Cliometrica (2019) 13:83–125
https://doi.org/10.1007/s11698-018-0171-7

ORIGINAL PAPER

Economic history goes digital: topic modeling the Journal of Economic History

Lino Wehrheim[1]

Received: 2 February 2018 / Accepted: 24 April 2018 / Published online: 11 May 2018

Abstract Digitization and computer science have established a completely new set of methods with which to analyze large collections of texts. One of these methods is particularly promising for economic historians: topic models, i.e., statistical algorithms that automatically infer the content from large collections of texts. In this article, I present an introduction to topic modeling and give an initial review of the research using topic models. I illustrate their capacity by applying them to 2675 articles published in the Journal of Economic History between 1941 and 2016. By comparing the results to traditional research on the JEH and to recent studies on the cliometric revolution, I aim to demonstrate how topic models can enrich economic historians' methodological toolboxes.

Keywords Economic history · Topic models · Latent Dirichlet allocation · Cliometrics · Digitization · Methodology

JEL Classification A12 · C18 · N01

1 Introduction

The shift of economic history toward economics and quantitative methods during the 1960s can at least be partially explained by the technological changes which facilitated the dissemination of computers (Haupert 2016). With digitization, economic history (as every other field in science) is again confronted with far-reaching technological changes. Despite the many uncertainties concerning the effects

✉ Lino Wehrheim
lino.wehrheim@ur.de

1 Department of History, Economic and Social History, Department of Economics, University of Regensburg, Universitätsstraße 31, 93040 Regensburg, Germany

of digitization on economic history, one factor appears to be indisputable[1]: Falling costs of digitization leading to large collections of digitized records (like *Chronicling America*) and advanced methods for analyzing them will change the way economic historians carry out their research in the future (Abramitzky 2015; Collins 2015; Mitchener 2015).

When looking at the rapidly developing field of the digital humanities, the narrative changes from the future into the present tense. Here, scholars have already adapted to the growing mass of digital resources by incorporating methods from computer science.[2] Standing at the forefront of these methods, so-called topic models enjoy increasing popularity (Meeks and Weingart 2012, p. 2). The term "topic model" refers to statistical algorithms which automatically infer themes, categories, or topics, i.e., content, from texts, and which are the state of the art in automated text analysis.[3]

The idea behind topic modeling is relatively simple. Instead of reading texts and manually recording their topics, which for some collections of texts can require a great amount of resources or may even be impossible, the distribution of words across documents is used to infer the inherent topical structure. In this way, content can be quantified, a process that allows integrating qualitative sources into quantitative research, which Abramitzky (2015, p. 1248) calls "turning books into data".

Although Abramitzky (2015) and Mitchener (2015) already mention them, to the best of my knowledge there has not yet been a paper published in an economic history journal that explicitly covers or uses topic models. In this paper, I intend to shed some light on an "exciting new trend" (Abramitzky 2015, p. 1248) and illustrate that topic models are a tool which promises to be of great utility, especially for economic historians with their affinity for quantitative analysis. One reason topic models are largely unknown outside the community of digital humanities[4] may be that this is a rather young discipline and much, possibly even most of its research is not published in traditional print journals. Instead, research is communicated on blogs and Web sites, especially when it comes to tutorials, which may function as a "barrier of entry" to scholars from other disciplines (Meeks and Weingart 2012, p. 3).[5] In this paper, I will provide a largely non-technical description of topic modeling as its statistical foundations are explained in detail by others.[6] Rather, the aim is to give an insight into the general principles of topic modeling from a user's perspective

[1] Questions on the future of economic history were discussed on a special panel at the 75th anniversary of the Economic History Association. See Journal of Economic History, Volume 75 Issue 4.

[2] For an assessment of the status quo in digital history, see the white paper "Digital History and Argument", the Arguing with Digital History working group, Roy Rosenzweig Center for History and New Media.

[3] Jockers (2013, p. 123) calls them the "mother of all collocation tools".

[4] As the literature review shows, there are several economists who have recently become aware of them.

[5] Scott Weingart blog gives a helpful overview of blogs on topic modeling. See http://www.scottbot.net/HIAL/index.html@p=19113.html. Although this overview may be somewhat outdated, it is still a good starting point for scholars who are unfamiliar with topic modeling.

[6] See Blei et al. (2003), Blei and Lafferty (2009), Griffiths and Steyvers (2004), and Steyvers and Griffiths (2007) for formal descriptions.

and to address the questions that must be considered before commencing a topic model project. For instance: how do the texts have to be processed in order to be analyzed by a topic model? Which parameters require specification? Which potential problems have to be addressed, especially when using topic models for historical research? I will provide an overview of the literature using topic models, which illustrates their disciplinary versatility, followed by a practical application. I use the most prominent topic model—*Latent Dirichlet Allocation*—to automatically extract topics from all articles published in the *Journal of Economic History* (JEH) between 1941 and 2016. The results will demonstrate that topic models are the right tool for research on publications trends, such as the work by Whaples (1991, 2002), who performed a topic analysis of the JEH in a more traditional fashion. Furthermore, I will show that in terms of methodology, topic models can contribute to current research by Diebolt and Haupert (2018) and Margo (2018) on the disciplinary shift in economic history known as the cliometric revolution.

2 The principles of topic modeling

Topic models are one tool among others in the field of text mining, which again is a melting pot of different disciplines such as data mining, computational linguistics, and machine learning (Miner 2012, pp. 31–34; Grimmer and Stewart 2013, p. 268).[7] Essentially, they are statistical algorithms that analyze word occurrences in a large collection of documents (the corpus) to discover groups of words that have a high probability of occurring together. Originally, they were developed in the field of computer science, machine learning, and information retrieval approximately 15 years ago (Meeks and Weingart 2012, p. 2), but meanwhile have expanded into a variety of other disciplines. To be precise, they should be called "probabilistic topic models", as they build on the assumption that a document can exhibit different topics and therefore work with probability distributions of words and topics (Steyvers and Griffiths 2007, pp. 430–432). There are different types of topic models, depending on the statistical assumptions of the algorithms (Steyvers and Griffiths 2007). The one most commonly used and "state of the art in topic modeling" (Lüdering and Winker 2016, p. 493) is *Latent Dirichlet Allocation* (LDA), which was introduced by Blei et al. (2003).[8]

What can we expect from topic models? In essence, we wish to gain a first impression of what our documents concern before actually reading them. If the volume of documents prevents us from reading them in their entirety and if there is a

[7] Describing the origins of topic modeling in the context of digital humanities is relatively challenging as this touches on several disciplines which all have different histories. For example, the 'history of humanities computing' can be traced back to Father Roberto Busa, who indexed the work of Thomas Aquinas in the late 1940 s. See Hockey (2004) and Jockers (2013). For a brief description of the recent development of topic models, see Lüdering and Winker (2016).

[8] There have been several extensions of the original model covering different assumption of LDA Blei (Blei 2012a, pp. 82–84). In the following, the terms LDA and topic model will be used synonymously.

lack of other guiding lines such as abstracts or keywords,[9] topic models help us to structure the documents and identify those that are relevant for our research question or simply for general interest. If we want to integrate the documents' content into quantitative analysis, e.g., investigating publication trends in a scientific journal, we need numerical representations of the content. Topic models can help us by providing the topic composition of every document in our corpus. But what exactly is a topic? Topic models treat topics as probability distributions over words. Thus, the second type of output is comprised of lists of words that the model has identified as having a high probability of occurring together. Examples of topics inferred from the *Journal of Economic History* are given in Fig. 1. In other words, topic models are a tool for producing descriptive statistics for the content of a plethora of documents.

Topic models are built on two basic assumptions. Firstly, they assume that the semantic meaning of a text is created by the joint occurrence of words, although not all word clusters produce what we would call meaning (for example, there can be clusters of prepositions and pronouns). These clusters can be interpreted as being topics "because terms that frequently occur together tend to be about the same subject" (Blei 2012b, p. 9). In other words, this assumption implies that meaning is relational, i.e., the meaning of one single word depends on its co-occurrence with other words (Mohr and Bogdanov 2013, pp. 546–547). For example, the word *table* can have at least two meanings, and it depends on other words such as *chair, sitting,* or *column* to determine which meaning is referred to in a given sentence. Topic models account for this polysemy by allowing a word to belong to different topics (Steyvers and Griffiths 2007, p. 429), so *table* could be found in a topic on furniture and a topic on spreadsheets at the same time.[10]

Secondly, topic models assume that a document is generated in a process which can be described by the following model (Blei 2012a, p. 80).[11] The corpus consists of D documents, each of which that consists of N_d words $w_{d,n}$, where $w_{d,n}$ is the nth word in document d. The overall vocabulary V is fixed. Documents exhibit a share of every topic k (although some might be infinitesimally small) with θ_d describing document d's distribution across topics.[12] The overall number of topics K is assumed to be fixed. As stated above, topics are treated as distributions over words with β_k representing the distribution of topic k. In other words, β_k corresponds to one of the word clouds depicted in Fig. 1. Every word in a document is assigned to either

9 Abstracts and keywords pose their own problems. With large collections, even the reading of abstracts can become too time-consuming. However, keywords may be too vague to be useful.

10 In this example, the word *table* could be accompanied by words like *chair, tablecloth,* or *leg* in the first topic, while in the second it could be words like *column, row,* or *cell*.

11 This is why topic models are also called generative models. See Steyvers and Griffiths (2007, p. 427).

12 More precisely, this nonzero probability follows from estimating the topic shares using Gibbs sampling, also used in this paper (see below). The share of topic k in document d is approximated by $\hat{\theta}_{d,k} = \frac{n_{d,k}+\alpha}{N_d+K\alpha}$ with $n_{d,k}$ being the number of times document d uses topic k and N_d equaling the total number of words in document d. Including the Dirichlet parameter α results in $\hat{\theta}_{d,k}$ always being nonzero. See Boyd-Graber et al. (2017, pp. 15–16) and Griffiths and Steyvers (2004, pp. 5229–30).

one or multiple topics, which is represented by topic assignment $z_{d,n}$ for word n in document d. A graphical representation of this model is provided in Fig. 2. The only observed variable is words, which is represented by a shaded node. All other variables are hidden.

The generative process of a document itself is assumed to be as follows (Blei and Lafferty 2009, pp. 73–75; Blei 2012a, pp. 78–82). First, choose a distribution over topics θ_d. From this, draw a topic k. Finally, choose a word $w_{d,n}$ from this topic. This is repeated for every word in every document. In other words, it is assumed that first the author decides what topics the text should be about by determining the topic shares (step one). The actual writing is interpreted as the choosing of words from a topic-specific vocabulary according to the topic shares (steps two and three). The reader cannot observe the generative process but only the output (the words).

The basic idea behind LDA is that the generative process produces a joint probability distribution of the hidden variables (topic vocabulary and topic shares) and the observed variables (words). This distribution is used to answer the question: "What is the likely hidden topical structure that generated my observed documents?" (Blei 2012b, p. 9). The conditional distribution of the hidden variables given the observed variables, called *posterior distribution*, is given by (Blei 2012a, p. 80):

$$p\left(\beta_{1:K}, \theta_{1:D}, z_{1:D} | w_{1:D}\right) = \frac{p\left(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}\right)}{p\left(w_{1:D}\right)} \tag{1}$$

This posterior is what we are searching for because it tells us the probabilities of topics and topic assignments of words from our corpus. Unfortunately, the conditional distribution cannot be computed directly (Griffiths and Steyvers 2004, p. 5229). There are several techniques with which to estimate the posterior (Blei and Lafferty 2009, pp. 76–78) and explaining all of them would go beyond the scope of this paper. The most common one, Gibbs sampling, can be outlined as follows[13]: Technically, LDA assumes that the two steps of generating the documents happen randomly (Blei 2012a, p. 78). Starting from a random topic assignment, Gibbs sampling resamples the topic assignment of a given word by asking two questions: Which topics can be found in the document and which topics is this word assigned to in other documents? It calculates the topic assignment with the highest probability given the assignments of the other words in the document and given the topic assignment of the word under consideration in other documents and updates the word's topic assignment accordingly. This is performed for every word in every document yielding an iterative process of probability updating. How many times this updating is carried out can be determined by the researcher with more iterations

[13] The following description is inspired by a lecture given by Mimno (2012b) and Ted Underwood description of topic modeling on his blog (available at https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/). For a technical description, see Griffiths and Steyvers (2004) and Steyvers and Griffiths (2007).

Fig. 1 Four examples of topics automatically inferred from the *Journal of Economic History*. The size of a term is proportional to its importance for the topic. *Source*: See text

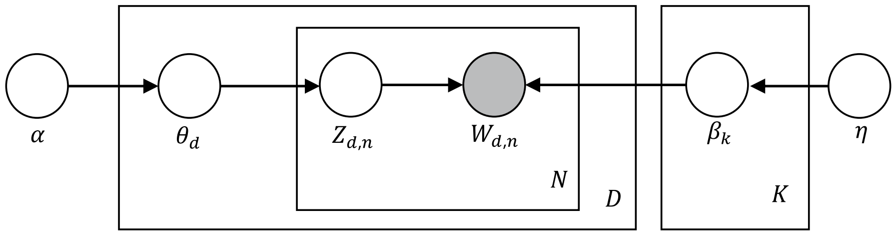

Fig. 2 Graphical representation of the Latent Dirichlet Allocation *Source*: Reproduced with permission from Blei (2012a)

leading to more coherent topics, although this effect will level off at some point (Jockers 2014, p. 147).[14]

[14] There is a trade-off between topic coherence and the time it takes to train the model. Finding many topics in large corpora can keep the computer busy for hours.

So far, the name LDA has not been explained. A document's distribution over topics in the first step θ_d is assumed to follow a *Dirichlet* distribution (Blei 2012a), which is a distribution over another distribution. It is specified by the *Dirichlet* parameter $\boldsymbol{\alpha}$, which is a vector over (α_1; α_2; $\ldots\alpha_K$), and which describes the shape of the distribution (Steyvers and Griffiths 2007, pp. 430–32; Wallach et al. 2009).[15] This parameter $\boldsymbol{\alpha}$ can be interpreted as being a concentration parameter that determines the distribution of topics over the corpus. It can be modeled as symmetric, i.e., $\alpha_1 = \alpha_2 = \cdots = \alpha_K = \alpha$, which implies that topics are distributed equally over the corpus. Alternatively, it can be estimated, implying that some topics are more prevalent than others. The higher α_k, the higher the relative importance of topic k. The distribution of a topic over words, β_k, is also assumed to follow a *Dirichlet* distribution, with the corresponding parameter being $\boldsymbol{\eta}$. Again, $\boldsymbol{\eta}$ can be modeled symmetrically or asymmetrically, implying that words are equally important for topics or that some words are more important for a topic than others. Finally, the model *allocates* words to different *latent* (i.e., not observable) topics (Blei 2012a).

3 Topic models in practice

As stated above, topic models treat topics as distributions over words. Accordingly, the results are groups of words that have a high probability of occurring together. However, these groups lack any kind of label (Blei 2012a, p. 79). They may or may not be recognizable as a theme at first glance. By anticipating the results from the topic models on the JEH, I will give two examples. The ten most probable words for topic 1 are *japanese, japan, china, chinese, rice, land, period, government, meiji,* and *tokugawa* (words are ranked in a decreasing order in terms of importance for the topic).[16] For topic 7, it is *bank, banks, banking, deposits, reserve, national, notes, system, state,* and *credit.* Both topics are relatively coherent, i.e., there is an obvious common meaning, and reasonable labels could be *Japan and China* and *Banking*. It is important to note that the topic model found these topics without any prior information on entities such as countries or financial institutions.

Although the topics are generated automatically by the model, they must be identified as such by the researcher, which is one of the most vital tasks in topic modeling. Identification means that the researcher is required to study the topics by reading the word lists, find the common meaning of the words grouped together by the model, and provide a suitable label.[17] This step is non-technical and solely based on the researcher's interpretation, which is why, in most cases, domain-specific knowledge is compulsory. Thus, while the topics and their distributions across documents are inferred by the model, their meaning is "inferred" by considering a plausible interpretation. This will not pose any problems in unequivocal cases as in

[15] This parameter is sometimes called "hyperparameter".

[16] One step in topic modeling frequently consists of removing capitalization (see below). In the following, words are kept in lower case if they constitute a topic.

[17] There are also attempts to automatically assign labels to topics. See Lau et al. (2011).

the examples given above; still, it can be quite challenging, as the model may also find topics that, at first glance, have no meaning. If this is the case, it may be that the model identified a merely linguistical pattern which indeed lacks any kind of useful meaning.[18] However, topics do not necessarily have to describe what the documents "are about". As I will show later in this paper, they can also be clusters of methodological words or days of weeks (see also Boyd-Graber et al. 2015, p. 240). If a topic lacks a straightforward interpretation, it can be helpful to read the documents that exhibit a large share of this topic. This can illustrate how to interpret the topic. In general, interpretability (or coherence) can be regarded as being the linchpin in topic modeling. We can only use a topic for further analysis if we are able to identify its meaning. The degree of topic coherence depends on model specification (especially the number of topics), the characteristics of the corpus, and the level of granularity in which one is interested (Jockers 2013, pp. 127–28). In general, decreasing the number of topics results in more coherent but also less specific topics. Running several topic models with different numbers of topics is the most practical solution for identification of the correct number for the given research question.

Interpreting and labeling the topics is, of course, relatively subjective as the interpretation of words can differ between one reader and another (Jockers 2013, p. 130). Nevertheless, subjectivity is a familiar problem. Individual judgments also must be made when coding a text manually. In other words, by using topic models, we can postpone the moment when subjective assessments become necessary: from the ex ante subjectivity of specifying categories to the ex post subjectivity of interpreting them. Still, the subjectivity of interpreting topics requires the words constituting the topics to be included in every publication using topic models. There are some metrics to diagnose the "quality" of the topics (Boyd-Graber et al. 2015), but in the end it depends on human interpretation to identify their common denominator. The aspect of human interpretation is discussed by Chang et al. (2009), who pursue an experimental approach to investigating how humans interpret topics.

Some remarks must be made on the data. To infer the topics and their distributions across documents, the algorithm requires a numerical representation of the texts. Therefore, texts are represented in a "vector space model for text data" (Newman and Block 2006, p. 753). Every document d is converted into a so-called word vector, whose entries are the counts of occurrences of each of the N different words in document d. The vectors of all documents taken together produce the document-term matrix (DTM).[19] The entries of this matrix are comprised by the counts of occurrences of every word type from the vocabulary V in every document d.[20] Accordingly, the row sums of the DTM equal the number of words in document

[18] For example, in a German newspaper corpus analyzed in a different project, one topic consisted mainly of modal verbs.

[19] Some algorithms operate with a term-document matrix, which is a transposed DTM.

[20] In general, the word vectors as well as TM can contain absolute word counts (occurrences) or relative word counts (term frequency or tf–idf. See below.). The conversion of text files into word vectors or a DTM can be carried out within the different topic model applications (see below) or independently from topic modeling. Tools are, e.g., the *tm* (text mining) package for *R* developed by Feinerer (2017) or programs like *RapidMiner*.

d, and the column sums equal the number of times a word can be found in all documents. For example, the corpus used in this paper consists of 2675 documents with an overall vocabulary of 288,416 unique terms, yielding a 2675 × 288,416 DTM. With a sparsity of 99%, this matrix is relatively sparse as most documents contain only a few different word types, which is the case in most corpora.[21] For creating the DTM, the documents must be split into single units which are called tokens (Boyd-Graber et al. 2015, p. 230). This process of separating a string of text into pieces is carried out by a tokenizer.[22] The simplest way of converting text into tokens is by splitting it at whitespace and punctuation marks, but this can also imply converting to lower case and removing numbers. Furthermore, working with word vectors and DTM implies that the order of words in a documents is irrelevant, which is why they are called "bag of words" representations of texts (Blei 2012a, p. 82). The sentences "France industrialized after Great Britain" and "Great Britain industrialized after France" are treated identically. This may look somewhat unrealistic, but both sentences suggest that their content is about industrialization, France, and Great Britain.[23]

There are several further steps which can be applied in order to preprocess the corpus and which influence the resulting topics (Boyd-Graber et al. 2015, pp. 227–31). It is common to remove words which occur frequently and have no semantic meaning (like *the, and, a, or*). These so-called *stopwords* are removed based on a fixed list, but sometimes it may be necessary to further remove corpus-specific words if they occur too often and therefore only produce noise (Jockers 2013, p. 131). A common measure of a word's relative importance is the *term frequency–inverse document frequency (tf–idf)*. This measure places higher weights on words that occur frequently in a single document (term frequency) but only rarely in the overall corpus (inverse document frequency), thereby emphasizing words that have a high level of importance for single documents (Blei and Lafferty 2009). Depending on where the documents are obtained, they may contain words that do not belong to the text itself, so-called *boilerplate* (Boyd-Graber et al. 2015, p. 228). This could include HTML tags, when the text has been directly received from a Web site, download signatures, or text fragments from other texts caused by missing page breaks. Another step is characterized by the normalization of the text itself. Removing capitalization, reducing the words to their stem, or lemmatizing (reducing words to their basic forms) can help to remove noise from the data (Boyd-Graber et al. 2015).[24] What kind of preprocessing steps should be taken depends on the corpus

[21] The dimension of this DTM will be reduced by removing certain words. See below.

[22] Ostensibly, a tokenizer is a computer program which cuts sentences into pieces called "tokens", based on predefined rules. For example, the sentence *"Mr. Smith's mother is seventy-nine years old, but she doesn't look her age."* could be tokenized most simply by cutting at each whitespace and punctuation mark: *"Mr|Smith|s|mother|is|seventy|nine|years|old|but|she|doesn|t|look|her|age"*. This simple rule could be modified, for example, so that it does not split numbers, to delete every *"s"* following an apostrophe, or to leave common expression like *doesn't* intact.

[23] For an extension relaxing the bag of words assumption, see Wallach (2006).

[24] For tools to carry out these steps, see Graham et al. (2016). For a discussion of the effect of stopword removal, see Schofield et al. (2017).

and the research question. For example, it can be helpful to concentrate only on nouns, which can be achieved by using so-called part-of-speech taggers which automatically identify a words' part of speech (Jockers 2013, p. 131).[25]

Topic models come with a caveat, which is especially important for historians. Their results crucially depend on the quality of the documents. Most texts used by historians are either transcriptions or optical character recognition (OCR) treated scans. As both are be prone to orthographical errors, one has to check the documents carefully before applying a topic model (otherwise it could be a typical case of garbage in, garbage out). In some cases, the road of digital scholarship can already end here as the text quality may be too poor and correcting the texts would be either too time consuming or too costly. In others, as Walker and Lund (2010) show, systematic errors such as repeated OCR mistakes can be treated, and a certain amount of random errors may be tolerated.[26]

Another technical issue is especially important for historical research. The standard LDA topic model does not capture changes in the use of language. For example, sources from the early eighteenth and the late nineteenth century may describe the same subject with different vocabularies, which would probably lead to two different topics. There are extensions of LDA accounting for this (Blei and Lafferty 2006), but this problem could theoretically be solved by combining topics covering the same subject in different "languages" or by creating sub-corpora. Changes in terminology are a potential problem in some cases; in others they may be what we are looking for,[27] so controlling for them depends on the corpus as much as on the research question.

Language touches upon another important aspect of topic modeling. In general, topic models can be applied to documents written in any language,[28] although the characteristics and complexity of some languages may pose different challenges. For example, personal experience with a German newspaper corpus suggests that topics in German documents appear to be less coherent and necessitate more stop-word removal than documents written in English.[29] Some corpora may contain multilingual sources which share topics without sharing the language, e.g., articles published in Wikipedia. As the corpus analyzed in this paper contains only documents written in English, and as multilingualism is a strand of its own in the topic model literature, it will be discussed only briefly. Studies addressing multilingualism provide extensions of the LDA model (Boyd-Graber and Blei 2009; Mimno et al. 2009). For example, Mimno et al. (2009) introduce the polylingual topic model (PLTM) and show that this model can detect topics in a corpus which consists of direct translations (proceedings of the European parliament) as well as in a corpus which consists of topically connected documents which are not direct translations

[25] For a general discussion of texts as data for economic research, see Gentzkow et al. (2017).

[26] OCR mistakes can build their own topic, see Jockers (2013).

[27] See, for example, McFarland et al. (2013).

[28] To name just two examples, Miller (2013) uses a Chinese corpus and Heiberger and Koss (2018) use documents written in German.

[29] This statement is only based on impressions based on work conducted in a different project and has not been verified empirically.

(Wikipedia articles). With PLTM, topics are modeled as a set of collections of words, each collection containing words in one language. For example, in the parliament proceedings Mimno et al. (2009) find a topic concerning children whose three most probable words are *children, family,* and *child* for the English, *kinder, kindern,* and *familie* for the German, and *enfants, famille,* and *enfant* for the French collection.[30] This invites the question of whether topics are invariant to translations, i.e., applying the same model on documents and translations separately yields the same topics.[31] To the author's knowledge, there is no study that investigates this issue, but from what Mimno et al. (2009) describe, it seems likely that direct translations will not substantially change the topic distribution. Yet, there may be minor variations, especially if the number of topics is high, i.e., there are many small and granular topics, and when the translator has a high degree of interpretational freedom.[32]

There are several applications for topic modeling with different degrees of options for users (Graham et al. 2016), inter alia a package for *R* (Grün and Hornik 2011). In this paper, I used the *Machine Learning for Language Toolkit* (MALLET), a user-friendly tool developed by Andrew McCallum in 2002 (McCallum 2002), which implements LDA and Gibbs sampling, as well as the *R* package.

To conclude this chapter, the main strengths of topic modeling shall be emphasized. Topic modeling is primarily concerned with reducing complexity by finding and applying categories. In this sense, topic models resemble the "old-fashioned" way of text coding. In the latter approach, features of documents are recorded according to predefined categories, like JEL codes or any other kind of coding scheme defined in a codebook.[33] After coding, documents can be grouped according to their exposure of certain categories. The codes are specified in advance by the researcher, which poses several problems. It is impossible to know in advance all the categories that will be found in text, so to a certain degree the categories must be updated during the coding process, which takes a considerable amount of time. In the end, only in the rarest case do the categories fit the data perfectly. This holds true especially for historical research, in which the usage of contemporary categories may miss the point due to not necessarily matching the historical sources. Furthermore, manual coding requires careful reading, which again may be too expensive for large text collections. Additionally, human coding is prone to imprecision and

[30] The authors also show how their model can be used for machine translations by generating bilingual lexica. For further applications of topic models in machine translation, see Eidelman et al. (2012) and Zhao and Xing (2007).

[31] I thank one of the anonymous referees for this intriguing question.

[32] For example, translations of scientific publications are similar to the parliament proceedings used in Mimno et al. (2009) in that they are rather direct. However, there can be quite considerable differences in translations of novels. Furthermore, some aspects of a text may become "lost in translation", as, for example, the German differentiation between the formal form of address *Sie* and the informal *Du* is lost in the English translation *you*. It appears unlikely, though, that issue of this kind pose serious problems to topic modeling.

[33] JEL codes are used by Abramitzky (2015), McCloskey (1976), and Whaples (2002). Another example of a comparable way of text coding can be found in the financial literature on sentiment following Tetlock (2007), which basically measures the tone of texts by counting negative and positive words based on predefined dictionaries.

mistakes, to which the computer is immune. Finally, when coding large collections of texts, humans may not be perfectly consistent in their decisions made over time.[34]

Despite the resemblance of topic modeling and manual coding in terms of producing numerical representations of texts, the crucial point of topic modeling is that no classification scheme requires specification in advance. Rather, the documents speak for themselves and define their own categories. So, the first strength is that topic models are objective, which touches some fundamental epistemological considerations. We as scientists approach our sources with a framework a priori in mind, which results from our prior knowledge, our personal interests, our socialization, our individual concepts of relevance, our theories, and so forth. For the sake of simplicity, we can call this "priming".[35] Priming itself is neither good nor bad as long as our reasoning is comprehensible. It concerns the selection and structuring of sources as well as the building of econometric models. Returning to the issue of categories, the choice of JEL codes implies the judgement that they reflect all relevant categories and that the definitions of the codes fit the sources, which may be primed by an education as an economic historian.[36] Furthermore, the risk of being biased in assigning documents to categories prevails, as, for instance, people subconsciously tend to choose information which confirms their beliefs.[37] If we have chosen a certain set of codes, confirmation bias poses the threat of our perception being selective, i.e., of us perceiving the texts in ways that favor our codes, although a different code set may be more appropriate. This confirmation bias will be reinforced when decisions are discretionary, which is usually the case when coding texts. Consequently, both the choice and the assignment of categories determine and potentially bias the results. In contrast, topic models are agnostic, i.e., they work without any a priori understanding of the sources. Instead, they identify the sources' inherent structure according to the statistical algorithm. They produce categories which are independent of the researcher's priming and assign documents accordingly.[38] In this respect, topic models are also different from so-called supervised learning approaches, in which algorithms are iteratively trained by human intervention. A practical benefit of this unbiased nature is that topic models can help to find categories that were not thought

[34] This should not create the impression that manual coding is regarded as being futile. On the contrary, there are many cases in which it is completely appropriate, and studies like Whaples (2002) as well as professionals' reliance on human coding, e.g., in media analysis, build a strong case for manual coding. Still, the advantages of one method are best illustrated when compared with the shortcomings of another, and in times of almost unlimited availability of textual sources, automatic methods like topic modeling will probably prevail.

[35] Although there may be some overlaps, this should not be confused with priming as it is understood in psychology.

[36] Of course, even the decision to use any type of quantitative representation of texts is based on the conviction that this can contribute something to our research. It could be that, for example, economic historians affiliated with history departments find this a less useful approach. Naturally, this argument holds true for the use of topic models as well.

[37] For further explanation of the confirmation bias, see, e.g., Oswald and Grosjean (2004).

[38] Of course, the preprocessing steps applied in topic modeling, such as the choice to remove certain stop words, can be regarded as being a priori decisions by the researcher that influence—and thus potentially bias—the output.

of, and to identify relevant sources, also including those that could be overlooked when using search terms. In this way, topic models also can be used for browsing databases.[39]

The second strength of topic models is their ability to process large collections of documents in a short amount of time,[40] by far exceeding the possibilities of traditional methods of quantification of texts. The size of the database is only limited by computing power. In fact, the model works even better if the corpus is larger.[41] Additionally, the possibility of words and documents being assigned to multiple categories enables a degree of granularity that would be infeasible in manual coding. Besides, topic models can be applied to input other than texts, such as images (Blei 2012a, p. 83), opening new possibilities for quantitative research.

Thirdly, topic models produce numerical representations of texts, allowing us to integrate textual sources into a quantitative research design. In this way, we can combine textual with traditional data. There are a myriad of conceivable applications for economic historians. For instance, topic models allow us to gain insights into the reasoning of economic agents as we can now use textual resources on a completely different scale. In particular, combining topic models with other text mining approaches like measuring sentiment appears to be very promising.[42] Minutes of central banks, ministries, cabinets, or executive boards seem to be ideal candidates for a topic modeling application. Furthermore, the ambiguous notion of impact can be investigated much more tangibly. To cite just one example from current research, we can study how decision-makers are influenced by economic policy advice. In this regard, the way topic models produce numerical representations of content provides another useful application. As stated above, topic models treat documents as distributions over topics (θ_d). There are various methods for comparing the difference (or divergence) between two distributions, such as the Kullback–Leibler divergence or the Jensen–Shannon divergence (Steyvers and Griffiths 2007, pp. 443–44).[43] These

39 For instance, applications of topic models for the use of databases are explored by JSTOR One example is the "text analyzer", an online tool which identifies documents in the JSTOR database that are similar to a search document in terms of topics. See http://www.jstor.org/analyze/.

40 Running one model on the 2675 documents of this paper took approximately 35 min using an ordinary computer.

41 There is no exact lower limit to the number of documents, but experience shows that it takes at least about two hundred scientific paper-sized documents to produce meaningful topics. If the corpus consists of a few long documents, such as books, these documents can be split, e.g., into single chapters. See Jockers (2013). The documents must not be too short either. For example, the length of a single tweet would not be sufficient in finding any meaningful topics, so in this case, several tweets can be aggregated, e.g., on a daily basis. See Lüdering and Tillmann (2016).

42 The analysis of textual sentiment, i.e., the tone of documents, is a second major approach in text mining, which is used particularly in finance and financial economics. For example, pessimism expressed in financial newspapers is found to influence stock returns and trading volume, see, e.g., Tetlock (2007) and García (2013). An example for a combination of topic modeling and sentiment analysis is given by Nguyen and Shirai (2015).

43 The Kullback–Leibler divergence (or distance) between two probability distributions p and q is defined as $\mathrm{KLD}(p,q) = \frac{1}{2}\left[D(p,q) + D(q,p)\right]$ with $D(p,q) = \sum_{j=1}^{T} p_j \log_2 \frac{p_j}{q_j}$. The Jensen–Shannon divergence is defined as $\mathrm{JSD}(p,q) = \frac{1}{2}\left[D\left(p, \frac{p+q}{2}\right) + D\left(q, \frac{p+q}{2}\right)\right]$. Both measures are equal to zero when p and

methods provide a numerical value of the extent of divergence between two distributions. In this way, they allow us to compare the similarity of documents in terms of content by comparing the corresponding topic distributions. For example, Mimno (2012a) uses the Jensen–Shannon divergence for comparing the content of different classics journals. The following section will provide an overview of further studies which use topic models.

4 Literature review

We now have a considerable amount of research using topic models. In Table 1, the literature of potential interest for economic historians is presented using the old-fashioned way of categorizing texts. For example, I record the authors' institutional affiliation, which shows that recently, topic modes have attracted the attention of economists. In the following, some of the literature will receive special attention.

Newman and Block (2006) were among the first to apply topic models to historical sources. They test several types of topic models, identifying themes in a colonial US newspaper, the *Pennsylvania Gazette*. Their analysis of 80,000 documents published between 1728 and 1800 is an impressive illustration of the potential of topic modeling for large scale historical research. They were able to identify the topics that moved eighteen-century Pennsylvania and described how this changed during the American Revolution. For example, they show that the *Gazette* mainly covered topics related to politics and economics, while religion accounts only for a minor topic. With the *Gazette* becoming more political in the 1760s, they found that many political topics contained references to the emergence of the Revolution. The coverage of crime declined after a high in the 1730s and rose again in the 1760s, showing a trend similar to that of religion. The largest individual topic relates to runaways and indentured servants, which, according to the authors, reveals the importance of servants in Pennsylvanian life.

Newspapers are also contained in the database for Bonilla and Grimmer (2013), who investigate the influence of several increases in the terror alert level under the Bush administration between 2002 and 2005 on public debate in the media. DiMaggio et al. (2013) apply topic models to newspapers in order to find how the shrinking public support of the arts in the US between 1986 and 1997 was framed by media coverage. Jacobi et al. (2015) examine the coverage of nuclear technology in the *New York Times* between 1945 and 2013. In his project "Mining the Dispatch" Robert Nelson applies topic models to the *Richmond Daily Dispatch*, a Confederate daily newspaper, between 1860 and 1865, to investigate social and political life in Civil War Richmond.[44]

Footnote 43 (continued)

q are completely identical and, with higher divergence, their value approaches one. See Steyvers and Griffiths (2007).

[44] Available at http://dsl.richmond.edu/dispatch/pages/home.

Table 1 Literature review summary

Paper	Database	Time	Topic	Department
Newman and Block (2006)*	Colonial newspaper (Pennsylvania Gazette)	1728–1800	Topic scouting/TM methodology	Computer Science, History
Blei and Lafferty (2007)*	Science	1990–1999	Topic scouting/TM methodology	Computer Science
Hall et al. (2008)	Association for Computational Linguistics (ACL) Anthology	1978–2006	Disciplinary History	Symbolic Systems, Linguistics, Computer Science
Grimmer (2010)*	US Senate press releases	2007	Identifying politicians' agendas	Government
Quinn et al. (2010)*	Speeches in the U.S. Senate	1995–2004	Measuring political attention	Law, Political Science, Computer Science
Yang et al. (2011)	Multiple Texan Newspapers	1989–2008	Topic scouting/TM methodology	Computer Science/Engineering, History
Mimno (2012a)*	Multiple Classics Journals	1850–2006	Topic scouting/TM methodology	Computer Science
Bonilla and Grimmer (2013)*	Multiple U.S. newspapers, transcripts of newscasts	2002–2005	Influence of terror alerts on public opinion	Political Science
DiMaggio et al. (2013)	Multiple U.S. newspapers	1986–1997	Coverage of US public financial assistance to arts in newspapers	Sociology, Computer Science
Jockers (2013)	Fiction from the U.S. and Great Britain	1750–1899	Topic scouting/topic analysis	English
Miller (2013)	Crime Reports from the Chinese Administration	1722–1911	Analysis of the nature of unrest and violence in Qing China	East Asian Languages and Civilizations
Riddell (2014)	Multiple US-based German Studies Journals	1928–2006	Topic scouting/Disciplinary History	Computational Science
Jacobi et al. (2015)	New York Times	1945–2013	Press coverage of nuclear technology	Communication Science
Larsen and Thorsrud (2015)	Business Newspaper (*Dagens Næringsliv*)	1988–2014	Forecasting macroeconomic data	Economics
Shirota et al. (2015)	Minutes of Meetings of the Bank of Japan	2014	Effects of consumption tax increase on monetary policy	Economics
Hansen and McMahon (2016)	Federal Open Market Committee statements	1998–2014	Effects of central bank communication on macroeconomic and financial variables	Economics

Table 1 (continued)

Paper	Database	Time	Topic	Department
Lüdering and Tillmann (2016)	Twitter messages referring to the Fed	2013	Measuring expectations of monetary policy and its effects on asset prices	Economics
Lüdering and Winker (2016)	Journal of Economics and Statistics	1949–2010	Time perspective of economic research/ disciplinary history	Economics
Thorsrud (2016a, b)	Business Newspaper (*Dagens Næring-sliv*)	1988–2014	Estimating business cycles based on news	Economics
Bellstam et al. (2017)	Analyst reports on S&P 500 firms	1990–2012	Measuring firms' inventive activities	Finance
Fligstein et al. (2017)	Federal Open Market Committee minutes	2000–2008	FOMC's perception of the financial crisis in 2008	Sociology
Grajzl and Murrell (2017)*	Multiple writings by Francis Bacon	ns	Identifying features and origins of Francis Bacon's ideas	Economics
Larsen and Thorsrud (2017)	Business Newspaper (*Dagens Næring-sliv*)	1988–2014	Effects of news on asset prices	Economics
Daniel et al. (2018)	International New York Times, Financial Times	2009–2015	Media effects on financial markets in the Greek debt crisis	Economics, Politics
Hansen et al. (2018)	Federal Open Market Committee minutes	1987–2006	Effects of transparency on monetary policy	Economics
Heiberger and Koss (2018)	Minutes of the German *Bundestag* (parliament)	1990–2013	Topic scouting	Sociology

"Topic scouting" refers to papers which apply topic models with the primary goal to that of identifying topics. "TM methodology" refers to papers that apply topic models to discuss methodological issues. Asterisks mark papers that use a different topic model than LDA. Department is recorded according to authors' affiliations

Jockers (2013) uses topic models for a corpus, which, at first glance, does not appear to be particularly relevant for economic historians (nineteenth-century novels from Great Britain and the US). Nevertheless, his work illustrates how topic models can be combined with the metadata of the documents and, in this way, be further refined. In particular, he records the authors' gender and nationality, which allows him to show that, for example, female authors write more about "Affection and Happiness" than their male counterparts.

Fligstein et al. (2017) use topic models to answer the question of why the Federal Reserve (Fed) failed to predict the financial crisis in 2008. They particularly use the topics found in the Federal Open Market Committee's (FOMC) minutes to measure how the Fed perceived the US economy between 2000 and 2008. They were able to show that the Fed was neither aware of a housing bubble nor of the entanglement of the housing and financial markets. Hansen et al. (2018) use the same database to investigate how transparency affects the deliberation of monetary policymakers.

That topic models can be combined with econometrics and economic data is shown by Hansen and McMahon (2016). Also studying the FOMC minutes, they investigate the effects of central bank communication on macroeconomic and financial variables. Another example of the integration of topic models into economic analysis is provided by Lüdering and Winker (2016). They study the question of whether economic research anticipates changes in the economy or merely looks at the economy from an ex post viewpoint. They apply a topic model on the *Journal of Economics and Statistics* and compare the temporal occurrence of topics connected to the inflation rate, net exports, debt, unemployment, and the interest rate to their corresponding economic indicators. Scientific journals also comprise the sources of Hall et al. (2008), Mimno (2012a), and Riddell (2014).

How topic models can be used for research in finance is shown by Larsen and Thorsrud (2015), Larsen and Thorsrud (2017), Thorsrud (2016a), and Thorsrud (2016b), who all build on the same corpus (articles published in a Norwegian business newspaper between 1988 and 2014). Here, the topics of the newspaper are used to predict asset prices (Larsen and Thorsrud 2017) and economic variables (Larsen and Thorsrud 2015). Furthermore, they are used to construct a real-time business cycle index for so-called nowcasting (Thorsrud 2016a).[45]

5 Topic modeling the JEH: Whaples reloaded

When testing something new, it can be helpful to know what the results should ideally look like. That is why in the following, topic models will be used to identify themes in the *Journal of Economic History*. The JEH is chosen as a case study because, including Whaples (1991, 2002), there are two works which deliver an invaluable benchmark. Whaples classified the content of the JEH according to a modified version of *Journal of Economic Literature's* (JEL) codes, counting the

[45] In finance, there appears to be an affinity toward text as data, which can be traced back to Tetlock (2007), who was the first to use textual analysis in order to measure market sentiment.

percentage of pages published in a given category. In contrast, the topic model works without an ex ante classification scheme, and it works automatically. In other words, the following can be understood as a kind of Turing test for topic extraction (Andorfer 2017).

A topic model is applied to two text samples. The first one includes all articles from Volume 1 to Volume 50, Number 2 using 41 topics just as in Whaples (1991).[46] The second sample extends the analysis into the present, consisting of all articles published between 1941 and 2016. Here, a topic model with 25 topics is used, which corresponds to the number of subjects in Whaples (2002).

In both samples, the topics were generated with MALLET, using 2000 iterations and allowing for hyperparameter optimization (that is, allowing topics and words to have different weights). The corpus was preprocessed in the following manner. Regular expressions from the header on the first page and the copyright section of every paper were deleted. The documents consist of bibliographical text to a large degree, which distorted the topics in the first trials. Therefore, the most frequent expressions related to bibliographical references were removed. This mainly concerns places of publications. For instance, each variation of “university press” was removed, as was every occurrence of New York, Cambridge, London, and Oxford in a bibliographical reference.[47] Names of universities were not removed as they may form part of a subject like disciplinary history. Furthermore, the expressions “per cent” and “New York” (if not in a reference) were merged into “percent” and “newyork” as “per” and “new” are part of the stoplist.[48] Furthermore, download signatures had to be removed.

For the stoplist, the MALLET built-in list was used, as was the built-in tokenizer which removed capitalization and numbers. Further preprocessing steps like stemming (reducing words to a common stem) were not applied in order to keep the process as transparent as possible. In total, the overall database consists of 2675 articles or 19.8 million tokens, which is approximately 35 times the amount of text in “War and Peace”.[49]

MALLET ostensibly provides two types of output. First, it produces the topic keys, which displays the most probable words for every topic (the number of words displayed can be varied by the user). Second, it generates a file containing the topic shares (or distributions) for every document which add up to one. This makes it possible to identify the most prominent topics for every article and to calculate average topic shares for every topic. In particular, by using the timestamp of every document, we can compute the time series of topic prevalence, allowing us to investigate publication trends.

[46] That is, all articles published in regular and Task issues except regular book reviews and dissertation summaries, see Whaples (1991).

[47] In the first trials, almost all topics contained the word *Cambridge*. Other cities that occur in the final topics were not found to appear regularly in bibliographical references except in combination with “university press”.

[48] As ‘york’ and ‘cent’ occurred in several early topics, it became clear that in fact New York and per cent was meant. Thus, this step was taken for reasons of clarity and esthetics.

[49] The stopwords can be received upon request. For sample one, the database consists of 1728 documents.

Table 2 Topics in the Journal of Economic History, 1941–1990

No.	Most probable words	Label	ATS
0	Company firms industry oil companies production firm industries coal industrial steel market american research competition standard u.s. electric manufacturing large sales small business corporation plant size plants largest petroleum gas	Other industry studies*	1.75%
1	Japanese japan–china chinese rice land period government meiji tokugawa tokyo development agricultural tax modern economic taiwan osaka irrigation samurai merchants rural century village modem population history han shanghai traditional	Country studies, Japan and China	0.9
2	Prices price trade demand goods exports market supply imports production export index period products terms years commodities consumption decline increase rise year real percent rose increased markets century domestic commodity	Prices*	2.7
3	Economic history historical work historians theory study analysis studies point question discussion problem research view questions problems book review fact historian data professor past approach time evidence general economics recent	History of economic history*	8.7
4	Railroad railroads canal transportation construction cost railway costs canals social miles western freight river railways pacific water road ohio rail roads lines erie traffic transport improvements line system central rates	Transportation*	1.3
5	Growth income capita economic real rate output population agricultural estimates national product labor percent sector agriculture economy consumption increase change share century productivity rates force farm gross distribution index relative	Economic growth*	2.6
6	Percent table data total average period year years estimates number rate annual rates source series figures figure sources time estimate large index based estimated statistics appendix percentage made increase ratio	Descriptive language/time series	7.9
7	Bank banks banking deposits reserve national notes system state credit financial money loans deposit federal commercial capital assets states reserves bankers private savings newyork specie country monetary bank's currency free	Banking*	1.6
8	Capital investment long series united growth depression british cycle fluctuations cycles migration movements states period economic american population swings business construction building emigration economy demand australia expansion unemployment net great	Business cycles*	1.5

Table 2 (continued)

No.	Most probable words	Label	ATS
9	States american united newyork state massachusetts boston america u.s philadelphia washington pennsylvania england war north journal early john national historical history dollars report americans james connecticut robert d.c william thomas	Country studies*, USA	2.1
10	Labor workers union unions strike national industrial employers strikes trade welfare insurance industry members work hours social wages association benefits organization committee management collective bargaining worker unemployment employer a.f local	Labor*, labor relations	0.9
11	Agricultural agriculture wheat grain farmers crops yields crop farm land production farming food output yield productivity bushels corn dairy harvest animals cattle livestock acre enclosure milk animal meat labor grains	Agriculture*	1.6
12	Social political society class theory capitalism wealth life men man revolution thought classes power human state marx labor economy keynes economics great capital economists ideas capitalist principles religious natural free	History of economic thought*	2.7
13	Law state public government rights laws property political private act interests legal court legislation policy power protection constitution regulation general economic acts courts support vote crown corporations interest cases issue	Law*	1.9
14	Capital interest market rates investment rate financial stock percent loans debt loan credit funds bonds securities return company companies million assets mortgage investments finance shares london markets money exchange investors	Finance	2.2
15	Trade british ships african slave africa ship coast vessels traders slaves voyage shipping century european liverpool west sailing profits freight cargo port voyages goods ports shipbuilding dutch gold herring sea	Trade*, slave trade	1.2
16	Italian italy genoese century medieval venice medici florence merchants del venetian bruges genoa merchant fourteenth della storia business florentine rome fifteenth milan ages roover commercial thirteenth wool insurance cloth branch	Country studies*, Italy	1.0
17	Economic countries foreign industrial development industry british world growth capital trade international europe country united european domestic investment britain production industries policy national states industrialization war great period france germany	International investment*	3.3

Table 2 (continued)

No.	Most probable words	Label	ATS
18	Slaves slave slavery labor free south contract southern servants north engerman sugar white fogel emancipation cost plantation indentured war civil negro work black costs servant planters history market freedom plantations	Slavery and servitude*	1.0
19	Iron steam steel power engine machine industry engines production coal patent patents invention water machines tons pig machinery fuel technology technological products diffusion inventions cost furnaces furnace early process technical	Power/energy industries*	1.5
20	Canadian canada mexico spain spanish latin america brazil madrid mexican brazilian ontario toronto quebec american century royal chile wool rio government mining castile reciprocity indigo percent seville del crown toledo	Country studies*, Colonies	0.8
21	Labor capital productivity output costs production cost factor change american rate technical relative manufacturing scale industry input efficiency inputs technology wage united technological prices british price higher states function economies	Manufacturing*	2.3
22	Gold money exchange monetary currency silver specie standard rate foreign price paper treasury coins inflation coin notes mint circulation dollar real supply bills international market interest prices period series rates	Money*	1.6
23	Time made part years large great fact found important small general make long system number end place good times people high early brought order period case means hand country set	Not specified	12.3
24	Colonial colonies trade tobacco british merchants england sugar american west london english indies shipping britain tonnage planters pounds maryland virginia vessels merchant chesapeake america exports great revolution north middle south	Trade*, imperialism/colonialism*, North Atlantic	1.4
25	Labor workers wage wages women force work earnings employment percent census men children occupations skilled age school female immigrants male job black unskilled immigration occupational occupation education jobs participation schooling	Labor*	2.2
26	German der germany und die des industry von berlin industrial austria hungary austrian hungarian prussian das zur growth deutschen monarchy tariffs geschichte habsburg protection steel marks iron customs development prussia	Country studies*, Germany/Austro-Hungary	0.9

Table 2 (continued)

No.	Most probable words	Label	ATS
27	Cotton farm farms farmers south agricultural agriculture labor land southern crop tenants production acreage farmer census acres plantation california tenant tenancy crops georgia counties size states contracts share acre county	Agriculture*, cotton	1.3
28	South regional regions region cities urban population city north west southern areas states growth central development state local eastern differences market western national east census differentials antebellum northeast interregional atlantic	Regional studies, geographic descriptions	1.9
29	Russian land russia peasant peasants labor century serfs serfdom serf europe population village moscow lord medieval estates rubles agricultural system rural estate agrarian feudal peasantry services demesne rent petersburg manorial	Country studies*, Russia	1.2
30	War government tax expenditures public state policy federal taxes military private income fiscal percent national revenue revenues million budget debt political finance administration controls inflation policies local civil program army	Public finance*, War*	1.9
31	Variables variable model results level equation data significant hypothesis coefficient regression effect coefficients test demand time income positive rate equations expected sample values analysis effects economic u.s evidence estimated function	Econometric language	3.8
32	Population age wealth mortality family fertility children life birth rates families death marriage demographic century number women sample rural living england health household deaths county social rate growth decline households	Demography*	2.0
33	England english century british poor london britain revolution relief wages eighteenth industrial irish history wage evidence counties law wales ireland population early parliamentary laborers nineteenth oxford scotland great parish parishes	Industrialization*, Great Britain	1.4
34	Economic development economy growth system change process social market political structure role institutions century institutional major organization systems production markets resources traditional conditions control analysis society problems important individual early	Economic growth*	7.1
35	French france paris century des les dutch revolution europe eighteenth amsterdam english seventeenth van francs annales histoire sur history archives england livres european economique crisis louis holland vols revue netherlands	Country studies*, France	1.3

Table 2 (continued)

No.	Most probable words	Label	ATS
36	Company committee march january papers june report december april office letter february october september august city november president john house congress business secretary board letters treasury year received directors plan	People	2.5
37	Business history research economic study enterprise american university men years field development entrepreneurs company committee enterprises entrepreneurial businessmen management group public general entrepreneur records individual published entrepreneurship corporation social institutions	Business*	2.5
38	Land lands india indian acres settlement iowa county illinois frontier acre western cattle price prairie federal settlers area speculators farm grant grants counties kansas large property state sales history west	Imperialism/colonialism*, westward movement	1.3
39	Cotton industry textile mills cloth factory spinning production mill firms workers england textiles looms quality factories silk machinery manufacturing manufacturers labor manufacture industrial weaving industries work woolen weavers learning yarn	Industrialization*, textile industry	1.3
40	Empire trade merchants greek ancient jewish roman ottoman century egypt merchant economic world balkan jews east greece byzantine commerce palestine traders state evidence silver arab b.c greeks goods mediterranean middle	Trade*, ancient trade	0.6

The table shows the 30 most probable words for each topic in descending order. ATS stands for average topic share over the corpus in percent, asterisks mark labels used by Whaples (1991). Topic numbers were randomly given by MALLET. *Source*: See text

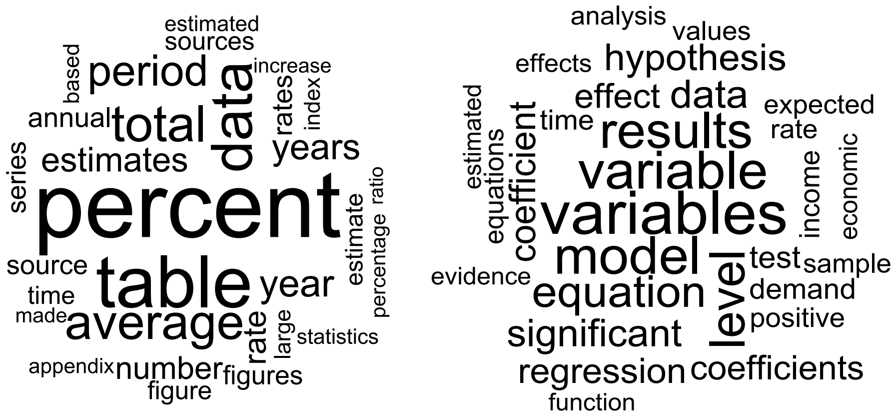

Fig. 3 Topic 6 "Descriptive language" (left) and topic 31 "Econometric language" (right). Both topics were inferred from sample 1. *Source*: See text

The topics of the first sample are shown in Table 2. The first column states the topic number randomly given by MALLET. In the second column, the 30 most probable words for every topic are shown in descending order. For example, in topic 1 *japanese* is the most probable word, followed by *japan*.[50] The relative importance of words for a topic may be better visualized using word clouds (as in Figs. 1, 3).[51]

In most cases, the topics appear to clearly exhibit what one would expect when thinking of topics and they show a great degree of coherence. For example, in topic 11, the words *agricultural, agriculture, wheat, grain, farmers,* and *crops* suggest that this topic is most likely to be about agriculture. That topic 18 can be labeled *Slavery and Servitude* is not only justified by words like *slaves, slave,* and *slavery* but also by the reference to Robert Fogel and Stanley Engerman.

A topic which stands out is topic 36, which (at least for the author) cannot be interpreted intuitively. When looking at the articles which show the highest share of topic 36, it becomes clear this topic covers research concerning people. They either cover individuals, like the article by Walters and Walters (1944) on David Parish (48%), or groups of people like the article by Freeman Smith (1963) on the international bankers committee on Mexico (48%). The numerous occurrences of months seem to derive from the fact that these articles are largely based on correspondence with references containing the date of the original letter.[52]

Topics 6 and 31 (Fig. 3) show that topics can also represent a different type of theme, in this case the use of technical expressions typical for quantitative methods. Topic 6 contains words which can be attributed to basic descriptive statistics. In

[50] If a stemmer had been used, these words would have been collapsed into *japan*.

[51] Depicting every topic as a word cloud would exceed the available space of this article.

[52] Footnote 14 in Walters and Walters (1944) may serve as an example: "[…] Parish to John Craig, March 1, 1806, to Villaneuva, March 18, 1806, to Robert and John Oliver, October 29, 1806, in Parish LB, I, 239, 290, 291; II, 5."

particular, words like *period, year(s), series, annual, time,* and *index* are terms connected to time series. Topic 31 contains words which could be found in the glossary of a textbook on econometrics. Obviously, the topic model differentiates between comparatively descriptive and econometric methods. Topic 23, having the highest average share of all topics, appears to contain general expressions which could be typical for an economic historian's jargon.

Wherever they seem appropriate, subjects from Whaples (1991) were added as labels.[53] If this was not the case, a new label was given.[54] The overall impression is that the topics seem to match the subjects used in Whaples (1991) quite well. From the 41 subjects, 26 can be identified, including nearly all the major ones.

In some cases, the topics seem to be more highly differentiated than the subjects. Topics such as *Japan and China* (1), *Germany* (26), and *France* (35) could, of course, be assigned to "Country Studies" but they are identified as independent subjects by the topic model.[55] The same holds true for "Trade": The topic model finds different subcategories like *Slave Trade* (15) or topic *North Atlantic* (24). The subject "Economic Growth" appears to be split into two topics, one describing growth (topic 5) and one explaining it (topic 34). Topic 38 could be attributed to "Imperialism/Colonialism", but a label like *Westward Movement* could be deemed more appropriate.

The topic model also differentiates between geographical and sectoral aspects of industrialization. Topic 33 contains words relating to Great Britain as the first country to industrialize, whereas topic 39 shows words referring to the textile industry as a central sector concerning industrialization. Furthermore, some topics are connected to different subjects. For example, topic 30 contains words which could belong to both "Public Finance" and to "War", which is not surprising as a common, major portion of public spending is on military purposes.

The topics about individual countries draw attention to the question of different languages. Words like *der, die, das,* or *des, les, sur* would be regarded as stopwords in a German or French corpus.[56] Of course, these words could be removed by expanding the stoplist. They facilitate the identification of documents which build on sources in languages other than English, which, for example, could support research concerning geographical coverage. In the topic on France, the words *annales* and *histoire* may be regarded as a reference to the *Annales School* and its major journal *Annales d'histoire économique et sociale* (Burguière 2009).

The topic model did not identify any subjects from Whaples (1991), which could have several explanations. The subject may just be too small compared with the corpus (like in the case of "Minorities/Discrimination"), which could possibly be

[53] These subjects are based on JEL codes. See Whaples (1991, pp. 289–90).

[54] Of course, this assignment is somewhat subjective, but it is no more subjective than assigning pages to subjects by hand.

[55] Except for Canada, which shares a topic with other countries (Topic 20), every country analyzed in Whaples (1991) is comprised of a separate topic. These countries are Britain (33), France (35), Italy (16), Germany (26), Japan (1), Russia/Soviet Union (29), and the United States (9).

[56] These words most probably stem from bibliographical references, which often remained untranslated.

solved by increasing the number of topics or by reducing the corpus into a subsample. Here, the agnostic nature of the model once again comes into play. Searching for a subject on minorities may be legitimate in a certain framework. However, the model did not identify this topic as being substantial at the given level of granularity, i.e., given the number of topics, the model assesses "Minorities" as irrelevant.

Another reason could be that different subjects share a similar type of vocabulary (or meaning) and therefore cannot be separated by the topic model (like "Business Cycles" and "Recessions/Depressions"), which again could imply that they are not clearly specified. Compared to his 1991 study, Whaples (2002) combines several subjects which may point in this direction. Another theoretical, although not very likely, reason could be that a subject does not have any specific vocabulary and therefore is untraceable for a topic model.

Expanding the analysis into the present, another topic model is run on all articles between 1941 and 2016, this time with 25 topics as in Whaples (2002).[57] The results are shown in Table 3 (the development of all topics can be found in "Appendix 1"). Again, the labels where chosen as in Whaples (2002), wherever they fit the topics. From the 24 subjects used in Whaples (2002), 17 can be attributed to topics.[58] In principle, the reduction of the number of topics will lead to more coherent but also more general topics. Of course, neither table can be compared directly as the reduction in topics does not happen ceteris paribus. Nevertheless, some general observations can be made. The topic of *Industrialization* in the second sample is again spread over two topics. One topic comprises references to Great Britain as the place of the first industrialization (topic 14). This time, the second one is much broader. Topic 3 contains references to the textile industry as well as other early industries. When looking at the documents with the highest share of topic 3, it becomes evident that their common theme is technology.

Reducing the numbers of topics creates a subject which is omitted in Whaples (2002) but was used in its predecessor. Topic 16 appears to cover several countries, bringing back the subject "Country Studies". Again, there is one topic containing econometric vocabulary (topic 6), although the words are slightly different. In the case of the *Descriptive Language* (8), there now appears to be a stain of words connected to economic growth.

The results of the topic model can be used to describe some general trends in the JEH. Judging by the development of topic shares in sample two (see "Appendix 1"), *Methodology and Disciplinary History* (20) has experienced a major decline from the very beginning (with the exception of 1959/60), a finding consistent with Whaples (1991, 2002).[59] The same holds true for *People* (11) and *Economic Growth* (21). The most prominent topic is *Economic Growth* (21) with an average topic share of 16.6%. During its heyday in the 1960s, *Economic Growth* achieved almost 27% (1967), which mirrors economic history's focus on economic development at the

---

[57] A direct comparison of the results in Whaples (2002) as in sample one could have been carried out as well but was relinquished due to space restrictions.

[58] The 25th subject in Whaples (2002) is the residual "Other".

[59] The peak of 1960 can be attributed mainly to the Task issue containing a nice punchline: the article with the highest share of topic 20 is Goodrich (1960), which discusses how the use of quantitative methods affects economic history.

Table 3 Topics in the Journal of Economic History, 1941–2016

No.	Most probable words	Label	ATS
0	Age wealth population percent family children migration women table household fertility income census families marriage immigrants number households sample rates men years states rural migrants inequality economic data total mortality	Demography*	3.7%
1	Japanese japan india china chinese indian rice government period development land asia tokyo century economic population modern history meiji economy asian price early tokugawa prices taiwan osaka agricultural cotton system	Asia	1.4
2	Education health mortality school percent height schools disease rates population schooling birth public age states human years high educational children data rate water united income malaria heights diseases life death	Standard of living and health*	1.7
3	Industry production cotton iron industries technology firms manufacturing power industrial textile mills costs steam machinery american output technological steel technical productivity cost machine british coal machines percent spinning capital cloth	Industrialization*, technology*	3.8
4	Cotton south slaves slave black southern slavery white labor blacks carolina north states free war plantation american whites georgia racial civil race antebellum history negro fogel engerman northern state percent	Slavery and servitude*	1.7
5	Land agricultural farm agriculture farmers wheat production farms labor percent grain crop acres prices crops yields farming acre productivity tenants output cattle harvest average acreage price corn yield number year	Agriculture and land*	3.7
6	Data table results variables variable significant sample effect model percent effects level economic time average coefficient number regression analysis equation coefficients change year estimates test estimated information evidence journal period	Econometric language	7.0
7	Bank banks banking loans credit financial state national reserve deposits capital newyork states interest percent rates loan market system federal notes assets funds deposit commercial money insurance rate total bankers	Banking and credit*	3.1
8	Growth percent income prices output table price estimates data rate series period real index productivity economic total capital labor average capita production rates relative national year industrial united consumption years	Descriptive language	7.8

Table 3 (continued)

No.	Most probable words	Label	ATS
9	Trade british colonial colonies ships slave percent west african slaves shipping tobacco dutch ship africa vessels merchants century prices price american sugar coast servants london eighteenth america english history north	Slave trade	2.9
10	Railroad states railroads regional west transportation cities south city newyork american region regions united construction costs canal cost state western north central railway ohio urban rates percent east railways miles	Transportation*	3.1
11	Made business years time american company general committee great government men part john papers found year public make war office March order fact trade william records good January June April	People	7.7
12	War tax government public state taxes percent military expenditures private fiscal revenues income soviet revenue spending total federal national policy control million local housing political years administration economy taxation budget	Public finance*	2.4
13	State states political law american canadian united federal congress government laws u.s act vote canada public legislation policy support voting interests party national regulation report reform politics economic bill power	Political*	2.3
14	England revolution century english history eighteenth poor london british wages population early britain industrial living economic evidence irish common enclosure europe ireland medieval modern review society parish wales towns relief	Industrialization*, Great Britain	2.7
15	French france paris century des italy empire les ottoman italian merchants roman trade europe medieval middle early eighteenth egypt merchant centuries venice commercial european history rome genoese livres greek histoire	Mediterranean relations	2.6
16	Mexico russian latin russia spanish spain mexican brazil america government percent century economic sugar colonial madrid brazilian opium cuba development del foreign rio moscow political land argentina serf peasant peru	Country studies	1.4
17	Gold money exchange rate monetary market interest price rates debt percent currency prices silver standard financial stock government foreign period policy real inflation bonds coins specie london paper war crisis	Money*	3.4
18	German germany patents der patent und die des berlin industrial von invention inventors economic austria patenting inventions industry prussia coal inventive hungary percent prussian das zur deutschen deutsche habsburg market	Germany/Austria-Hungary	1.3

Table 3 (continued)

No.	Most probable words	Label	ATS
19	Property rights law land legal contracts costs contract court institutions private cases trade institutional system crown common courts case enforcement rules company apprenticeship apprentices pay claims bay cost masters political	Industrial organization*	2.5
20	Economic history social work historical political business theory research historians society development world industrial study economics economy american studies class economists life great science institutions revolution production capitalism knowledge growth	Methodology and disciplinary history*	7.5
21	Economic growth period system development capital change time important case general fact made economy century part point conditions demand problem paper large analysis question long major evidence process discussion market	Economic growth*	16.6
22	Trade british countries united world foreign states exports britain international domestic economic imports tariff european prices price europe country export american goods percent france war kingdom markets market germany import	Trade*	3.5
23	Firms company companies market business firm stock capital investment industry percent oil shares corporate price corporations large financial sales investors profits competition corporation share ownership number information limited insurance private	Business*	2.8
24	Labor workers wage wages work employment earnings unemployment force women percent hours worker census working rates employers industrial men industry skilled jobs job unions average number employed time report manufacturing	Labor and migration*	3.2

The table shows the 30 most probable words for each topic in descending order. ATS stands for average topic share over the corpus in percent, asterisks mark labels used by Whaples (2002). Topic numbers were randomly given by MALLET. *Source*: See text

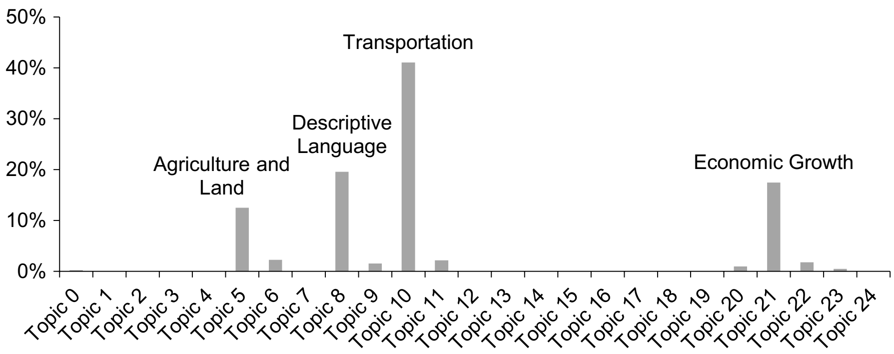

Fig. 4 Topic distribution of Fogel (1962). *Source*: See text

time (Haupert 2016). In other words, every article in 1967 consisted, on average, of more than one quarter of words connected to this topic.[60] Since then, academic interest has constantly declined (see "Appendix 1"), a finding that is consistent with Whaples (1991, 2002).

Technically, every document comprises a share of every topic, even though it may be vanishingly small. Defining a topic as "substantial" if it has a share of 10% or more, articles in the JEH contain an average of 3.2 topics, which, since 1941, has changed only marginally.[61] The same continuity can be stated for topic concentration with an average Herfindahl index of 0.24 per article. These findings are most likely due to the nature of the JEH as a specialists' journal. In general, looking at the topic distribution of a document can provide an insight into what it is about. To give a prominent example, the topic distribution of Fogel's (1962) railroad paper is provided in Fig. 4. Not very surprisingly, the most prominent topic is *Transportation* with a topic share of 41%.

Topic shares can be used to investigate topic correlation. Calculating the topic correlation based on individual documents mainly yields uncorrelated topics, except for *Econometric Language*, which shows some negative correlation with *People*, *Methodology and Disciplinary History* and *Economic Growth*.[62] The low correlation probably results from to the low number of topics within the documents. Nevertheless, topics may correlate across time, in terms of several topics occurring together, resulting from larger topical trends. Computing the correlation coefficients based on annual topic shares increases the number of correlated topics and confirms what can reasonably be expected. For example, there is a high correlation between *Standard of Living and Health* (2) and *Econometric Language* (6). The topics *People* and *Methodology & Disciplinary History* stand out as they are correlated negatively to almost every other topic. On the positive side, *Econometric Language* is the topic

[60] The use of annual means is of course prone to outliers. If one is interested in the long-term development, a moving average would probably be more appropriate. However, the outliers could be what we are looking for if we are interested in identifying special events.

[61] The same continuity holds true at a 5 and 20% threshold.

[62] The correlation matrix is available upon request.

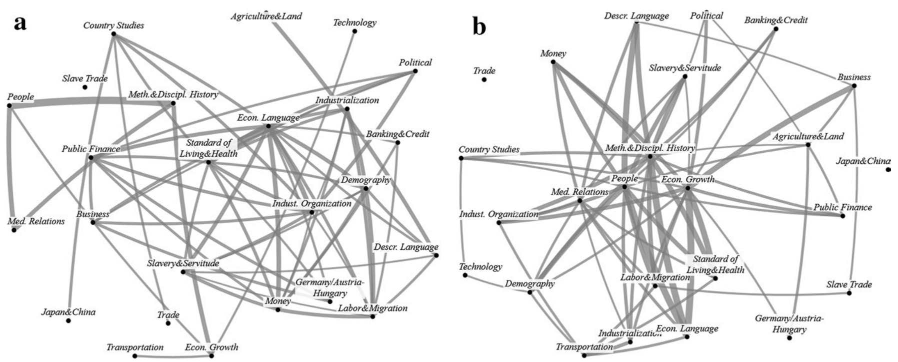

Fig. 5 **a** Positive topic correlation. **b** Negative topic correlation. *Notes*: Width of connecting lines is proportionate to the value of the corresponding correlation coefficient, including only coefficients with absolute values of at least 0.3 and which are significant at the 5% level. Correlation coefficients are computed based on the annual means of topic shares of sample 2. *Source*: See text

most correlated to others, which confirms its quality as a meta topic. A network representation of topic correlation can be found in Fig. 5a, b, illustrating the connection between topics based on their correlation. The question of correlation should be at the heart of further research, e.g., by using a type of topic model which explicitly accounts for correlation (Blei and Lafferty 2007). Another option for future research could be the inclusion of article metadata such as author information (as in Whaples 1991, 2002) and the comparison with other economic history journals.

As stated above, the crucial step in topic modeling is setting the right number of topics *K*. In this paper, this is solved through its purpose of comparison. Yet, what is the "natural" number of topics in the JEH? There are several metrics with which to identify the correct number of topics, which, ostensibly, work by running multiple topic models with different *Ks*, computing a measure for every topic model, and then identifying the extremum.[63] Again, these metrics give no exact number, but instead present a range within which the optimal *K* can be found.[64] Using the *R* package *ldatuning* developed by Murzintcev Nikita (2016),[65] the optimal number of topics in the JEH (including Task issues) appears to be somewhere in the region of 80 (see "Appendix 2"). Running a topic model with 80 topics still mainly produces

[63] See Arun et al. (2010), Cao et al. (2009), Deveaud et al. (2014), and Griffiths and Steyvers (2004) for detailed explanations of each metric. It is important to note that these metrics only deliver the optimal number of topics from a technical point of view. In the end, the optimal *K* depends on the research question.

[64] Computing these metrics takes a considerable amount of time. For the articles in sample 2, it took 4 days per metric on a standard computer. It is therefore necessary to retain the span of *K*s at a manageable level, resulting in a coarse span of topics. It is important to note that these metrics only deliver the optimal number of topics in a technical sense. In the end, the optimal *K* depends on the research question.

[65] See https://cran.r-project.org/web/packages/ldatuning/ldatuning.pdf.

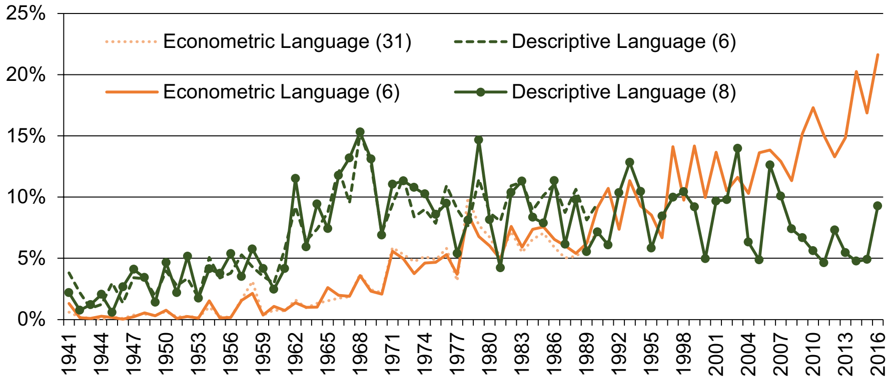

Fig. 6 Topic shares of quantitative topics. Dotted lines mark topics from sample 1 and solid lines mark topics from sample 2; annual means. *Source*: See text

easily interpretable topics. Naturally, they become much more granular. For example, countries which were previously grouped together now form their own topics.[66]

6 The cliometric revolution in topics

The methodological topics lead us to a subject recently addressed by Diebolt and Haupert (2018) and Margo (2018), which is also covered by Whaples (1991, 2002): the shift in economic history toward economic theory and quantitative/econometric methods during the 1960s, known as the cliometric revolution.[67] Can this shift be observed in the topics? When looking at the distribution of the methodological topics over time, the answer is clearly yes.

Figure 6 shows the annual average topic shares of the methodological topics in both samples.[68] There is a continuous rise in the econometric topics, beginning in the 1960s, a finding which is completely consistent with Diebolt and Haupert (2018), although the rise in the econometric topics is not as steep as in their measure, and their first peak in the early 1970s cannot be found.[69] This may be due to the fact that they do not include the Task issues, which, until the late 1960s, contained more disciplinary reflections and less cliometrics than regular issues (Whaples 1991, p. 293).[70]

[66] Due to space restrictions, the topics are not presented here; they are available upon request.

[67] For a comprehensive history of cliometrics see Haupert (2016) and the cited literature.

[68] The econometric language topics exhibit almost identical shares in both samples indicating that they are relatively congruent. The descriptive topics exhibit a degree of difference, because, in sample 2 this topic appears to be less coherent than it does in sample 1.

[69] Diebolt and Haupert (2018) count equations, tables, and graphs per page. See Fig. 8.

[70] Until 1996, papers presented at the annual meetings of the *Economic History Association* were published in a fourth issue, which was devoted to the “Tasks of Economic History”, see Diebolt and Haupert (2018, p. 22) and Margo (2018, p. 12).

The integration of economic history into economics has recently been studied by Margo (2018). He finds that the expansion of econometric language in the JEH was delayed compared with general economic journals like the *American Economic Review* or journals concerning labor economics. Furthermore, he finds that for the JEH, the level of his index measuring econometric language is below those of every other journal in his study.[71] In using econometric language as a proxy for the use of econometrics, Margo's approach is relatively similar to the topic model approach in this paper. Consequently, it comes as no surprise that his observation, being that the JEH's language became more econometric in the 1960s, can be confirmed. The development of his index of econometric language, which is based on six search terms,[72] is akin to the development of topic 6 and topic 31 shown in Fig. 6. Still, it is important to note that by using a topic model we are not required to specify the search terms ex ante. Instead, the model identified this theme without any prior knowledge, which again emphasizes the agnostic nature of the model. Therefore, the topic could provide candidates for "an exhaustive list of words and phrases that objectively characterize what is meant by 'econometric language'" which Margo (2018, p. 10) misses.[73] There is another difference between topic models and dictionary approaches. With a dictionary, it is necessary to use unambiguous terms, which limits the search list. By allowing for polysemy (see chapter 2) the topic model additionally includes words which also have a non-econometric meaning (like *test* or *significant*).

Remaining is the question posed by Margo (2018, p. 3) regarding why the JEL lagged economic journals in the use of econometrics. As it covers only the JEH, this question cannot be answered thoroughly in this paper. Still, the topic model could shed some light on this question as it suggests that focusing only on the spread of econometrics is not sufficient in capturing the entire extent of the methodological developments of the JEH. Margo (2018, pp. 18–19) points out that, although early cliometric work did apply econometric methods, it discussed the results only briefly. Accordingly, only a low degree of econometric terminology can be anticipated.[74] The low share of econometric vocabulary before the mid-1960s does not necessarily imply that papers published in the JEH did not use quantitative methods. On the contrary, the topic model identified a second, more descriptive topic. The development depicted in Fig. 6 can be interpreted as a gradual integration of ever more advanced quantitative methods over time mirrored in a linguistic shift. The descriptive topic was present well before 1960, indicating that, although at a low level, quantitative methodology was used before the arrival of econometrics, which is also consistent

[71] The journals analyzed in his study are: the *American Economic Review, Explorations in Economic History,* the *Journal of Economic History,* the *Industrial and Labor Relations Review,* and the *Journal of Human Resources*.

[72] Margo (2018) uses an index based on the terms *regression, logit, probit, maximum likelihood, coefficient,* and *standard error*.

[73] Another candidate for words and expressions that characterize econometric language are the indices and glossaries of econometric textbooks, which we use in an ongoing project. This provides the advantage that levels of methodological advancement can be differentiated between by using indices from introductory and advanced textbooks.

[74] This touches on the issue of changes in the use of language discussed earlier.

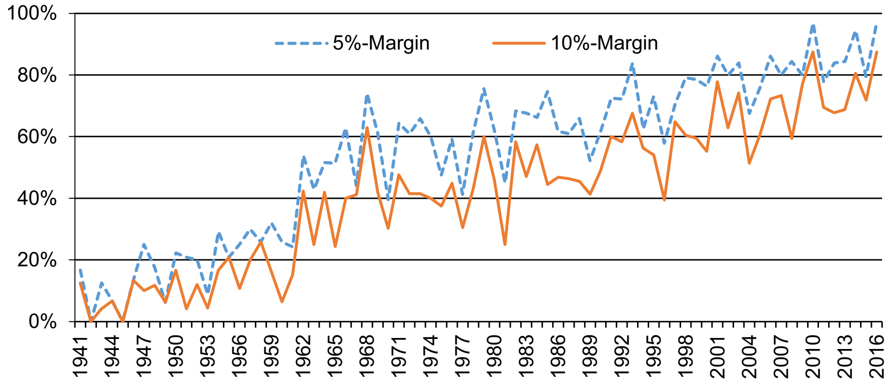

Fig. 7 Share of quantitative articles. Number of quantitative documents per year divided by the total number of documents per year. Documents are classified as quantitative if their share of topic 6 or 8 amounts to 5% (10%) or more. *Source*: See text

with Diebolt and Haupert (2018).[75] In other words, the JEH had become quantitative before it became econometric, which, from a methodological point of view, appears to be a natural course of events.[76] Considering this and the fact that before Douglass North and William Parker became editors in 1961, the JEH was dominated by "old" economic historians (Diebolt and Haupert 2018), the delay of econometrics may be less surprising. It might have been just the consequence of a delay in the use of quantitative methods in general, which had to become established before more advanced methods could be disseminated. This point certainly requires further research, e.g., by also extending the topic model to economic and historical journals.

Figure 6 shows the intensity of the use of quantitative methods. On average, papers became more cliometric during the 1960s. But was this development accompanied by an increase in the number of cliometric articles? To measure the extent of the cliometric revolution, another feature of topic models is applied. These models can be used to classify articles according to their content. An example is given in Fig. 7. Articles were classified as being "quantitative" if their share of either topic 6 or 8 (the two topics related to economic methods) amounted to at least 5%. In total, there are 1583 papers classified as being quantitative, which equals a share of 59.2% of all papers in the corpus. Compared with an average topic share of 4% in the overall corpus (median 0.04%), the 5% threshold appears to be appropriate. Additionally, a narrower definition of "quantitativeness" was used by increasing the threshold up to 10%, which yields 1216 quantitative articles (45.5%).

In both cases, the development of the 1960s now bears for more similarity to the one described by Diebolt and Haupert (2018). Still, we can note a continuous rise after the 1970s, a difference that could again result from the different database. To

[75] To cite just one example: with a share of 32%, Kuznets' (1952) study on US national income before 1870 is among the papers with the highest share of the descriptive topic 6.

[76] I thank one anonymous referee for the remark concerning the fact that this distinction between quantitative (in the sense of mere counting) and econometric approaches was made already by early cliometricians. See, e.g., McCloskey (1978, 1987).

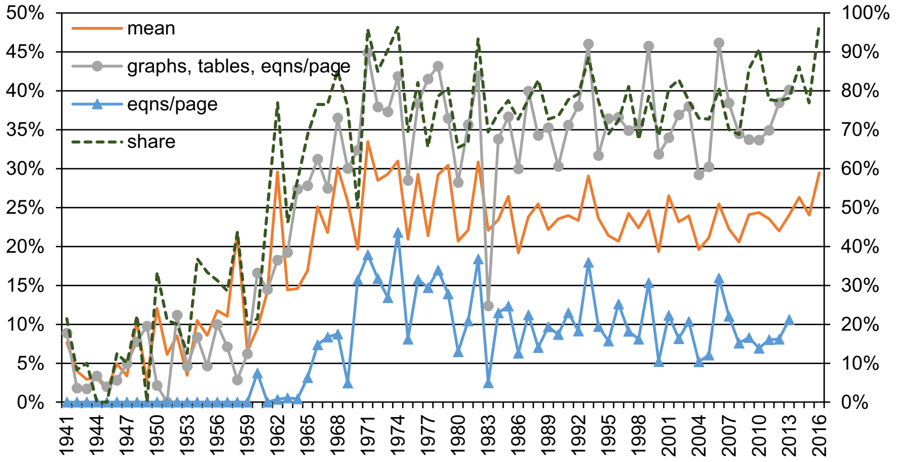

Fig. 8 Four different measures of quantification. The dotted line refers to the right scale. Share refers to the number of quantitative documents at the 10% threshold per year divided by the overall number of documents per year. Mean (per annum) is calculated based on the sum of shares of the two quantitative topics in every article. All numbers are computed based on the corpus excluding Task issues. *Sources*: Author's own computations, Diebolt and Haupert (2018)

account for the Task issues effect, another topic model is run on all articles excluding the Task issues, which again delivers two quantitative/methodological topics (the words constituting these topics can be found in "Appendix 3"). The results presented in Fig. 8 show that the share of quantitative articles reached a peak of 95% at the 10%-level in 1971, since then remaining above 70% most of the time.

Following Diebolt and Haupert (2018), a further topic for future research could be addressed by the question of whether topics are, to some degree, dependent on the JEH's varying editorship regimes. Judging by the topic models developed in this paper, changes in topics can hardly be connected to changing editors. This could be explained by the fact that, especially since the mid-1970s onward, the two co-editors' terms have often overlapped.

7 Conclusion

In this article, I present a state-of-the-art method from digital humanities: topic models, which are statistical algorithms that extract themes (or, more generally, categories) from large collections of texts. I introduce the basic principles of topic modeling, give an initial review of the existing literature, and illustrate the capability of topic models by decomposing 2675 papers published in the *Journal of Economic History* between 1941 and 2016. By comparing my results to traditional scholarship on the JEH and to current research on the cliometric revolution, I have been able to show that topic models are a sophisticated alternative to established classification approaches. Without any prior specification, the topic model identifies two topics containing terms connected to quantitative research. By using the temporal distribution of these topics, the model can retrace economic history's shift toward economics during the 1960s. Further research could include a topic model analysis of purely economic and historical journals in order to infer topical reference points, and, of course, of other journals from economic history, to gain a more comprehensive perspective on the discipline.

For economic historians, the three main strengths of topic models are efficiency, objectivity and quantification. They provide the means for analyzing a myriad of documents in a short amount of time; they are agnostic in terms of waiving ex ante classification schemes such as JEL codes, thereby avoiding the risk of human biases; and they deliver quantitative representations of texts which can be integrated into existing econometric frameworks.

Especially the latter point makes topic models a worthwhile approach for economic historians. As part of the wider approach of distant reading (Moretti 2013), they provide the opportunity to reintegrate textual sources into economic historians' research. One conceivable application could be the generation of historical data. As the research in finance described in the literature review has shown, topic models can be used to predict developments on financial markets and short-term economic development. Instead of predicting the future, this approach could be transferred to settings with a lack of historical data. For instance, applying topic models to historical newspapers could yield surrogates for financial and macroeconomic data.

Topic models provide a useful tool for reducing complexity, identifying relevant sources, and generating new research questions. By their very nature, they possess the unifying potential of interdisciplinary scholarship. As "the future of economic history must be interdisciplinary" (Lamoreaux 2015, p. 1251), topic models represent one step toward securing the significance of economic history. If it is true that "our tribe has been particularly adept at drawing on metaphors, tools, and theory from a variety of disciplines" (Mitchener 2015, p. 1238), economic history should use this ability and integrate digital tools such as topic models in its toolkit. Building on the distinct propensity to work empirically, digitization will not be a threat but rather an opportunity for economic history to become a role model in uniting traditional quantitative analysis, digital methods, and, by a return to some "old" economic historians' virtues, thorough study of narrative sources. As Collins (2015, p. 1232) phrases it: "It may […] improve the economic history that we write by ensuring our exposure to state-of-the-art methods and theory." This article hopes to provide some of this exposure.

Acknowledgements I am grateful to Claude Diebolt and Michael Haupert for generously sharing their data, Robert Whaples and Ann Carlos for insights concerning the JEH, and two anonymous referees for invaluable comments and suggestions on the manuscript. I thank Manuel Burghardt for patiently answering my technical questions on topic modeling, and Mark Spoerer, Tobias Jopp, and Katrin Kandlbinder for their continued support. Finally, I am very grateful to the participants in the research seminar in economic history as well as the lecture series on Digital Humanities at Universität Regensburg.

Appendix 1

See Fig. 9.

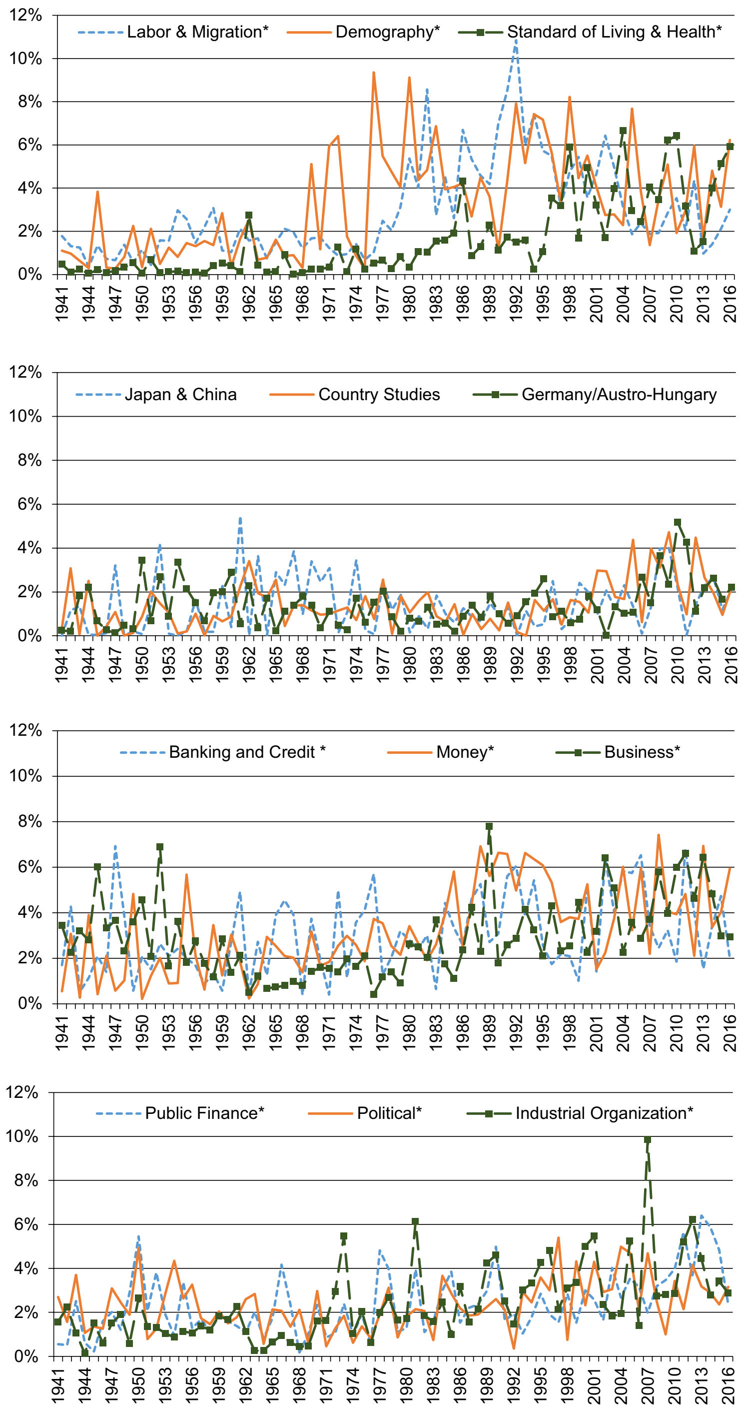

Fig. 9 Topic development of sample 2. Asterisks mark labels used in Whaples (2002); annual means. *Source*: See text

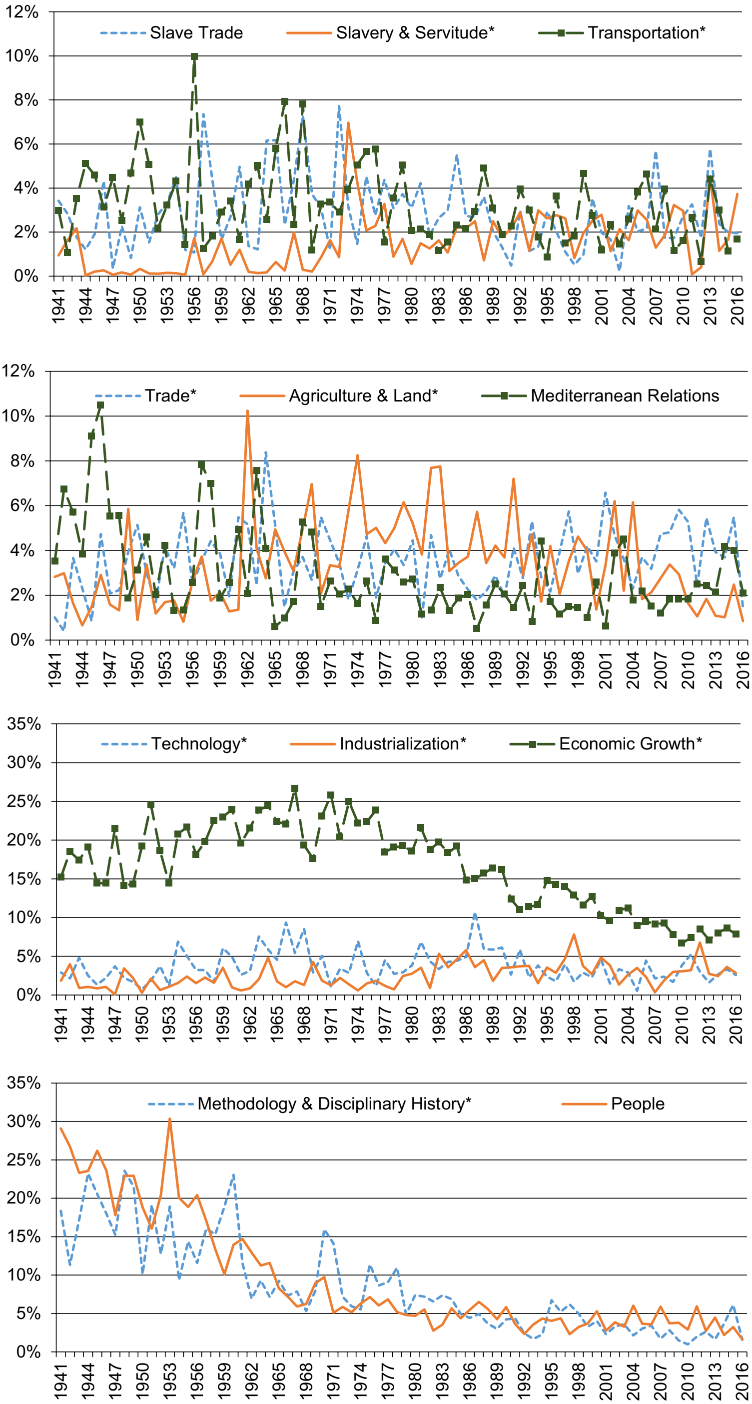

Fig. 9 (continued)

Appendix 2

See Fig. 10.

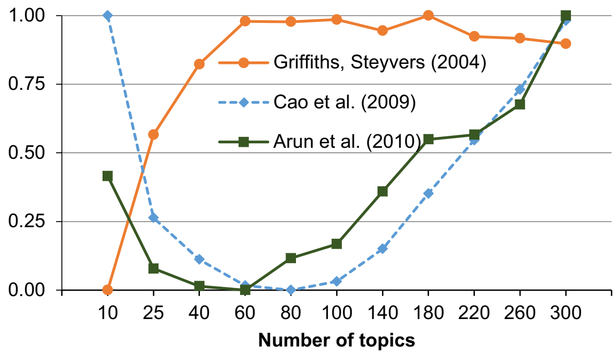

Fig. 10 Optimal number of topics. Measures are normalized with 0 (1) referring to the series' minimum (maximum). For the measures proposed by Arun et al. (2010) and Cao et al. (2009), the optimal number of topics can be found at the minimum, for Griffiths and Steyvers (2004) it is the maximum. Measures (Arun et al. 2010) and Cao et al. (2009) indicate that the optimal number of topics lies between 60 and 80, while Griffiths and Steyvers (2004) is somewhat ambiguous. Still, as the line of Griffiths and Steyvers (2004) levels off between 60 and 80, the latter seems to be a plausible compromise. Deveaud et al. (2014) is not computed due to computational limitations. *Source*: See text

Appendix 3: Excluding task issues

A topic model with 25 topics is applied on all articles published between 1941 and 2016 excluding Task issues as identified by Diebolt and Haupert (2018) which reduces the corpus from 2675 to 1885 documents. Again, the topic model identifies two topics which can be interpreted as representing quantitative methods. The 15 most probable words of the quantitative topics are shown in Table 4.

Table 4 Quantitative topics without Task issues

Percent price prices table period rate data average total years rates year increase series estimates	Data results variables variable table significant effects effect model economic sample level time percent coefficient

Table shows 15 most probable terms in descending order. *Source*: See text

References

Abramitzky R (2015) Economics and the modern economic historian. J Econ Hist 75(4):1240–1251
Andorfer P (2017) Turing Test für das Topic Modeling. Von Menschen und Maschinen erstellte inhaltliche Analysen der Korrespondenz von Leo von Thun-Hohenstein im Vergleich. Zeitschrift für digitale Geisteswissenschaften. https://doi.org/10.17175/2017_002
Arguing with Digital History working group Digital History and Argument. White paper, Roy Rosenzweig Center for History and New Media (13 Nov 2017). https://rrchnm.org/argument-white-paper/
Arun R, Suresh V, Veni Madhavan CE, Narasimha Murthy MN (2010) On finding the natural number of topics with latent Dirichlet allocation: some observations. In: Zaki MJ, Yu JX, Ravindran B, Pudi V (eds) Advances in knowledge discovery and data mining, vol 6118. Springer, Berlin
Bellstam G, Sanjai B, Cookson JA (2017) A text-based analysis of corporate innovation. SSRN working paper no. 2803232, May 2017
Blei DM (2012a) Probabilistic topic models. Commun ACM 55(4):77–84
Blei DM (2012b) Topic modeling and digital humanities. J Digit Human 2(1):8–11
Blei DM, Lafferty JD (2006) Dynamic topic models. In: Proceedings of the 23rd international conference on machine learning, pp 113–120
Blei DM, Lafferty JD (2007) A correlated topic model of science. Ann Appl Statist 1(1):17–35
Blei DM, Lafferty JD (2009) Topic models. In: Srivastava AN, Sahami M (eds) Text mining: classification, clustering, and applications. CRC Press, Boca Raton
Blei D, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022
Bonilla T, Grimmer J (2013) Elevated threat levels and decreased expectations: how democracy handles terrorist threats. Poetics 41(6):650–669
Boyd-Graber J, Blei D (2009) Multilingual topic models for unaligned text. In: UAI '09 Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence, pp 75–82
Boyd-Graber J, Mimno D, Newman DJ (2015) Care and feeding of topic models. In: Blei DM, Erosheva EA, Fienberg SE, Airoldi EM (eds) Handbook of mixed membership models and their applications. Taylor and Francis, Boca Raton
Boyd-Graber J, Hu Y, Mimno D (2017) Applications of topic models. Foundations and Trends in Information Retrieval, Boston
Burguière A (2009) The Annales school: an intellectual history. Cornell University Press, Ithaca
Cao J, Xia T, Li J, Zhang Y, Tang S (2009) A density-based method for adaptive LDA model selection. Neurocomputing 72:1775–1781. https://doi.org/10.1016/j.neucom.2008.06.011
Chang J, Boyd-Graber J, Wang C, Gerrish S, Blei DM (2009) Reading tea leaves: how humans interpret topic models. Adv Neural Inf Process Syst 2009:288–296
Collins WJ (2015) Looking forward: positive and normative views of economic history's future. J Econ Hist 75(4):1228–1233
Daniel V, Neubert M, Orban A (2018) Fictional expectations and the global media in the Greek debt crisis: a topic modeling approach. Working papers of the Priority Programme 1859 "Experience and Expectation. Historical Foundations of Economic Behaviour" No 4, Mar 2018
Deveaud R, Sanjuan E, Bellot P (2014) Accurate and effective latent concept modeling for ad hoc information retrieval. Doc Numérique 17:61–84. https://doi.org/10.3166/dn.17.1.61-84
Diebolt C, Haupert M (2018) A cliometric counterfactual: what if there had been neither Fogel nor North? Cliometrica. https://doi.org/10.1007/s11698-017-0167-8
DiMaggio P, Nag M, Blei D (2013) Exploiting affinities between topic modeling and the sociological perspective on culture: application to newspaper coverage of U.S. Government arts funding. Poetics 41(6):570–606
Eidelman V, Boyd-Graber J, Resnik P (2012) Topic models for dynamic translation model adaptation. In: ACL '12 proceedings of the 50th annual meeting of the association for computational linguistics
Feinerer I (2017) Introduction to the tm Package: Text Mining in R. https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf. Accessed 27 Mar 2018
Fligstein N, Brundage JS, Schultz M (2017) Seeing like the fed: culture, cognition, and framing in the failure to anticipate the financial crisis of 2008. Am Sociol Rev 82(5):879–909
Fogel R (1962) A quantitative approach to the study of railroads in American economic growth: a report of some preliminary findings. J Econ Hist 22(2):163–197
Freeman Smith R (1963) The formation and development of the International Bankers Committee on Mexico. J Econ Hist 23(4):574–586

García D (2013) Sentiment during recessions. J Finance 68(3):1267–1300
Gentzkow M, Kelly BT, Taddy M (2017) Text as data. NBER working paper no. 23276, Cambridge, MA, Mar 2017
Goodrich C (1960) Economic history: one field or two? J Econ Hist 20(4):531–538
Graham S, Milligan I, Weingart SB (2016) Exploring big historical data: the Historian's macroscope. Imperial College Press, London
Grajzl P, Murrell P (2017) A structural topic model of the features and the cultural origins of Bacon's ideas. CESifo working paper no. 6643, Oct 2017
Griffiths TL, Steyvers M (2004) Finding scientific topics. PNAS 101(1):5228–5235
Grimmer J (2010) A Bayesian hierarchical topic model for political texts: measuring expressed agendas in senate press releases. Polit Anal 18(1):1–35. https://doi.org/10.1093/pan/mpp034
Grimmer J, Stewart BM (2013) Text as data: the promise and pitfalls of automatic content analysis methods for political texts. Polit Anal 21(3):267–297. https://doi.org/10.1093/pan/mps028
Grün B, Hornik K (2011) Topicmodels: an R package for fitting topic models. J Stat Softw 40(13):1–30
Hall D, Jurafsky D, Manning CD (2008) Studying the history of ideas using topic models. In: Proceedings of the conference on empirical methods in natural language processing, pp 363–371
Hansen S, McMahon M (2016) Shocking language: understanding the macroeconomic effects of central bank communication. J Int Econ 99(1):S114–S133
Hansen S, McMahon M, Prat A (2018) Transparency and deliberation within the FOMC: a computational linguistics approach. Q J Econ 133:801–870
Haupert M (2016) History of cliometrics. In: Diebolt C, Haupert M (eds) Handbook of cliometrics. Springer, Berlin
Heiberger RH, Koss C (2018) Computerlinguistische Textanalyse und Debatten im Parlament: Themen und Trends im Deutschen Bundestag seit 1990. In: Brichzin J, Krichewsky D, Ringel L, Schank J (eds) Soziologie der Parlamente: Neue Wege der politischen Institutionenforschung. Springer VS, Wiesbaden
Hockey S (2004) The history of humanities computing. In: Schreibman S, Siemens R, Unsworth J (eds) A companion to digital humanities. Blackwell, Malden
Jacobi C, van Atteveldt W, Welbers K (2015) Quantitative analysis of large amounts of journalistic texts using topic modelling. Digit J 4(1):89–106
Jockers ML (2013) Macroanalysis: digital methods and literary history. University of Illinois Press, Urbana
Jockers ML (2014) Text analysis with R for students of literature. Quantitative methods in the humanities and social sciences. Springer, Cham
JSTOR Text analyzer. http://www.jstor.org/analyze/. Accessed 29 Mar 2018
Kuznets S (1952) National income estimates for the United States prior to 1870. J Econ Hist 121(2):115–130
Lamoreaux N (2015) The future of economic history must be interdisciplinary. J Econ Hist 75(4):1251–1257
Larsen VH, Thorsrud LA (2015) The value of news. CAMP working paper no. 6/2015, Oslo, Oct 2015
Larsen VH, Thorsrud LA (2017) Asset returns, news topics, and media effects. CAMP working paper no. 5/2017, Oslo, Sept 2017
Lau JH, Grieser K, Newman DJ, Baldwin T (2011) Automatic labelling of topic models. In: ACL '11 Proceedings of the 49th annual meeting of the association for computational linguistics, pp 1536–1545
Lüdering J, Tillmann P (2016) Monetary policy on Twitter and its effect on asset prices: evidence from computational text analysis. Joint discussion paper series in economics no. 12-2016, Marburg, Mar 2016
Lüdering J, Winker P (2016) Forward or backward looking? The economic discourse and the observed reality. J Econ Stat 236(4):483–515
Margo RA (2018) The integration of economic history into economics. Cliometrica. https://doi.org/10.1007/s11698-018-0170-8
McCallum A (2002) MALLET: a machine learning for language toolkit. http://mallet.cs.umass.edu/index.php. Accessed 19 Mar 2018
McCloskey D (1976) Does the past have useful economics. J Econ Lit 14(2):434–461
McCloskey D (1978) The achievements of the cliometrics school. J Econ Hist 38(1):13–28
McCloskey D (1987) Econometric history. Studies in economic and social history. Palgrave, Basingstoke
McFarland DA, Ramage D, Chuang J, Heer J, Manning CD, Jurafsky D (2013) Differentiating language usage through topic models. Poetics 41(6):607–625

Meeks E, Weingart SB (2012) The digital humanities contribution to topic modeling. J Digit Human 2(1):2–6
Miller IM (2013) Rebellion, crime and violence in Qing China, 1722–1911: a topic modeling approach. Poetics 41(6):626–649
Mimno D (2012a) Computational historiography: data mining in a century of classics journals. ACM J Comput Cult Herit 5(1):1–19
Mimno D (2012b) Lecture held at the Maryland Institute for technology in the humanities (topic modeling workshop). https://vimeo.com/53080123. Accessed 19 Mar 2018
Mimno D, Wallach HM, Naradowsky J, Smith DA, McCallum A (2009) Polylingual topic models. EMNLP 2009:880–889
Miner G (2012) Practical text mining and statistical analysis for non-structured text data applications. Elsevier/Academic Press, Amsterdam
Mitchener KJ (2015) The 4D future of economic history: digitally-driven data design. J Econ Hist 75(4):1234–1239
Mohr JW, Bogdanov P (2013) Introduction—topic models: what they are and why they matter. Poetics 41(6):545–569
Moretti F (2013) Distant reading. Verso, London, New York
Nelson RK mining the dispatch: digital Scholarship Lab, University of Richmond. http://dsl.richmond.edu/dispatch/pages/home. Accessed 19 Mar 2018
Newman DJ, Block S (2006) Probabilistic topic decomposition of an eighteen-century American newspaper. J Am Soc Inform Sci Technol 57(6):753–767
Nguyen TH, Shirai K (2015) Topic modeling based sentiment analysis on social media for stock market prediction. In: Proceedings of the 53rd annual meeting of the association for computational linguistics, pp 1354–1364
Nikita M (2016) ldatuning (R package). https://cran.r-project.org/web/packages/ldatuning/ldatuning.pdf. Accessed 19 Mar 2018
Oswald ME, Grosjean S (2004) Confirmation bias. In: Pohl R (ed) Cognitive illusions: a handbook on fallacies and biases in thinking, judgement and memory, 1st edn. Psychology Press, Hove
Quinn KM, Monroe BL, Colaresi M, Crespin MH, Radev DR (2010) How to analyze political attention with minimal assumptions and costs. Am J Polit Sci 54(1):209–228
Riddell AB (2014) How to read 22,198 Journal Articles: studying the history of German studies with topic models. In: Erlin M, Tatlock L (eds) Distant readings: topologies of German culture in the long nineteenth century. Boydell & Brewer, Suffolk
Schofield A, Magnusson M, Mimno D (2017) Pulling out the stops: rethinking stopword removal for topic models. In: Proceedings of the 15th conference of the European chapter of the association for computational linguistics, pp 432–436
Shirota Y, Hashimoto T, Sakura T (2015) Topic extraction analysis for monetary policy minutes of Japan in 2014: effects of the consumption tax hike in April. In: Perner P (ed) Advances in data mining: applications and theoretical aspects. Springer, Cham
Steyvers M, Griffiths T (2007) Probabilistic topic models. In: Landauer TK, McNamara DS, Dennis S, Kintsch W (eds) Handbook of latent semantic analysis. Taylor and Francis, Hoboken
Tetlock PC (2007) Giving content to investor sentiment: the role of media in the stock market. J Finance 62(3):1139–1168
Thorsrud LA (2016a) Nowcasting using news topics. Big data versus big bank. Norges Bank working paper 20/2016, Oslo, Dec 2016
Thorsrud LA (2016b) Words are the new numbers: a newsy coincident index of business cycles. Norges Bank working paper 21/2016, Oslo, Dec 2016
Underwood T (2018) The stone and the shell (blog). https://tedunderwood.com/. Accessed 19 Mar 2018
Walker DD, Lund WB (2010) Evaluating models of latent document semantics in the presence of OCR errors. In: Proceedings of the 2010 conference on empirical methods in natural language processing, pp 240–250
Wallach HM (2006) Topic modeling: beyond bag of words. In: Proceedings of the 23rd international conference on machine learning, pp 977–987
Wallach HM, Mimno D, McCallum A (2009) Rethinking LDA: why priors matter. Adv Neural Inf Process Syst 22:1973–1981
Walters PG, Walters R (1944) The American career of David Parish. J Econ Hist 2(2):149–166
Weingart SB (2018) The scottbot irregular (blog). http://www.scottbot.net/HIAL/index.html@p=19113.html. Accessed 19 Mar 2018

Whaples R (1991) A quantitative history of the journal of economic history and the cliometric revolution. J Econ Hist 51(2):289–301

Whaples R (2002) The supply and demand of economic history: recent trends in the journal of economic history. J Econ Hist 62(2):524–532

Yang T-I, Torget AJ, Mihalcea R (2011) Topic modeling on historical newspapers. In: Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities, pp 96–104

Zhao B, Xing EP (2007) HM-BiTAM: bilingual topic exploration, word alignment, and translation. In: NIPS'07 Proceedings of the 20th international conference on neural information processing systems, pp 1689–1696