), pp. 61–82.

[7] C. E. Chaski, *Empirical evaluation of language-based author identification techniques*, Forensic Linguistics, 8 (2001), pp. 1–65.

[8] I. Chevyreva and A. Kormilitzin, *A primer on the signature method in machine learning*, arXiv:1603.03788, (2016).

[9] H. Chiang, Y. Ge, and C. Wu, *Classification of book genres by cover and title*, 2015. Class report, Computer Science 229, Stanford University, available at `http://cs229.stanford.edu/proj2015/127_report.pdf`.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York City, NY, USA, 1991.

[11] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley and Sons, New York City, NY, USA, 2001.

[12] R. S. Forsyth, *Stylochronometry with substrings, or: A poet young and old*, Literary and Linguistic Computing, 14 (1999), pp. 467–477.

[13] H. W. Fowler and F. G. Fowler, *The King's English*, Oxford University Press, Oxford, UK, 1906.

[14] M. Gerlach and F. Font-Clos, *A standardized Project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics*, arXiv:1812.08092, (2018).

[15] M. Gerlach, F. Font-Clos, and E. G. Altmann, *Similarity of symbol frequency distributions with heavy tails*, Physical Review X, 6 (2016), 021009.

[16] J. Grieve, *Quantitative authorship attribution: An evaluation of techniques*, Literary and Linguistic Computing, 22 (2007), pp. 251–270.

[17] M. S. Hart, *Project Gutenberg*, 1971. Available at `https://www.gutenberg.org`.

[18] C. O. Hartman, *Verse: An Introduction to Prosody*, Wiley Blackwell, Hoboken, NJ, USA, 2015.

[19] D. I. Holmes, *The evolution of stylometry in humanities scholarship*, Literary and Linguistic Computing, 50 (1998), pp. 111–117.

[20] M. Honnibal, spaCy, 2017. Available at `https://spacy.io`.

[21] J. M. Hughes, N. J. Foti, D. C. Krakauer, and D. N. Rockmore, *Quantitative patterns of stylistic influence in the evolution of literature*, Proceedings of the National Academy of Sciences of the United States of America, 109 (2012), pp. 7682–7686.

[22] M. P. Jackson, *Pause patterns in Shakespeare's verse: Canon and chronology*, Literary and Linguistic Computing, 17 (2002), pp. 37–46.

[23] B. Kessler, G. Nunberg, and H. Schutze, *Automatic detection of text genre*, in Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, 1996.

[24] B. Kjell, *Authorship attribution of text samples using neural networks and bayesian classi-*

fiers, in Proceedings of IEEE International Conference on Systems, Man and Cybernetics, vol. 2, Oct 1994, pp. 1660–1664.

[25] S. Kullback and R. A. Leibler, *On information and sufficiency*, The Annals of Mathematical Statistics, 22 (1951), pp. 79–86.

[26] S. Lai, L. Xu, K. Liu, and J. Zhao, *Recurrent convolutional neural networks for text classification*, in Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI'15, AAAI Press, 2015, pp. 2267–2273.

[27] J. Lawler, *Punctuation*, The International Encyclopedia of Language and Linguistics, (2006). Available at `https://doi.org/10.1016/B0-08-044854-2/04573-9`.

[28] A. Lesne, *Mathematical structures in computer science*, Shannon entropy: A rigorous notion at the crossroads between probability, information theory, dynamical systems and statistical physics, 24 (2014), e240311.

[29] T. Lewis, *'Notes on Punctuation', in* The Medusa and the Snail: More Notes of a Biology Watcher, Viking Press, New York City, NY, USA, 1979.

[30] W. Ebeling and T. Pöschel, *Entropy and long-range correlations in literary english*, Europhysics Letters, 26 (1994), pp. 241–246.

[31] J. Lin, *Divergence measures based on the Shannon entropy*, IEEE Transactions on Information Theory, 37 (1991), pp. 145–151.

[32] T. Lyons, *Rough paths, signatures and the modelling of functions on streams*, Proceedings of the International Congress of Mathematicians 2014, Korea, (2014). Available at `http://www.icm2014.org/download/Proceedings_Volume_IV.pdf`.

[33] T. C. Mendenhall, *The characteristic curves of composition*, Science, 9 (1887), pp. 237–249.

[34] F. Mosteller and D. L. Wallace, *Inference and Disputed Authorship: The Federalist*, Addison-Wesley, Reading, MA, USA, 1964.

[35] T. Neal, K. Sundararajan, A. Fatima, and D. Woodard, *Surveying stylometry techniques and applications*, ACM Computing Surveys, 50 (2018), 86.

[36] L. Neidorf, M. S. Krieger, M. Yakubek, P. Chaudhuri, and J. P. Dexter, *Large-scale quantitative profiling of the Old English verse tradition*, Nature Human Behaviour, 3 (2019), pp. 560–567.

[37] G. Nunberg, *The Linguistics of Punctuation*, Center for the Study of Language and Information, Stanford, CA, USA, 1990.

[38] M. B. Parkes, ed., *Pause and Effect: An Introduction to the History of Punctuation in the West*, University of California Press, Berkeley, CA, USA, 1992.

[39] G. Pullum and R. Huddleston, *The Cambridge Grammar of the English Language*, Cambridge University Press, The Other Place, UK, 2001.

[40] C. Qian, T. He, and R. Zhang, *Deep learning based authorship identification* (2017). Class report, Computer Science 224, Stanford University, available at `https://pdfs.semanticscholar.org/ab0e/be094ec0a44fb0013d640b344d8cfd7adc81.pdf?_ga=2.215953495.1190289256.1578845031-6826891.1578845031`.

[41] M. Santini, *A shallow approach to syntactic feature extraction for genre classification*, in Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics, 2004.

[42] ———, *State-of-the-art on automatic genre identification*, Information Technology Research Institute (ITRI) Technical Report Series 04-03, University of Brighton, UK, (2004). Available at `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.5.7680`.

[43] C. E. Shannon, *A mathematical theory of communication*, The Bell System Technical Journal, (1948), pp. 379–423, 623–656.

[44] J. Shlens, *Notes on Kullback–Leibler divergence and likelihood theory*, arXiv:1404.2000, (2014).

[45] E. Stamatatos, *A survey of modern authorship attribution methods*, Journal of the American Society for Information Science and Technology, 60 (2009), pp. 538–556.

[46] C. Stamou, *Stylochronometry: Stylistic development, sequence of composition, and relative dating*, Literary and Linguistic Computing, 23 (2008), pp. 181–199.

[47] L. Truss, *Eats, Shoots and Leaves: The Zero Tolerance Approach to Punctuation*, Profile Books, London, UK, 2004.

[48] D. S. Vieira, S. Picoli, and R. S. Mendes, *Robustness of sentence length measures in written texts*, Physica A, 506 (2018), pp. 749–754.

[49] C. Watson, *Semicolon: The Past, Present, and Future of a Misunderstood Mark*, Ecco Press, New York, NY, USA, 2019.

[50] C. Whissell, *Traditional and emotional stylometric analysis of the songs of Beatles Paul McCartney and John Lennon*, Computers and the Humanities, 30 (1996), pp. 257–265.

[51] A. C.-C. Yang, C.-K. Peng, H.-W. Yien, and A. Goldberger, *Information categorization approach to literary authorship disputes*, Physica A, 329 (2003), pp. 473–483.

[52] Y. Zhao, J. Zobel, and P. Vines, *Using relative entropy for authorship attribution*, in Proceedings of the Third Asia Conference on Information Retrieval Technology (AIRS '06), 2006, pp. 92–105.