Studying Literary History with Latent Feature Models

by

Allen Beye Riddell

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
David Dunson, Supervisor

_____
Surya Tokdar

_____
Katherine Hayles

Thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science in the Department of Statistical Science
in the Graduate School of Duke University
2013

Abstract

Studying Literary History with Latent Feature Models

by

Allen Beye Riddell

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
David Dunson, Supervisor

_____
Surya Tokdar

_____
Katherine Hayles

An abstract of a thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in the Department of Statistical Science in the Graduate School of Duke University
2013

Abstract

Novelistic genres—such as gothic novels, epistolary novels, and *Bildungsromane*—were an abiding feature of literary production in the nineteenth century. Their appearance, disappearance, and transmission across national and linguistic boundaries continues to be an object of interest for scholars in literary history and sociology of culture. This thesis considers two non-parametric latent feature models of a corpus of British literary fiction and compares the models' representations with the judgments of literary historians. I find that the models agree with expert classifications of novelistic genre better than chance. This thesis contributes to efforts to validate latent feature models against human judgments and offers further confirmation that probabilistic models of text collections can support historical scholarship.

Contents

List of Tables

List of Figures

1

Introduction

Gothic, epistolary, and historical novels flourished in the British Isles during the late eighteenth and early nineteenth century. The share of literary production claimed by these novelistic genres is considerable.[1] Gothic novels, for example, accounted for thirty percent of new novels published in 1795 (fig. 1.1). Literary historians have documented the rise and fall of these and other novelistic genres—national tale, silver fork, and Newgate novels also number among significant genres from the period (Lévy 1968; Adburgham 1983; Hollingsworth 1963; Trumpener 1998). Recently, Moretti (2005) has revived interest in these categories by aggregating information about genres' periods of popularity and studying apparent regularities in the arrival and disappearance of genres.

While there is no consensus among historians on a definition of novelistic genre, many of the genres identified by historians were recognized by writers, readers, and publishers at the time. The clearest evidence comes from (sub)titles—e.g., *The Baron's Daughter: A Gothic Romance*; *Durston Castle: Or, the Ghost of Eleonora.*

1. Following Moretti (2005), I will refer to these categories as novelistic genres. If context makes clear the discussion is limited to novels, the qualifier "novelistic" may be dropped. In discussions of eighteenth- and nineteenth-century literature generally, "genre" is used in a variety of ways—e.g., to distinguish epic and tragic narratives, or among poetry, plays, and novels.

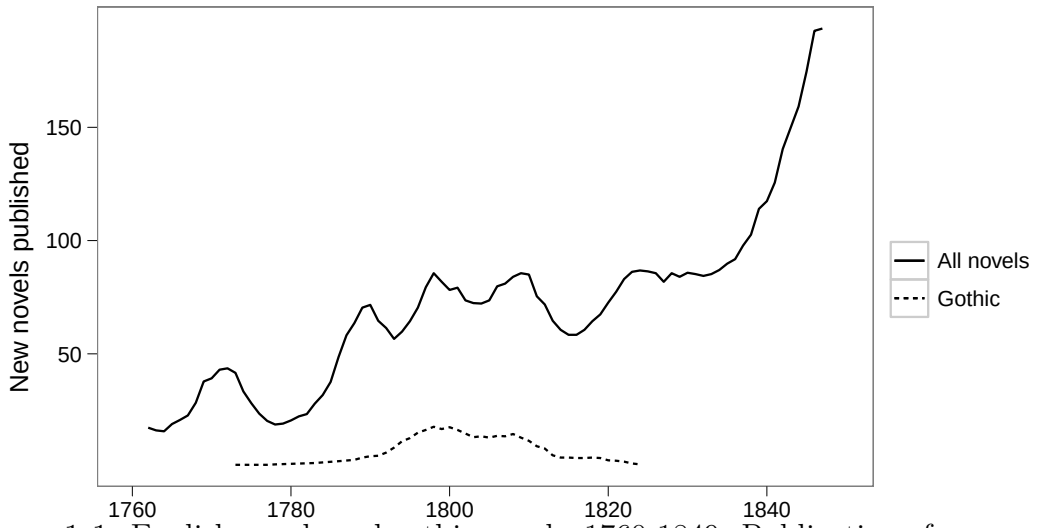

FIGURE 1.1: English novels and gothic novels, 1760-1849. Publication of new novels and those classified as gothic novels (five year moving average). Sources: 1770-1836 from Garside and Schöwerling (2000) and Garside et al. (2006); 1837-1849 from Block (1961); gothic novels from Lévy (1968).

A Gothic Story; *The Wild Irish Girl: A National Tale*; and *Caledonia: A National Tale*. Literary historians provide rich descriptions of individual genres, often by reference to shared features or "codes" (Moretti 2005; Cohen 2002). For instance, gothic novels are described as "a group of novels, set somewhere in the past, that exploit the possibilities of mystery and terror in sullen, craggy landscapes; decaying mansions with dank dungeons, secret passages, and stealthy ghosts; chilling supernatural phenomena; and often, sexual persecution of a beautiful maiden by an obsessed and haggard villain" (Abrams and Greenblatt 2000, 19). The features characteristic of novelistic genres need not be limited to settings, or indeed to anything found in the text of novels. A book's binding may be an important signal to readers, as it was for gothic novels, earning them the moniker "bluebooks" (Koch 2002). Narrative voice and plot structure have been suggested as distinguishing morphology (Elson, Dames, and McKeown 2010; Allison et al. 2011). As is the case today, a small number of publishers and authors were often strongly associated with specific genres (Adburgham 1983; Trumpener 1998).

Scholars in the human and social sciences have made considerable use of novelistic genres. Countless monographs and journal articles have been devoted to the genres of the eighteenth- and nineteenth-century European novel. The association of a specific genre's period of popularity with political and social events is not uncommon. The concern for youth and the process of development found in the *Bildungsroman* has been read as symptomatic of the situation in Europe after the upheaval of the French Revolution. The favorable portrayal of high society found in many silver fork novels has been connected with Regency Era social aspirations (Moretti 2000; Adburgham 1983). Novelistic genres have also been interpreted as offering a record of social relations. That is, by identifying a novel with a genre, writers situate their work in relation to existing novels, writers, conventions, and institutions. In this way chronicling a novelistic genre offers information about the internal dynamics of novelistic production and a window into the wider cultural field (Cohen 2002; Bourdieu 1988, 1996). Novelistic genres have also been used in sociology. Recently Isaac (2009) investigated the relationship between the publication rate of early twentieth-century "labor problem novels" and the historical record of labor strikes in the United States.

For the salient facts about a novelistic genre—such as its period of popularity and lists of associated novels, authors, and publishers—researchers often rely on the work of one or two literary historians who have studied a category in depth. I refer to these historians as "genre experts." There are three cases in which it would be particularly desirable to have an alternative means of identifying novels associated with a genre. First, for many lesser-known novels, no expert classification exists at all. When describing a genre, experts often mention only a handful of novels regarded as exemplary. Even when a list of novels belonging to a genre is provided, the list is rarely exhaustive. A list frequently covers a genre's period of popularity and omits titles published after the genre ceased to be prominent.[2] Second, experts

2. Adburgham (1983) stops listing silver fork novels after 1842 even though there are a several

have been known to disagree on the genre membership of individual novels. The status of very early and very late novels in a genre is particularly difficult to decide, as the characteristics and conventions associated with the genre during its period of popularity may be attenuated or otherwise difficult to identify. Disagreements about early and late novels are of particular concern as they are likely to affect periodizations of genres (Shalizi 2011). Novels may also be independently claimed as a member of more than one genre—e.g., Lady Morgan's *Florence Macarthy* (silver fork and national tale), Bulwer-Lytton's *Paul Clifford* (silver fork and Newgate), and Roche's *Tradition of the Castle* (gothic and national tale). When this occurs, additional evidence would be useful in understanding the reasons for competing classifications.[3] Finally, and most importantly, it is desirable to have an alternative to relying on the authority of one or two experts for the list of novels associated with a genre. Because expert classifications rely on tacit background knowledge and familiarity with a broad range of novels, it is difficult for other researchers to reproduce classifications or understand in detail the assumptions behind expert classifications. Having an alternative means of grouping novels together, particularly one that is reproducible, would inspire more confidence in the comprehensiveness and accuracy of any classification.

Literary historians have considered the challenge of inferring novelistic genre without relying on expert classifications. While studying the titles of novels published between 1740 and 1850, Moretti (2009) observed regularities in title word frequencies and phrases that correlated with novels being classified as gothic. Allison et al. (2011) took up the problem of unsupervised classification explicitly and examined

novels published after that date that would clearly qualify, such as novels by Catherine Gore, including *Castles in the Air* (1847).

3. Scholars have identified novels that they believe borrow morphology from novels in a genre but do so in a way that obscures their origins (Garside 1991). An alternative method of classification might help substantiate claims about such "cryptic" novels.

whether or not patterns in selected word and punctuation frequencies might be associated with genres. Allison et al. used principal components analysis (PCA) to characterize the dissimilarity between novels and examined the separation among novels visually using multidimensional scaling.[4] Allison et al. found that novels from certain genres did separate visually whereas others did not.[5]

The comparison of unsupervised latent feature models of text collections with classifications provided by human readers is common in computer science and computational linguistics. One study relevant for the present discussion is that of Chang et al. (2009), who evaluate several latent feature models of a corpus with assessments provided by human readers. Chang et al. used human readers, recruited online, to evaluate several models of a corpus of newspaper and Wikipedia articles by presenting different aspects of the topic models to human readers. In the study a reader's ability to identify an "intruder" word (a word strongly associated with a different topic) was taken as a proxy for the coherence of a topic. Notably Chang et al. show that models performing better on automatic measures (such as predictive likelihood of a held-out sample) do not necessarily perform better when measured by human readers' assessments. Whereas Chang et al. focused on the characteristics of the associations of latent features with words—the composition of "topic" distributions—this

4. Allison et al. cite Cavalli-Sforza, Menozzi, and Piazza (1994) as influential in their approach to the problem and in their use of PCA. The traffic between population genetics, cultural evolution, and quantitative literary history deserves some mention. An afterword to the widely discussed *Graphs, Maps, Trees* is written by Alberto Piazza, a coauthor of Cavalli-Sforza, Menozzi, and Piazza (1994). That work, in turn, is in conversation with the paper by Pritchard, Stephens, and Donnelly, which independently developed a mixed-membership model of allele frequencies that is essentially identical to Latent Dirichlet Allocation (Blei 2012). Novembre and Stephens (2008) is also a notable point of contact that concerns the use of PCA.

5. Allison et al. also provide a valuable discussion of challenges facing unsupervised clustering of eighteenth- and nineteenth-century novels on the basis of word frequencies alone. Specifically, Allison et al. identify two important challenges confronting unsupervised classification. First, certain genres may be marked by narrative structure rather than lexical features. The *Bildungsroman*'s episodic structure is the example provided: "discussions with old mentors and young friends, false starts, disappointments, the discovery of one's vocation …" (15). Second, authors may switch genres (or write in several) and so the lexical "signature" of the genre may be confounded by authorial style. For example, Lady Sydney Morgan (née Owenson) is a good example; she wrote national tale novels and silver fork novels.

thesis seeks empirical validation of the distribution of latent features over documents in a corpus. The source of the human judgments is also different; this work uses the expert judgments of literary historians who have spent years rather than minutes forming their judgments. The documents (novels) in the present work are also considerably longer.

In this thesis, I evaluate two non-parametric latent feature models as ways of identifying novels with shared features in a large collection of novels. Both models, the Dirichlet Process (DP) mixture of unigrams and the Hierarchical Dirichlet Process (HDP), bring with them different assumptions about how novels share morphology (Blei, Ng, and Jordan 2003; Teh et al. 2006). I focus on three genres—gothic, silver fork, and national tale—for which substantial bibliographies exists. These are also genres that are thought to be identifiable by the use of characteristic vocabulary. To the extent that expert classifications are uncontroversial—many novels are indeed formulaic and derivative—this experiment also contributes to a body of work validating topic models against human judgments.

The remainder of this thesis is organized as follows. Chapter 2 contains a description of the corpus of novels, chapter 3 presents the models used, chapter 4 describes and discusses the results of modeling the corpus, and chapter 5 offers a concluding discussion of the work and directions for future research in quantitative literary history.

2

Data: Three Novelistic Genres

The corpus of 93 novels consists of a random sample of novels published between 1800 and 1836 and a representative selection of gothic, silver fork, and national tale novels. The characteristics of gothic novels have been mentioned before. Abrams and Greenblatt describe them as novels "exploit[ing] the possibilities of mystery and terror in sullen, craggy landscapes; decaying mansions with dank dungeons, secret passages, and stealthy ghosts; chilling supernatural phenomena; and often, sexual persecution of a beautiful maiden by an obsessed and haggard villain" (Abrams and Greenblatt 2000, 19). Silver fork novels are typically set in London and portray Regency-era high society (Adburgham 1983). (A small silver fork was a culinary accessory found at dinner tables among the wealthy.) Adburgham lists the "essential facets" of a silver fork novel: "there are some politics, some gambling scenes and a duel; there are dazzling balls in the London season, and country-house parties in the winter; the characters include a dandy, a toad-eater, a scheming high-society villain, a pair of lovers ill-starred until towards the end of the third volume. There are social climbers clambering towards Almack's, provincial belles at a race meeting ball in Doncaster Assembly Rooms; there is satire at the expense of the middle class and the

rich roturiers. But above all, there are semi-flirtatious drawing-room conversations and dinner-table repartee" (Adburgham 1983, 92-3). Trumpener describes the basic plot shared by early national tale novels as follows: "[A] young hero or heroine, raised in England or on the continent, travels to Ireland or Scotland expecting to find barbarism. Instead, the protagonist falls in love with his or her new surroundings and with the aristocratic native guide who has helped him or her understand the region's beauty and cultural interest. The novel ends with the marriage of the lovers—and thus also with the allegorical union of Britain and its constituent 'national characters" (Trumpener 1998, 910).

The corpus contains 35 gothic novels, 22 silver fork novels, 18 national tale novels, and 18 randomly selected novels. The random sample of novels is drawn from the exhaustive survey of novelistic production in Garside and Schöwerling (2000). The genre-specific samples are drawn principally from random samples from the genre-specific bibliographies of Adburgham (1983), Lévy (1968), and Trumpener (1998). Also included are a small number of well-known novels associated with the three genres. These novels were used previously in Allison et al. (2011). The number of these familiar novels—including such gothic novels *The Mysteries of Udolpho* by Ann Radcliffe—is small and they have been included to connect this present work with existing research. Scans and machine-readable text versions of the novels were gathered from a number of repositories, including the Internet Archive—in particular, the University of Illinois at Urbana-Champaign's nineteenth-century novels collection—as well as Project Gutenberg, University of Adelaide, and the Corvey Collection.[1] The random sample originally included twenty-four novels. Scans of two novels falling in the random sample could not be located although all novels in the sample do survive to the present day in library collections. Four of the novels in

1. All the novels in the corpus were published before 1924 and are therefore in the public domain in the United States. The corpus will be made available online.

the random sample were also listed in Adburgham's bibliography of silver fork novels and are counted among those novels.[2]

To facilitate computation, I have "stemmed" words in the corpus: inflected forms of words are reduced to their stem. For example, "abandoned" and "abandoning" are counted as instances of "abandon." From this initial corpus I removed a selection of frequent English and French words ("stop words"), words having fewer than three characters, words occurring in fewer than fifteen novels, and words corresponding to character names and their capitalized forms of address ("Mr," "Miss," "Captain," etc.). The final corpus of 93 novels contains 3,515,515 words. There are 9,104 unique words in the corpus.

Figure 2.1 shows the lengths of the novels after preprocessing. The shortest and longest novels in the corpus are both by Maria Edgeworth. The shortest is *Castle Rackrent* (1800) and the longest is *Tales of Fashionable Life* (1809), a collection of stories.

2. There are 99 silver fork novels mentioned in Adburgham's bibliography the population of novels published during this period is 2,903. The probability of finding four or more such novels in a sample of twenty-four is quite low, roughly 0.01. To verify that nothing had gone wrong during sampling, I counted the number of novels appearing in the first 100 novels in the sample (with replacement) that also appeared in the silver fork bibliography. Six silver fork novels were among the first 100 sampled novels. Finding six or more in 100 trials is expected to occur more than ten percent of the time.

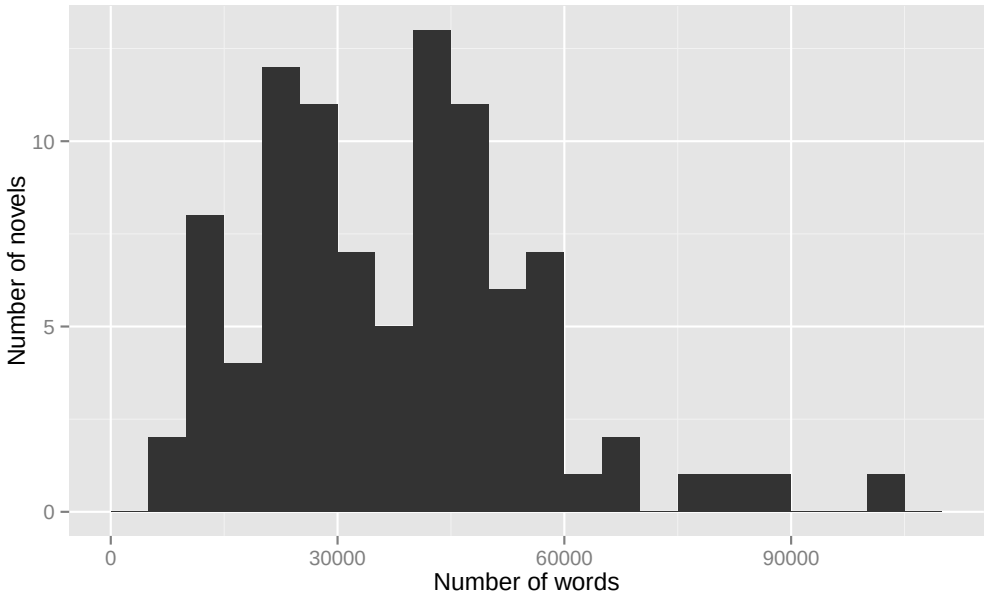

FIGURE 2.1: Length of the 93 novels in the corpus in words after preprocessing.

3

Models for Text Analysis

Characterizing novels in a genre by shared morphology is common (Moretti 2005, 14). Moretti appeals to shared morphology as a provisional definition of novelistic genre: "morphological arrangements that *last* in time, but always for *some* time" (14). We have already seen such a definition of the gothic novels relying on shared features—"sullen, craggy landscapes; decaying mansions with dark dungeons, secret passages, and stealthy ghosts…." Moving from morphology inferred by a human reader to individual words is not overly problematic for the three genres considered in this thesis. For example, given the description of the gothic novels, we anticipate a set of words—e.g., "ghosts," "dungeon," "cell," "manor"—being more likely in gothic novels than in non-gothic novels. This does not mean that the only way for a novel to feature "stealthy ghosts" is for the novel to contain a word referring to ghosts, such as "ghost" or "projection." A novel's narrative may feature ghosts in its storyline without using the word "ghost" or any synonyms. In the event that a novel does makes extensive use of words and phrases associated with the characteristic features mentioned (ghosts, dungeons, old mansions), we should anticipate

the novel being classified as a gothic novel. For the three genres considered in this thesis, the descriptions provided by genre experts and a passing familiarity with a handful of novels associated with the genres support the hypothesis that individual words (unigrams), in addition to being features of novels in their own right, provide information about morphology that may be described more generally.

Defining a group of novels by reference to explicitly shared morphology—such as a handful of distinctive words—works well as a preliminary strategy, but it has important limitations. The following comparison of ten novels in the corpus provides an illustration of an approach that focuses on the presence and absence of individual words. Comparing five gothic novels with a random sample of five non-gothic novels, we find that the presence of a few characteristic words does indeed distinguish gothic novels from non-gothic novels. "Depraved," "inhuman," "monstrous," "mouldering," and "turbulent," are unique to the gothic novels and the words come as no surprise given descriptions of the genre (fig. 3.1). Attempting to generalize an approach relying a fixed dictionary of features, however, runs into two difficulties. First, what counts as relevant morphology is in important respects arbitrary and, second, even when those doing the classifying agree on relevant features they may disagree on their measurement. One group of literary historians may believe plot structure to be more relevant than vocabulary for determining a novel's genre. Another group may put weight on "paratext"—e.g., frontispieces, illustrations, binding, paper, and typeface. Yet another group may stress particular aspects of the narrative, such as focalization, presence of indirect discourse, or absolute number of characters (Elson, Dames, and McKeown 2010; Moretti 2005).[1] And even within these groups there may be disagreement about how to measure features. If one regards the number of distinct characters in a novel as relevant— 紅樓夢 (*Dream of the Red Chamber*) has more

1. Of course, specific elements of narratives may be of interest, as in Vladimir Propp's *Morphology of the Folktale.*

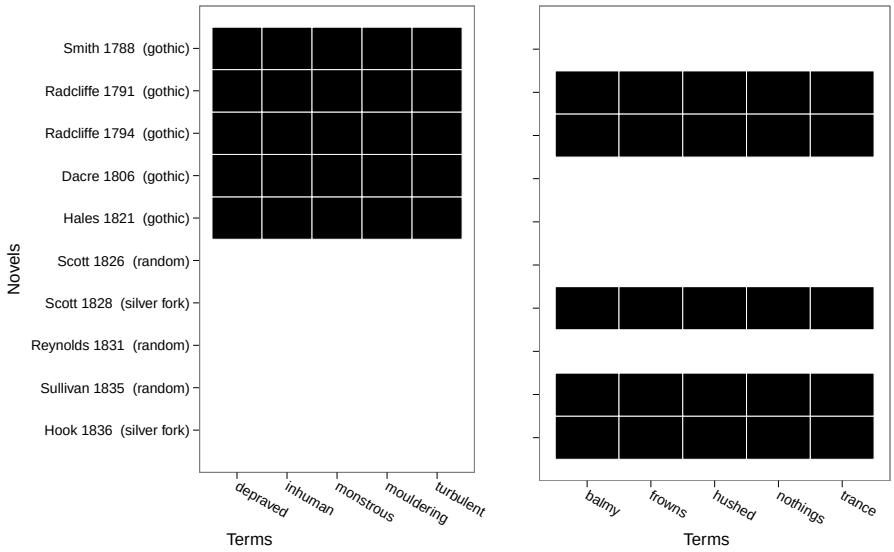

FIGURE 3.1: Words characteristic of five gothic novels (left) and words characteristic of a random partition of the same ten novels (right).

than 400 characters—does one count the total number of distinct characters or does one account for the novel's length and consider the average number of characters per thousand words? Different measurements may give rise to different categorizations. These two challenges are not purely theoretical. If we consider the same group of ten novels and randomly assign half to one group, it is not difficult to find words that distinguish the first group as a distinct category: "balmy," "frowns," "hushed," "nothings," and "trance" (fig. 3.1). With countless features available to describe any given novel and countless interpretations of those features, it will be possible to locate a handful of properties that, taken in isolation, support almost any classification.[2]

An alternative strategy for defining a novelistic genre would welcome a wide range of morphology, define a way to measure similarity between novels given a set of features, and insist novels of the same genre will tend to be more similar to each other than to novels not associated with the genre. We might describe this approach

2. There are no guarantees of agreement on relevant features and measurements. It is unlikely that radically different conceptions of morphology could result in shared categories. The range of morphology one might consider is endless: number of vowels, chemical composition of the ink, month of publication, and so forth.

as looking for "family resemblances" among novels. Just as it is often possible to guess at familial relationships despite the absence of any one trait that all family members share, such as eye or hair color, it may be possible to group novels together despite there being no feature that all members of the group share. Such an approach would avoid problems associated with fixing a set of features. The thought that a model might accommodate new features is also reassuring. That is, a method of grouping novels together that maintains its groupings even when new features are added—e.g., binding, city of publication, writer's social connections—seems more reliable than a method that ignores or cannot accommodate new features.

One strategy fitting this description would use a feature set consisting of word frequencies and measure similarity with Jaccard or cosine distance. The problem with Jaccard and cosine distance is that they make no consideration of polysemy pervasive in human language. For example, they do not distinguish among the "hook" in "Theodore Hook" (a silver fork novelist), "coat hook," "right hook," and so forth. Latent feature models of novels' word frequencies offer an approach that is attentive to the problem of polysemy (Blei, Ng, and Jordan 2003). Non-parametric models are particularly appealing as they permit inference about the number of latent features in addition to their distribution among novels in the corpus.

3.1 Representing Texts

The models considered in this chapter rely on a particular representation of the words appearing on the pages of the novels. This representation is known informally as the bag-of-words or vector space representation. The moniker "bag-of-words" captures what is left after discarding word order in a document—an unordered list or "bag" of words.[3] A convenient way of organizing these lists is in a table of word frequencies.

3. Formally, we might consider a bag in the context of the following three concepts: set, bag, and sequence. A set is an unordered list of elements that ignores order and duplicates, $S = \{4, 4, 5\} =$

	elizabeth|lizzy	darcy
Chapter 1	3	0
Chapter 2	3	0
Chapter 3	6	7
Chapter 4	4	6
Chapter 5	2	3
⋮	⋮	⋮
Chapter 31	9	14
Chapter 32	10	8
Chapter 33	4	17
Chapter 34	7	9
Chapter 35	2	4
⋮	⋮	⋮
Chapter 60	6	6
Chapter 61	9	8

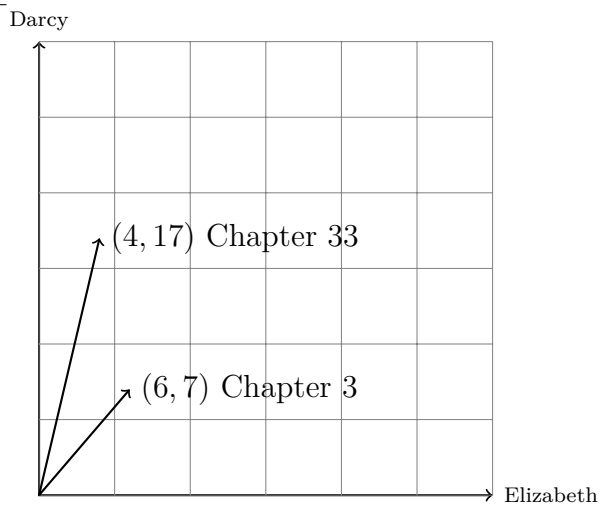

FIGURE 3.2: Chapters of *Pride and Prejudice* represented as vectors in $\mathbb{R}^2$.

If I collected the bag-of-words for each novel in the corpus and collected them into a table, a tiny portion of that table resembles table 3.1.

Table 3.1: Word frequency table for five novels and four words.

	abandon	abandoned	abandoning	abandonment
Dacre 1806	10	13	4	0
Morgan 1806	2	1	0	0
Bury 1837	1	1	0	0
Normanby 1825	2	6	2	3
Mosse 1816	1	7	1	0

A smaller corpus with a two-word vocabulary illustrates the origins of the name vector-space model. If we limit ourselves to the names of two characters in the novel *Pride and Prejudice*, we can visually compare the chapter vectors (fig. 3.2). In our Elizabeth-Darcy space it is straightforward, given rudimentary knowledge of the novel's plot, to see that the vectors reflect how much the two characters figure in each chapter. Chapters featuring one character but not the other point in different

$\{4, 5\}$. A bag is an unordered list that takes into account repeated elements, $B = \{4, 4, 4, 5\} = \{5, 4, 4, 4\}$. A sequence considers both order and repeated elements, $Q = \{4, 4, 5\} \neq \{5, 4, 4\}$.

directions. This notion of "pointing" in the same direction can be made precise by calculating the angle between vectors, the "cosine distance" (Manning and Schüzte 1999, 296–303).

The representation of texts used by the models considered in this thesis makes use of a slightly modified version of the bag-of-words schema. A novel is represented as a sequence of words $\mathbf{w}$, where each word is drawn from a vocabulary $\mathcal{V}$. If the word "it" is the third word in the vocabulary and the first word in a document is "it" then the first element of $\mathbf{w}$ would be the integer 3. Equivalently, the first element of $\mathbf{w}$ could also be a vector of length $|\mathcal{V}|$ whose components are all zero except for the third, which would be 1.

3.2 Latent Feature Models of Texts

The latent feature models used in this thesis share a number of characteristics. Both models, the Dirichlet Process (DP) mixture and the Hierarchical Dirichlet Process (HDP), have an interpretation as positing a latent set of discrete probability distributions (or "topic" distributions), characterized by a vector of probabilities over words, that generate the words observed in the corpus. The models differ in how they associate topics with words in the corpus. Both models are non-parametric, allowing for an unbounded number of topic distributions. The use of the word "topic" to describe a probability distribution over a vocabulary of words appears in the description of Latent Dirichlet Allocation (LDA) in Blei, Ng, and Jordan (2003). The models considered in this thesis might also be called latent class or latent cluster models. A firm distinction between latent class and feature models seems unnecessary, as latent class models can be considered as a kind of latent feature model (Broderick, Pitman, and Jordan 2013).

The most significant difference between the assumptions of the models based on the DP and the HDP is familiar from the models' parametric counterparts, mixture

of unigrams and LDA. Whereas the DP mixture associates every word in a document with a single topic, the HDP mixed-membership model allows words in a document to be associated with a range of topics. In other words, whereas the DP mixture of unigrams models the corpus as a mixture of topic distributions, the HDP models each document as a distinct mixture. In the context of LDA, Blei, Ng, and Jordan demonstrated that the mixed-membership model performs better than the simple mixture by a number of measures, including held-out likelihood (Blei, Ng, and Jordan 2003, 1008-1012).

The Dirichlet distribution is used in all the models considered in this thesis. The distribution is the multivariate extension of the Beta distribution and describes a distribution over the $K-1$ simplex. The density of a $Dirichlet(\alpha)$ distribution is written

$$p(\alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \beta^{\alpha_k - 1} = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \beta^{\alpha_k - 1}$$

If the Dirichlet distribution is parameterized by a single scalar value ($\alpha_1 = \ldots = \alpha_K$) it is referred to as a symmetric Dirichlet distribution.

In many models of text collections the Dirichlet distribution occurs as prior on the parameters of a multinomial distribution. Integrating out the Dirichlet distribution yields the Dirichlet compound multinomial distribution or Pólya distribution. The combination of the multinomial distribution with a Dirichlet prior on its parameters is given as

$$\theta \sim Dirichlet(\alpha_1, \ldots, \alpha_K)$$
$$x_i \sim Multinomial(1, \theta),\ i \in \{1, \ldots, N\}$$

The marginal probability of $\mathbf{x}$ is found by integrating over θ,

$$\begin{aligned} p(\mathbf{x}|\alpha_1, \ldots, \alpha_K) &= \int Dirichlet(\theta|\alpha_1, \ldots, \alpha_K) \prod_{i=1}^{N} Multinomial(x_i|\theta) d\theta \\ &= \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \int \prod_{k=1}^{K} \theta^{\alpha_k + n_k - 1} d\theta \\ &= \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \frac{\prod_{k=1}^{K} \Gamma(n_k + \alpha_k)}{\Gamma(N + \sum_{k=1}^{K} \alpha_k)} \\ &= \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\Gamma(N + \sum_{k=1}^{K} \alpha_k)} \prod_{k=1}^{K} \frac{\Gamma(n_k + \alpha_k)}{\Gamma(\alpha_k)} \end{aligned}$$

where n_k is the number of times k appears in the values taken by $\{x_i\}_{i=1}^{N}$.

The two models share the following notation: D is the number of documents in the corpus, w_{di} is the ith word of document d, and $n_k^{(d)}$ is the number of words in document m associated with topic k. $m_k^{(v)}$ is the number of times words corresponding to the index v are associated with topic k. A dot in the sub- or superscript in the $n_k^{(d)}$ term expresses summation. For example, the number of words in document d is $n_\cdot^{(d)} = \sum_k n_k^{(d)}$.

Topic models as a group make reference to Latent Dirichlet Allocation (LDA), which may be seen as a Bayesian recasting of probabilistic Latent Semantic Analysis (Blei, Ng, and Jordan 2003; Hofmann 1999). A virtue of LDA is that it provides a

generative description of a corpus, permitting predictions to be made about unseen documents. LDA assumes the following generative model for the corpus:

1. For $k = 1, \ldots, K$
   (a) draw topic distribution over words $\beta_k \sim Dirichlet(\eta)$.
2. For $d = 1, \ldots, D$
   (a) draw document-specific mixture weights $\theta_d \sim Dirichlet(\alpha_1, \ldots, \alpha_K)$
   (b) for $i = 1, \ldots, n_{\cdot}^{(d)}$
      i. draw topic index $z_{di} \sim Discrete(\theta_d)$
      ii. draw word $w_{di} \sim Discrete(\beta_{z_{di}})$

As the notation indicates, the prior distribution for each β_k is a symmetric Dirichlet distribution with scalar parameter η. Using an asymmetric prior distribution for the document-specific topic proportions θ_d has been discussed in Wallach, Mimno, and McCallum (2009).

The joint probability of the LDA model has the following factorization

$$p(\mathbf{w}, \mathbf{z}, \theta_{1:D}, \beta_{1:K} | \boldsymbol{\alpha}, \eta) = \prod_{k=1}^{K} Dir(\beta_k | \eta) \times \prod_{d=1}^{D} Dir(\theta_d | \boldsymbol{\alpha}) \times \prod_{d=1}^{D} \prod_{i=1}^{n_{\cdot}^{(d)}} Discrete(x_{di} | \beta_{z_{di}})$$

Taking advantage of the conjugacy between the Dirichlet prior distribution and the multinomial distribution discussed previously, we may integrate over $\theta_{1:D}$ and $\beta_{1:K}$

$$p(\mathbf{w}, \mathbf{z}|\boldsymbol{\alpha}, \eta) = \prod_{d=1}^{D} \left(\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\Gamma(n_{\cdot}^{(d)} + \sum_{k=1}^{K} \alpha_k)} \prod_{k=1}^{K} \frac{\Gamma(n_k^{(d)} + \alpha_k)}{\Gamma(\alpha_k)} \right) \times$$
$$\prod_{k=1}^{K} \left(\frac{\Gamma(V\eta)}{\Gamma(m_k^{(\cdot)} + V\eta)} \prod_{v=1}^{V} \frac{\Gamma(m_k^{(v)} + \eta)}{\Gamma(\eta)} \right)$$

The HDP model benefits from comparison with LDA and integrating out the Dirichlet distribution is required for posterior sampling in both models.

3.2.1 Dirichlet Process Mixture of Unigrams

The first and simplest topic model is a DP mixture of unigrams, a non-parametric counterpart to the mixture of unigrams model (Nigam et al. 1999). This primitive topic model may be expressed as follows

$$\begin{aligned} G|\alpha, G_0 &\sim DP(\alpha, G_0) \\ \theta_m|G &\sim G \quad m \in \{1, 2, \ldots, M\} \\ w_{mi}|\theta_m &\sim Discrete(\theta_m)\, i \in \{1, 2, \ldots, n_{\cdot}^{(m)}\} \end{aligned}$$

where DP denotes a Dirichlet Process (Ferguson 1973). The base distribution G_0 is a symmetric Dirichlet distribution with parameter η. Inference for Dirichlet Process mixture models can be performed by Gibbs sampling. Moreover, a simple auxiliary variable sampling method is available if a Gamma hyperprior is used to characterize uncertainty about α (Escobar and West 1995).

3.2.2 Hierarchical Dirichlet Process

The second model uses the Hierarchical Dirichlet Process (HDP) to model the words in the corpus. Each document is modeled as a mixture of the latent topic distributions

in the corpus. The generative model described by the HDP is given by

$$
\begin{aligned}
G_0|\gamma, H &\sim DP(\gamma, H) \\
G_j|\alpha_0, G_0 &\sim DP(\alpha_0, G_0) \quad j \in \{1, \ldots, M\} \\
\theta_{ij} &\sim G_j \quad i \in \{1, \ldots, n_i\} \\
w_{ij}|\theta_{ij} &\sim Discrete(\theta_{ij})
\end{aligned}
$$

where the base distribution H is a symmetric Dirichlet distribution with parameter η. Posterior inference using Gibbs sampling is described in Teh et al. (2006).

4

Experimental Results

In this chapter the two models are compared in terms of predictive accuracy and in terms of the alignment of their distributions of latent features with the classifications provided by literary historians.

4.1 Prior Specification

Both models share a set of topic distributions, $\{\beta_1, \beta_2, \ldots\}$. The prior specification holds that these distributions are drawn from a symmetric Dirichlet distribution with parameter η. The value assigned to η is typically interpreted as either a parameter for a probability distribution describing prior beliefs about the topic distributions or as "prior data"—a smoothing pseudo-count. That the Dirichlet distribution associated with η is symmetric captures the belief that no word is more or less likely to occur under a given topic a priori. Using η as an expression of prior belief is complicated in non-parametric models by the tight coupling between the parameter η and the number of topics in the corpus. Higher values of η permit each topic to capture more variability, leading to topics that can account for a greater number of words observed in the corpus. Assignments of 0.1 and 0.01 are common in the literature

and have been reported as performing well in applications (Griffiths and Steyvers 2004). For both models, η is modeled with a weakly informative $Gamma(0.1, 0.1)$ prior distribution. As this is not a conjugate prior, inference about η makes use of a slice sampler (Neal 2003).

In the DP mixture of unigrams model, the parameter α influences the prior number of topics. With α fixed and n draws from the DP, the expected number of mixture components is given by

$$\sum_{i=1}^{n} \frac{\alpha}{\alpha + i - 1}$$

This result follows from the fact that the probability of a new mixture component—i.e., a draw from the DP being a draw from the base distribution—after n draws is $\alpha/(\alpha + n)$. The role played by α in the DP is analogous to the role is played by the parameter γ in the HDP model. The corpus under consideration here contains 93 novels. If the value of α were 10 then the expected number of distinct mixture distributions would be approximately 24. Given prior work in topic modeling and the intuition that there need to be a reasonable number of clusters—typically well in excess of ten—to represent a heterogeneous corpus, I use a broad $Gamma(1, 0.1)$ hyperprior to reflect prior beliefs about α. Such a distribution is also used for γ in the HDP model.

An additional DP figures in the HDP model, $G_j \sim DP(\alpha_0, G_0)$ and the role played by α_0 is that of a precision parameter and influences the fidelity of draws from the DP to the base distribution G_0. A vague $Gamma(0.1, 0.1)$ hyperprior on α_0 reflects the prior belief that the corpus-level DP distribution G_0 should minimally influence the topics appearing in any given document.

4.2 Comparing Predictive Accuracy

The per-word predictive probability of a held-out portion of a dataset is a standard measure of how well a model generalizes to new data. Higher predictive probability indicates the model is better at predicting unseen words.

Calculating the predictive probability of a held-out set of novels is challenging under the HDP model because the number of possible topic assignments grows exponentially with the number of words in the held-out corpus. A range of approximation strategies exist and are a topic of study in their own right (Wallach et al. 2009). I have chosen to pursue a document-completion strategy. While the strategy performs worse than the sequential strategies detailed in Wallach et al. (2009) and Buntine (2009), it is preferable to a number of alternatives—such as the harmonic mean estimator—and relatively easy to implement. It has been used elsewhere in the topic modeling literature (Newman et al. 2009). Under a document-completion strategy, the corpus is randomly partitioned into five "folds" of roughly equal size. In the present case, the share of novels from each genre is also held approximately constant. Each fold contains roughly twenty percent of the novels and each, in turn, serves as a held-out test corpus. Under the document-completion strategy the half the words (selected at random) from each novel in the test corpus are removed and added back into the training corpus in order to estimate the topic proportions for that novel. The words remaining in the test corpus are not used to fit the model.

For each fold the training corpus is fit using the models starting from three different random initializations. For each run, the per-word log predictive probability of each document is calculated over 600 iterations of the Markov Chain Monte Carlo (MCMC) chain, where the first 900 iterations are discarded and samples are taken every thirty iterations of the chain. The results of these calculations are shown in figure 4.1. The HDP model performs considerably better than the DP mixture model

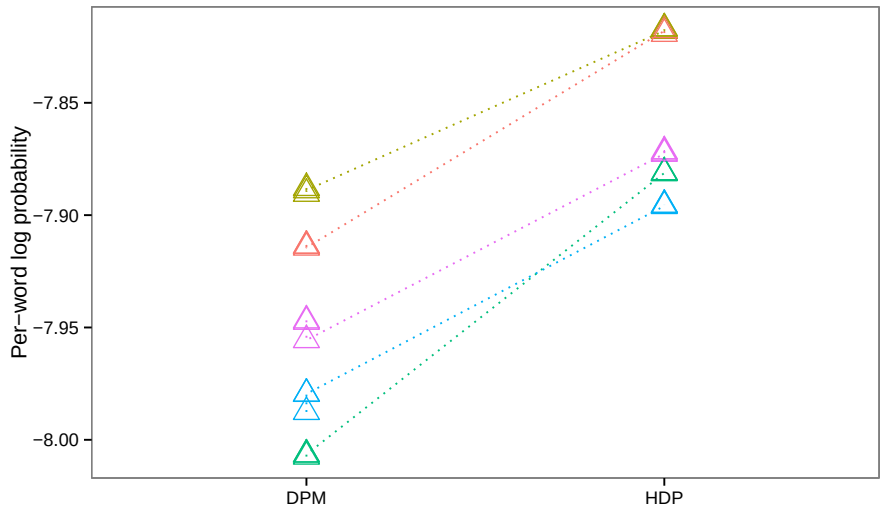

FIGURE 4.1: Comparison of predictive performance on held-out portions of novels. Points represent the log predictive probability of the test corpus for that fold. Dotted lines connect folds, permitting comparison of models' performance on the same held-out portions of novels.

at predicting the held-out portions of novels in the test corpus.

4.3 Comparing Clusterings

With the HDP model the topic weights associated with each document may be considered as a form of soft classification and compared directly with the expert classifications. A similar comparison may be made between the expert classifications and the distribution over mixture components for the DP mixture of unigrams. Normalized mutual information provides a convenient metric for these comparisons. If C is the class of expert classifications ("gothic," "silver fork," and "national tale") and C' is the set of topics inferred by a topic model, then the mutual information $MI(C, C')$ between the expert classifications and a model is given by

$$MI(C, C') = \sum_{c_i \in C, c'_j \in C'} p(c_i, c'_j) \log_2 \frac{p(c_i, c'_j)}{p(c_i)p(c'_j)}$$

where $p(c_i)$ and $p(c'_j)$ denote the probability that a novel selected at random from the corpus falls into c_i and c'_j respectively. $p(c_i, c'_j)$ is the probability that a document selected at random falls into both c_i and c'_j. Using a normalized variant of mutual information facilitates comparison of different models. Normalized mutual information (NMI) is calculated as follows

$$NMI(C, C') = \frac{MI(C, C')}{\max(H(C), H(C'))}$$

where $H(C)$ and $H(C')$ are the respective entropies of the two clusterings being compared. An approximation averaging different samples from the posterior distribution of latent features for each model can be made despite the problem of non-identifiability in topic models ("label switching") because mutual information is invariant to changes in class labels.

The conceptual switch between "topic" and class (or cluster) has an important precedent. Formally the same as LDA, the mixed-membership model developed independently by Pritchard, Stephens, and Donnelly (2000) was used to model population structure of biological species. In the application discussed by Pritchard, Stephens, and Donnelly individual organisms were associated with one another based on genetic similarity. In the present application, the association between the latent features of the model (topics) and (sub)populations is more explicit than often is the case with LDA and other topic models, where topics are often used as a form of document summarization.

When asking how closely the latent space of topics derived from one of the topic models aligns with the expert classifications, it is useful to have a baseline for comparison. A natural baseline in this case is a classification arrived at by randomly assigning novels to one of K categories, where K is the number of categories used by the experts. Because distinguishing between a gothic and a randomly selected

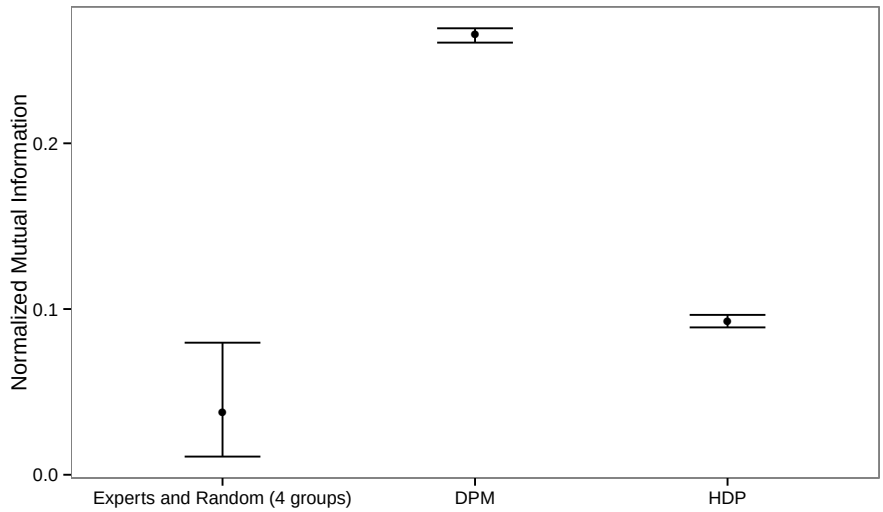

FIGURE 4.2: Similarity of expert genre classifications and the clustering given by the two models. Similarity is measured in terms of normalized mutual information, with one indicating the clusterings are indistinguishable and zero indicating no or minimal overlap. Error bars indicate 95% credible intervals.

(non-gothic) novel is equally important as distinguishing between a gothic and, for instance, a silver fork novel, the randomly selected novels are counted as their own category. By repeatedly generating a random clustering of the novels and calculating the normalized mutual information between that clustering and the expert classifications, we can establish a baseline that any model will need to beat if it is to be said to align with the expert classifications better than chance. It is worth recalling that this is a severe test of the models, as it makes use of information about the number of expert classifications. This information is not used by the models.

In both cases the mutual information between the respective model and the expert classifications is calculated over 600 iterations of the Markov Chain Monte Carlo (MCMC) chain, where the first 900 iterations are discarded and samples are taken every thirty iterations of the chain. As figure 4.2 shows, the DP mixture of unigrams aligns with expert classifications better than chance and the HDP model aligns with expert classifications better than chance in more than .99 of the cases.

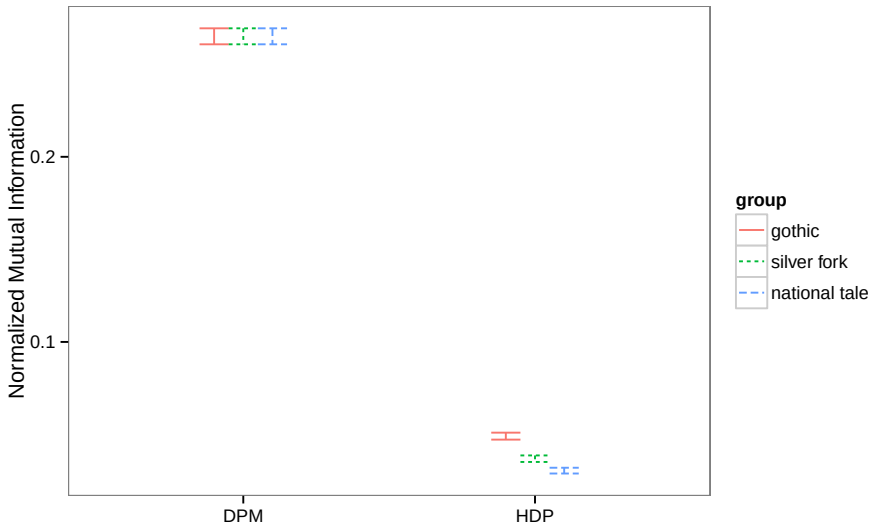

FIGURE 4.3: Similarity of expert genre classifications and the clustering given by the two models, each genre considered separately. Similarity is measured in terms of normalized mutual information, with one indicating the clusterings are indistinguishable and zero indicating no or minimal overlap. Error bars indicate 95% credible intervals.

Comparing the clusterings of the models with expert classifications one genre at a time also provides information about the general characteristics of the genres. As has been discussed above, literary historians have suggested that novels in some genres may be characterized more by common vocabulary, whereas others may exhibit distinctive plot structure or paratextual features. As both of the models infer latent features from word frequency data alone, the results shown in figure 4.3 suggest that the gothic novels in this corpus are more readily identified by their vocabulary use than the other two genres.

4.4 Discussion

In the HDP mixed membership model topics emerge that capture preconceptions about shared vocabulary for the three genres. For example, one topic that occurs frequently among silver fork novels is characterized by the words: "family," "society," "influence," and "marriage" (Topic 6). A topic found frequently in gothic novels

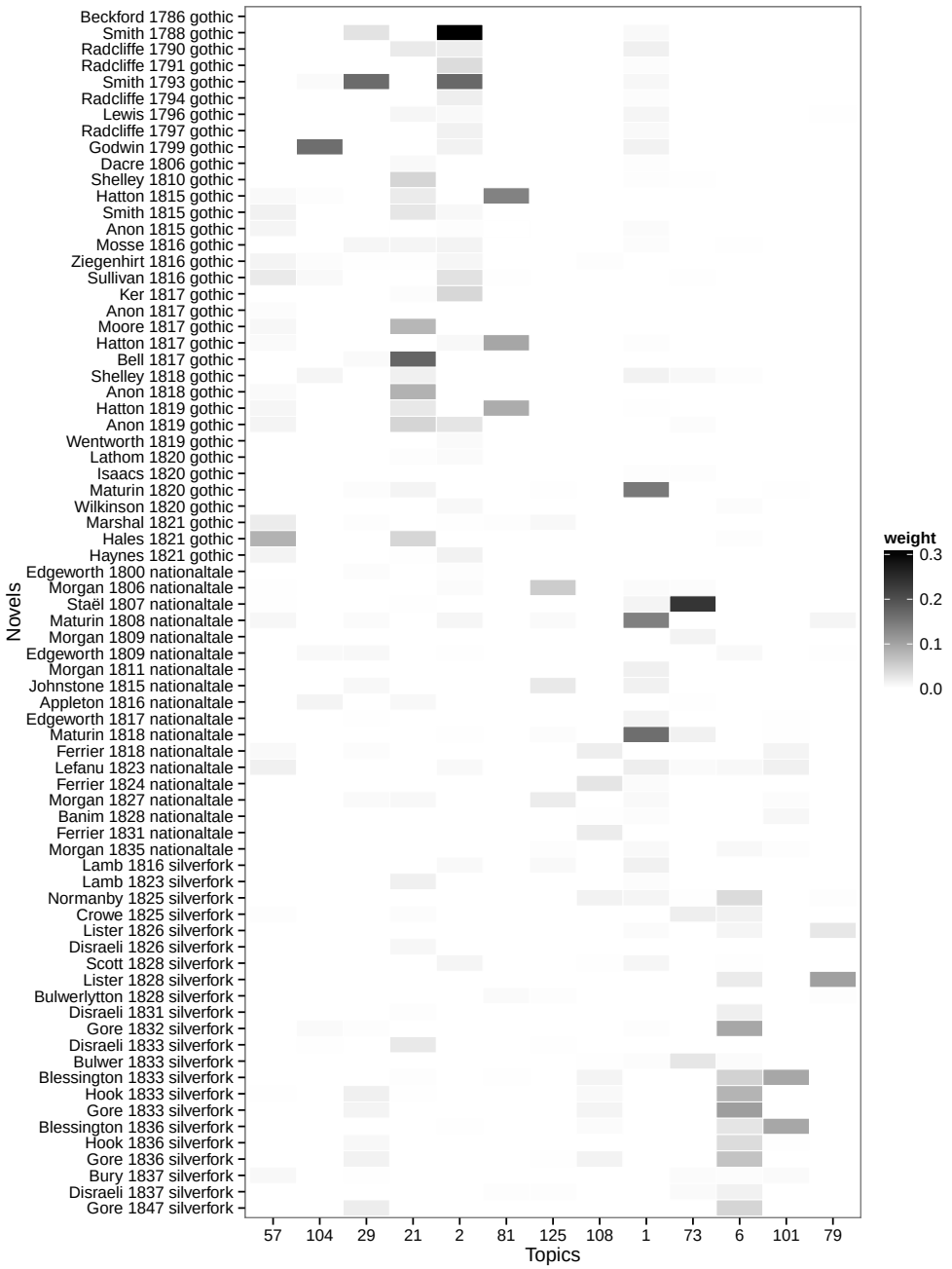

FIGURE 4.4: Selected HDP topic concentrations for gothic, national tale, and silver fork novels. Topics have been selected to highlight areas of alignment between expert classifications and latent features.

features words such as “fear,” “murder,” “dark,” “power,” and “terror” (Topic 21). A topic featuring words such as “ancestor,” “national,” “harp,” and “desolate” appears frequently among national tale novels (Topic 125). This informal inspection of topic distributions and their association with documents (fig. 4.4) while it yields results that appear to confirm the salience of the model is in other respects akin to reading tea leaves: we may be selecting out of the various high probability words those that match our expectations. It is easy to overlook that the topic distributions also assigns high probability to words less consonant with our preconceptions, such as “hitherto” for national tale, “summer” for silver fork, and “instruction” for gothic. For this reason, it is the evaluation in terms of mutual information (sec. 4.3) that counts as systematic evidence that the model is picking up on characteristics across the entire corpus that align with expert judgments.

That there is no simple mapping between genre and latent features seems more likely to be an important reminder of the heterogeneity in novelistic production than any failing on the part of the models under consideration. There are no unambiguous lexical signals of genre. Some novels assigned to a genre conform closely to stereotypes, whereas others do not. That the distribution of latent features nonetheless aligns with expert judgments better than chance suggests that it may still be appropriate to speak of “family resemblance” of novels associated with the same genre. It is also important to recall that the foregoing study has only made use of word frequencies. Distinctions between novels judged to be members of different genres may become much sharper (or less sharp) when additional features are considered, such as narrative structure.

4.4.1 How Topic Models Interpret Literary Historians

Comparing topic models to the standard of expert judgment is complicated by the fact that the expert classifications are not guaranteed to provide a “gold standard”

or "ground truth." Experts have been known exaggerate the similarity between cultural artifacts as well as the distinctiveness of types (Sperber 1996, 32–55). Expert classifications may have been made hastily or may be inaccurate for any number of reasons (Kahneman, Slovic, and Tversky 1982). Even cases where it seems clear that a genre classification is appropriate, such as one made in the light of explicit identification with the genre by the publisher or author—e.g., the novel has a subtitle that includes "national tale"—need to be treated with care. Both in the nineteenth century and in the present there are examples of "cryptic" genre novels. The clearest cases are those involving publishers or writers seeking to disassociate their novels with a "popular" genre. Margaret Atwood's disavowal of the label "science fiction" for her novels—such as *Oryx and Crake* and *The Year of the Flood*—is a contemporary example (Canavan and Wald 2011). Walter Scott's novels provide an example from the early nineteenth century. Scott's *Waverly* novels, commonly referred to as historical novels, have been described as cryptic national tale novels (Garside 1991).

For these reasons, a topic model of a corpus of novels that performs well by the standard of posterior predictive checks or held-out predictive probability may deserve attention even if the model does not align best with human readers' judgments. The topic space inferred by a predictively accurate topic model would merit attention especially in cases where there were evidence that the predictive accuracy in question was attributable to a successful modeling of word frequencies (or other features) clearly connected to morphology believed to be relevant for classification. In the present case, if topics inferred by the model assigned high probability to specific words unambiguously tied to a genre in question (such as "dungeon" or "ghost" in the case of gothic novels) and yet the latent space nevertheless did not align with the expect classifications, it would be appropriate to consider the possibility that the expert classifications were incomplete. To the extent that the model and the expert are believed to be "reasoning" about the same thing (genre classification)

using different kinds of evidence, a disagreement between the two should prompt further investigation rather than dismissal of one of the two approaches.

4.4.2 Persistent Gaps

There are important gaps between the inferred "topics" of topic models and topics familiar to human readers. Many of these gaps are familiar, as is the caution to avoid reifying what are discrete probability distributions over words (Blei, Ng, and Jordan 2003, 996fn1). Frequent "non-informative" words (stop words), when they are not removed, naturally appear among the words most strongly associated with topics. Few human readers, however, would associate such words with a recurring topic in a collection of texts. Proper nouns strongly linked with topics are another instance where topic models focus attention on words that human readers would likely not. In their study of how human readers evaluate topic models of news and Wikipedia articles, Chang et al. (2009) found it necessary to remove all proper nouns "for the benefit of the human subjects" as "success in early experiments required too much encyclopedic knowledge." Proper nouns in the form of character names present a similar problem with this corpus. For instance, along with words typical of gothic novels, the names of characters in gothic novels are naturally associated with topics inferred by the topic models. Forms of address (frequently capitalized) such as "Mr.," "Miss," "Captain," and "Lord" also number among those words associated with inferred topics but seem unlikely to appear in any reader's lists of characteristic words. The full names of characters also provide an example of a familiar feature of literary texts that cannot be accommodated in the generative description of topic models: intentionally unique words. Idiosyncratic character names, such as "Mr. Simpkinson" or "Captain Hobkirk," are often specific to a novel (or novels) by a single author.[1] (The character names in the novels of Charles Dickens also provide

1. Names from *Gilbert Gurney* by Theodore Hook.

a host of such examples.) Such names are inappropriately modeled as being drawn from any distribution shared by other novels.

The models considered in this thesis also make specific assumptions known to be incorrect. The assumption that the appearance of a topic in a document is uncorrelated with the appearance of other topics is one example. Other examples include the assumption that topic distributions do not change over time and that, in the case of the HDP, the prevalence of a topic in a corpus is correlated with the proportion of words attributed to the topic in documents in which the topic appears. Topic models exist that permit these assumptions to be relaxed (Blei and Lafferty 2006, 2007; Williamson et al. 2010).

5

Conclusion

One feature of topic models that invites their use in literary historical work is that there is superficial agreement between the assumptions of mixed-membership models and literary-historical accounts of genre. Even the most commodified corners of novelistic production are home to a range of writers, each with their own particular style, favored settings, and plots. Such novels might be expected to be comprised of features drawn from a distribution over features (again, "topic" seems a useful shorthand) specific to the genre as well as from other distributions reflecting the idiosyncrasies of the author's style. This expectation roughly fits the assumptions of a mixed-membership model such as one arrived at via LDA or HDP. Mixed-membership models also accommodate the idea of novels explicitly "sampling" from the conventions of more than one genre. A contemporary example is Philip K. Dick's *Do Androids Dream of Electric Sheep* (1967) (also known by its film adaption, *Blade Runner* (1982)), which borrows from conventions in science fiction and detective stories (Kerman 1997). An example closer to our period would be Edward Bulwer-Lytton's bestselling *Paul Clifford* (1830), a story of a prosperous gentleman who also

leads a life as a criminal.[1] The novel has been justifiably classified as both a silver fork novel and an early Newgate novel (Adburgham 1983; Hollingsworth 1963).

It may be the case that certain latent feature models are better suited to specific novelistic genres. While literary historians comfortably mention Newgate novels, nautical tales, and national tale novels as instances of novelistic genres—that is, as instances of the same "kind" of thing—we do not know if the novels within each genre resemble each other in comparable ways. It is possible that one model of latent structure may model one novelistic genre whereas a different model proves better suited to other genres. The diversity of non-parametric models—phylogenetic, nested, correlated, dynamic, and so forth—presents amble opportunity for studying a remarkably well-preserved and well-documented part of cultural history.

The success of topic models—measured by their widespread use and their often uncanny ability to accurately summarize semantic themes within a large corpus—should not blind us to the fact that most models do not incorporate any theory about why novels share the features that they share. The original example of modeling allele frequencies via a mixed-membership model provides the best context to articulate this concern (Pritchard, Stephens, and Donnelly 2000). While a mixed-membership model may successfully recover salient aspects of the distribution of genetic features among related species, the model makes no acknowledgment about why species are related. Organisms share allele frequencies because they share ancestors. Given an exhaustive survey of the organisms in a population of interest, it would, I suspect, the ancestry that would be of interest, rather than the mere fact that many of the organisms were similar. For novels the question is roughly the same: Why do novels exhibit similar features? It is this question that is ultimately of interest to literary historians and sociologists of culture. Models that are able to incorporate and explore hypotheses about the relations among novels and other "influences," for

1. *Paul Clifford* is also remembered for its opening line: "It was a dark and stormy night."

lack of a better word, would command considerable attention.

In this thesis, I evaluated latent feature models of a corpus of novels by comparing the distribution of latent features with judgments of similarity made by literary historians. I found that the models align with expert classifications better than chance, demonstrating their suitability for use in literary-historical research.

Appendix: Novels Used

The following list of novels maintains the formatting found in Garside and Schöwerling (2000) and supplements (Garside et al. 2006).

1786	gothic	[BECKFORD, William]	[VATHEK]. AN ARABIAN TALE, FROM AN UNPUBLISHED MANUSCRIPT: WITH NOTES CRITICAL AND EXPLANATORY.
1788	gothic	SMITH, Charlotte	EMMELINE, THE ORPHAN OF THE CASTLE. BY CHARLOTTE SMITH. IN FOUR VOLUMES
1790	gothic	[RADCLIFFE, Ann]	A SICILIAN ROMANCE. BY THE AUTHORESS OF THE CASTLES OF ATHLIN AND DUNBAYNE. IN TWO VOLUMES.
1791	gothic	[RADCLIFFE, Ann]	THE ROMANCE OF THE FOREST: INTERSPERSED WITH SOME PIECES OF POETRY. BY THE AUTHORESS OF "A SICILIAN ROMANCE," &C. IN THREE VOLUMES.
1793	gothic	SMITH, Charlotte	THE OLD MANOR HOUSE. A NOVEL, IN FOUR VOLUMES. BY CHARLOTTE SMITH.
1794	gothic	RADCLIFFE, Ann	THE MYSTERIES OF UDOLPHO, A ROMANCE; INTERSPERSED WITH SOME PIECES OF POETRY. BY ANN RADCLIFFE, AUTHOR OF THE ROMANCE OF THE FOREST, ETC. IN FOUR VOLUMES
1796	gothic	LEWIS, M[atthew] G[regory]	THE MONK: A ROMANCE. IN THREE VOLUMES. BY M. G. LEWIS, ESQ. M.P.
1797	gothic	RADCLIFFE, Ann	THE ITALIAN, OR THE CONFESSIONAL OF THE BLACK PENTITENTS. A ROMANCE. BY ANN RADCLIFFE, AUTHOR OF THE MYSTERIES OF UDOLPHO, &C. &C. IN THREE VOLUMES.
1799	gothic	GODWIN, William	ST. LEON: A TALE OF THE SIXTEENTH CENTURY. BY WILLIAM GODWIN. IN FOUR VOLUMES
1800	nationaltale	Maria EDGEWORTH	CASTLE RACKRENT, AN HIBERNIAN TALE. TAKEN FROM FACTS, AND FROM THE MANNERS OF THE IRISH SQUIRES, BEFORE THE YEAR 1782
1801	random	ANON	MYSTERIOUS FRIENDSHIP: A TALE. IN TWO VOLUMES
1801	random	Mary CHARLTON	THE PIRATE OF NAPLES. A NOVEL. IN THREE VOLUMES. BY MARY CHARLTON, AUTHOR OF ROSELLA, ANDRONICA, PHEDORA, &C
1805	random	Isaac D'ISRAELI	FLIM-FLAMS! OR, THE LIFE AND ERRORS OF MY UNCLE, AND THE AMOURS OF MY AUNT! WITH ILLUSTRATIONS AND OBSCURITIES, BY MESSIEURS TAG, RAG, AND BOBTAIL. WITH AN ILLUMINATING INDEX! IN THREE VOLUMES, WITH NINE PLATES
1806	random	ANON	FORRESTI; OR, THE ITALIAN COUSINS. A NOVEL. IN THREE VOLUMES. BY THE AUTHOR OF VALAMBROSA [sic]
1806	gothic	Charlotte DACRE	ZOFLOYA; OR, THE MOOR: A ROMANCE OF THE FIFTEENTH CENTURY. IN THREE VOLUMES. BY CHARLOTTE DACRE, BETTER KNOWN AS ROSA MATILDA, AUTHOR OF THE NUN OF ST. OMERS, HOURS OF SOLITUDE, &C
1806	nationaltale	Sydney OWENSON [afterwards MORGAN, Lady Sydney]	THE WILD IRISH GIRL; A NATIONAL TALE. BY MISS OWENSON, AUTHOR OF ST. CLAIR, THE NOVICE OF ST. DOMINICK, &C. &C. &C. IN THREE VOLUMES
1807	nationaltale	Anne Louise Germaine de STAËL-HOLSTEIN	CORINNA; OR, ITALY. BY MAD. DE STAËL HOLSTEIN. IN THREE VOLUMES

1808	nationaltale	Charles Robert MATURIN	THE WILD IRISH BOY. IN THREE VOLUMES. BY THE AUTHOR OF MONTORIO
1808	random	Ellen Rebecca WARNER	HERBERT-LODGE; A NEW-FOREST STORY. IN THREE VOLUMES. BY MISS WARNER, OF BATH
1809	nationaltale	Maria EDGEWORTH	TALES OF FASHIONABLE LIFE, BY MISS EDGEWORTH, AUTHOR OF PRACTICAL EDUCATION, BELINDA, CASTLE RACKRENT, ESSAY ON IRISH BULLS, &C. IN THREE VOLUMES
1809	nationaltale	Sydney OWENSON [afterwards MORGAN, Lady Sydney]	WOMAN: OR, IDA OF ATHENS. BY MISS OWENSON, AUTHOR OF THE "WILD IRISH GIRL," THE "NOVICE OF ST. DOMINICK," &C. IN FOUR VOLUMES
1810	gothic	Percy Bysshe SHELLEY	ZASTROZZI, A ROMANCE. BY P. B. S
1811	nationaltale	Sydney OWENSON [afterwards MORGAN, Lady Sydney]	THE MISSIONARY: AN INDIAN TALE. BY MISS OWENSON. WITH A PORTRAIT OF THE AUTHOR. IN THREE VOLUMES
1815	gothic	Anne Julia Kemble HATTON	SECRET AVENGERS; OR, THE ROCK OF GLOTZDEN. A ROMANCE. IN FOUR VOLUMES. BY ANNE OF SWANSEA, AUTHOR OF CAMBRIAN PICTURES; SICILIAN MYSTERIES; CONVICTION, &C. &C
1815	gothic	ANON	THERESA; OR, THE WIZARD'S FATE. A ROMANCE. IN FOUR VOLUMES. BY A MEMBER OF THE INNER TEMPLE
1815	gothic	ANON	DANGEROUS SECRETS. A NOVEL. IN TWO VOLUMES
1815	gothic	Catherine SMITH	BAROZZI; OR THE VENETIAN SORCERESS. A ROMANCE OF THE SIXTEENTH CENTURY. IN TWO VOLUMES. BY MRS. SMITH, AUTHOR OF THE CALEDONIAN BANDIT, &C. &C
1815	nationaltale	Christian Isobel JOHNSTONE	CLAN-ALBIN: A NATIONAL TALE. IN FOUR VOLUMES
1816	nationaltale	Elizabeth APPLETON	EDGAR: A NATIONAL TALE. BY MISS APPLETON, AUTHOR OF PRIVATE EDUCATION, &C. IN THREE VOLUMES
1816	gothic	Henrietta Rouviere MOSSE	CRAIGH-MELROSE PRIORY; OR, MEMOIRS OF THE MOUNT LINTON FAMILY. A NOVEL. IN FOUR VOLUMES
1816	silverfork	Lady Caroline LAMB	GLENARVON. IN THREE VOLUMES
1816	gothic	Mary Ann SULLIVAN	OWEN CASTLE, OR, WHICH IS THE HEROINE? A NOVEL. IN FOUR VOLUMES. DEDICATED BY PERMISSION TO THE RIGHT HONOURABLE LADY COMBERMERE, BY MARY ANN SULLIVAN, LATE OF THE THEATRES ROYAL, LIVERPOOL, MANCHESTER, NEWCASTLE, BIRMINGHAM, AND NORWICH
1816	gothic	Sophia F. ZIEGENHIRT	THE ORPHAN OF TINTERN ABBEY. A NOVEL. IN THREE VOLUMES. BY SOPHIA F. ZIEGENHIRT, AUTHOR OF SEABROOK VILLAGE, AND SEVERAL HISTORICAL ABRIDGEMENTS
1817	gothic	Anne Julia Kemble HATTON	GONZALO DE BALDIVIA; OR, A WIDOW'S VOW. A ROMANTIC LEGEND. IN FOUR VOLUMES. INSCRIBED, BY PERMISSION, TO WILLIAM WILBERFORCE, ESQ. BY THE AUTHOR OF CAMBRIAN PICTURES, SICILIAN MYSTERIES, CONVICTION, SECRET AVENGERS, CHRONICLES OF AN ILLUSTRIOUS HOUSE, &C. &C
1817	gothic	Anne KER	EDRIC, THE FORESTER: OR, THE MYSTERIES OF THE HAUNTED CHAMBER. AN HISTORICAL ROMANCE, IN THREE VOLUMES. BY MRS. ANNE KER, OF HIS GRACE THE DUKE OF ROXBURGH'S FAMILY, AUTHOR OF THE HEIRESS DI MONTALDE—ADELINE ST. JULIAN—EMMELINE, OR THE HAPPY DISCOVERY—MYSTERIOUS COUNT—AND MODERN FAULTS
1817	gothic	ANON	HOWARD CASTLE; OR A ROMANCE FROM THE MOUNTAINS. IN FIVE VOLUMES. BY A NORTH BRITON
1817	gothic	Edward MOORE	THE MYSTERIES OF HUNGARY. A ROMANTIC HISTORY, OF THE FIFTEENTH CENTURY. IN THREE VOLUMES. BY EDWARD MOORE, ESQ. AUTHOR OF SIR RALPH DE BIGOD, &C. &C
1817	nationaltale	Maria EDGEWORTH	HARRINGTON, A TALE; AND ORMOND, A TALE. IN THREE VOLUMES. BY MARIA EDGEWORTH, AUTHOR OF COMIC DRAMAS, TALES OF FASHIONABLE LIFE, &C. &C

1817	gothic	Nugent BELL	ALEXENA; OR, THE CASTLE OF SANTA MARCO, A ROMANCE, IN THREE VOLUMES. EMBELLISHED WITH ENGRAVINGS
1818	gothic	ANON	THE BANDIT CHIEF; OR, LORDS OF URVINO. A ROMANCE. IN FOUR VOLUMES
1818	nationaltale	Charles Robert MATURIN	WOMEN; OR, POUR ET CONTRE. A TALE. BY THE AUTHOR OF "BERTRAM," &C. IN THREE VOLUMES
1818	gothic	Mary Wollstonecraft SHELLEY	FRANKENSTEIN; OR, THE MODERN PROMETHEUS. IN THREE VOLUMES
1818	nationaltale	Susan Edmonstone FERRIER	MARRIAGE, A NOVEL. IN THREE VOLUMES
1819	random	Adelaide O'KEEFFE	DUDLEY. BY MISS O'KEEFFE, AUTHOR OF PATRIARCHAL TIMES, OR THE LAND OF CANAAN; ZENOBIA, QUEEN OF PALMYRA; &C. IN THREE VOLUMES
1819	gothic	Anne Julia Kemble HATTON	CESARIO ROSALBA; OR, THE OATH OF VENGEANCE. A ROMANCE. IN FIVE VOLUMES. BY ANN OF SWANSEA, AUTHOR OF SICILIAN MYSTERIES, CONVICTION, GONZALO DE BALDIVIA, SECRET AVENGERS, SECRETS IN EVERY MANSION, CAMBRIAN PICTURES, CHRONICLES OF AN ILLUSTRIOUS HOUSE, &C
1819	gothic	ANON	THE CASTLE OF VILLA-FLORA. A PORTUGUESE TALE, FROM A MANUSCRIPT LATELY FOUND BY A BRITISH OFFICER OF RANK IN AN OLD MANSION IN PORTUGAL. IN THREE VOLUMES
1819	random	Elizabeth BENNETT	EMILY, OR, THE WIFE'S FIRST ERROR; AND BEAUTY & UGLINESS, OR, THE FATHER'S PRAYER AND THE MOTHER'S PROPHECY. TWO TALES. IN FOUR VOLUMES. BY ELIZABETH BENNET, AUTHOR OF FAITH AND FICTION, &C. &C
1819	gothic	Zara WENTWORTH	THE RECLUSE OF ALBYN HALL. A NOVEL. IN THREE VOLUMES. BY ZARA WENTWORTH
1820	gothic	Charles Robert MATURIN	MELMOTH THE WANDERER: A TALE. BY THE AUTHOR OF "BERTRAM," &C. IN FOUR VOLUMES
1820	gothic	Francis LATHOM	ITALIAN MYSTERIES; OR, MORE SECRETS THAN ONE. A ROMANCE. IN THREE VOLUMES. BY FRANCIS LATHOM, AUTHOR OF THE MYSTERIOUS FREEBOOTER; LONDON; THE UNKNOWN; MEN AND MANNERS; ROMANCE OF THE HEBRIDES; HUMAN BEINGS; FATAL VOW; MIDNIGHT BELL; IMPENETRABLE SECRET; MYSTERY; &C. &C
1820	gothic	Mrs ISAACS	EARL OSRIC; OR, THE LEGEND OF ROSAMOND. A ROMANCE. BY MRS. ISAACS, AUTHOR OF "TALES OF TO-DAY,"—"WANDERINGS OF FANCY," &C. &C. &C. IN THREE VOLUMES
1820	gothic	Sarah Scudgell WILKINSON	THE SPECTRE OF LANMERE ABBEY, OR THE MYSTERY OF THE BLUE AND SILVER BAG; A ROMANCE. BY SARAH WILKINSON: AUTHORESS OF THE BANDIT OF FLORENCE, FUGITIVE COUNTESS, WHEEL OF FORTUNE, &C. IN TWO VOLUMES
1821	gothic	J. M. H. HALES	DE WILLENBERG; OR, THE TALISMAN. A TALE OF MYSTERY. IN FOUR VOLUMES. BY I. M. H. HALES, ESQ. AUTHOR OF THE ASTROLOGER
1821	gothic	Miss C. D. HAYNES [afterwards GOLLAND, Mrs C. D.]	ELEANOR; OR, THE SPECTRE OF ST. MICHAEL'S. A ROMANTIC TALE. IN FIVE VOLUMES. BY MISS C. D. HAYNES, AUTHOR OF CASTLE LE BLANC; FOUNDLING OF DEVONSHIRE; AUGUSTUS AND ADELINA, &C. &C
1821	gothic	Thomas Henry MARSHAL	THE IRISH NECROMANCER; OR, DEER PARK. A NOVEL. IN THREE VOLUMES. BY THOMAS HENRY MARSHAL
1822	random	Isabel HILL	CONSTANCE, A TALE. BY ISABEL HILL, AUTHOR OF 'THE POET'S CHILD,' A TRAGEDY
1822	random	Jean Charles Léonard SIMONDE DE SISMONDI	JULIA SEVERA; OR THE YEAR FOUR HUNDRED AND NINETY-TWO; TRANSLATED FROM THE FRENCH OF J. C. L. SIMONDE DE SISMONDI, AUTHOR OF NEW PRINCIPLES OF POLITICAL ECONOMY; THE HISTORY OF FRANCE; THE ITALIAN REPUBLICS OF THE MIDDLE AGE; THE LITERATURE OF THE SOUTH OF EUROPE, &C. IN TWO VOLUMES

1823	nationaltale	Alicia LEFANU	TALES OF A TOURIST. CONTAINING THE OUTLAW, AND FASHIONABLE CONNEXIONS. IN FOUR VOLUMES. BY MISS LEFANU, AUTHOR OF STRATHALLAN, LEOLIN ABBEY, HELEN MONTEAGLE, &C
1823	random	George JONES	TEMPTATION. A NOVEL. BY LEIGH CLIFFE, AUTHOR OF "THE KNIGHTS OF RITZBERG,"—"PARGA," "SUPREME BON TON," &C. IN THREE VOLUMES
1823	silverfork	Lady Caroline LAMB	ADA REIS, A TALE. IN THREE VOLUMES
1824	random	Hannah Maria JONES	THE FORGED NOTE: OR, JULIAN AND MARIANNE. A MORAL TALE, FOUNDED ON RECENT FACTS. BY MRS. H. M. JONES, AUTHORESS OF GRETNA GREEN,—WEDDING RING,—BRITISH OFFICER, &C
1824	nationaltale	Susan Edmonstone FERRIER	THE INHERITANCE. BY THE AUTHOR OF MARRIAGE. IN THREE VOLUMES
1825	silverfork	Constantine Henry, Marquis of Normanby PHIPPS	MATILDA; A TALE OF THE DAY
1825	silverfork	Eyre Evans CROWE	THE ENGLISH IN ITALY. IN THREE VOLUMES
1826	silverfork	Benjamin, Earl of Beaconsfield DISRAELI	VIVIAN GREY
1826	random	Sir Walter SCOTT	WOODSTOCK; OR, THE CAVALIER. A TALE OF THE YEAR SIXTEEN HUNDRED AND FIFTY-ONE. BY THE AUTHOR OF "WAVERLEY, TALES OF THE CRUSADERS," &C. IN THREE VOLUMES
1826	silverfork	Thomas Henry LISTER	GRANBY. A NOVEL. IN THREE VOLUMES
1827	random	Sarah Wilmot WELLS	TALES; MOURNFUL, MIRTHFUL, AND MARVELLOUS. BY MRS. WILMOT WELLS, OF MARGATE. IN THREE VOLUMES
1827	nationaltale	Sydney OWENSON [afterwards MORGAN, Lady Sydney]	THE O'BRIENS AND THE O'FLAHERTYS; A NATIONAL TALE. BY LADY MORGAN. IN FOUR VOLUMES
1828	silverfork	Edward George BULWER LYTTON	PELHAM; OR, THE ADVENTURES OF A GENTLEMAN. IN THREE VOLUMES
1828	nationaltale	John BANIM	THE ANGLO-IRISH OF THE NINETEENTH CENTURY. A NOVEL. IN THREE VOLUMES
1828	silverfork	Lady Caroline Lucy SCOTT	A MARRIAGE IN HIGH LIFE. EDITED BY THE AUTHORESS OF 'FLIRTATION.' IN TWO VOLUMES
1828	silverfork	Thomas Henry LISTER	HERBERT LACY. BY THE AUTHOR OF GRANBY. IN THREE VOLUMES
1831	silverfork	[DISRAELI, Benjamin, Earl of Beaconsfield]	THE YOUNG DUKE. BY THE AUTHOR OF "VIVIAN GREY." IN THREE VOLUMES.
1831	nationaltale	Ferrier	DESTINY; OR, THE CHIEF'S DAUGHTER. BY THE AUTHOR OF "MARRIAGE," AND "THE INHERITANCE."
1831	random	REYNOLDS, Frederick	A PLAYWRIGHT'S ADVENTURES
1832	silverfork	[GORE, Catharine Grace Frances]	THE OPERA: A NOVEL. BY THE AUTHOR OF "MOTHERS AND DAUGHTERS." IN THREE VOLUMES.
1833	silverfork	[BULWER LYTTON, Edward George]	GODOLPHIN. A NOVEL. IN THREE VOLUMES.
1833	silverfork	[DISRAELI, Benjamin, Earl of Beaconsfield]	THE WONDROUS TALE OF ALROY. THE RISE OF ISKANDER. BY THE AUTHOR OF "VIVIAN GREY," "CONTARINI FLEMING," &C.; IN THREE VOLUMES.
1833	silverfork	[GARDINER, Marguerite], Countess of Blessington	THE REPEALERS. A NOVEL. BY THE COUNTESS OF BLESSINGTON. IN THREE VOLUMES
1833	silverfork	[GORE, Catherine Grace Frances]	THE SKETCH BOOK OF FASHION. BY THE AUTHOR OF "MOTHERS AND DAUGHTERS." IN THREE VOLUMES.
1833	silverfork	[HOOK, Theodore Edward]	LOVE AND PRIDE. BY THE AUTHOR OF "SAYINGS AND DOINGS," ETC. IN THREE VOLUMES.
1833	random	[TONNA], Charlotte Elizabeth	DERRY, A TALE OF THE REVOLUTION. BY CHARLOTTE ELIZABETH, AUTHORESS OF OSRIC, THE ROCKITE, THE SYSTEM, &C.; &C.;
1835	random	CAUNTER, J[ohn] Hobart	POSTHUMOUS RECORDS OF A LONDON CLERGYMAN. EDITED BY THE REV. HOBART CAUNTER, B.D., AUTHOR OF THE ORIENTAL ANNUAL.
1835	random	[DEACON, William Frederick]	THE EXILE OF ERIN; OR, THE SORROWS OF A BASHFUL IRISHMAN. IN TWO VOLUMES.
1835	nationaltale	MORGAN, Lady [Sydney] [née OWENSON, Sydney]	THE PRINCESS; OR, THE BEGUINE. BY LADY MORGAN, AUTHOR OF "O'DONNEL," &C.; IN THREE VOLUMES.

1835	random	[SULLIVAN, Arabella Jane]; DACRE, Lady [Barbarina] (editor)	TALES OF THE PEERAGE AND PEASANTRY. EDITED BY LADY DACRE. IN THREE VOLUMES.
1836	silverfork	[GARDINER, Marguerite], Countess of Blessington	THE CONFESSIONS OF AN ELDERLY GENTLEMAN. ILLUSTRATED BY SIX FEMALE PORTRAITS, FROM HIGHLY FINISHED DRAWINGS BY E. T. PARRIS. BY THE COUNTESS OF BLESSINGTON.
1836	silverfork	[GORE, Catherine Grace Frances]	MRS. ARMYTAGE; OR, FEMALE DOMINATION. BY THE AUTHORESS OF "MOTHERS AND DAUGHTERS." IN THREE VOLUMES.
1836	silverfork	[HOOK, Theodore Edward]	GILBERT GURNEY. BY THE AUTHOR OF ""SAYINGS AND DOINGS,"" ""LOVE AND PRIDE,"" ETC. IN THREE VOLUMES.
1837	silverfork	[BURY, Lady Charlotte Susan Maria]	THE DIVORCED. BY LADY CHARLOTTE BURY, AUTHORESS OF FLIRTATION, &c. &c IN TWO VOLUMES.
1837	silverfork	[DISRAELI, Benjamin, Earl of Beaconsfield]	VENETIA. BY THE AUTHOR OF "VIVIAN GREY" AND "HENRIETTA TEMPLE." IN THREE VOLUMES.
1847	silverfork	[GORE, Catherine Grace Frances]	CASTLES IN THE AIR. A NOVEL. BY MRS. GORE. IN THREE VOLUMES.

Bibliography

Abrams, M. H., and Stephen J. Greenblatt, eds. 2000. *The Norton Anthology of English Literature.* 7th ed. Vol. 2. New York: Norton.

Adburgham, Alison. 1983. *Silver Fork Society.* London: Constable.

Allison, Sarah, Ryan Heuser, Matthew L. Jockers, Franco Moretti, and Michael Witmore. 2011. *Quantitative Formalism: An Experiment.* Pamphlet 1. Stanford Literary Lab: Stanford University.

Antoniak, Charles E. 1974. "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems" [in EN]. Mathematical Reviews number (MathSciNet): MR365969; Zentralblatt MATH identifier: 0335.60034, *The Annals of Statistics* 2 (6): 1152–1174. doi:10.1214/aos/1176342871.

Blei, David. 2012. "Introduction to Probabilistic Topic Models." *Communications of the ACM* 55 (4): 77–84. doi:10.1145/2133806.2133826.

Blei, David M., and John D. Lafferty. 2006. "Dynamic Topic Models." In *Proceedings of the 23rd International Conference on Machine Learning,* 113–120. Pittsburgh, PA: ACM.

———. 2007. "A Correlated Topic Model of Science." *The Annals of Applied Statistics* 1 (1): 17–35. doi:10.1214/07-AOAS114.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3:993–1022.

Block, Andrew. 1961. *The English Novel, 1740-1850.* 2nd ed. London: Dawsons.

Bourdieu, Pierre. 1988. "Flaubert's Point of View." Translated by Priscilla Parkhurst Ferguson. *Critical Inquiry* 14 (3): 539–562.

———. 1996. *The Rules of Art: Genesis and Structure of the Literary Field.* Stanford: Stanford University Press.

Broderick, Tamara, Jim Pitman, and Michael I. Jordan. 2013. "Feature Allocations, Probability Functions, and Paintboxes." *arXiv:1301.6647* (January 28). Accessed March 2, 2013.

Buntine, Wray. 2009. "Estimating Likelihoods for Topic Models." In *Advances in Machine Learning, First Asian Conference on Machine Learning,* 51–64. doi:`http://dx.doi.org/10.1007/978-3-642-05224-8_6`.

Canavan, Gerry, and Priscilla Wald. 2011. "Preface." *American Literature* 83 (2): 237–249.

Cavalli-Sforza, Luigi Luca, Paolo Menozzi, and Alberto Piazza. 1994. *The History and Geography of Human Genes.* Princeton University Press, July 5.

Chang, Jonathan, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David Blei. 2009. "Reading Tea Leaves: How Humans Interpret Topic Models." In *Advances in Neural Information Processing Systems 22,* edited by Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, 288–296.

Cohen, Margaret. 2002. *The Sentimental Education of the Novel.* Princeton: Princeton University Press.

Elson, David, Nicholas Dames, and Kathleen McKeown. 2010. "Extracting Social Networks from Literary Fiction." In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics,* 138–147. Uppsala, Sweden: Association for Computational Linguistics.

Escobar, Michael D., and Mike West. 1995. "Bayesian Density Estimation and Inference Using Mixtures." *Journal of the American Statistical Association* 90 (430): 577–588.

Ferguson, T. S. 1973. "A Bayesian analysis of Some Nonparametric Problems." *Annals of Statistics* 1 (2): 209–230.

Garside, Peter. 1991. "Popular Fiction and National Tale: Hidden Origins of Scott's Waverley." *Nineteenth-Century Literature* 46 (1): 30–53.

Garside, Peter, Anthony Mandal, Verena Ebbes, Angela Koch, and Rainer Schöwerling. 2006. "The English Novel, 1830-36: A Bibliographic Survey of Fiction Published in the British Isles." The English Novel, 1830–36. January 26. Accessed August 12, 2011. `http://www.cardiff.ac.uk/encap/journals/corvey/1830s/index.html`.

Garside, Peter, and Rainer Schöwerling. 2000. *The English Novel, 1770-1829: A Bibliographical Survey of Prose Fiction Published in the British Isles.* Edited by

Peter Garside, James Raven, and Rainer Schöwerling. Vol. 2. 2 vols. Oxford: Oxford University Press.

Griffiths, Thomas L., and Mark Steyvers. 2004. "Finding Scientific Topics." *Proceedings of the National Academy of Sciences* 101:5228–5235. doi:10.1073/pnas.0307752101.

Heinrich, Gregor. 2008. *Parameter Estimation for Text Analysis.* Version 2.9. vsonix GmbH + University of Leipzig, August.

———. 2011. *"Infinite LDA"—Implementing the HDP with Minimum Code Complexity.* Accessed January 7, 2013.

Hofmann, Thomas. 1999. "Probabilistic Latent Semantic Indexing." In *Proceedings of the Twenty-Second Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* 50–57. New York: ACM. doi:10.1145/312624.312649.

Hollingsworth, Keith. 1963. *The Newgate Novel, 1830-1847; Bulwer, Ainsworth, Dickens & Thackeray.* Detroit: Wayne State University Press.

Hope, Jonathan, and Michael Witmore. 2004. "The Very Large Textual Object: A Prosthetic Reading of Shakespeare." *Early Modern Literary Studies* 9 (3). Accessed September 20, 2011.

Isaac, Larry. 2009. "Movements, Aesthetics, and Markets in Literary Change: Making the American Labor Problem Novel." *American Sociological Review* 74 (6): 938–965. doi:10.1177/000312240907400605.

Kahneman, Daniel, Paul Slovic, and Amos Tversky, eds. 1982. *Judgment Under Uncertainty: Heuristics and Biases.* Cambridge: Cambridge University Press.

Kerman, Judith B. 1997. *Retrofitting Blade Runner: Issues in Ridley Scott's Blade Runner and Phillip K. Dick's Do Androids Dream of Electric Sheep?* Bowling Green: Bowling Green State Univ. Popular Press.

Koch, Angela. 2002. "Gothic Bluebooks in the Princely Library of Corvey and Beyond." *Cardiff Corvey: Reading the Romantic Text* (9).

Lévy, Maurice. 1968. *Le Roman gothique anglais, 1764-1824.* Toulouse: Association des publications de la Faculté des lettres et sciences humaines.

Manning, Christopher D., and Hinrich Schüzte. 1999. *Foundations of Statistical Natural Language Processing.* Cambridge, MA: MIT Press.

Meilă, Marina. 2002. *Comparing Clusterings.* UW Statistics Technical Report 418. University of Washington.

Mimno, D., M. Hoffman, and D. Blei. 2012. "Sparse stochastic inference for latent Dirichlet allocation." In *International Conference on Machine Learning.*

Moretti, Franco. 2000. *The Way of the World: The Bildungsroman in European Culture.* 2nd ed. London: Verso.

———. 2003. "Graphs, Maps, Trees: Abstract Models for Literary History-1." *New Left Review* (24): 67–93.

———. 2005. *Graphs, Maps, Trees: Abstract Models for Literary History.* London: Verso.

———. 2009. "Style, Inc. Reflections on Seven Thousand Titles (British Novels, 1740–1850)." *Critical Inquiry* 36 (1): 134–158. doi:10.1086/606125.

Neal, Radford M. 2003. "Slice Sampling." *The Annals of Statistics* 31 (3): 705–767. doi:10.1214/aos/1056562461.

Newman, David, Arthur Asuncion, Padhraic Smyth, and Max Welling. 2009. "Distributed Algorithms for Topic Models." *Journal of Machine Learning Research* 10:1801–1828.

Nigam, Kamal, Andrew Kachites Mccallum, Sebastian Thrun, and Tom Mitchell. 1999. "Text Classification from Labeled and Unlabeled Documents using EM." In *Machine Learning,* 103–134.

Novembre, John, and Matthew Stephens. 2008. "Interpreting Principal Component Analyses of Spatial Population Genetic Variation." *Nature Genetics* 40 (5): 646–649. doi:10.1038/ng.139.

Pritchard, Jonathan K., Matthew Stephens, and Peter Donnelly. 2000. "Inference of Population Structure Using Multilocus Genotype Data." *Genetics* 155 (2): 945–959.

Propp, Vladimir. 1968. *Morphology of the Folktale.* 2nd ed. Edited by Louis A. Wagner. Translated by Laurence Scott. Austin: University of Texas Press.

Rogers, Deborah S., Marcus W. Feldman, and Paul R. Ehrlich. 2009. "Inferring Population Histories Using Cultural Data." *Proceedings of the Royal Society B: Biological Sciences* 276 (1674): 3835–3843. doi:10.1098/rspb.2009.1088.

Shalizi, Cosma. 2006. "Graphs, Trees, Materialism, Fishing." The Valve - A Literary Organ. January 24. Accessed May 3, 2008. http://www.thevalve.org/go/valve/article/graphs_trees_materialism_fishing/.

Shalizi, Cosma. 2011. "Graphs, Trees, Materialism, Fishing." In *Reading Graphs, Maps, and Trees: Responses to Franco Moretti,* edited by John Holbo and Jonathan Goodwin. Parlor Press.

Sperber, Dan. 1996. *Explaining Culture: A Naturalistic Approach.* Oxford: Blackwell.

Teh, Yee Whye, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. "Hierarchical Dirichlet Processes." *Journal of the American Statistical Association* 101 (476): 1566–1581.

Trumpener, Katie. 1998. "National Tale." In *Encyclopedia of the Novel,* edited by Paul E. Schellinger, 910–11. Chicago: Fitzroy Dearborn.

Vinh, N.X., J. Epps, and J. Bailey. 2010. "Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance." *Journal of Machine Learning Research* 11:2837–2854.

Wallach, Hanna M., Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. "Evaluation methods for topic models." In *Proceedings of the 26th Annual International Conference on Machine Learning,* 1105–1112. ICML '09. New York, NY, USA: ACM. Accessed December 7, 2011. doi:10.1145/1553374.1553515.

Wallach, Hanna, David Mimno, and Andrew McCallum. 2009. "Rethinking LDA: Why Priors Matter." In *Advances in Neural Information Processing Systems 22,* edited by Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, 1973–1981.

White, Hayden. 2003. "Commentary: Good of Their Kind." *New Literary History* 34 (2): 367–376.

Williamson, S., C. Wang, K. Heller, and D. Blei. 2010. "The IBP Compound Dirichlet process and its Application to Focused Topic Modeling." In *Proceedings of the 27th International Conference on Machine Learning,* edited by Thorsten Joachims and Johannes Fürnkranz, 1151–1158. Haifa, Israel: Omnipress, June.