




ARTICLE



<https://doi.org/10.1057/s41599-024-02933-6>

OPEN

# Computational thematics: comparing algorithms for clustering the genres of literary fiction

Oleg Sobchuk<sup>1</sup>✉ & Artjoms Šeļa<sup>2</sup> 

What are the best methods of capturing thematic similarity between literary texts? Knowing the answer to this question would be useful for automatic clustering of book genres, or any other thematic grouping. This paper compares a variety of algorithms for unsupervised learning of thematic similarities between texts, which we call “computational thematics”. These algorithms belong to three steps of analysis: text pre-processing, extraction of text features, and measuring distances between the lists of features. Each of these steps includes a variety of options. We test all the possible combinations of these options. Every combination of algorithms is given a task to cluster a corpus of books belonging to four pre-tagged genres of fiction. This clustering is then validated against the “ground truth” genre labels. Such comparison of algorithms allows us to learn the best and the worst combinations for computational thematic analysis. To illustrate the difference between the best and the worst methods, we then cluster 5000 random novels from the HathiTrust corpus of fiction.

<sup>1</sup>Department of Human Behavior, Ecology and Culture, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany. <sup>2</sup>Institute of Polish Language, Polish Academy of Sciences, Kraków, Poland. ✉email: [oleg\\_sobchuk@eva.mpg.de](mailto:oleg_sobchuk@eva.mpg.de)

## Introduction

Computational literary studies have rapidly grown in prominence over the recent years. One of the most successful directions of inquiry within this domain, in terms of both methodological advances and empirical findings, has been computational stylometry, or computational stylistics: a discipline that develops algorithmic techniques for learning *stylistic similarities* between texts (Bories et al., 2023; Burrows, 1987; Eder et al., 2016). For this purpose, computational stylometrists extract linguistic features specifically associated with authorial style or individual authorial habits. Often, these features are the most frequent words from the analyzed literary texts—they tend to be function words (“a”, “the”, “on”, etc.)—to which various measures of similarity (e.g., Euclidean distance) are applied. The most common goal of computational stylistics is attributing the authorship of texts where it is disputed like the authorship of Molière’s plays (Cafiero and Camps, 2019), the Nobel Prize-winning novel *And Quiet Flows the Don* (Iosifyan and Vlasov, 2020), or Shakespeare and Fletcher’s play *Henry VIII* (Plecháč, 2021). Thanks to numerous systematic comparisons of various approaches to computational stylometry, we now have a fairly good idea of which procedures and textual features are the most effective ones—depending on the goal of stylometric analysis, the language of texts, or their genre (Evert et al., 2017; Neal et al., 2017; Plecháč et al., 2018).

At the same time, we lack such systematic comparisons in the research area that might be called “computational thematics”: the study of *thematic similarities* between texts. (Thematic similarities: say, that novels A and B both tell a love story or have a “fantasy” setting.) Why is learning about thematic similarities important? Genre—a group of texts united by broad thematic similarities (e.g., fantasy, romance, or science fiction) is a central concept in literary studies, necessary not only for categorizing and cataloging books but also for the historical scholarship of literature. The definition of genres used in this study is more specific: following a long tradition of evolutionary thinking in the humanities (Fowler, 1971; Moretti, 2005), we understand genres as evolving populations of texts that emerge at certain moments of time, spread across the field of literary production, and then disappear in their original form—usually becoming stepping stones for subsequent genres (Fowler, 1971). For example, the genre of “classical” detective fiction crystallized in the 1890–1930s and then gave birth to multiple other genres of crime fiction, such as “hardboiled crime fiction”, “police procedural”, “historical detective”, and others (Symons, 1985). This understanding of cultural phenomena as populations, consisting of items (in our case, books) similar to each other due to common descent (that is, shared influences) is also common in the research on cultural evolution (Baraghith, 2020; Houkes, 2012).

Studying the historical dynamics of genres—not only in literature, but also in music or visual arts—is an important task of art history and sociology, and digital archives allow doing so on a much larger scale (Allison et al., 2011; Klimek et al., 2019; Sigaki et al., 2018). But to gain the most from this larger scale, we must identify the best, most reliable algorithms for detecting the thematic signal in books—similarly to how computational stylometrists have learned the most effective algorithms for detecting the signal of authorship.

Quantitative analysis of genres usually takes one of these forms. The first one is the *manual tagging* of books by genre—often, through large-scale crowdsourced efforts, like the Goodreads website (Thelwall, 2019). This approach is prone to human bias, it is laborious and also based on the idea that the differences between genre populations are qualitative, not quantitative (e.g., a certain book is either a “detective” or “romance”, or both, but not 0.78 detectives and 0.22 romance, which, we think, would be a more informative description). The second approach is an extension of

manual tagging: *supervised machine learning* of book genres using a training dataset with manually tagged genres (Piper et al., 2021; Underwood, 2019). This approach has important strengths: it is easily scalable and it provides not qualitative but quantitative estimates of a book’s belongingness to a genre. Still, it has a problem: it can only assign genre tags included in the training dataset, and it cannot find new, unexpected book populations—which is an important component of the historical study of literature. The third approach is *unsupervised clustering* of genres: algorithmic detection of book populations based on their similarity to each other (Calvo Tello, 2021; Schöch, 2017). This approach is easily scalable, allows quantitative characterization of book genres, and does not require a training dataset with manually assigned tags, thus allowing to detection of previously unknown book populations. All these features of unsupervised clustering make it highly suitable for historical research, and this is why we will focus on it in this paper.

Unsupervised clustering can be conducted in a variety of ways. For example, texts can be lemmatized or not lemmatized; as text features, simple word frequencies can be used or some higher-level units, such as topics of a topic model; to measure the similarity between texts, a host of distance metrics can be applied. Hence, the question is: what are the best computational methods for detecting thematic similarities in literary texts? This is the main question of this paper. To answer it, we will compare various combinations of (1) pre-processing (which, in this study, we will also call “thematic foregrounding”), (2) text features, and (3) the metrics used for measuring the distance between features. To assess the effectiveness of these combinations, we use a tightly controlled corpus of four well-known genres—detective fiction, science fiction, fantasy, and romance—as our “ground truth” dataset. To illustrate the significant difference between the best and the worst combinations of algorithms for genre detection, we later cluster genres in a much larger corpus, containing 5000 works of fiction.

## Materials and methods

**Data: The “ground truth” genres.** Systematic research on computational stylistics is common, while research on computational thematics is still rare (Allison et al., 2011; Schöch, 2017; Šeĵa et al., 2022; Underwood, 2016; Wilkens, 2016). Why? Computational stylistics has clear “ground truth” data against which various methods of text analysis can be compared: authorship. The methods of text analysis in computational stylistics (e.g., Delta distance or Cosine distance) can be compared as to how well they perform in the task of classifying texts by their authorship. We write “ground truth” in quotes, as authorship is no more than a convenient proxy for stylistic similarity, and, as any proxy, it is imprecise. It assumes that texts written by the same author should be more similar to each other than texts written by different authors. However, we know many cases when the writing style of an author would evolve significantly over the span of their career or would be deliberately manipulated (Brennan et al., 2012). Authorship as a proxy for “ground truth” is a simplification—but a very useful one.

The lack of a widely accepted “ground truth” proxy for thematic analysis leads to the comparisons of algorithms that are based on nothing more than subjective judgment (Egger and Yu, 2022). Such subjective judgment cannot lead us far: we need quantitative metrics of the performance of different algorithms. For this, an imperfect “ground truth” is better than none at all. What could become such an imperfect, but still useful, ground truth in computational thematics? At the moment, these are genre categories. They capture, to a different degree, the thematic similarities between texts. To a different degree, as genres can be organized according to several principles, or “axes of categorization”: e.g., they can be based on the similarity of storylines

**Table 1 Examples of books in each genre corpus (full list in Supplementary materials).**

Genre	Examples
Detective fiction 🕵️	Josephine Tey, <i>The Daughter of Time</i> , 1951 Agatha Christie, <i>At Bertram's Hotel</i> , 1965 Colin Dexter, <i>Last Bus to Woodstock</i> , 1975 Peter Lovesey, <i>The False Inspector Dew</i> , 1982 Sue Grafton, <i>M is for Malice</i> , 1996
Fantasy fiction 🧚	J. R. R. Tolkien, <i>The Fellowship of the Ring</i> , 1954 Michael Moorcock, <i>Stormbringer</i> , 1965 Ursula K. Le Guin, <i>The Tombs of Atuan</i> , 1970 Terry Pratchett, <i>The Color of Magic</i> , 1983 J. K. Rowling, <i>Harry Potter and the Philosopher's Stone</i> , 1997
Romance fiction ❤️	Barbara Cartland, <i>Love is the Enemy</i> , 1952 Jackie Collins, <i>The World is Full of Married Men</i> , 1968 Gordon Merrick, <i>The Lord Won't Mind</i> , 1970 Danielle Steel, <i>A Perfect Stranger</i> , 1981 Diana Gabaldon, <i>Outlander</i> , 1991
Science fiction 🚀	Robert A. Heinlein, <i>Double Star</i> , 1956 Arthur C. Clarke, <i>2001: A Space Odyssey</i> , 1968 Frank Herbert, <i>Children of Dune</i> , 1976 C. J. Cherryh, <i>Downbelow Station</i> , 1981 Kim Stanley Robinson, <i>Red Mars</i> , 1992

(adventure novel, crime novel, etc.), settings (historical novel, dystopian novel, etc.), emotions they evoke in readers (horror novel, humorous novel, etc.), or their target audience (e.g., young adult novels). It does seem that these various “axes of categorization” correlate: say, “young adult” novels are appreciated by young adults because they often have particular topics or storylines. Or, horror novels usually have a broad, but consistent, arsenal of themes and settings that are efficient at evoking pleasant fear in readers (like the classical haunted house). Still, some axes of genre categorization are probably better for comparing the methods of computational thematics than others. Genres defined by their plots or settings may provide a clearer thematic signal than genres defined by their target audience or evoked emotions.

We have assembled a tightly controlled corpus of four genres (50 texts in each) based on their plots and settings:

- Detective fiction (recurrent themes: murder, detective, suspects, investigation).
- Fantasy fiction (recurrent themes: magic, imaginary creatures, quasi-medieval setting).
- Romance fiction (recurrent themes: affection, erotic scenes, love triangle plot).
- Science fiction (recurrent themes: space, future, technology).

Our goal was to include books that would be rather uncontroversial members of their genres. Thus, we picked canonical representatives of each genre, winners of genre-based prizes (e.g., Hugo and Nebula awards for science fiction, the Gold Dagger award for detective fiction), and books with the largest numbers of ratings in respective genres on the Goodreads website ([www.goodreads.com](http://www.goodreads.com)). We took several precautions to remove potential confounds. First, these genres are situated on a similar level of abstraction: we are not comparing rough-grain categories (say, romance or science fiction) to fine-grain ones (historical romance or cyberpunk science fiction). Second, we limited the time span of the book publication year to a rather short period of 1950–1999: to make sure that our analysis is not affected too much by language change (which would inevitably happen if we compared, for example, 19th-century gothic novels to 20th-

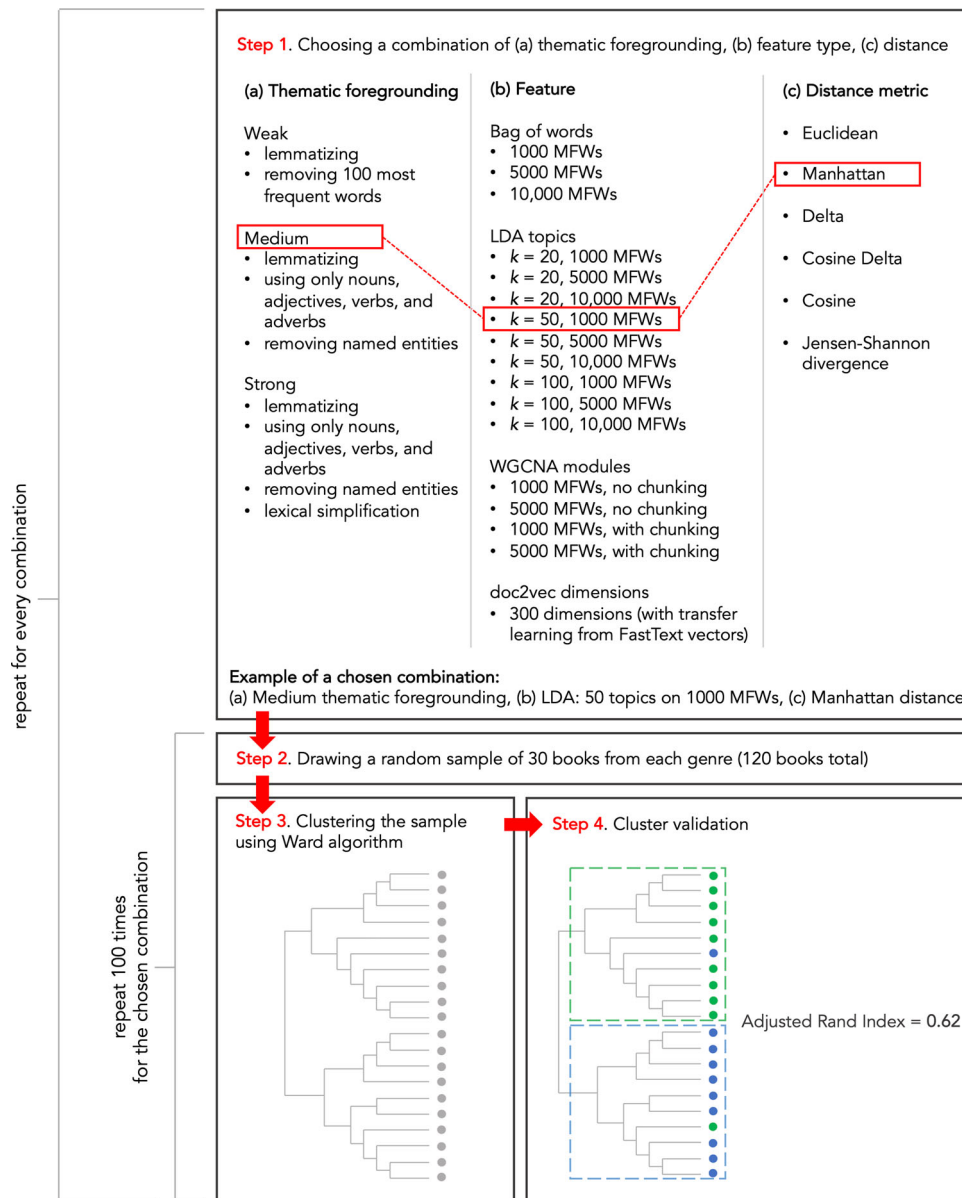
century science fiction). Third, each genre corpus has a similar number of authors (29–31 authors), each represented by 1–3 texts. Several examples of books in each genre are shown in Table 1. The complete list of books is in Supplementary Materials. Before starting our analysis, we pre-registered this list on Open Science Framework's website (<https://osf.io/rce2w>).

**Analysis: The race of algorithms.** To compare the methods of detecting thematic signals, we developed a workflow consisting of four steps—see Fig. 1. Same as our corpus, all the detailed steps of the workflow were pre-registered.

*Step 1. Choosing a combination of thematic foregrounding, features, and distance.* As a first step, we choose a combination of (a) the level of thematic foregrounding, (b) the features of analysis, and (c) the measure of distance.

By *thematic foregrounding* (Step 1a on Fig. 1) we mean the extent to which the thematic aspects of a text are highlighted (and the stylistic aspects—backdropped). With *weak* thematic foregrounding, only the most basic text pre-processing is done: lemmatizing words and removing 100 most frequent words (MFWs)—the most obvious carriers of strong stylistic signal. 100 MFWs roughly correspond to function words (or closed-class words) in English, routinely used in authorship attribution (Chung and Pennebaker, 2007; Stamatatos, 2009) beginning with the classical study of *Federalist Papers* (Mosteller and Wallace, 1963). With *medium* thematic foregrounding, in addition to lemmatizing, we also remove entities (named entities, proper names, etc.) using SpaCy tagger (<https://spacy.io/>). Additionally, we perform part-of-speech tagging and remove all the words that are not nouns, verbs, adjectives, or adverbs, which are the most content-bearing parts of speech. With *strong* thematic foregrounding, in addition to all the steps of the medium foregrounding, we also apply lexical simplification. We simplify the vocabulary by replacing less frequent words with their more frequent synonyms—namely, we replace all words outside of 1000 MFWs with their more common semantic neighbors (out of 10 closest neighbors), with the help of pre-trained FastText model that includes 2 million words and is trained on English Wikipedia (Grave et al., 2018).

Then, we transform our pre-processed texts into lists of features (Step 1b on Fig. 1). We vary both the type of features and the length of lists. We consider four types of features. The simplest features are the most frequent words as used in the bag-of-words approach (1000, 5000, or 10,000 of them)—a common choice for thematic analysis in computational literary studies (Hughes et al., 2012; Underwood, 2019). The second type of feature is topic probabilities generated with the Latent Dirichlet Allocation (LDA) algorithm (Blei et al., 2003)—another common choice (Jockers, 2013; Liu et al., 2021). LDA has several parameters that can influence results, such as the predefined  $k$  of topics or the number of most frequent words used. Plus, a long text like a novel is too large for meaningful LDA topic modeling, and the typical solution is dividing the text into smaller chunks. We use an arbitrary chunk size of 1000 words. The third type of feature is modules generated with weighted correlation network analysis, also known as weighted gene co-expression network analysis (WGCNA)—a method of dimensionality reduction that detects clusters (or “modules”) in networks (Langfelder and Horvath, 2008). WGCNA is widely used in genetics (Bailey et al., 2016; Ramírez-González et al., 2018), but also showed promising results as a tool for topic modeling of fiction (Elliott, 2017). We used it with either 1000 or 5000 most frequent words. Typically, WGCNA is used without chunking data, but, since chunking leads to better results in LDA, we decided to try using WGCNA with and without chunking, with the chunk size of 1000 words. All the parameters of WGCNA were kept at defaults. Finally,



**Fig. 1 Four steps of the analysis.** The workflow includes two loops. Big loop goes through various combinations of thematic foregrounding (Step 1a), feature type (1b), and distance metric (1c). For each such combination, a smaller loop is run: it randomly draws a genre-stratified sample of 120 novels (Step 2), clusters the novels using the Ward algorithm (Step 3), and validates the clusters on the dendrogram using the Adjusted Rand Index (Step 4). As a result, each combination receives an ARI score: a score of its performance in detecting genres.

as the fourth type of feature, we use document-level embeddings doc2vec (Lau and Baldwin, 2016; Le and Mikolov, 2014) that directly position documents in a latent semantic space defined by a pre-trained distributional language model—FastText (Grave et al., 2018). Document representations in doc2vec depend on the features of the underlying model: in our study, each document is embedded in 300 dimensions of the original model. Doc2vec and similar word embedding methods are increasingly used for assessing the similarity of documents (Dynomant et al., 2019; Kim et al., 2019; Pranjic et al., 2020). As a result of Step 1b, we obtain a document-term matrix formed of texts (rows) and features (columns).

Finally, we must learn the similarity between the texts represented with the chosen lists of features—by using some metric of distance (Step 1c on Fig. 1). There exist a variety of metrics for this purpose: Euclidean, Manhattan, Delta, Cosine, Cosine Delta distances and Jensen–Shannon divergence (symmetrized Kullback–Leibler divergence) for features that are

probability distributions (in our case, this can be done for LDA topics and bag-of-words features).

Variants of Step 1a, 1b, and 1c, can be assembled in numerous combinations. In our “race of algorithms”, each combination is a competitor—and a potential winner. Say, we could choose a combination of weak thematic foregrounding, LDA topics with 50 topics on 5000 most frequent words, and Euclidean distance. Or, medium thematic foregrounding, simple bag-of-words with 10,000 most frequent words, and Jensen–Shannon divergence. Some of these combinations are researchers’ favorites, while others are underdogs—used rarely, or not at all. Our goal is to map out the space of possible combinations—to empirically test how each combination performs in the task of detecting the thematic signal. In total, there are 291 competing combinations.

*Step 2. Sampling for robust results.* A potential problem with our experiment is that some combinations might perform better or



worse simply because they are more suitable to our specific corpus—for whatever reason. To reduce the impact of individual novels in our corpus, we use cross-validation: instead of analyzing the corpus as a whole, we analyze smaller *samples* from it multiple times. Each sample contains 120 novels: 30 books from each genre. Altogether, we perform the analysis for each combination on 100 samples. For each sample, all the models that require training—LDA, WGCNA, and doc2vec—are trained anew.

**Step 3. Clustering.** As a result of Step 2, we obtain a matrix of text distances. Then, we need to cluster the texts into groups—our automatically generated genre clusters, which we will later compare to the “true” clusters. For this, we could have used a variety of algorithms (e.g., *k*-means). We use hierarchical clustering with Ward’s linkage (Ward, 1963): it clusters two items when the resulting clusters maximize variance across the distance matrix. Despite being originally defined only for Euclidean distances, it was empirically shown that Ward’s algorithm outperforms other linkage strategies in text-clustering tasks (Ochab et al., 2019). We assume that novels from four defined genres should roughly form four distinct clusters (as the similarity of texts within the genre is greater than the similarity of texts across genres). To obtain the groupings from a resulting tree we cut it vertically by the number of assumed clusters (which is 4).

**Step 4. Cluster validation.** How similar are our generated clusters to the “true” genre populations? To learn this, we compare the clusters generated by each chosen combination to the original genre labels. For this, we use a measure of cluster validation called the adjusted Rand index (ARI) (Hubert and Arabie, 1985). The ARI score of a particular combination shows how well this combination performs in the task of detecting genres—and thus, in picking the thematic signal. Steps 1–4 are performed for every combination so that every combination receives its ARI score. At the end of the analysis, we obtained a dataset of 29,100 rows (291 combinations, each tested on 100 random samples).

## Results

Figure 2 shows the average performance of all the combinations of thematic foregrounding, features, and distance metrics. Our first observation: the average ARI of the best-performing algorithms ranges from 0.66 to 0.7, which is rather high for the complicated, noisy data that is literary fiction. This gives additional support to the idea that unsupervised clustering of fiction genres is possible. Even a cursory look at the 10 best-performing combinations immediately reveals several trends. First, none of the top combinations have weak thematic foregrounding. Second, 6 out of 10 best-performing features are LDA topics. Third, 8 out of 10 distances on this list are Jensen–Shannon divergence.

But how generalizable are these initial observations? How shall we learn the average “goodness” of a particular kind of thematic foregrounding, feature type, or distance metric? To learn this, we need to control for their influence on each other, as well as for additional parameters, such as the number of most frequent words and chunking. Hence, we have constructed five Bayesian linear regression models (see Supplement 5.1). They answer questions about the performance of various combinations of thematic foregrounding, features, and distances, helping us reach conclusions about the performance of individual steps of thematic analysis. All the results of this study are described in detail in Supplement 5.1. Below, we focus only on key findings.

**Conclusion 1. Thematic foregrounding improves genre clustering.** The goal of thematic foregrounding was to highlight the contentful parts of the texts and to backdrop the stylistic parts. So,

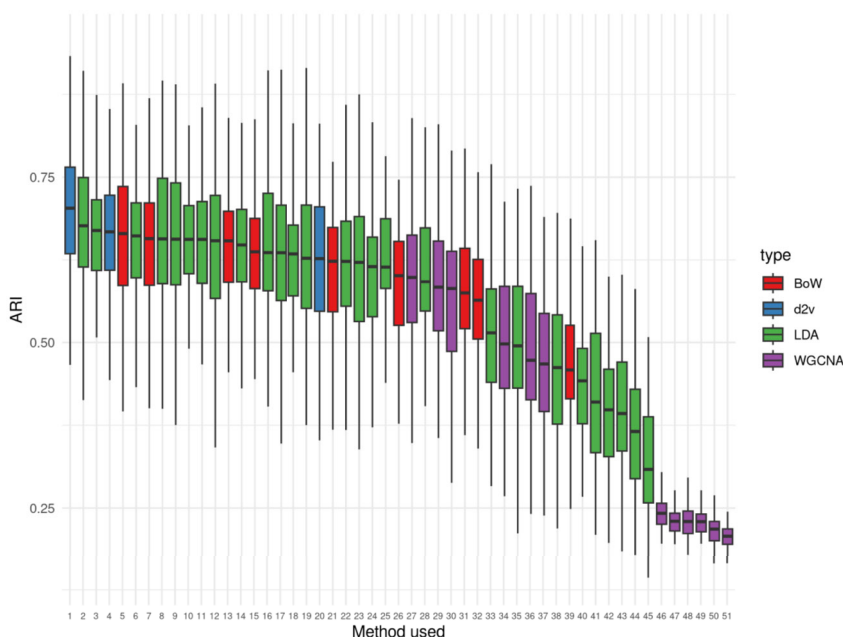
does larger thematic foregrounding improve genre clustering? As expected, we have found that low thematic foregrounding shows the worst performance across all four feature types (see Fig. 3). For LDA and bag-of-words, it leads to drastically worse performance. At the same time, we do not see a large difference between the medium and the strong levels of thematic foregrounding. The major difference in the strong level of thematic foregrounding is the use of lexical simplification. However, lexical simplification has not led to a noticeable improvement in genre recognition. The gains of using strong thematic foregrounding for document embeddings, LDA, and bag-of-words are marginal and inconsistent.

**Conclusion 2. Various feature types show similarly good performance.** Does the choice of feature type matter for the performance of genre clustering? We have found that almost all feature types can perform well. As shown in Fig. 2, three out of four feature types—doc2vec, LDA, and bags of words—when used in certain combinations, can lead to almost equally good results. But how good are they on average? Fig. 4 shows the posterior distributions of ARI for each type of feature—in each case, for a high level of thematic foregrounding.

As we see, doc2vec shows the best average performance, but this study has not experimented enough with other parameters of this feature type. It might be that another number of dimensions (e.g., 100 instead of 300) would worsen its performance. More research is needed to better understand the performance of doc2vec. LDA is the second-best approach—and interestingly, the variation of parameters in LDA (such as *k* of topics or *n* of MFWs) does not increase the variance compared to doc2vec. The bag-of-words approach, despite being simplest, proves to be surprisingly good. It does not demonstrate the best performance, but it is not far behind LDA. At the same time, bags of words have a powerful advantage: simplicity. They are simpler to use and require fewer computational resources, meaning that in many cases they can still be a suitable choice for thematic analysis. Finally, WGCNA shows the worst ARI scores on average.

**Conclusion 3. The performance of LDA does not seem to depend on *k* of topics and *n* of most frequent words.** LDA modeling depends on parameters, namely *k* of topics and *n* of most frequent words, which should be decided, somewhat arbitrarily, before modeling. There exist algorithms for estimating the “good” number of topics, which help assess how many topics are “too few” and how many are “too many” (Sbalchiero and Eder, 2020). In our study, however, we find no meaningful influence of either of these choices on learning the thematic signal (Fig. 5). The single most important factor making a massive influence on the effectiveness of thematic classification is thematic foregrounding. Weak thematic foregrounding (in our case, only lemmatizing words and removing the 100 most frequent words) proves to be a terrible choice that noticeably reduces ARI scores. Our study points towards the need for further systematic comparisons of various approaches to thematic foregrounding, as it plays a key role in the solid performance of LDA.

**Conclusion 4. Bag-of-words approach requires a balance of thematic foregrounding and *n* of most frequent words.** Bags of words are the simplest type of feature in thematic analysis, but still an effective one, as we have demonstrated. But how does one maximize the chances that bags of words perform well? We have varied two parameters in the bag-of-words approach: the level of thematic foregrounding and the number of MFWs used. Figure 6 illustrates our findings: both these parameters influence performance. Using 5000, instead of 1000, MFWs drastically improve ARI scores. Similarly, using medium, instead of weak, thematic foregrounding, makes



Rank	Combination			Median ARI	Median absolute deviation
1	Strong foregr.	doc2vec (300 dimensions)	cosine	0.703	0.095
2	Strong foregr.	LDA (k=50, 5000 MFWs)	Jensen-Shannon	0.677	0.104
3	Strong foregr.	LDA (k=100, 1000 MFWs)	Jensen-Shannon	0.670	0.085
4	Medium foregr.	doc2vec (300 dimensions)	cosine	0.668	0.084
5	Strong foregr.	bag-of-words (10,000 MFWs)	Jensen-Shannon	0.665	0.107
6	Medium foregr.	LDA (k=50, 5000 MFWs)	Jensen-Shannon	0.661	0.092
7	Strong foregr.	bag-of-words (5000 MFWs)	Jensen-Shannon	0.657	0.089
8	Strong foregr.	LDA (k=20, 10,000 MFWs)	Jensen-Shannon	0.657	0.121
9	Strong foregr.	LDA (k=20, 5000 MFWs)	Jensen-Shannon	0.656	0.120
10	Strong foregr.	LDA (k=100, 5000 MFWs)	Jensen-Shannon	0.656	0.076

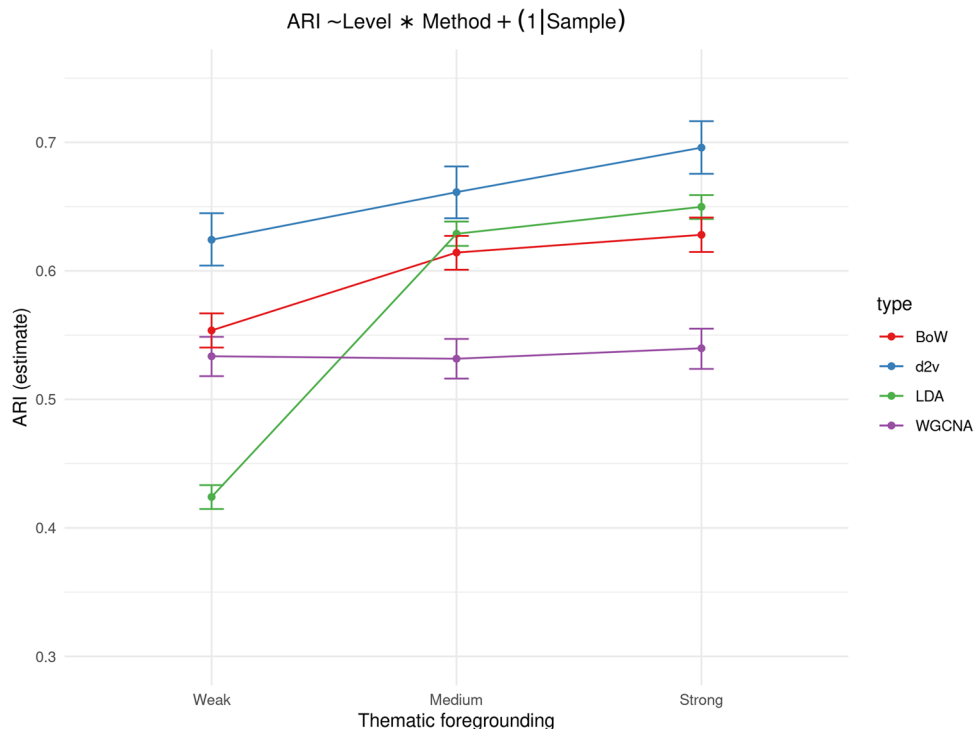
**Fig. 2 Raw distributions of ARI scores for all the combinations of thematic foregrounding, feature type, and distance metric.** Boxplots are colored by feature type. Numbers on the horizontal axis correspond to the names of combinations in the table to the right, showing 10 best-performing combinations (see all the combinations in Supplement, Table S7).

a big difference. At the same time, pushing these two parameters further—using 10,000 MFWs and strong thematic foregrounding—brings only marginal, if any, improvement in ARI scores.

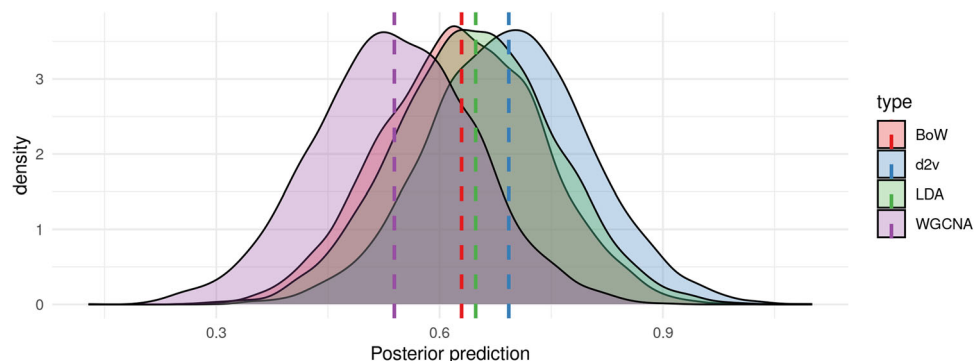
**Conclusion 5. Jensen–Shannon divergence is the best distance metric for genre recognition, Euclidean—the worst.** Choosing the right distance metric is crucial for improving genre clustering. Figure 7 shows the performance of various distances for each type of feature (note that Jensen–Shannon divergence, which was formulated for probability distributions, could not be applied to doc2vec dimensions and WGCNA module weights). For LDA and bag-of-words, Jensen–Shannon divergence is the best distance, with Delta and Manhattan distances being highly suitable too. For doc2vec, the choice of distance matters less. Interestingly, Euclidean distance is the worst-performing distance for LDA, bag-of-words, and WGCNA. This is good to know, as this distance is often used in text analysis, also in combination with LDA (Jockers, 2013; Schöch, 2017; Underwood et al., 2022), while our study

suggests that this distance should be avoided in computational thematic analysis. Cosine distance is known to be useful for authorship attribution when combined with bag-of-words as a feature type. At the same time, cosine distance is sometimes used to measure the distances between LDA topic probabilities, and our study shows that it is not the best combination.

**Comparison of algorithms on a larger dataset.** How well does this advice apply to clustering other corpora, not just our corpus of 200 novels? A common problem in statistics and machine learning is overfitting: tailoring one’s methods to a particular “sandbox” dataset, without making sure that these methods work “in the wild”. In our case, this means: would the same combinations of methods work well/poorly on other genres and other books than those included in our analysis? One precaution that we took to deal with overfitting was sampling from our genre corpus: instead of analyzing the full corpus just once, we analyzed smaller samples from it. But, additionally, it would be useful to



**Fig. 3** The effect of thematic foregrounding (weak, medium, or strong) on clustering genres, stratified by feature type.



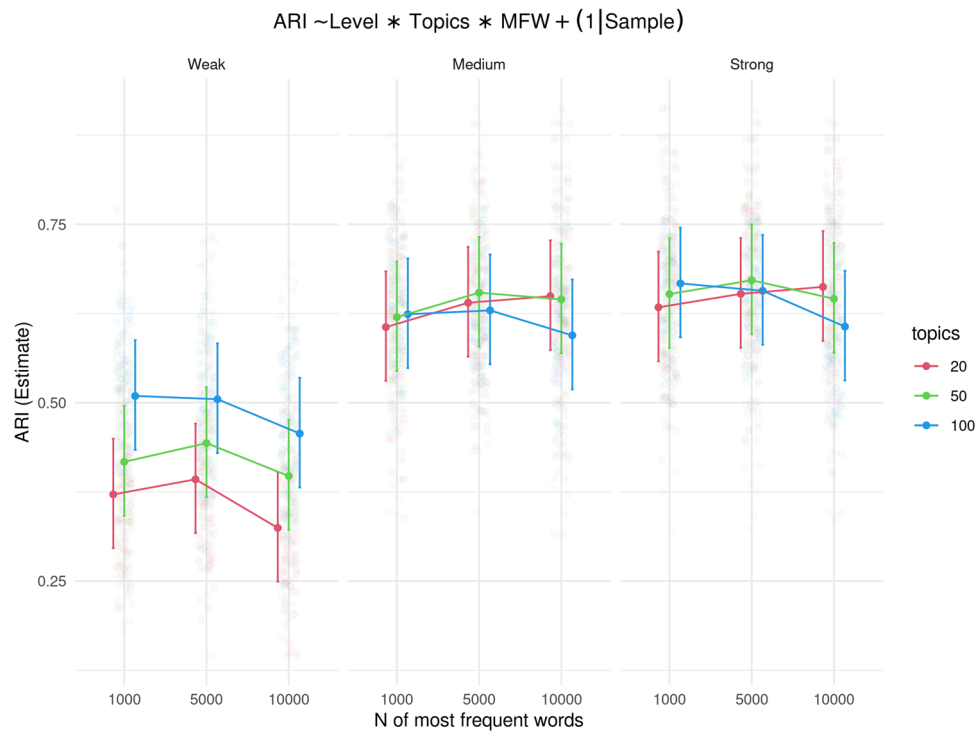
**Fig. 4** Posterior distributions of ARI scores for four feature types, at a high level of thematic foregrounding.

compare the best-performing and the worst-performing methods against a much larger corpus of texts.

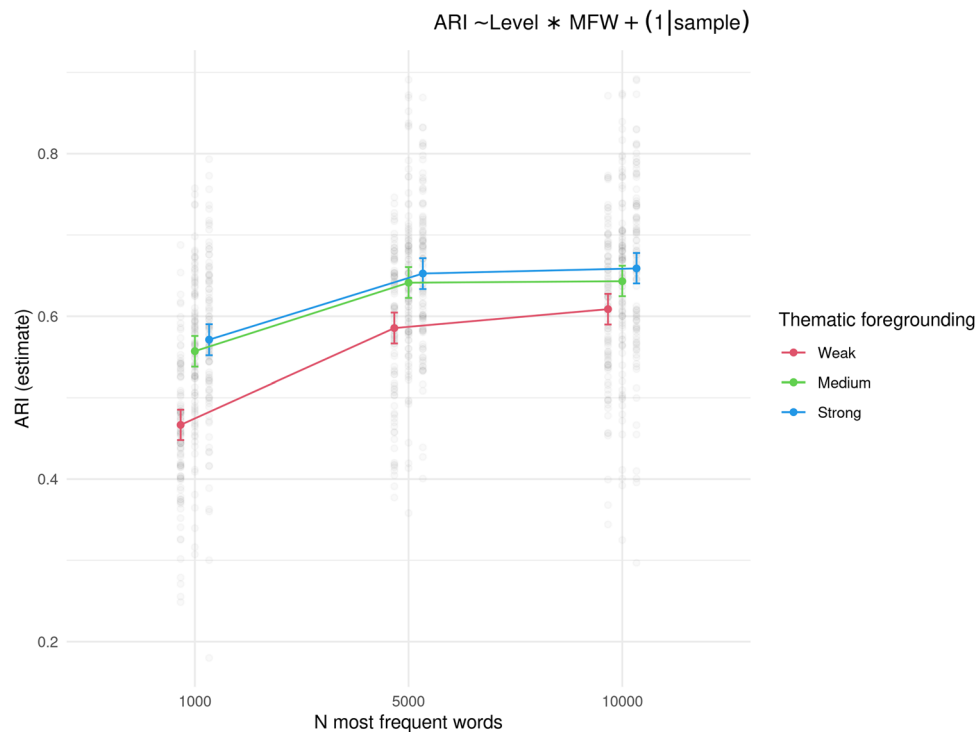
For this purpose, we use a sample of 5000 books of the NovelTM dataset of fiction, built from HathiTrust corpus (Underwood et al., 2020). Unlike our small corpus of four genres, these books do not have reliable genre tags, so we could not simply repeat our study on this corpus. Instead, we decided to inspect how a larger sample of our four genres (detective, fantasy, science fiction, and romance) would cluster within the HathiTrust corpus. For this, we included all the books in these four genres that we could easily identify and seeded them into a random sample of 5000 works of fiction. Then we clustered all these books using one of the best combinations of methods for identifying genres (medium thematic foregrounding, LDA on 1000 words with 100 topics, clustered with Delta distance). The result, visualized with a UMAP projection (McInnes et al., 2018), is shown in Fig. 8. This is just an illustration of the possible first step toward further testing various algorithms of computational thematics “in the wild”. The assessment of the accuracy of clustering of this larger corpus, and the comparison of it with the clustering using one of the worst-performing combinations of methods, are given in the Supplement (section 5.2).

**Discussion**

This study aimed to answer the question: how good are various techniques of learning thematic similarities between works of fiction? In particular, how good are they at detecting genres—and are they good at all? For this, we tested various techniques of text mining, belonging to three consecutive steps of analysis: pre-processing, extraction of features, and measuring distances between the lists of features. We used four common genres of fiction as our “ground truth” data, including a tightly controlled corpus of books. Our main finding is that unsupervised learning can be effectively used for detecting thematic similarities, but algorithms differ in their performance. Interestingly, the algorithms that are good for computational stylometry (and its most common task, authorship attribution) are not the same as those good for computational thematics. To give an example, one common approach to authorship attribution—using limited pre-processing, with a small number of most frequent words as features, and cosine distance—is one of the least accurate approaches for learning thematic similarities. How important are these differences in the real-world scenario, not limited to our small sample of books? To test this, we have contrasted one of the



**Fig. 5** Posterior probabilities of the effects of  $k$  of topics on ARI, stratified by the level of thematic foregrounding and  $n$  of most frequent words used in LDA. Error bars show 95% credible intervals.



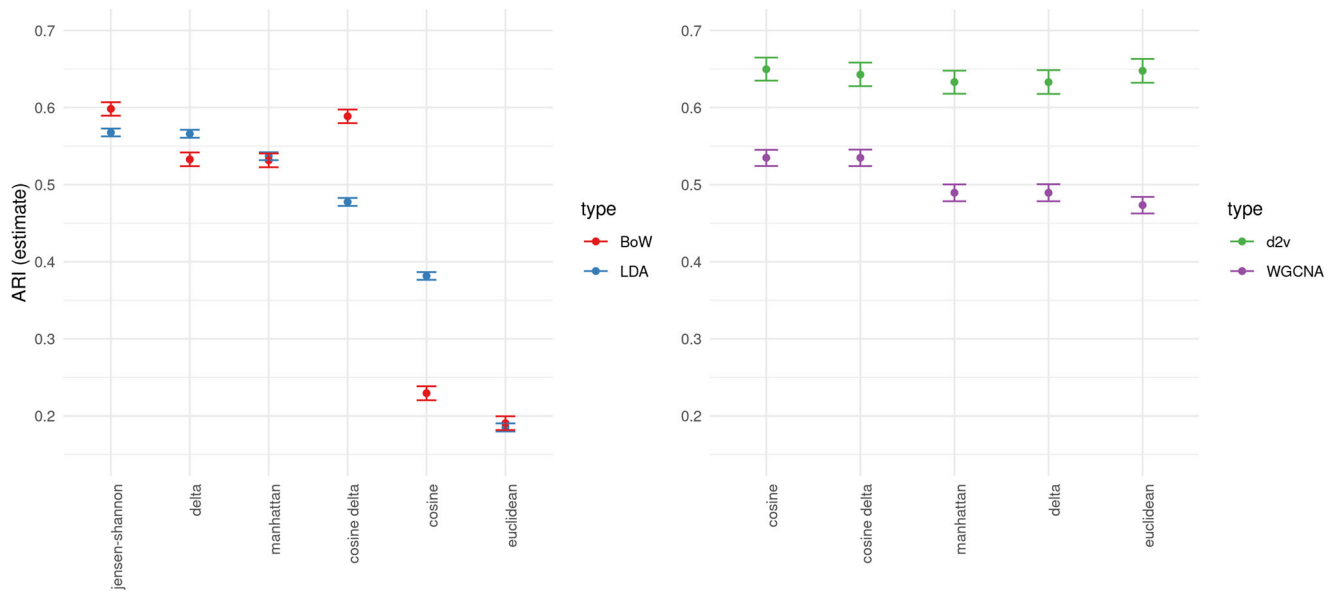
**Fig. 6** The influence of the number of most frequent words, used as text features, on learning the thematic signal, measured with ARI. There is a positive relationship between the  $n$  of words and ARI, as well as between the level of thematic foregrounding and ARI. However, the middle parameter values of both (5000 MFWs and medium foregrounding) should be enough for most analyses.

worst-performing combinations of algorithms, and one of the best-performing combinations, using a large sample of the HathiTrust corpus of books.

Systematic comparisons between various algorithms for computational thematic analysis will be key for a better understanding

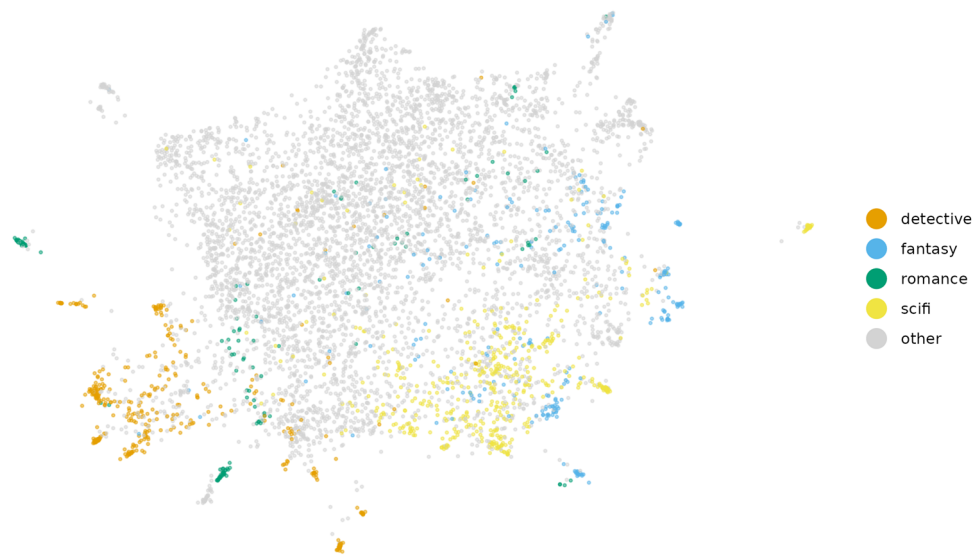
of which approaches work and which do not—a requirement for assuring reliable results in the growing research area which we suggest calling “computational thematics”. Using a reliable set of algorithms for thematic analysis would allow tackling several large problems that remain not solved in the large-scale analysis of





**Fig. 7 The influence of distance metrics on ARI scores, separately for each feature type.** Note that Jensen-Shannon divergence could not be combined with WGCNA and doc2vec.

LDA 1k words, 100 topics, Delta



**Fig. 8 A UMAP projection of 5000 novels, randomly sampled from the NovelTM HathiTrust corpus and all the novels by the authors included in the original corpus of four genres, found in NovelTM.** The figure is based on one of the best-performing combinations.

books. One such problem is creating better *genre tags* for systematizing large digital libraries of digitized texts. Manual genre tags in corpora such as HathiTrust are often missing or are highly inconsistent, which leads to attempts to use supervised machine learning, trained on manually tagged texts, to automatically learn the genres of books in the corpus overall. However, this approach, by design, allows capturing only the genres we already know about, not the genres we do not know exist: “latent” genres. Unsupervised thematic analysis can be used for this task. Another important problem that unsupervised approaches to computational thematics may be good at is the historical analysis of *literary evolution*. So far, we are lacking a comprehensive “map” of literary influences, based on the similarity of books. Such a map would allow creation of a

computational model of literary macroevolution, similar to phylogenetic trees (Bouckaert et al., 2012; Tehrani, 2013) or rooted phylogenetic networks (Neureiter et al., 2022; Youngblood et al., 2021) used in cultural evolution research of languages, music, or technologies. Having reliable unsupervised algorithms for measuring thematic similarities would be crucial for any historical models of this sort. Also, measuring thematic similarities may prove useful for creating *book recommendation* systems. Currently, book recommendation algorithms are mostly based on the analysis of user behavior: ratings or other forms of interaction (Duchen, 2022). Such methods are highly effective in cases when user-generated data is abundant, like songs or brief videos. However, for longer content types, which take more time to consume, the

amount of user-generated data is much smaller. Improving the tools for content-based similarity detection in books would allow recommending books based on their content—as it is already happening to songs: projects such as Spotify’s *Every Noise at Once* (<https://everynoise.com/>) combine user behavior data with the acoustic features of songs themselves to learn the similarity between songs and recommend them to listeners.

This study is a preliminary attempt at systematizing various approaches to computational thematics. More work is needed to further test the findings of this paper and to overcome its limitations. First, any clustering approach, be it a dendrogram or a projection, is a simplification of the relationships described by a distance matrix, which, in turn, is based on aggregating information from imperfect proxies (e.g., words and topics). This poses two major limitations to the unsupervised analysis of fiction: (1) the ability to generalize over an observed clustering that arises from a very specific set of choices and parameters, (2) the lack of control over textual features and the representation of novels. Clustering limitations are often overcome by using bootstrap approaches (Eder, 2017), or meta-analysis of clusters (“clustering of clusters”) for the detection of stable, previously unidentifiable subgroups (Calvo Tello, 2021); in recent years, there were major advances in Bayesian clustering methods that provide non-parametric solutions to data partitioning and allow to associate uncertainty with data membership in clusters, or with number of clusters in a dataset (Wade and Ghahramani, 2018). These appear particularly suitable to cultural data, yet come with their own challenges in presentation and reporting.

Another limitation of our study is its inability to distinguish between *thematic* and *formal* elements of texts. A narrative can be told from the first-person perspective or from the standpoint of the omniscient narrator; events in a story can follow the chronological order or may be re-ordered, with flashbacks, flashforwards, and various other temporal patterns. Formal structures like these are not really “thematic” and require separate examination. Ideally, one would have to extract them as a separate set of textual features and use them in the analysis alongside themes. Unfortunately, we only have a limited set of tools for finding such formal patterns—though novel computational methods offer a promise of capturing some of them (Langlais, 2023; Piper and Toubia, 2023). Also, one may argue, that the methods of feature extraction used in this paper do capture not only topics as such but also some of the formal variation in texts, even though it is not clear where the line between forms and themes should be drawn. For example, the narrative perspective is partly captured by the distribution of pronouns and verb forms, pushing the first-person and third-person books to form different groups.

Yet another limitation is the concept of “ground truth” genres. It may be noted—rightly—that there are no “true” genres and that genre tags overall may not be the best approach for testing thematic similarities. As further steps, we see using large-scale user-generated tags from Goodreads and similar websites as a proxy for “ground truth” similarity.

Finally, this study has certainly not exhausted all the possible techniques for text analysis that can be used for computational thematics. For example, much wider testing of vector models, like doc2vec, but also BERTopic (Grootendorst, 2022) or Top2Vec (Angelov, 2020) is an obvious next step, as well as testing other network-based methods for community detection (Gerlach et al., 2018). Text simplification could have a large potential for thematic analysis—it must be explored further. Possibly, the most straightforward way to test our findings would be to attempt to replicate them on other genre corpora, containing more books or other genres. Testing these methods on books in other languages is also critical. The approach taken in this paper offers a simple analytical pipeline—and we encourage other researchers to use it for testing

other computational approaches. Such a communal effort will be key for assuring robust results in the area of computational thematics.

### Data availability

R scripts used in our analysis, together with pre-registration documents, can be found on Open Science Framework’s website <https://osf.io/rtvb6/>. Due to copyright law, we cannot share the corpus of books used in this study, instead, we share the document-term matrices based on the samples of this corpus.

Received: 27 March 2023; Accepted: 4 March 2024;

Published online: 20 March 2024

### References

- Allison S, Heuser R, Jockers M, Moretti F, Witmore M (2011) Quantitative formalism: an experiment. Stanford Literary Lab, Pamphlet 1. <https://litlab.stanford.edu/LiteraryLabPamphlet1.pdf>
- Angelov D (2020) Top2Vec: distributed representations of topics. arXiv. <https://doi.org/10.48550/arXiv.2008.09470>
- Bailey P, Chang DK, Nones K, Johns AL, Patch A-M, Gingras M-C, Miller DK, Christ AN, Bruxner TJC, Quinn MC, Nourse C, Murtaugh LC, Harliwong I, Idrisoglu S, Manning S, Nourbakhsh E, Wani S, Fink L, Holmes O, Grimmond SM (2016) Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature* 531(7592):47–52. <https://doi.org/10.1038/nature16965>
- Baraghith K (2020) Investigating populations in generalized Darwinism. *Biol Philos* 35(1):19. <https://doi.org/10.1007/s10539-020-9735-6>
- Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. *J Mach Learn Res* 3:993–1022
- Bories A-S, Plecháč P, Ruiz Fabo P (eds.) (2023) Computational stylistics in poetry, prose, and drama. De Gruyter. <https://doi.org/10.1515/9783110781502>
- Bouckaert R, Lemey P, Dunn M, Greenhill SJ, Alekseyenko AV, Drummond AJ, Gray RD, Suchard MA, Atkinson QD (2012) Mapping the origins and expansion of the Indo-European language family. *Science* 337(6097):957–960. <https://doi.org/10.1126/science.1219669>
- Brennan M, Afroz S, Greenstadt R (2012) Adversarial stylometry: circumventing authorship recognition to preserve privacy and anonymity. *ACM Trans Inf Syst Secur* 15(3):12:1–12:22. <https://doi.org/10.1145/2382448.2382450>
- Burrows JF (1987) Computation into criticism: a study of Jane Austen’s novels and an experiment in method. Clarendon Press
- Cafiero F, Camps J-B (2019) Why Molière most likely did write his plays. *Sci Adv* 5(11):eaax5489
- Calvo Tello J (2021) The novel in the Spanish Silver Age: a digital analysis of genre using machine learning. Bielefeld University Press. <https://doi.org/10.1515/9783839459256>
- Chung C, Pennebaker J (2007) The psychological functions of function words. In: *Social communication*. Psychology Press. pp. 343–359
- Duchen H (2022) A comparative study of various book recommendation algorithms for public libraries. *Tech Serv Q* 39(4):369–380. <https://doi.org/10.1080/07317131.2022.2125676>
- Dynomant E, Lelong R, Dahamna B, Massonnaud C, Kerdelhué G, Grosjean J, Canu S, Darmoni SJ (2019) Word embedding for French natural language in healthcare: a comparative study. In: L Ohno-Machado & B Séroussi (eds.) *MEDINFO 2019: Health and Wellbeing e-Networks for All—Proceedings of the 17th World Congress on Medical and Health Informatics*, Lyon, France, 25–30 August 2019 (vol. 264). IOS Press. pp. 118–122
- Eder M (2017) Visualization in stylometry: cluster analysis using networks. *Digit Scholarsh Humanit* 32(1):50–64
- Eder M, Rybicki J, Kestemont M (2016) Stylometry with R: a package for computational text. *Anal R J* 8(1):107
- Egger R, Yu J (2022) A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Front Sociol* 7. <https://doi.org/10.3389/fsoc.2022.886498>
- Elliott J (2017) Whole genre sequencing. *Digital Scholarsh Humanit* 32(1):65–79
- Evert S, Proisl T, Jannidis F, Reger I, Pielström S, Schöch C, Vitt T (2017) Understanding and explaining Delta measures for authorship attribution. *Digital Scholarsh Humanit* 32:ii4–ii16
- Fowler A (1971) The life and death of literary forms. *N Lit Hist* 2(2):199–216. <https://doi.org/10.2307/468599>
- Gerlach M, Peixoto TP, Altmann EG (2018) A network approach to topic models. *Sci Adv* 4(7):eaq1360. <https://doi.org/10.1126/sciadv.aq1360>
- Grave E, Bojanowski P, Gupta P, Joulin A, Mikolov T (2018) Learning word vectors for 157 languages. *Proceedings of the Eleventh International Conference on*

- Language Resources and Evaluation (LREC 2018). LREC 2018, Miyazaki, Japan. <https://aclanthology.org/L18-1550>
- Grootendorst M (2022) BERTopic: neural topic modeling with a class-based TF-IDF procedure. <https://doi.org/10.48550/arXiv.2203.05794>
- Houkes W (2012) Population thinking and natural selection in dual-inheritance theory. *Biol Philos* 27(3):401–417. <https://doi.org/10.1007/s10539-012-9307-5>
- Hubert L, Arabie P (1985) Comparing partitions. *J Classif* 2(1):193–218. <https://doi.org/10.1007/BF01908075>
- Hughes JM, Foti NJ, Krakauer DC, Rockmore DN (2012) Quantitative patterns of stylistic influence in the evolution of literature. *Proc Natl Acad Sci USA* 109(20):7682–7686
- Iosifyan M, Vlasov I (2020) And quiet flows the Don: the Sholokhov-Kryukov authorship debate. *Digit Scholarsh Humanit* 35(2):307–318. <https://doi.org/10.1093/lc/fqz017>
- Jockers ML (2013) Macroanalysis: digital methods and literary history. University of Illinois Press
- Kim D, Seo D, Cho S, Kang P (2019) Multi-co-training for document classification using various document representations: TF-IDF, LDA, and Doc2Vec. *Inf Sci* 477:15–29. <https://doi.org/10.1016/j.ins.2018.10.006>
- Klimek P, Kreuzbauer R, Thurner S (2019) Fashion and art cycles are driven by counter-dominance signals of elite competition: quantitative evidence from music styles. *J R Soc Interface* 16(151):20180731. <https://doi.org/10.1098/rsif.2018.0731>
- Langfelder P, Horvath S (2008) WGCNA: An R package for weighted correlation network analysis. *BMC Bioinforma* 9(1):559
- Lau JH, Baldwin T (2016) An empirical evaluation of doc2vec with practical insights into document embedding generation. Proceedings of the 1st Workshop on Representation Learning for NLP, pp. 78–86. <https://doi.org/10.18653/v1/W16-1609>
- Le Q, Mikolov T (2014) Distributed representations of sentences and documents. Proceedings of the 31st International Conference on Machine Learning, pp. 1188–1196. <https://proceedings.mlr.press/v32/le14.html>
- Liu L, Dehmamy N, Chown J, Giles CL, Wang D (2021) Understanding the onset of hot streaks across artistic, cultural, and scientific careers. *Nat Commun* 12(1):5392. <https://doi.org/10.1038/s41467-021-25477-8>
- McInnes L, Healy J, Saul N, Großberger L (2018) UMAP: uniform manifold approximation and projection. *J Open Source Softw* 3(29):861. <https://doi.org/10.21105/joss.00861>
- Moretti F (2005) Graphs, maps, trees: abstract models for literary history. Verso
- Mosteller F, Wallace DL (1963) Inference in an authorship problem. *J Am Stat Assoc* 58(302):275–309
- Neal T, Sundararajan K, Fatima A, Yan Y, Xiang Y, Woodard D (2017) Surveying stylometry techniques and applications. *ACM Comput Surv* 50(6):86:1–86:36. <https://doi.org/10.1145/3132039>
- Neureiter N, Ranacher P, Efrat-Kowalsky N, Kaiping GA, Weibel R, Widmer P, Bouckaert RR (2022) Detecting contact in language trees: a Bayesian phylogenetic model with horizontal transfer. *Humanit Soc Sci Commun* 9(1):1. <https://doi.org/10.1057/s41599-022-01211-7>
- Ochab J, Byszuk J, Pielström S, Eder M (2019) Identifying similarities in text analysis: hierarchical clustering (linkage) versus network clustering (community detection). ADHO 2019, Utrecht. <https://dh-abstracts.library.cmu.edu/works/10014>
- Langlais P-C (2023) Brahe. *Hugging Face*. <https://huggingface.co/Pclanglais/Brahe>
- Piper A, Bagga S, Monteiro L, Yang A, Labrosse M, Liu, YL (2021) Detecting narrativity across long time scales. CHR 2021: Computational Humanities Research Conference. CEUR Workshop Proceedings, pp. 319–332. [https://ceur-ws.org/Vol-2989/long\\_paper49.pdf](https://ceur-ws.org/Vol-2989/long_paper49.pdf)
- Piper A, Toubia O (2023) A quantitative study of non-linearity in storytelling. *Poetics* 98:101793. <https://doi.org/10.1016/j.poetic.2023.101793>
- Plecháč P (2021) Relative contributions of Shakespeare and Fletcher in Henry VIII: An analysis based on most frequent words and most frequent rhythmic patterns. *Digital Scholarsh Humanit* 36(2):430–438. <https://doi.org/10.1093/lc/fqaa032>
- Plecháč P, Bobenhausen K, Hammerich B (2018) Verification and authorship attribution. A pilot study on Czech, German, Spanish, and English poetry. *Stud Metret Poetica* 5(2):29–54
- Pranjic M, Podpečan V, Robnik-Sikonja M, Pollak S (2020) Evaluation of related news recommendations using document similarity methods. Conference on Language Technologies & Digital Humanities, pp. 81–86
- Ramírez-González RH, Borrill P, Lang D, Harrington SA, Brinton J, Venturini L, Davey M, Jacobs J, van Ex F, Pasha A, Khedikar Y, Robinson SJ, Cory AT, Florio T, Concia L, Juery C, Schoonbeek H, Steuermagel B, Xiang D, Uauy C (2018) The transcriptional landscape of polyploid wheat. *Science* 361(6403):eaar6089. <https://doi.org/10.1126/science.aar6089>
- Sbalchiero S, Eder M (2020) Topic modeling, long texts and the best number of topics. Some problems and solutions. *Qual Quant* 54(4):1095–1108. <https://doi.org/10.1007/s11135-020-00976-w>
- Schöch C (2017) Topic modeling genre: an exploration of French classical and enlightenment drama. *Digit Humanit* Q 011:2
- Šeļa A, Plecháč P, Lassche A (2022) Semantics of European poetry is shaped by conservative forces: the relationship between poetic meter and meaning in accentual-syllabic verse. *PLoS ONE* 17(4):e0266556. <https://doi.org/10.1371/journal.pone.0266556>
- Sigaki HYD, Perc M, Ribeiro HV (2018) History of art paintings through the lens of entropy and complexity. *Proc Natl Acad Sci USA* 115(37):E8585–E8594. <https://doi.org/10.1073/pnas.1800083115>
- Stamatatos E (2009) A survey of modern authorship attribution methods. *J Am Soc Inf Sci Technol* 60(3):538–556. <https://doi.org/10.1002/asi.21001>
- Symons J (1985) Bloody murder: from the detective story to the crime novel: a history. Viking
- Tehrani JJ (2013) The phylogeny of Little Red Riding Hood. *PLoS ONE* 8(11):e78871
- Thelwall M (2019) Reader and author gender and genre in Goodreads. *J Librariansh Inf Sci* 51(2):403–430. <https://doi.org/10.1177/0961000617709061>
- Underwood T (2016) The life cycles of genres. *Journal of Cultural Analytics*, 2(2). <https://doi.org/10.22148/16.005>
- Underwood T (2019) Distant horizons. The University of Chicago Press
- Underwood T, Kiley K, Shang W, Vaisey S (2022) Cohort succession explains most change in literary culture. *Sociol Sci* 9:184–205. <https://doi.org/10.15195/v9.a8>
- Underwood T, Kimutis P, Witte J (2020) NovelTM datasets for English-language fiction, 1700–2009. *J Cult Analyt* 5(2). <https://doi.org/10.22148/001c.13147>
- Wade S, Ghahramani Z (2018) Bayesian cluster analysis: point estimation and credible balls (with discussion). *Bayesian Anal* 13(2):559–626. <https://doi.org/10.1214/17-BA1073>
- Ward JH (1963) Hierarchical grouping to optimize an objective function. *J Am Stat Assoc* 58(301):236–244. <https://doi.org/10.2307/2282967>
- Wilkins M (2016) Genre, computation, and the varieties of twentieth-Century U.S. fiction. *J Cult Analyt* 1(1):11065. <https://doi.org/10.22148/16.009>
- Youngblood M, Baraghith K, Savage PE (2021) Phylogenetic reconstruction of the cultural evolution of electronic music via dynamic community detection (1975–1999). *Evol Hum Behav* 42(6):573–582. <https://doi.org/10.1016/j.evolhumbehav.2021.06.002>

## Acknowledgements

OS and AŠ were supported by the project “Phylogenies of Literature” from the Cultural Evolution Society Transformation Fund granted by the John Templeton Foundation (grant #61913). AŠ was supported by the project “Large-Scale Text Analysis and Methodological Foundations of Computational Stylistics” (SONATA-BIS 2017/26/E/H52/01019).

## Author contributions

OS conceived the study. OS and AŠ jointly prepared the data and conducted the analysis. OS wrote the manuscript and reviewed the supplement. AŠ wrote the supplement and reviewed the manuscript. Both authors have read and agreed to the published version of the paper.

## Funding

Open Access funding enabled and organized by Projekt DEAL.

## Competing interests

The authors declare no competing interests.

## Ethical approval

The study included no human or non-human participants and thus required no ethical approval.

## Informed consent

Since secondary data were used in the study, informed consent was not required.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1057/s41599-024-02933-6>.

**Correspondence** and requests for materials should be addressed to Oleg Sobchuk.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024