AMERICAN MARKETING ASSOCIATION

Article

Journal of Marketing Research
2021, Vol. 58(6) 1142–1158

DOI: 10.1177/0022243720943209
journals.sagepub.com/home/mrj
SAGE

A Poisson Factorization Topic Model for the Study of Creative Documents (and Their Summaries)

Olivier Toubia

Abstract

The author proposes a topic model tailored to the study of creative documents (e.g., academic papers, movie scripts), which extends Poisson factorization in two ways. First, the creativity literature emphasizes the importance of novelty in creative industries. Accordingly, this article introduces a set of residual topics that represent the portion of each document that is not explained by a combination of common topics. Second, creative documents are typically accompanied by summaries (e.g., abstracts, synopses). Accordingly, the author jointly models the content of creative documents and their summaries, and captures systematic variations in topic intensities between the documents and their summaries. This article validates and illustrates the model in three domains: marketing academic papers, movie scripts, and TV show closed captions. It illustrates how the joint modeling of documents and summaries provides some insight into how people summarize creative documents and enhances understanding of the significance of each topic. It shows that the model described produces new measures of distinctiveness that can inform the perennial debate on the relation between novelty and success in creative industries. Finally, the author shows how the proposed model may form the basis for decision support tools that assist people in writing summaries of creative documents.

Keywords

creativity, entertainment, topic models, Poisson factorization, variational inference

Online supplement: https://doi.org/10.1177/0022243720943209

With the digitization of the economy, people are both producing and consuming more creative content. On the supply side, according to Florida (2014), more than 40 million Americans (or approximately one-third of the employed population) belong to the "creative class." This class includes people in science and engineering, education, arts, and entertainment whose primary economic function is to create new ideas, new technology, and new creative content. On the demand side, the average American spends approximately 12 hours per day consuming media (Statista 2017), and the media and entertainment industry alone is valued at approximately $2 trillion globally (Statista 2018).

This article uses the term "creative document" to refer to any written document that describes the output of a creative process. Examples include academic papers, fiction books, movie and TV show scripts, plays, business models, and new product descriptions. In contrast, noncreative documents include news articles, instruction manuals, and so on. In addition to being managerially relevant, creative documents have captured the interest of academics. Several studies have attempted to identify correlates of success in creative industries—in particular, the link between the distinctiveness of a creative document and its success (e.g., the link between the distinctiveness of an academic paper and its number of citations).

Studying creative documents on a large scale in a scientific manner has been historically challenging, due to the unstructured nature of the data contained in these documents. With the development of natural language processing tools such as latent Dirichlet allocation (LDA) (Blei et al. 2003) and Poisson factorization (Canny 2004), it has become possible to systematically extract text-based topics and features from creative documents. Although some studies have applied variations of traditional topic models to the study of creative documents (e.g., Berger and Packard 2018; Eliashberg, Hui, and Zhang 2007, 2014; Toubia et al. 2018), I argue that these models fail to capture at least two essential aspects of creative documents.

First, the creativity literature has shown that *novelty* is a key construct when it comes to creative content. Traditional topic

Olivier Toubia is the Glaubinger Professor of Business, Columbia Business School, Columbia University (email: ot2107@gsb.columbia.edu).

models perform dimensionality reduction by approximating each document using a set of topics, which are common across all documents in the corpus. With a traditional topic model, the distinctiveness of a document may be measured by the distinctiveness of its combination of topics. However, traditional topic models fail to capture another aspect of distinctiveness: the extent to which a document may *not* be explained by common topics. As such, I argue that traditional topic models are limited in their ability to provide rich measures of distinctiveness, which may inform the debate on the link between novelty and success in creative industries.

Second, creative documents are often accompanied by *summaries*. For example, academic papers are accompanied by abstracts, books and movies by synopses, new products by short descriptions, business plans by executive summaries, and so on. Summaries play a key role in the market by helping consumers extract information from creative products more efficiently and decide which products to consume. For example, a consumer may be enticed to buy a book or watch a movie by a synopsis, or to buy a new product by its short description. One may argue that summaries serve as "lubricant" in the market for creative content and soften competition by making it easier for consumers to decide which products to consume.[1] Traditional topic models do not capture the relation between a document and its summary. I argue that modeling and quantifying the process by which humans summarize creative documents not only is interesting from an academic perspective but also offers practical benefits. From the perspective of extracting meaningful, interpretable topics from a corpus of creative documents, summaries may be viewed as shorter documents produced by people who invested time and effort to determine which topics in a creative document are "essential" enough to be included in its summary. As such, summaries have the potential to improve our understanding of the significance of each topic. Moreover, modeling the summarization process opens the door for the development of computer-based tools to assist authors and marketers in creative industries in writing summaries of creative documents. For example, by identifying characteristics of summaries that correlate with success in a specific creative industry, it is possible to advise authors to emphasize certain topics in their summaries.

Motivated by these two characteristics of creative documents, I propose a topic model tailored to the study of creative documents. The contribution of this research is primarily methodological. The model extends Poisson factorization in two ways. First, it accounts for not only the portion of a document that may be explained by topics that are common across documents, but also the "residual" (or "outside the cone"; see the geometric interpretation) portion that is not explained by combinations of these common topics. Second, I jointly model the content of creative documents and their summaries. The model represents systematic variations in the extent to which each common topic, as well as each "residual" topic, appears in summaries compared with full documents.

While topic models have been applied to creative documents, to the best of my knowledge this model is the first topic model specifically tailored for creative documents. The model offers at least three benefits that traditional topic models cannot provide, for both academics and practitioners. First, each topic estimated by the model comes with a variable that quantifies the extent to which the topic was deemed "summary worthy" by the people who wrote the summaries of the documents in the corpus. I illustrate how this additional layer of information provides some insight into the process by which people summarize creative documents in a particular domain and enhances our understanding of the significance of each topic. Second, for academics and practitioners interested in participating in the ongoing debate on the link between distinctiveness and success of creative products, I show that the model provides various measures of distinctiveness, which have the potential to uncover new insight into correlates of success in creative industries. I use three data sets to empirically explore the relation between three measures of distinctiveness (and various success measures, i.e., number of citations of academic papers, movie and TV show ratings, and movie return on investment). Third, I show that the model may serve as the basis for interactive decision support tools that assist people in writing summaries of creative documents. The development of such tools may be informed by an empirical analysis of correlates of success in the target industry. For example, I find that marketing academic papers whose abstracts put relatively more emphasis on the "outside the cone" content in the paper tend to have more citations. Accordingly, the model can help authors identify the "outside the cone" content in their paper and emphasize it in their abstract. I develop a proof of concept for such a tool.

Relevant Literatures

The study of creativity in various domains, from scientific discovery (e.g., Uzzi et al. 2013) to linguistics (e.g., Giora 2003) and innovation (Toubia and Netzer 2017), has suggested that creativity lies in the optimal balance between novelty and familiarity. For example, Ward (1995, p. 166) argues that "truly useful creativity may reflect a balance between novelty and a connection to previous ideas." Furthermore, building on previous research from a wide range of domains (e.g., Finke, Ward, and Smith 1992; Mednick 1962), Toubia and Netzer (2017) show that when attempting to quantify familiarity and novelty in a document using text analysis, researchers should focus on novel versus familiar *combinations* of words, rather than words that themselves appear more or less frequently.

These insights inform the modeling approach used herein. I adopt a natural language processing approach, which captures topics defined as combinations of words. The model nests and extends previous applications of Poisson factorization to the study of text documents, such as Canny (2004) and Gopalan, Charlin, and Blei (2014). For example, Gopalan, Charlin, and Blei (2014) study how researchers rate academic papers by

[1] I thank Anthony Dukes for this insight.

modeling documents and researcher preferences as latent vectors in a topic space. The model proposed herein builds on Gopalan, Charlin, and Blei's (2014) model, though it differs in a few important ways. First, I model the content of full documents and their summaries, rather than modeling the content of documents and consumers' preferences for these documents. These different objectives give rise to very different data, model specifications, and data-generating processes. Moreover, I jointly model the content of documents and their summaries, I explicitly model "residual" topics, and I model how residual topics are represented in summaries, none of which is performed by Gopalan, Charlin, and Blei's (2014) model. Finally, I use offset variables in a novel way to capture systematic variations in topic intensities in full documents versus summaries. As noted in the introduction, several papers have used extant topic models to study creative documents (e.g., Berger and Packard 2018; Eliashberg, Hui, and Zhang 2007, 2014; Toubia et al. 2018). However, to the best of my knowledge the model developed here is the first topic model tailored to the study of creative documents.

Note that most applications of topic modeling in the marketing literature have used latent Dirichlet allocation (LDA; Blei et al. 2003) or extensions thereof (e.g., Büschken and Allenby 2016; Liu and Toubia 2018; Puranam, Narayan, and Kadiyali 2017; Tirunillai and Tellis 2014; Toubia et al. 2018; Zhong and Schweidel 2020). The basic LDA model shares many similarities with the basic Poisson factorization model, although previous research has suggested that Poisson factorization tends to fit data better (Canny 2004; Gopalan, Hofman, and Blei 2013, Gopalan, Charlin, and Blei 2014). My choice of Poisson factorization was primarily driven by the attractive conjugacy property of this approach. Indeed, the model remains conditionally conjugate, despite the additional complexities resulting from jointly modeling the content of documents and summaries while explicitly capturing residual content.[2]

Despite the importance of summaries in the commercialization of creative content, summarization has received very little attention in the marketing literature. In contrast, it is a substantial subfield of computer science (see, e.g., Allahyari et al. 2017; Nenkova and McKeown 2012; Radev, Hovy, and McKeown 2002; Yao, Wan, and Xiao 2017). However, computer scientists have focused mostly on *automatic* text summarization, in which a summary is produced without any human intervention. This process is typically done by identifying and selecting a subset of the sentences in the original document, a process called *extractive summarization* (Allahyari et al. 2017). Such text summarization tools are useful for summarizing large numbers of documents (e.g., news articles) on a regular basis, quickly and efficiently (McKeown and Radev 1995; Radev and McKeown 1998). In contrast, I focus on situations in which summaries provide additional content written by humans, from which valuable insights might be learned. In terms of practical applications, I envision computers not as a replacement for, but rather as an aid to humans, and consider decision support tools that assist them in writing summaries of creative documents. The different perspective on summarization adopted here also translates into methodological differences. Some studies have applied topic modeling to text summarization, sometimes introducing document-specific topics that capture unique content in each document, which should be included in the summary (Daumé and Marcu 2006; Delort and Alfonseca 2012; Haghighi and Vanderwende 2009). These document-specific topics are similar in spirit to the residual topics in the model. However, given their focus on extractive summarization, unlike my model, these models do not consider summaries as an additional source of information, they do not model the content of summaries, and they do not include summaries in their training data.

Proposed Model

Model Foundation: Poisson Factorization

I index creative documents by $d = 1, ...D$, and words in the vocabulary by $v = 1, ...L$, denoting as w_{dv} the number of times word v appears in document d. In *standard* Poisson factorization (Canny 2004; Gopalan, Charlin, and Blei 2014), the assumed data generating process would be as follows:

1. For each regular topic $k = 1, ...K$, for each word v, draw $\beta_{kv} \sim \text{Gamma}(\alpha_1, \alpha_2)$
2. For each document $d = 1, ...D$,
   - For each topic, draw topic intensity $\theta_{dk} \sim \text{Gamma}(\alpha_3, \alpha_4)$, and
   - For each word v, draw word count $w_{dv} \sim \text{Poisson}(\sum_k \theta_{dk}\beta_{kv})$.

To gain intuition for this base model, recall that the sum of independent Poisson-distributed random variables is a Poisson variable. Hence, according to Poisson factorization, the number of occurrences of word v in document d, $w_{dv} \sim \text{Poisson}(\sum_k \theta_{dk}\beta_{kv})$, may be thought of as the sum of K independent Poisson variables (called auxiliary variables; e.g., Gopalan, Charlin, and Blei 2014): $z_{dv,k} \sim \text{Poisson}(\theta_{dk}\beta_{kv})$. These variables capture the number of occurrences of word v in document d associated with each topic k, such that $w_{dv} = \sum_k z_{dv,k}$. The distribution of each auxiliary variable $z_{dv,k}$ is influenced by the product of two terms: θ_{dk} represents the intensity of topic k in document d; β_{kv} represents the weight of word v in topic k.

One can also interpret Poisson factorization geometrically. (To the best of my knowledge, the following geometric interpretation of Poisson factorization is new to the literature.)

[2] Word embedding (Mikolov et al. 2013, 2017) has recently emerged as another popular natural language processing approach. Word embedding typically does not extract topics from text and does not assign topic intensities to documents. Hence, it is not directly relevant to my goal of developing a topic model tailored to the study of creative documents. However, by capturing the context around each word, word embedding may be better suited for studying the structure of creative document, which I leave for future research.

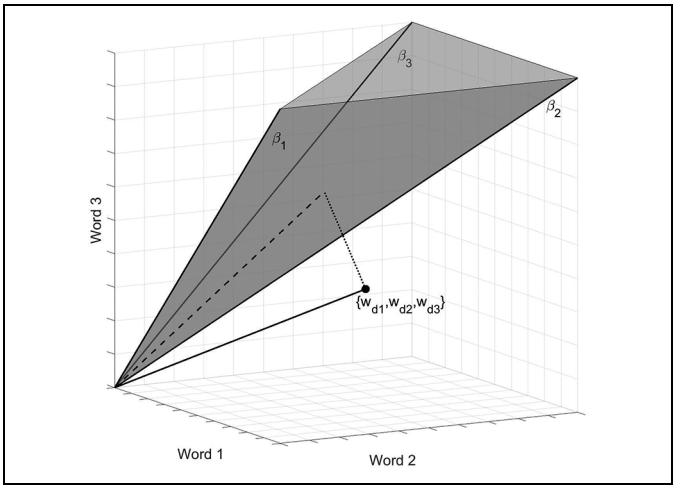

Figure 1. Geometric interpretation of "inside the cone" versus "outside the cone" content.

Notes: In this example with three words in the vocabulary and three topics, each vector β_k represents the weights of each word on topic k. The grey cone contains all positive combinations of the three topics. The black dot represents a vector that contains the number of occurrences of each word in a document d: $\{w_{d1}, w_{d2}, w_{d3}\}$. The dashed line represents the projection of this vector on the cone defined by the three topics ("inside the cone" content, captured by standard Poisson factorization). The dotted line represents the residual ("outside the cone" content, captured by the proposed model but not by standard Poisson factorization).

Topics and documents may be represented in the Euclidean space defined by the words in the vocabulary. That is, topic k may be represented by a $L \times 1$ vector $\beta_k = \{\beta_{kv}\}_v$ that captures the weights on each word in the topic. Similarly, document d may be represented by a $L \times 1$ vector $w_d = \{w_{dv}\}_v$ that contains the number of occurrences of each word observed in the document. According to Poisson factorization, the expected value of this vector is given as $E(w_d) = \sum_k \theta_{dk}\beta_k$. (Recall that the expected value of a variable with a Poisson distribution is the rate of the distribution.) In other words, the expected number of occurrences of words in the document may be written as a positive combination of the vectors $\{\beta_k\}_k$ that represent topics in the word space, where the weights are the topic intensities $\{\theta_{dk}\}_k$. In this illustration, for simplicity I focus on expected values.

Mathematically, the positive combinations of the set of topic vectors, $\{\sum_k \theta_{dk}\beta_k,\ \theta_{dk} \geq 0\}$, form a cone in the Euclidean space defined by the words in the vocabulary. This means that Poisson factorization may be viewed as approximating each document by projecting it onto the cone defined by the topic vectors. Consider the illustration of this geometric interpretation in Figure 1. For illustration purposes, this figure uses a vocabulary that consists of three words and assumes three topics (in practice, the number of topics should be much smaller than the number of words in the vocabulary). This figure illustrates the cone defined by positive combinations of the three topics. It also shows an example of one document represented by a vector in the same space and how Poisson factorization projects this document vector onto the cone defined by the topics.[3] (Note that in reality, the projection is not orthogonal, due to the prior on the parameters.)

In summary, the primary focus of traditional topic models such as Poisson factorization is to understand topics that are common across documents in a corpus and to quantify the intensity with which each topic is featured in each document. In doing so, Poisson factorization approximates each document as a positive combination of common topics.

Residual Topics

My model extends Poisson factorization in two ways. First, it captures "outside the cone" content by introducing one "residual topic" associated with each document (to my knowledge, a novel way of using Poisson factorization). For each document d, I introduce a topic β_d^{res} that is unique to this document. The weight of this topic on each word v is assumed to have a gamma prior, similar to the other "regular" topics: $\beta_{dv}^{res} \sim \text{Gamma}(\alpha_1, \alpha_2)$. I model the number of occurrences of word v in document d as follows: $w_{dv} \sim \text{Poisson}(\sum_k \theta_{dk}^{reg}\beta_{kv}^{reg} + \beta_{dv}^{res})$, where the superscript reg refers to the regular topics (β_{kv}^{reg} is common across all documents in the corpus).[4] The residual topic represents the residual content in document d.

The introduction of this residual topic was motivated by the creativity literature, in an attempt to account for distinct content in the document. One may wonder whether the residual topic is simply "noise." To address this issue, the "Empirical Applications" section empirically tests whether the residual topic indeed relates to the success of creative documents in ways that are predicted by the creativity literature. If this topic were "just noise," no systematic relation with the success of creative documents should be present. Theoretically, the model still includes "noise," above and beyond the residual topics. Indeed, the number of occurrences of each word remains stochastic and governed by a Poisson distribution. In addition, the prior induces sparsity and trades off fit with the complexity of the model. As a result, the expected value of the number of occurrences of each word according to the model does not perfectly fit the observed value, even in the presence of residual topics.

Figure 1 illustrates geometrically how the vector corresponding to a document is decomposed into two vectors: the "inside the cone" component that projects the document vector onto the cone defined by the regular topics and the "outside the cone" component that closes the gap between the original vector and the projection. (Again, this simple illustration focuses on expected values and ignores the effect of the prior; the actual model produces a *distribution* of word occurrences, and fit is not perfect due to the sparsity-inducing prior.)

[3] Note that in the case of LDA, documents are approximated by convex combinations of the topics. Hence, cones would be replaced with simplexes in this geometric interpretation.

[4] Note that the intensity of the residual topic θ_d^{res} is implicitly set to 1 for identification purposes.

Offset Variables

The second way in which the model extends Poisson factorization is that it jointly models the content of creative documents and their summaries. To that end, I introduce a set of "offset" variables that capture how topics are weighed in summaries, compared with full documents. The topic intensities in the summary of a creative document may not be the same as the topic intensities in the full document. First, some regular topics may be typically judged by the authors of summaries as being more or less worthy of being featured in a document's summary, which should translate into systematic differences across regular topics in how they are weighed in summaries versus full documents. For example, topics that relate to data analysis (substantive findings) may be relatively under-weighed (over-weighed) in the abstracts of academic papers compared with the full papers. To capture and quantify such phenomenon, I allow each regular topic k to have its own "offset" variable, ε_k^{reg}. Second, "inside the cone" and "outside the cone" content may be weighed differently in summaries versus full documents. Accordingly, I also introduce an offset variable for each residual topic, ε_d^{res}. More precisely, I model the number of occurrences of word v in the summary of document d as $w_{dv}^{summary} \sim \text{Poisson}(\sum_k \theta_{dk}^{reg}\beta_{kv}^{reg}\varepsilon_k^{reg} + \beta_{dv}^{res}\varepsilon_d^{res})$. That is, the topic intensity of regular topic k in the full document, θ_{dk}^{reg}, is multiplied by the offset variable ε_k^{reg} in the summary. Similarly, the intensity of the residual topic is multiplied by the offset variable ε_d^{res}. By specifying gamma priors on the offset variables, I preserve the conditional conjugacy of the model (i.e., the posterior distribution of each variable conditional on the other variables and the data is given in closed form).

A perennial issue with traditional topics models is the difficulty of interpreting topics, resulting from the unsupervised nature of these models. Offset variables provide an additional layer of information that helps users understand the significance of each topic by giving it a "score" that captures the extent to which people decide to include this topic when writing summaries of creative documents in the domain under study. Although offset variables have been used for different purposes in previous applications of Poisson factorization (e.g., Gopalan, Charlin, and Blei 2014), to the best of my knowledge this article, as the first to use Poisson factorization to jointly model documents and their summaries, is also the first to use offset variables to capture how the intensities of topics vary between documents and summaries. Web Appendix F further explores the impact of introducing offset variables by estimating an alternative version of the model that does not include these variables, showing that the topics learned by this alternative model are substantively different from the topics learned by the proposed model. In the proposed model, topics are defined as groups of words that tend to not only appear together but also appear with the same relative frequency in summaries compared with full documents. Accordingly, the presence of offset variables affects the topics learned from the model.

Data-Generating Process

Putting all these pieces together, the data generating process for the model is as follows:

1. For each regular topic $k = 1, ...K$, for each word v, draw $\beta_{kv}^{reg} \sim \text{Gamma}(\alpha_1, \alpha_2)$ and draw offset variable $\varepsilon_k^{reg} \sim \text{Gamma}(\alpha_5, \alpha_6)$.
2. For each residual topic $d = 1, ...D$, for each word v, draw $\beta_{dv}^{res} \sim \text{Gamma}(\alpha_1, \alpha_2)$, and draw offset variable $\varepsilon_d^{res} \sim \text{Gamma}(\alpha_5, \alpha_6)$.
3. For each document $d = 1, ...D$, for each regular topic, draw topic intensity $\theta_{dk}^{reg} \sim \text{Gamma}(\alpha_3, \alpha_4)$, and for each word v, draw word count $w_{dv} \sim \text{Poisson}(\sum_k \theta_{dk}^{reg}\beta_{kv}^{reg} + \beta_{dv}^{res})$.
4. For each document summary $d = 1, ...D$, for each word v, draw word count $w_{dv}^{summary} \sim \text{Poisson}(\sum_k \theta_{dk}^{reg}\beta_{kv}^{reg}\varepsilon_k^{reg} + \beta_{dv}^{res}\varepsilon_d^{res})$.

Estimation Using Variational Inference

To estimate the model, I start by defining auxiliary variables that allocate the occurrences of each word v in each document d across the various topics: $z_{dv,k}^{reg} \sim \text{Poisson}(\theta_{dk}^{reg}\beta_{kv}^{reg})$; $z_{dv}^{res} \sim \text{Poisson}(\beta_{dv}^{res})$, such that $w_{dv} = \sum_k z_{dv,k}^{reg} + z_{dv}^{res}$. Similar variables are defined for the summaries: $z_{dv,k}^{sum,reg} \sim \text{Poisson}(\theta_{dk}^{reg}\beta_{kv}^{reg}\varepsilon_k^{reg})$, and $z_{dv}^{sum,res} \sim \text{Poisson}(\beta_{dv}^{res}\varepsilon_d^{res})$, such that $w_{dv}^{summary} = \sum_k z_{dv,k}^{sum,reg} + z_{dv}^{sum,res}$. With the addition of these auxiliary variables, the model has the attractive property of being conditionally conjugate (i.e., the posterior distribution of each parameter conditional on the other parameters and the data are given in closed form). Although the model could be estimated using Gibbs sampling, to speed up computations and improve scalability, I estimate it using variational inference (Blei, Kucukelbir, and McAuliffe 2016). Details are provided in Web Appendix B.

Selecting the Number of Topics

Although the number of topics could be selected using cross-validation to achieve minimum perplexity, I use a simpler approach advocated by Gopalan, Charlin, and Blei (2014): I set the number of topics K to a large number (like these authors, I use $K = 100$), with the realization that some of these topics will be "flat," such that all topic weights β_{kv}^{reg} are very small and similar across words and all topic intensities θ_{dk}^{reg} are very small and similar across documents. I set the same value of $K = 100$ for all benchmarks. These flat topics emerge as a result of the gamma priors on topic weights and topic intensities, which induce sparsity. In other words, the model automatically attempts to explain the data with a few topics and corrects for values of K that are larger than needed. This means that the number of nonflat topics is influenced by the prior parameters $\{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6\}$. Herein, I follow Gopalan, Charlin, and Blei (2014) and set $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4\alpha_5 = \alpha_6 = 0.3$. Web Appendix E tests a more/less diffuse prior and reports how the number of nonflat topics (as well as the distinctiveness

measures introduced subsequently) vary in each data set when the prior is changed.

Extension: Dynamic Topics

Web Appendix G introduces a dynamic extension of this model, inspired by Blei and Lafferty (2006). I model each topic as having a base version and introduce a set of time-specific offset variables that capture the evolution of each topic over discrete time periods. In each time period, the weights of each topic are assumed to be equal to the weights in the previous period, plus a set of offset variables specific to that topic and that time period. This extension is also estimated using variational inference. I apply it to the marketing academic paper data set, which contain all papers published in a set of journals over six years. I find that the introduction of dynamics does not change the conclusions from the empirical analysis.

Empirical Applications

Data Sets

I apply the model to three data sets. In each data set, all documents were preprocessed following standard steps in natural language processing: eliminate non-English characters and words, numbers, and punctuation; tokenize the text (i.e., break each document into individual words or tokens); remove common stop words; and remove tokens (words) that contain only one character. No stemming or lemmatization was performed. In each data set, I randomly split the set of documents into two samples: a calibration set with 75% of the documents and a validation set with 25% of the documents.

I constructed the vocabulary of words in each data set based on the full documents in the calibration set only (i.e., I did not use the summaries and the validation documents to select the vocabulary). I computed the term frequency (tf) for each word (i.e., the total number of occurrences of the word across all training documents). I removed words that appear fewer than 100 times across documents (for movies, given the smaller sample size, I used a cutoff of $tf<65$). Next, I computed the tf-idf of each word w, defined as: $\text{tf-idf}(w) = \text{tf}(w) \times \log(\frac{N}{df(w)})$, where $df(w)$ is the document frequency for word w, defined as the number of documents in which word w appears at least once. The final vocabulary consists of the 1,000 words with the highest tf-idf (i.e., I removed words that appear too frequently and words that appear too infrequently (Blei and Lafferty 2009). Web Appendix H runs all models with vocabularies of 500 words and with vocabularies of 2,000 words (still selecting words based on tf-idf); it shows that results are qualitatively similar to the ones obtained with 1,000 words, although changing the vocabulary size does change the estimated measures of distinctiveness introduced subsequently. A simulation study, reported in Web Appendix H, confirms that distinctiveness measures should indeed be affected by vocabulary size.

The first data set consists of the full texts (excluding the abstracts, bibliographies, and appendices) and the abstracts of all 1,333 research papers published in *Journal of Consumer Research*, *Journal of Marketing*, *Journal of Marketing Research*, and *Marketing Science* between 2010 and 2015. Most of the papers were downloaded in PDF format. Some spelling errors occurred while converting PDF files to text files; hence, a spelling corrector was trained based on the autocorrection package in Python and applied before preprocessing the data. Table 1 reports descriptive statistics for all data sets after preprocessing.

The second data set consists of the scripts and synopses of 858 movies released in the United States for which scripts were available on the Internet Movie Script Database (imsdb.com) and synopses were available on the Internet Movie Database (IMDB; imdb.com). Words corresponding to names of locations, people, and organizations were identified using the Stanford Named Entity Recognition classifier and removed from the data before preprocessing.

For the third data set, I collaborated with a major global media company interested in creating a "knowledge graph" for its extensive library of TV content (i.e., identifying a set of meaningful, interpretable topics that describe each TV show episode to classify its content). The company made available the collection of closed captions for 26,561 unique TV show episodes, which constitute most of the company's catalog of U.S.-based, English-language TV show episodes. The company decided to work with closed captions because they are available systematically and consistently for all episodes, as they are required by the Federal Communication Commission. The company also made available the synopses of all TV show episodes, which are part of its internal programming system. As in the previous data set, words corresponding to names of locations, people, and organizations were removed from the data before preprocessing.

Fit and Predictive Performance

Benchmarks

The proposed model extends Poisson factorization in two ways. First, it models "residual" topics that are unique to each document. Second, it allows the topic intensities in summaries to differ from the topic intensities in main documents. To determine the benefits of these two extensions, I tested a series of nested models. All benchmarks are estimated using variational inference, with the same convergence criterion and hyperparameters. The first benchmark considered is a nested model that does not include residual topics. This benchmark is a nested version of the proposed model, in which $\{\beta_d^{res}, \varepsilon_d^{res}\}_d$ are constrained to be 0. This benchmark still includes offset variables for the regular topics $\{\varepsilon_k^{reg}\}_k$, which allows exploring the benefit of modeling "outside the cone" content using residual topics. The second benchmark includes residual topics but constrains all offset variables, $\{\varepsilon_k^{reg}\}_k$ and $\{\varepsilon_d^{res}\}$, to be equal to each other. That is, this benchmark assumes that the relative

Table 1. Descriptive Statistics.

	Metric	Unit of Analysis	Mean	SD	Range
Marketing academic papers	Number of word occurrences	Paper	2,110.26	647.31	[12;5,016]
	Number of word occurrences	Abstract	41.39	15.15	[4;125]
	Number of words with at least one occurrence	Paper	269.74	56.54	[7;409]
	Number of words with at least one occurrence	Abstract	23.44	7.56	[3;61]
	Number of occurrences across full texts	Word	2,812.97	4,020.34	[188;44,091]
	Number of occurrences across abstracts	Word	55.18	98.96	[0;1,420]
	Number of full texts with at least one occurrence	Word	359.57	268.50	[1;1,216]
	Number of abstracts with at least one occurrence	Word	31.25	48.00	[0;624]
Movies	Number of word occurrences	Script	1,490.86	580.36	[0;7,489]
	Number of word occurrences	Synopsis	91.26	80.93	[1;748]
	Number of words with at least one occurrence	Script	310.17	69.05	[0;533]
	Number of words with at least one occurrence	Synopsis	46.78	31.18	[1;219]
	Number of occurrences across scripts	Word	1,279.16	2,016.88	[89;33,633]
	Number of occurrences across synopses	Word	78.30	122.36	[0;1,322]
	Number of scripts with at least one occurrence	Word	266.12	196.20	[1;834]
	Number of synopses with at least one occurrence	Word	40.14	52.26	[0;426]
TV show episodes	Number of word occurrences	Closed caption	797.77	405.76	[0;3,819]
	Number of word occurrences	Synopsis	8.20	10.76	[0;261]
	Number of words with at least one occurrence	Closed caption	289.70	80.95	[0;718]
	Number of words with at least one occurrence	Synopsis	7.35	7.39	[0;155]
	Number of occurrences across closed captions	Word	21,189.56	41,049.97	[1,469;693,406]
	Number of occurrences across synopses	Word	217.72	283.79	[0;2,889]
	Number of closed captions with at least one occurrence	Word	7,694.59	5,526.92	[24;26,148]
	Number of synopses with at least one occurrence	Word	195.22	248.20	[0;2,498]

Notes: There are 1,333 papers in the academic paper data set, 858 movies in the movie data set, and 26,561 TV show episodes in the TV show data set. There are 1,000 words in the vocabulary for each data set. The first column (vertically aligned) contains the data set, the second the metric of interest, the third the unit of analysis, and the remaining columns report the mean, standard deviation, min, and max of the correspond metric across the units of analysis. For example, the first row indicates that in the marketing academic paper data set, papers have on average 2,110.26 word occurrences.

intensities of topics in summaries are the same as in the main documents. This benchmark is implemented by replacing the offset variables with a single variable ε. This benchmark enables exploring the benefit of allowing the relative topic intensities in summaries to differ from those in the main documents. The third nested benchmark does not include residual topics ($\{\beta_d^{res}, \varepsilon_d^{res}\}_d$ are set to 0) and constrains all offset variables on the regular topics $\{\varepsilon_k^{reg}\}_k$ to be equal. That is, this benchmark is similar to a basic Poisson factorization model that would assume that documents and their summaries have the same relative topic intensities. The fourth and final nested benchmark does not contain any regular topics, only residual topics. That is, this benchmark does not attempt to learn topics that are shared across all documents; rather, it treats each document as completely unique and learns one residual topic for each document. This benchmark is a special case of the proposed model, in which the number of regular topics K is set to 0.

Finally, I consider LDA, a nonnested benchmark, due to its popularity. Because LDA does not include offset variables, the topic intensities in the summary of a document are assumed to be the same as in the full document. In addition, LDA does not include residual topics. Web Appendix D provides details of the LDA benchmark.

Measures of Fit

I estimate each model on the full texts and summaries of the calibration documents in each data set. The output from the model and any of its nested benchmark may be summarized by computing a vector of fitted Poisson rates $\tilde{\lambda}_d = \{\tilde{\lambda}_{dv}\}_v$ for each document, which govern the number of occurrences of each word in the document:

$$\tilde{\lambda}_d = \sum_k \theta_{dk}^{reg} \beta_k^{reg} + \beta_d^{res}. \quad (1)$$

In addition, for each document, fitted Poisson rates can be constructed for the number of occurrences of words in the document's summary:

$$\tilde{\lambda}_d^{summary} = \sum_k \theta_{dk}^{reg} \beta_k^{reg} \varepsilon_k^{reg} + \beta_d^{res} \varepsilon_d^{res}. \quad (2)$$

To compare the model with LDA, I transform these Poisson rates into multinomial distributions $\tilde{\phi}_d$ and $\tilde{\phi}_d^{summary}$, where $\tilde{\phi}_{dv} = \frac{\tilde{\lambda}_{dv}}{\Sigma_{v'}\tilde{\lambda}_{dv'}}$ captures the probability that a given word in the document is equal to word v, and similarly for $\tilde{\phi}_d^{summary}$.

I measure fit using the standard measure of perplexity (Blei et al. 2003). Given a set of full documents D_{test} with a total of N words, where the word distribution of each document d is fitted by the $1 \times L$ vector $\tilde{\phi}_d$ and where {obs} represent the indices of the words observed in the documents, the perplexity score is given as follows:

$$\text{Perplexity} = \exp\left(-\frac{\sum_{d\in D_{test}}\sum_{obs\in d}\log(\tilde{\phi}_{d,obs})}{N}\right). \quad (3)$$

Perplexity is defined similarly for the document summaries:

$$\text{Perplexity}^{summary} = \exp\left(-\frac{\sum_{d^{summary}\in D_{test}}\sum_{obs\in d^{summary}}\log(\tilde{\phi}_{d,obs}^{summary})}{N^{summary}}\right), \quad (4)$$

where $N^{summary}$ is the total number of words in the summaries and $obs \in d^{summary}$ refers to the words observed in the summary of document d. Note that perplexity is equivalent to the inverse of the geometric mean of the per-word likelihood. Lower scores indicate better fit.

For each model, I also estimate the intensities on regular topics $\{\theta^{reg}_{d^{val}k}\}_k$ and the residual topic weights $\{\beta^{res}_{d^{val}v}\}_v$ for each validation document d^{val}, based on the text of this document and the parameters estimated from the calibration sample. Details are provided in Web Appendix C. Following Equation 3, I compute a perplexity score for the full texts of the validation documents.

Therefore, the in-sample fit measures consist of the perplexity scores for the full texts of the calibration documents, the summaries of the calibration documents, and the full texts of the validation documents. In addition, Web Appendix H reports the deviance information criterion (DIC) for each benchmark, showing that it is lowest for the full model in all three data sets.

Measure of Predictive Performance

The predictive task considered herein is that of predicting the content of the summary of a validation document, given the full text of this document and the model parameters estimated on the set of calibration documents. Consider a validation document d^{val} for which the intensities on the regular topics and the residual topic weights are estimated (as described previously and detailed in Web Appendix C) and for which the task is to predict the content of the summary. Poisson rates for the summary of document d^{val} are predicted according to Equation 2.[5] These rates capture the occurrences of words in the summary, predicted based on the full text of the document, given the model. Following Equation 4, I compute a perplexity score for the summaries of the validation documents, which then serves as the measure of predictive performance.

Results

Table 2 reports the performance of the proposed model, the nested benchmarks, and LDA on each of the three data sets. The comparisons between benchmarks are similar across data sets. It is evident that the proposed model performs best in terms of fitting the summaries of calibration documents and predicting the summaries of validation documents and that the "No residual topic" benchmark usually performs worse than the "ε constant" benchmark. This finding suggests that the better performance of the full model is driven primarily by the inclusion of "residual" topics rather than by allowing various topics to be weighed differently in summaries compared to the full texts. One exception is the TV shows data set, in which the "No residual topic" benchmark performs better than "ε constant" at predicting the content of the summaries of validation documents. As shown subsequently, this data set is the one that features the highest variation in the offset variables $\{\varepsilon^{reg}_k\}$ across regular topics. It is therefore not surprising that assuming that ε is constant is more detrimental in that data set.

The "Residual topics only" benchmark, not surprisingly, performs best in terms of fitting the full documents. This benchmark does not attempt to learn any topic across documents; that is, it does not generate any substantive insight. In addition, the fit on the full documents comes at the expense of fitting or predicting the content of the summaries of documents. Interestingly, this benchmark performs similarly to the "ε constant" benchmark at predicting the content of validation summaries. Both of these benchmarks include residual topics, and they both ignore differences across topics in their propensity to be featured in summaries versus full documents. This is particularly detrimental in the TV show data sets, in which offset variables vary the most across topics. Finally, LDA performs very similarly to the benchmark that has no residual topic and constant offset variables. This benchmark is equivalent to traditional Poisson factorization, which has many similarities with LDA.

Web Appendix H tests an alternative measure of predictive performance in which I randomly held out a subset of the word occurrences in each validation document that are predicted based on the parameter estimates and the other words in the document. In this scenario, the content of validation summaries is predicted based only on a subset of the words in the full

[5] I use the average ε^{res}_d from the validation documents in Equation 2 instead of $\varepsilon^{res}_{d^{val}}$ when predicting summaries of out-of-sample documents based on their full texts, as the estimation of $\varepsilon^{res}_{d^{val}}$ would require access to the very summary to be predicted.

Table 2. Fit and Predictive Performance.

		Fit			Predictive Performance
		Calibration Documents	Calibration Summaries	Validation Documents	Validation Summaries
Marketing academic papers	Full model	104.54	67.60	109.89	82.50
	No residual topic	197.04	141.39	227.73	169.90
	ε constant	104.46	73.13	109.92	85.74
	No residual topic and ε constant	196.91	146.39	227.54	176.83
	Residual topics only	101.83	70.84	106.68	84.30
	LDA	197.13	145.73	227.12	176.34
Movies	Full model	168.72	177.28	179.11	290.04
	No residual topic	265.80	279.76	324.51	358.17
	ε constant	169.07	213.13	179.25	348.25
	No residual topic and ε constant	265.50	346.99	324.54	428.55
	Residual topics only	163.82	204.87	172.39	355.07
	LDA	267.29	343.28	328.05	424.35
TV show episodes	Full model	241.61	246.36	241.08	410.90
	No residual topic	361.10	416.28	358.90	439.27
	ε constant	241.58	311.23	240.85	577.77
	No residual topic and ε constant	360.29	633.89	358.39	686.96
	Residual topics only	234.86	294.33	233.68	574.21
	LDA	360.56	643.63	358.33	696.76

Notes: Fit and predictive performance are measured using perplexity (lower values indicate better fit).

document. I find that the full model performs best in terms of predicting the held-out portion of validation documents and the summaries of validation documents, with the exception of the marketing academic paper data set, in which the "Residual topics only" benchmark performs slightly better at predicting the held-out portion of validation documents.

In summary, these results suggest it is reasonable to extend Poisson factorization to study creative documents and their summaries, by capturing residual content and systematic differences in topic intensities in summaries versus full documents using offset variables. The following three sections illustrate three benefits offered by the proposed model over traditional topic models, as listed in the introduction. First, the joint modeling of creative documents and their summaries sheds light on the process by which people summarize creative documents and enhances understanding of the significance of the topics estimated by the model. Second, the model may be used to construct various measures of distinctiveness for creative documents, which can inform the debate on the link between distinctiveness and success in creative industries. Third, I present a proof of concept of an online tool based on the model, which can assist humans in writing summaries of creative documents. The remainder of the article focuses on the results based on estimating the model on the calibration sample in each data set.

Model Output: Topics and Offset Variables

As mentioned previously, I set the number of regular topics K to 100, expecting only some of the topics to be nonflat. Indeed, I find that the number of regular topics that have meaningful variations in their weights $\{\beta_{kv}^{reg}\}_v$ and intensities $\{\theta_{dk}^{reg}\}_d$ is 30

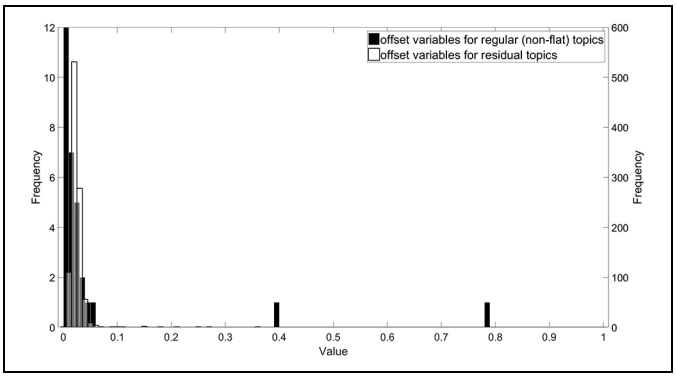

Figure 2. Distribution of offset variables: marketing academic papers.

for the marketing academic paper data set, 24 for the movie data set, and 19 for the TV show data set.[6] These regular topics are not defined merely as groups of words that tend to appear together in documents, but rather as groups of words that tend to appear together in documents *and* that tend to have similar weights in summaries relative to documents. In addition, each topic comes with an offset variable, which quantifies the extent to which the topic was deemed "summary worthy" by the people who wrote the summaries of the documents in the corpus. Figure 2 plots, for the marketing academic paper data set, the

[6] I identify nonflat topics using the standard deviation of topic weights across words. There is always a mass of topics that have very low standard deviation. Because the exact value of this low standard deviation varies slightly across data sets, I do not apply a fixed cutoff; rather, I identify the mass of flat topics on a case by case basis, by inspection.

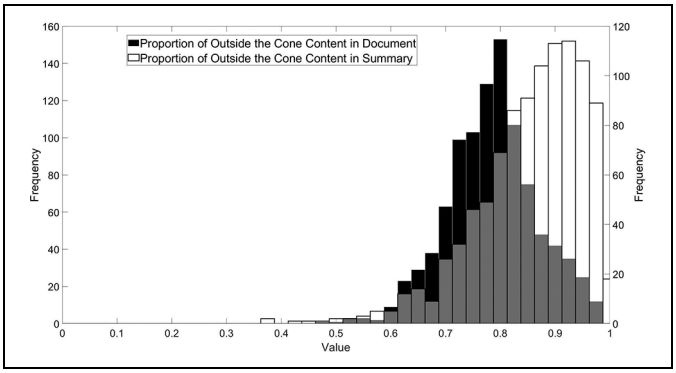

Figure 3. Distribution of the proportion of fitted "outside the cone" content in documents and summaries" marketing academic papers.

Notes: the proportion of fitted "outside the cone" content in document d is measured as $\frac{\sum_v \beta_{dv}^{res}}{\sum_v \left[\sum_k \theta_{dk}^{reg} \beta_{kv}^{reg} + \beta_{dv}^{res}\right]}$. The proportion of fitted "outside the cone" content in the summary of document d is measured as $\frac{\sum_v \beta_{dv}^{res} \varepsilon_d^{res}}{\sum_v \left[\sum_k \theta_{dk}^{reg} \beta_{kv}^{reg} \varepsilon_k^{reg} + \beta_{dv}^{res} \varepsilon_d^{res}\right]}$. The correlation between the two proportions is .66 across documents in the calibration sample ($p < .01$).

distribution of the offset variables across the nonflat regular topics, ε_k^{reg}, together with the distribution of the offset variables for the residual topics, ε_d^{res}, across documents in the calibration sample. Some variation is present in offset variables across regular topics, confirming that there is value in allowing each regular topic to be weighed differently in summaries versus full documents. In particular, two of the regular topics are outliers with very large offset variables, and there appears to be a mass of regular topics with very low offset variables. The corresponding distributions for the other data sets are reported in Web Appendix A. The standard deviation of ε_k^{reg} across regular topics is smallest in the marketing academic paper data set (std = .16), followed by the movie (std = 1.99) and the TV show (std = 11.41) data sets. This may be interpreted as suggesting that the difference in content between synopses and dialogues is greater than the difference between synopses and scripts, which itself is greater than the difference between academic abstracts and papers, which has good face validity. Note that the introduction of residual topics reduces the number of nonflat regular topics and changes their content. Indeed, the nested version without residual topics finds 100 nonflat topics in all three data sets. In addition, the regular topics identified by the nested version without residual topics have less variation in ε_k^{reg}: the standard deviation of ε_k^{reg} across regular topics is decreased respectively to .01, 1.07, and 2.51 in the marketing academic paper, movie, and TV show data sets.

Figure 3 reports the distribution of the proportion of fitted content assigned to the residual topic ("outside the cone") in documents and summaries, for the academic paper data set. The proportion of fitted "outside the cone" content in document d is measured as $\frac{\sum_v \beta_{dv}^{res}}{\sum_v \left[\sum_k \theta_{dk}^{reg} \beta_{kv}^{reg} + \beta_{dv}^{res}\right]}$, and the proportion of fitted "outside the cone" content in the summary of document d is measured as $\frac{\sum_v \beta_{dv}^{res} \varepsilon_d^{res}}{\sum_v \left[\sum_k \theta_{dk}^{reg} \beta_{kv}^{reg} \varepsilon_k^{reg} + \beta_{dv}^{res} \varepsilon_d^{res}\right]}$. In this data set, these two proportions have a correlation of .66 across documents ($p < .01$).

I next report descriptions of the nonflat regular topics and illustrate the type of insight offered by estimating offset variables for these topics. Web Appendix A reports the offset variables, the average topic intensities across documents, and the words with the highest topic weights for all nonflat regular topics in each data set. I also visualize some of these topics by creating word clouds using randomly drawn words according to a discrete probability distribution with weights proportional to the topic weights β. The position of words in these word clouds has no meaning, but the size of each word indicates its frequency in the simulated data (i.e., its weight on that topic). Figure 4 shows word clouds for the two regular topics with the smallest offset variables ε_k^{reg} in the marketing academic paper data set. These are topics that tend to be underrepresented in summaries when compared with full documents. For example, one of these topics has large weights on words like "participants," "people," and "manipulation," which can be interpreted as providing details related to experiments. The other topic has larger weights on words like "model," "table," "parameters," and "estimates," which can be interpreted as providing details related to data analysis. Figure 5 shows word clouds corresponding to the two regular topics that have the largest offset variables (i.e., that tend to be overrepresented in the abstracts of marketing academic papers compared with full papers). Note that one of them has a disproportionately large weight on the word "find" and the other has a disproportionately large weight on the word "firm"; these topics might be interpreted as describing the findings of a paper, and its implications for firms. In summary, the results suggest that when writing abstracts of marketing academic papers, authors tend to emphasize the paper's findings and its implications for firms, and underemphasize details related to data collection and data analysis. Such findings have good face validity.

Web Appendix A displays similar information for the movie and TV show data sets. In the movie data set, the topics with the lowest offset variables appear to relate to the setting of various scenes in the movie. In the TV show data set, the two topics with the smallest offset variables appear to relate to standard dialogues. The topic with the largest offset variable appears to relate to actions (e.g., "gets," "takes," "finds," "comes"), and relationships (e.g., "friends," "family"). The topic with the second largest offset variable appears to relate to the appearance of guest stars and other special events in the episode.

The figures and tables reported in this section illustrate the additional layer of information provided by the joint modeling of creative documents and their summaries. Offset variables provide insight into the process by which people summarize creative documents in a particular domain and enhance understanding of the significance of each topic. As noted previously, the introduction of residual topics reduces the number of nonflat regular topics estimated by the model. "Rare" topics, those that are shared by only a small number of documents, are likely to be reflected in residual topics rather than regular topics.

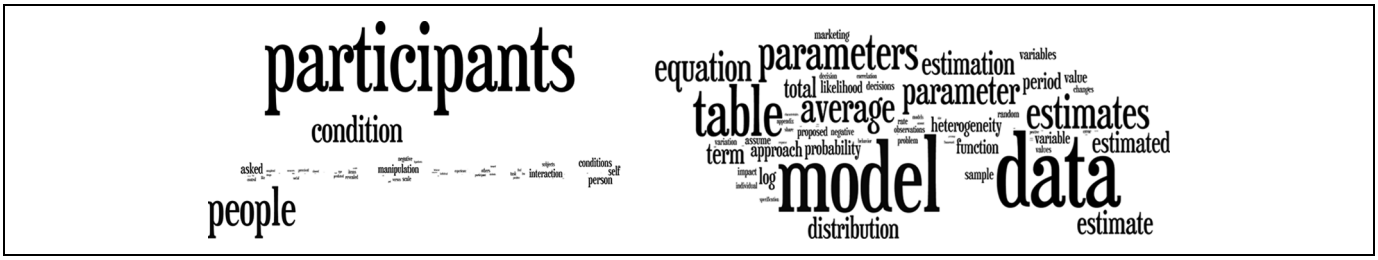

Figure 4. Word clouds for regular topics with smallest offset variables: marketing academic papers.
Notes: These topics tend to be underrepresented in summaries compared with full texts.

Figure 5. Word clouds for regular topics with largest offset variables: marketing academic papers.
Notes: These topics tend to be overrepresented in summaries compared with full texts.

Hence, if a researcher's goal is to identify such rare topics, the version of the model that does not include residual topics may be preferred. Including residual topics, in contrast, greatly improves the model's ability to fit and predict the content of documents and summaries and allows researchers to develop a rich set of distinctiveness measures that may be linked to success. Indeed, when residual topics are not present, two of the distinctiveness measures defined in the next section become unavailable.

Measuring the Distinctiveness of Creative Documents

Some debate has occurred in the literature on the relationship between distinctiveness and success in creative industries. This section reviews some of the empirical studies that have contributed to this debate and shows that the proposed model may be used to estimate various measures of distinctiveness, which may help researchers paint a more nuanced picture of the relationship between the distinctiveness and success of creative documents.

Distinctiveness Measures Based on Proposed Model

I consider three distinctiveness measures, each of which relies on different aspects of the model. The first, directly based on Berger and Packard (2018), measures the distinctiveness of the *combination* of *regular* topics in a document. Given a reference group g, I measure the distinctiveness of a document d with intensities on nonflat regular topics $\{\theta_{dk}^{reg}\}_k$, as $1 - \Sigma_k \frac{|\theta_{dk}^{reg} - \theta_{gk}^{reg}|}{\theta_{dk}^{reg} + \theta_{gk}^{reg} + 0.0001}$. (Berger and Packard [2018], use the same equation, based on the topic intensities provided by LDA.) I refer to this measure of distinctiveness as "inside the cone distinctiveness," because it is based on how the intensities of a document on the regular topics differ from the average in a reference group. I use the journal in which the paper was published, the genre of the movie,[7] and the TV series to which the episode belongs as the reference groups in the respective data sets.

The second measure, which is novel and not available from traditional topic models, is based on the "outside the cone" content in the document. I compute, for each creative document, the proportion of fitted content allocated to the residual topic: $\frac{\sum_v \beta_{dv}^{res}}{\sum_v \left[\sum_k \theta_{dk}^{reg} \beta_{kv}^{reg} + \beta_{dv}^{res}\right]}$. I refer to this measure as "outside the cone distinctiveness." The third measure, which is also novel and not available from traditional topic models, is based on how "outside the cone" content is weighted in the document's summary, relative to the full document. I simply consider the offset variable on the document's residual topic, ε_d^{res}, as a measure of how strongly the document's summary emphasizes the "outside the cone" content from the document. I refer to this measure as "outside the cone emphasis in summary."

In these analyses, I standardize all three measures across documents for interpretability. Table 3 reports the correlations between the three distinctiveness measures in each data set. The lack of consistently high correlation between any pair of measures suggests that these three measures indeed capture different aspects of creative documents.

Distinctiveness Versus Success in Academic Papers

The extant literature has found a general positive relationship between distinctiveness and number of citations in academic papers (Uzzi et al. 2013).[8] The number of citations of each paper in the data set is extracted using the application programming interface (API) offered by Crossref (www.crossref.org).[9] For the papers in the calibration data set, I regress the log of 1 plus the number of citations (I take the log due to the skewness of the number of citations) on the three distinctiveness

[7] If a movie belongs to multiple genres, I average the distinctiveness measure across these genres.
[8] Uzzi et al. (2013) measure distinctiveness based on the combinations of journals cited by the paper.
[9] Using a random subsample of 100 papers, I find that the number of citations provided by Crossref correlates highly with Google Scholar ($\rho = .964$) and ISI ($\rho = .979$), while offering the benefit of being publicly available via an API.

Table 3. Correlation Between Distinctiveness Measures.

		"Outside the Cone Distinctiveness"	"Outside the Cone Emphasis in Summary"
Marketing academic papers	"Inside the cone distinctiveness"	–.03	–.16**
	"Outside the cone distinctiveness"		.15**
Movies	"Inside the cone distinctiveness"	–.48**	–.14**
	"Outside the cone distinctiveness"		.05
TV shows	"Inside the cone distinctiveness"	.25**	.03**
	"Outside the cone distinctiveness"		.09**

*Significant at $p < .10$.
**Significant at $p < .05$.

Table 4. Link Between Distinctiveness Measures and Citations: Marketing Academic Papers.

Covariates	DV = log(1+#citations)
Journal fixed effects	✓
Publication year fixed effects	✓
Intensities of paper on nonflat regular topics	✓
Number of pages in paper	.040**
"Inside the cone distinctiveness" (from journal)	.113**
"Outside the cone distinctiveness"	.140**
"Outside the cone emphasis in summary"	.059**
Number of parameters	43
Number of observations	1,000
R^2	.353

*Significant at $p < .10$.
**Significant at $p < .05$.
Notes: Ordinary least squares regression. All three distinctiveness measures are standardized across papers for interpretability.

measures.[10] I control for journal fixed effects, publication year fixed effects, the paper's intensities on (nonflat) regular topics $\{\theta_{dk}^{reg}\}_k$, and the paper's number of pages. Table 4 provides results. Note that all three measures of distinctiveness are significantly and positively related to a paper's number of citations. That is, in the data the number of citations received by a marketing academic paper tends to be higher when the paper uses an unusual combination of intensities on regular topics, when the paper features more "outside the cone" content, and when the abstract weighs this content disproportionately. The magnitudes of the regression coefficients suggest that the strongest relation is with "outside the cone distinctiveness" (recall all distinctiveness measures are standardized).[11]

These results are purely correlational. Moreover, I was not able to include all the variables from all previous analyses of the factors of citations of marketing academic articles (e.g., Stremersch, Verniers, and Verhoef 2007; Stremersch et al. 2015, who do not focus on distinctiveness). My goal is not to make definitive claims on the causal relation between distinctiveness and number of citations of marketing academic papers; rather, it is to illustrate how the distinctiveness measures derived from the proposed model may be used by researchers interested in contributing to that literature. Interestingly, at least in this data set, the content of summaries appears to be related to the success of creative documents. This echoes Pryzant, Chung, and Jurafsky (2017), who study the link between the presence of certain phrases in the description of products in e-commerce platforms (e.g., including references to authority or seasonality) and product sales. Given the ubiquity of summaries across creative industries, further research may be conducted that links the success of creative products to variations in the content of their summaries.

Distinctiveness Versus Success in Entertainment Products

While the extant literature makes a clear prediction on the link between distinctiveness and citations in academic papers, the literature is not as clear on the link between distinctiveness and success in the context of entertainment products. On the one hand, Berger and Packard (2018) show that songs whose lyrics are more different from their genres are ranked higher in digital downloads. Danescu-Niculescu-Mizil et al. (2012) and Askin and Mauskapf (2017) also find that distinctiveness is an attractive feature of entertainment products. On the other hand, according to Salganik, Dodds, and Watts (2006), the content of entertainment products has little impact on the success of these products, echoing previous research by Bielby and Bielby (1994), who also report a quote from a past president of CBS entertainment that "all hits are flukes," and Hahn and Bentley (2003).

[10] I also tested specifications that include a square term for each measure, to allow for diminishing returns to distinctiveness. However, I find no evidence of significant diminishing returns to distinctiveness in any of the data sets.

[11] I repeated the analysis on a sample of 632 papers published between 2010 and 2015 in top sociology journals (*European Sociological Review*, *American Sociological Review*, and *American Journal of Sociology*; the data were graciously made available to us by Boghrati, Berger, and Packard [2020]). On this smaller data set, I also find that the strongest relation is with "outside the cone distinctiveness," although the coefficient is only marginally significant. The coefficients for the other two measures are not significant. Future research could explore potential commonalities and differences across academic fields.

Table 5. Link Between Distinctiveness Measures and Performance: Movies.

Covariates	DV = Movie Rating	DV = Log(Return on Investment)
MPAA rating fixed effects		
Genre fixed effects		
Intensities of script on nonflat regular topics		
Movie duration (in min)	.003*	–.003
Log(inflation-adjusted production budget)	–.093**	–.329**
Movie rating	–	.451**
"Inside the cone distinctiveness" (from genre)	–.090**	–.027
"Outside the cone distinctiveness"	.253**	–.109
"Outside the cone emphasis in summary"	.050	.069
Number of parameters	54	55
Number of observations	596	581
R^2	.357	.262

*Significant at $p < .10$.
**Significant at $p < .05$.
Notes: Each column corresponds to one regression estimated separately using ordinary least squares. All three distinctiveness measures and movie ratings are standardized across movies for interpretability. Observations in the first (second) regression are limited to movies for which production budget was available (production budget and box office performance were available).

Table 6. Link Between Distinctiveness Measures and Performance: TV Episodes.

Covariates	DV = Episode Rating
TV series fixed effects	✓
Intensities of script on nonflat regular topics	✓
"Inside the cone distinctiveness" (from TV series)	–.005
"Outside the cone distinctiveness"	.074**
"Outside the cone emphasis in summary"	–.008
Number of parameters	340
Number of observations	9,285
R^2	.687

*Significant at $p < .10$.
**Significant at $p < .05$.
Notes: Ordinary least squares regression. All three distinctiveness measures and episode ratings are standardized across episodes for interpretability.

Ratings. I first analyze the link between the three measures of distinctiveness and the ratings of movies and TV shows. For each movie in the calibration data set, I collect the average rating from IMDB (based on the ratings of IMDB users), which I standardize across movies for interpretability. I include fixed effects for the movie's MPAA rating, fixed effects for the movie's genre(s), the movie's intensities on the (nonflat) regular topics, the movie's duration (in min), and the log of the movie's production budget (in U.S. dollars, adjusted for inflation, using the tool available at https://data.bls.gov/cgi-bin/cpicalc.pl). All these control variables (with the exception of the intensities on regular topics) are obtained from IMDB. Results are provided in the first column of Table 5. I find that "outside the cone distinctiveness" is positively related to the movie's rating. Interestingly, "inside the cone distinctiveness" is actually negatively related to the movie's rating in the data set (i.e., movies whose regular topic intensities deviate more from the mean of their genre tend to receive lower ratings). I also find that "outside the cone emphasis in summary" is not significantly related to ratings. This is not surprising, given that the role played by synopses in the movie industry is more restricted than the role played by abstracts in academia.

For TV shows, I obtained IMDB ratings for 9,358 of the episodes in the calibration data set (some episodes were not found on IMDB, and IMDB reports ratings only for episodes that were rated by at least five users). For the analysis, I only kept episodes from TV series for which ratings on at least two episodes were available, so that I could include fixed effects for each TV series. This resulted in 9,285 observations and 318 fixed effects. In addition, I control again for the episode's intensities on the (nonflat) regular topics. Results are provided in Table 6. In this data set, consistent with the analysis of movie ratings, I find that "outside the cone distinctiveness" is positively related to the TV show's rating. The coefficients for the other two measures of distinctiveness are not statistically significant.

Return on investment. Finally, for movies, I analyze the link between distinctiveness and financial success, measured as the log of the movie's return on investment, defined as in Eliashberg, Hui, and Zhang (2014) as the ratio of the movie's domestic box office performance (also obtained from IMDB) to its production budget. In addition to the controls included in the first regression reported in Table 5, I control for the movie's rating. Results, again based on the calibration data set, are provided in the second column of Table 5. This data set shows that none of the distinctiveness measure is significantly related to financial success.

Discussion

The analysis provided herein suggests that "inside the cone distinctiveness," "outside the cone distinctiveness," and "outside the cone emphasis in summary" provide meaningful and useful measures of distinctiveness, which may have different relations to success, depending on the context and on how success is defined and measured. Across three data sets, "outside the cone distinctiveness" (a novel measure introduced here) is robustly and positively associated with success. In contrast, "inside the cone distinctiveness" (which is directly based on extant research) is positively related to the number of citations of marketing academic papers but negatively related to movie ratings. This is not inconsistent with the literature, which suggests that distinctiveness should be positively related to success for academic papers, but which is more ambivalent on the link between distinctiveness and success in entertainment industries. Finally, in the context of marketing

academic papers, I find that putting more emphasis in an academic paper's abstract on the "outside the cone" content from the paper is associated with a larger number of citations.

Note that the measures of distinctiveness are based on the entire set of training documents and thus do not capture novelty with respect to *contemporaneous* documents. In particular, some documents may have been novel when they were released/published and may have become influential, leading to similar future documents. Such novel documents may not score high on the distinctiveness measures despite being novel, due to the presence of similar documents in the corpus. The dynamic version of the model described in Web Appendix G addresses this issue by allowing topics to evolve over time, hence measuring the distinctiveness of a document with respect to the topics defined at the time this document was published. I apply this dynamic extension of the model to the marketing academic paper data sets, which contains all papers published in a set of journals over six years, and find that "inside the cone distinctiveness," "outside the cone distinctiveness," and "outside the cone emphasis in summary" are still all positively and significantly related to the number of citations.

Web Appendix H also tests various alternative measures of distinctiveness and alternative ways to explore the link between distinctiveness and success. I find that as the vocabulary size changes, the significance of some of the coefficients associated with distinctiveness measures may change, although I observe no reversal (i.e., a coefficient that is significant in one direction under one vocabulary size is never significant in the other direction under a different vocabulary size). A simulation study conducted to illustrate how measures of distinctiveness are affected as relevant words are omitted from the vocabulary or as irrelevant words are included in the vocabulary confirms that the selection of the vocabulary size is bound to have some impact on the output of the model. While this is not an attractive feature, unfortunately this is a characteristic of any topic model, not just the one presented here. Using alternative specifications that link distinctiveness to financial success in the movie data set yields similar results to those reported in Table 5. Measuring "inside the cone distinctiveness" using the entire set of training documents as the reference group, rather than documents in the same journal / genre / TV series, produces results similar to those reported in Tables 4–6. I perform an analysis that reflects the fact that measures of distinctiveness are constructed from model parameters that are estimated with uncertainty rather than measured precisely. I run each regression 1,000 times using different draws from the posterior distribution of the model parameters and report the average coefficients as well as whether the 90% and 95% credible intervals include 0. Results are consistent with those reported in Tables 4–6. Finally, I measure distinctiveness using standard topic models (LDA and Poisson factorization), rather than the proposed model. "Inside the cone distinctiveness" is the only distinctiveness measure available from these models, and it is never statistically significantly related to success in any of the regression, with the exception of "inside the cone distinctiveness" estimated based on the standard Poisson factorization, which is marginally related to return of investment of movies.

Computer-Assisted Summary Writing

As mentioned in the literature review, the traditional approach in the computer science literature would be to attempt to completely automate the summarization of documents, typically via sentence extraction. I argue that this approach is less relevant in the context of *creative* documents. In particular, the nature of creative documents is such that the stakes are usually high enough for people to be motivated and available to write summaries. For example, the author or publisher of a new book typically has enough motivation to write a synopsis for this document and may not find as much value in a tool that would automatically generate a summary. Similar comments may be made about the publisher of a new movie or play, the author of an academic paper, the developer of an innovative product, the author of a business plan, and so on. This is in contrast to the traditional text summarization literature that typically deals with the summarization of large volumes of documents such as news articles, where automation has significant cost saving implications. Moreover, sentence extraction is likely to be an inappropriate text summarization approach in many creative contexts. For example, an abstract of a scientific paper made exclusively of sentences from the paper, or a TV show synopsis made exclusively of sentences from the show's dialogues, may be unacceptable to the relevant audience. Accordingly, I argue that in the creative context, it is more useful to develop decision support tools that assist humans in writing summaries of creative documents, rather than developing automatic text summarization tools featuring sentence extraction.

I have built a proof of concept for such a decision support tool, using php and a mysql database. The tool allows a user to upload a creative document that was not necessarily part of the corpus on which the model was estimated. When the user submits a new document d^{out}, the text of this document is tokenized on the fly (using custom-built php code developed by the author), and the number of occurrences of each word in the vocabulary is computed for that document. Intensities on the regular topics $\{\theta^{reg}_{d^{out}k}\}_k$ and the residual topic $\beta^{res}_{d^{out}}$ for the new document d^{out} are estimated in real time using variational inference, given the other model parameters.[12]

As output, the tool reports representative words for the five regular topics with the largest intensities ($\theta^{reg}_{d^{out}k}$) and

[12] Variational inference is performed on the fly within php, using code developed by the author. To speed up computations, the di-gamma function $\Psi(x)$ is approximated as follows. If $x < 2$, x is rounded to the nearest thousandth, and $\Psi(x)$ is obtained from a look-up table. If $x \geq 2$, $\Psi(x)$ is approximated by its asymptotical expansion, with precision $O(\frac{1}{x^{16}})$. Also, to speed up computations, the ascent mean-field variational inference algorithm is run for 100 iterations systematically, rather than checking convergence at each iteration. These approximations are made only in the online tool. With the current implementation, all computations for a new document are typically performed within 5 seconds.

representative words for the residual topic. In addition, the tool reports representative words for the five regular topics that the model predicts should have the largest intensities in the summary of the new document (i.e., the regular topics with the largest values of $\theta^{reg}_{d^{out}k} \times \varepsilon^{reg}_{k}$). In the current implementation, for each topic I report the 10 words with the highest weights on the topic (β_{kv}) as representative words.[13]

Because the model should be run separately in each domain, I customize the tool for each domain of application. I have created one version of the online tool corresponding to each corpus studied herein (marketing academic papers, movies, and TV shows). This proof of concept is publicly available at http://creativesummary.org.[14]

Importantly, such a decision support tool may also leverage analysis such as the one reported in the previous section, to help users improve the effectiveness of their summaries. For example, I found that marketing academic papers in which the abstract puts more emphasis on the "outside the cone" content in the paper (i.e., higher ε^{res}_{d} or "outside the cone emphasis in summary") tend to have more citations. Without making unfounded causal claims, the online tool can report this correlational finding to the user. Accordingly, in the proof of concept of the tool tailored for marketing academic papers, I include the following statement next to the representative words from the "outside the cone" topic: "Our research suggests that a paper whose abstract puts more emphasis on the paper's 'outside the cone' topic tends to receive more citations."

Conclusions

The contribution of this article is primarily methodological. I develop and apply a new topic model designed specifically for the study of creative documents. Guided by the creativity literature, this model nests and extends Poisson factorization in two ways. First, I explicitly model residual, "outside the cone" content and how it is represented in summaries versus documents. Second, I jointly model the content of documents and their summaries, and quantify (using offset variables) how the intensity of each topic differs systematically in summaries compared with full documents. I validate the model using three data sets containing marketing academic papers (summarized by abstracts), movie scripts (summarized by synopses), and TV show closed captions (summarized by synopses). The proposed model offers the standard benefits of topic models; that is, it extracts topics from a corpus of documents and assigns intensities on each topic for each document (although the introduction of residual topics changes the number and content of the nonflat regular topics). This article illustrates three additional benefits the model provides for academics and practitioners. First, the offset variables estimated by the model, which quantify the extent to which each topic was deemed "summary worthy" by the humans who wrote the summaries of the documents in the corpus, shed light into the process by which humans summarize creative documents and identify the significance of each topic. Second, I illustrate how the model may be used to construct new measures of distinctiveness for creative documents, which have the potential to shed new light on the relation between distinctiveness and success in creative industries. Third, I develop an online, interactive, freely accessible tool based on the model, which provides a proof of concept for using the model's output to assist humans in writing summaries of creative documents.

I close by highlighting additional areas for future research. First, it would be interesting to introduce covariates into the model that influence the topic intensities and/or the offset variables. In the context of entertainment products, such covariates might include genres, country of origin, and so on. In the context of academic papers, these covariates may include subfields, whether the paper is based on a dissertation, and so on. Second, alternative topic models may capture the *structure* of creative documents (e.g., different sections, scenes, acts). Third, it would be worthwhile to study how the content of summaries varies systematically based on the objectives of the summary. For example, in some cases summaries serve primarily as "teasers" for creative products, while in others they serve more as "substitutes" for the products. For example, the offset variables might differ systematically between spoilers and synopses, or between abstracts written for conferences versus journal articles.

Acknowledgments

Yanyan Li, Ahmed Mrad, and Sibel Sozuer Zorlu provided outstanding research assistance on this project.

Associate Editor

Vrinda Kadiyali

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Allahyari, Mehdi, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut (2017), "Text Summarization Techniques: A Brief Survey," arXiv preprint arXiv:1707.02268.

[13] I only show words that have sufficient weights on the topic: $\frac{\beta_{kv}}{\sum_{v'} \beta_{kv'}} > .01$. If the topic is flat and no word satisfies this criterion, I do not report the topic.

[14] The php code is common across domains. The vocabulary for each domain, the weights of the regular topics $\{\beta^{reg}_{k}\}_{k}$, and the offset variables on regular topics $\{\varepsilon^{reg}_{k}\}_{k}$, which are all obtained from estimating the model on the corresponding corpus, are stored in the database that supports the tool. Creating a version of the tool for a new domain (e.g., business plans) only requires running the model on a corpus from this domain, and uploading the results onto the database.

Askin, Noah and Michael Mauskapf (2017), "What Makes Popular Culture Popular? Product Features and Optimal Differentiation in Music," *American Sociological Review*, 820 (5), 910–44.

Berger, Jonah and Grant Packard (2018), "Are Atypical Things More Popular?" *Psychological Science*, 290 (7), 1178–84.

Bielby, William T. and Denise D. Bielby (1994), "'All Hits are Flukes': Institutionalized Decision Making and the Rhetoric of Network Prime-Time Program Development," *American Journal of Sociology*, 990 (5), 1287–313.

Blei, David M. and John D. Lafferty (2006), "Dynamic Topic Models," in *Proceedings of the 23rd International Conference on Machine Learning*. New York: Association for Computing Machinery, 113–20.

Blei, David M. and John D. Lafferty (2009), "Topic Models," *Text Mining: Classification, Clustering, and Applications*, 100 (71), 34.

Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe (2016), "Variational Inference: A Review for Statisticians," arXiv preprint arXiv:1601.00670.

Blei, David M., Andrew Y. Ng, Michael I. Jordan, and John Lafferty (2003), "Latent Dirichlet Allocation," *Machine Learning*, 30 (4/5), 993–1022.

Boghrati, Reihane, Jonah Berger, and Grant Packard (2020), "How Writing Style Shapes the Impact of Scientific Findings," working paper, The Wharton School, University of Pennsylvania.

Büschken, Joachim and Greg M Allenby (2016), "Sentence-Based Text Analysis for Customer Reviews," *Marketing Science*, 350 (6), 953–75.

Canny, John (2004), "Gap: A Factor Model for Discrete Data," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM, 122–29..

Danescu-Niculescu-Mizil, Cristian, Justin Cheng, Jon Kleinberg, and Lillian Lee (2012), "You Had Me at Hello: How Phrasing Affects Memorability," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers, Volume 1*. Stroudsburg, PA: Association for Computational Linguistics, 892–901.

Daumé, Hal, III, and Daniel Marcu (2006), "Bayesian Query-Focused Summarization," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics, 305–12.

Delort, Jean-Yves and Enrique Alfonseca (2012), "Dualsum: A Topic-Model Based Approach for Update Summarization," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics, 214–23.

Eliashberg, Joshua, Sam K. Hui, and John Z. Zhang (2007), "From Story Line to Box Office: A New Approach for Green-Lighting Movie Scripts," *Management Science*, 530 (6), 881–93.

Eliashberg, Joshua, Sam K. Hui, and Z. John Zhang (2014), "Assessing Box Office Performance using Movie Scripts: A Kernel-Based Approach," *IEEE Transactions on Knowledge and Data Engineering*, 260 (11), 2639–48.

Finke, Ronald A., Thomas B. Ward, and Steven M. Smith (1992), *Creative Cognition: Theory, Research, and Applications*. Cambridge, MA: The MIT Press.

Florida, Richard (2014), *The Rise of the Creative Class—Revisited: Revised and Expanded*. New York: Basic Books.

Giora, Rachel (2003), *On Our Mind: Salience, Context, and Figurative Language*. Oxford, UK: Oxford University Press.

Gopalan, Prem K., Laurent Charlin, and David Blei (2014), "Content-Based Recommendations with Poisson Factorization," in *Advances in Neural Information Processing Systems*. New York: Association for Computing Machinery, 3176–84.

Gopalan, Prem, Jake M. Hofman, and David M. Blei (2013), "Scalable Recommendation with Poisson Factorization," arXiv preprint arXiv:1311.1704.

Haghighi, Aria and Lucy Vanderwende (2009), "Exploring Content Models for Multi-Document Summarization," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics, 362–70.

Hahn, Matthew W. and R. Alexander Bentley (2003), "Drift as a Mechanism for Cultural Change: An Example from Baby Names," *Proceedings of the Royal Society of London B: Biological Sciences*, 2700 (Suppl 1), S120–23.

Liu, Jia and Olivier Toubia (2018), "A Semantic Approach for Estimating Consumer Content Preferences from Online Search Queries," *Marketing Science*, 37 (6), 855–1052.

McKeown, Kathleen and Dragomir R. Radev (1995), "Generating Summaries of Multiple News Articles," in *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: Association for Computing Machinery, 74–82.

Mednick, Sarnoff (1962), "The Associative Basis of the Creative Process," *Psychological Review*, 690 (3), 220.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013), "Efficient Estimation of Word Representations in Vector Space," arXiv preprint arXiv:1301.3781.

Mikolov, Tomas, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin (2017), "Advances in Pre-Training Distributed Word Representations," arXiv preprint arXiv:1712.09405.

Nenkova, Ani and Kathleen McKeown (2012), "A Survey of Text Summarization Techniques," in *Mining Text Data*. Berlin/Heidelberg: Springer, 43–76.

Pryzant, Reid, Youngjoo Chung, and Dan Jurafsky (2017), "Predicting Sales from the Language of Product Descriptions," http://sigir-ecom.weebly.com/uploads/1/0/2/9/102947274/paper_3.pdf.

Puranam, Dinesh, Vishal Narayan, and Vrinda Kadiyali (2017), "The Effect of Calorie Posting Regulation on Consumer Opinion: A Flexible Latent Dirichlet Allocation Model with Informative Priors," *Marketing Science*, 360 (5), 726–46.

Radev, Dragomir R., Eduard Hovy, and Kathleen McKeown (2002), "Introduction to the Special Issue on Summarization," *Computational Linguistics*, 280 (4), 399–408.

Radev, Dragomir R. and Kathleen R. McKeown (1998), "Generating Natural Language Summaries from Multiple On-line Sources," *Computational Linguistics*, 240 (3), 470–500.

Salganik, Matthew J., Peter Sheridan Dodds, and Duncan J. Watts (2006), "Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market," *Science*, 3110 (5762), 854–56.

Statista (2017), "Average Time Spent with Major Media Per Day in the United States as of April 2016 (in minutes)," https://www.statista.com/statistics/276683/media-use-in-the-us/.

Statista (2018), "Value of the Global Entertainment and Media Market from 2011 to 2021 (in trillion U.S. dollars)," https://www.statista.com/statistics/237749/value-of-the-global-entertainment-and-media-market/.

Stremersch, Stefan, Nuno Camacho, Sofie Vanneste, and Isabel Verniers (2015), "Unraveling Scientific Impact: Citation Types in Marketing Journals," *International Journal of Research in Marketing*, 320 (1), 64–77.

Stremersch, Stefan, Isabel Verniers, and Peter C. Verhoef (2007), "The Quest for Citations: Drivers of Article Impact," *Journal of Marketing*, 710 (3), 171–93.

Tirunillai, Seshadri and Gerard J. Tellis (2014), "Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation," *Journal of Marketing Research*, 51 (4), 463–79.

Toubia, Olivier, Garud Iyengar, Renée Bunnell, and Alain Lemaire (2018), "Extracting Features of Entertainment Products: A Guided LDA Approach Informed by the Psychology of Media Consumption," *Journal of Marketing Research*, 56 (1), 18–36.

Toubia, Olivier and Oded Netzer (2017), "Idea Generation, Creativity, and Prototypicality," *Marketing Science*, 360 (1), 1–20.

Uzzi, Brian, Satyam Mukherjee, Michael Stringer, and Ben Jones (2013), "Atypical Combinations and Scientific Impact," *Science*, 3420 (6157), 468–72.

Ward, Thomas B. (1995), "What's Old About New Ideas," in *The Creative Cognition Approach*, S. M. Smith, T. B. Ward, and R. A. Finke, eds. Cambridge, MA: The MIT Press, 157–78.

Yao, Jin-ge, Xiaojun Wan, and Jianguo Xiao (2017), "Recent Advances in Document Summarization," *Knowledge and Information Systems*, 530 (2), 297–336.

Zhong, Ning and David A. Schweidel (2020), "Capturing Changes in Social Media Content: A Multiple Latent Changepoint Topic Model," *Marketing Science*, 39 (4), 827–46.