On Provenance in Topic Models

Misha Sharma
Wheeler High School
Marietta, Georgia, USA
misha.sharma@wheelermagnet.com

Arthur Choi
Kennesaw State University
Marietta, Georgia, USA
achoi13@kennesaw.edu

ABSTRACT

Topic models are statistical models that can be used to discover the latent topical structure of a text dataset. In this paper, we consider the provenance of topics found by topic models. In particular, we seek to answer questions such as “Why does this topic appear in this dataset?” and “Why does this topic appear in this document?” We propose to answer these questions by formulating them as probabilistic queries in a topic model. The ability to answer such questions helps us to better understand the topic models that we learn from data, and further providing insights on the text datasets that they represent. Moreover, they provide trust in the topic model, and trust in the topics learned from the data. To facilitate the analysis of provenance in topic models, we contribute a new open-source software system for navigating topic models, called SPOT (Seeking the Provenance Of Topics). Using our new tool, we provide some case studies where we seek the provenance of topics that were discovered from datasets of Wikipedia articles.

CCS CONCEPTS

• **Computing methodologies** → **Topic modeling**; **Bayesian network models**.

KEYWORDS

topic models, latent Dirichlet allocation, Bayesian networks, explainable artificial intelligence (XAI), provenance

ACM Reference Format:
Misha Sharma and Arthur Choi. 2024. On Provenance in Topic Models. In *2024 ACM Southeast Conference (ACMSE 2024), April 18–20, 2024, Marietta, GA, USA*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3603287.3651210

1 INTRODUCTION

The past decade has seen rapid advances in the field of artificial intelligence (AI). As AI systems become more pervasive, the need to understand and explain the behavior of such systems has risen correspondingly. This need has been emphasized recently by an executive order signed by President Biden in October 2023, which was titled *Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. As a response to such concerns, a new subfield of AI, called eXplainable Artificial Intelligence (XAI) has also gained prominence over the past decade [3, 12, 17, 18].

In this paper, we are interested in explaining the behavior of *topic models*. Topic models are statistical models that are used to discover the latent topical structure of a text dataset [2, 4, 9, 11]. For example, topic models are used to automatically discover the topics appearing in Wikipedia [10], or to discover the topics appearing in web pages on the internet; see, e.g., Google’s rephil system [14].

More specifically, we consider the *provenance* of topics found by topic models. As an example, suppose that we are analyzing a dataset of articles about Harry Potter, and that we have learned the topics of this dataset. Say that, to our surprise, we find a topic about religion and Christianity, leading us to ask questions like: *Why does this topic appear in this dataset?* We propose to answer such questions by tracing the source of a topic back to the data points that gave rise to it.[1] By answering such *why?* questions, we build trust in a topic model, and trust in the topics that it learned from the data, one of the goals of XAI [17]. Answering such questions can also provide additional insights about the dataset that a topic model was learned from (in our example, by revealing the connections between Harry Potter and topics such as religion and Christianity).

Our paper is organized as follows. We start in Section 2 by reviewing topic models. Next, in Section 3, we motivate the search for the provenance of topics, and propose specific ways to explain why a topic was discovered in a dataset or document. In Section 4, we introduce a new tool for visualizing topic models, which we call SPOT (**S**eeking the **P**rovenance **O**f **T**opics). We provide two case studies based on datasets collected from Wikipedia: in Section 5, we analyze a dataset on Harry Potter, and in Section 6 we analyze a dataset on 1984. We conclude in Section 7.

2 TOPIC MODELS

Topic models [2, 9], and related models such as Latent Dirichlet Allocation (LDA) [4] and Probabilistic Latent Semantic Analysis (PLSA) [11], are statistical models that seek to learn the underlying topical structure of data, typically of text corpora.

Say we have a dataset of documents, where each document is a set of words. A topic model views each document as a mixture-of-topics, where in turn, a topic is viewed as a mixture-of-words. In Figure 1, we depict a topic model graphically as a Bayesian network (BN), as in [2, 11],[2] along with a corresponding text dataset.

Here, each row of the dataset consists of a single word, where each word is associated with a document D (the index of the document that the word belongs to), a latent topic T of the word (that we are trying to learn), and the word itself W. Given such a dataset,

ACMSE 2024, April 18–20, 2024, Marietta, GA, USA

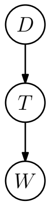
ACM ISBN 979-8-4007-0237-2/24/04…$15.00
https://doi.org/10.1145/3603287.3651210

[1] We note a distinction between topic provenance and data provenance. Data provenance typically refers to the history or lineage of data [6].

[2] More typically, topic models are depicted using plate notation, as in [4].

Table 1: Five (out of Ten) Selected Topics About `Harry Potter`

topic 5		topic 6		topic 7		topic 8		topic 9	
w	$Pr(w)$	w	$Pr(w)$	w	$Pr(w)$	w	$Pr(w)$	w	$Pr(w)$
harry	3.4421	jesus	0.7549	city	1.4491	new	0.7134	book	1.0748
potter	1.4275	new	0.4979	london	1.4095	bbc	0.7062	books	0.8249
voldemort	1.1734	first	0.4705	california	0.8097	times	0.6097	novel	0.7128
hogwarts	0.8396	one	0.4550	los	0.6121	news	0.6049	fiction	0.6745
dumbledore	0.8006	also	0.4468	angeles	0.5982	also	0.5737	series	0.6590
death	0.6898	english	0.4437	moscow	0.5461	company	0.5407	children	0.6551
hermione	0.5658	century	0.4407	world	0.5025	first	0.4699	first	0.5962
rowling	0.5349	bible	0.4111	manchester	0.4971	hbo	0.4603	story	0.5949
order	0.5257	books	0.3503	san	0.4841	service	0.3923	published	0.5627
snape	0.5121	kipling	0.3456	also	0.4646	playstation	0.3918	stories	0.5067



D	T	W
1	?	there's
1	?	a
1	?	lady
1	?	who's
1	?	sure
⋮	⋮	⋮

Figure 1: A Bayesian Network Representation of a Topic Model and a Corresponding Dataset

we can learn the distribution of the BN depicted in Figure 1 (right):

$$Pr(D, T, W) = Pr(W \mid T)Pr(T \mid D)Pr(D)$$

where we assume a word W is independent of the document D that it comes from, when given the topic T of the word. When we learn this BN from data, we obtain:

- the *topic-given-document* distributions $Pr(T \mid D)$, which summarizes each document D as a mixture of topics T; and
- the *word-given-topic* distributions $Pr(W \mid T)$, which represents a topic T as a distribution over words W.

We also obtain $Pr(D)$, which gives us the relative size (number of words) of each document. If we have n documents over a vocabulary of m words, and we assume a total of k topics, then $Pr(W \mid T)$ corresponds to an $(m \times k)$-matrix, $Pr(T \mid D)$ corresponds to a $(k \times n)$-matrix, and $Pr(D)$ corresponds to an $(n \times 1)$-vector.[3]

Topic models can concisely summarize a large dataset of text, as a collection of the topics that it contains. A topic model can also summarize the individual documents of a dataset, as a mixture-of-topics. Consider, for example, google's Rephil system, which is another type of topic model that can represent millions of topics, and tens of millions of words [14]. Rephil can be used, for example, to determine what topics a given web page is about, which can in turn be used to automatically serve a set of appropriate ads.

As an example, consider Table 1, which highlights five topics learned from a small dataset of articles relating to Harry Potter. Each topic is summarized by its top ten most prevalent words. For example, the words of Topic 5 are indicative of a topic about the plot and characters of Harry Potter, and the words of Topic 9 are indicative of a topic about books and literature. We will discuss this example in more depth in our case study of Section 5

3 ON PROVENANCE IN TOPIC MODELS

Beyond summarizing a large dataset of text, topic models (and Bayesian networks in general) facilitate more sophisticated types of analysis and reasoning, that can bring additional insights about a dataset [7, 8, 15]. In particular, topic models are probabilistic models, where the answers to questions may be naturally formulated as probabilistic queries. In this section, we consider questions about the provenance of topics, where our goal is to provide insights about the source of a topic discovered by a topic model.

Say that we are inspecting the topics discovered by a topic model, and we encounter an unexpected topic. We may ask the question:

> *Why does this topic appear in this dataset?*

A topic discovered by a topic model is, intuitively, a collection of words that frequently co-occurred in a sufficiently large number of documents. Thus, one may posit that a topic exists because enough documents discuss that topic. Thus, to explain *why* a topic model discovered a particular topic, one may point to the documents that are most relevant to the topic. That is, we ask: given a topic, what are its most likely documents? More formally, given a topic T, we look for the documents D that maximize the probability $Pr(D \mid T)$:

$$Pr(D \mid T) = \frac{Pr(T \mid D)Pr(D)}{Pr(T)} \propto Pr(T \mid D)Pr(D)$$

Here, $Pr(T \mid D)$ is our topic-given-document distribution, where each document is a weighted mixture of each of the k topics. Moreover, $Pr(D)$ is proportional to the number of words in each document. Hence, for a given topic, a document that maximizes $Pr(D \mid T)$, will maximize a balance between two components:[4]

- $Pr(T \mid D)$: the prominence of the topic in the document; and
- $Pr(D)$: the size of the document (i.e., the number of words).

Note that, by this measure, a document that is very specialized to a topic (where $Pr(T \mid D)$ is close to 1), may be penalized if the document is very short (where $Pr(D)$ is close to 0).[5]

[3] We can view the word-given-document distribution $Pr(W \mid D)$ as an $(m \times n)$-matrix representation of a text dataset where each document is represented by a (column) vector of word counts. A topic model can be viewed as learning a low-rank matrix factorization: $Pr(W \mid D) = \sum_T Pr(W \mid T)Pr(T \mid D)$ where $k < n < m$ [20].

[4] Note that if we are maximizing $Pr(D \mid T)$ with respect to D, then $Pr(T)$, which is independent of D, is effectively a constant, and we can ignore it.

[5] For comparison, the Topic Model Visualization Engine (TMVE) of [5] ranks the most relevant documents of a topic by $Pr(T \mid D)$ alone.

Next, suppose that for a specific topic, we are now inspecting its most likely documents. We may encounter a document that is, unexpectedly, related to the given topic. We may ask the question:

Why does this topic appear in this document?

Again, a topic is (roughly) a collection of words that frequently co-occur in a sufficiently large subset of documents. One may thus posit that a topic appears in a document because enough of the words in the document belong to that topic. Thus, to explain *why* a topic exists in a document, one may point to the set of words in the document that belong to that topic. That is, we ask: given a document, which words are most likely about a given topic? More formally, given a document D, and for each word W in the document, we compute the probability $Pr(T \mid W, D)$ of its topic T:

$$Pr(T \mid W, D) = \frac{Pr(W \mid T)Pr(T \mid D)Pr(D)}{Pr(W, D)} \propto Pr(W \mid T)Pr(T \mid D)$$

Here, $Pr(W \mid T)$ is our word-given-topic distribution, where each topic is viewed as a weighted mixture of words. We also have $Pr(T \mid D)$, which is again our topic-given-document distribution. Hence, for a given word W in a document D, the probability $Pr(T \mid W, D)$ that it belongs to a topic T is a balance of two components:[6]

- $Pr(W \mid T)$: the significance of the word W to the topic T;
- $Pr(T \mid D)$: the significance of the topic T to the document D.

Note that a word may be highly indicative of a topic (and $Pr(W \mid T)$ is relatively large), but if that topic is not otherwise apparent in the document (and $Pr(T \mid D)$ is relatively small), then this indicates that the word may belong to a different topic.[7] For example, the words `tree`, `branch`, and `pruning` may be highly indicative of a topic related to botany or horticulture, but depending on the other words appearing in the document, these same words could also be indicative of a topic about computer science and data structures.

Many other questions naturally arise when one navigates a topic model.[8] One may encounter an unexpected word appearing in a topic, leading one to ask: *Why does this word appear in this topic?* One may encounter a topic that, unexpectedly, does not appear in a document, leading one to ask: *Why doesn't this topic appear in this document?* Both are (possibly) more challenging questions about the provenance of topics, which we propose as future work.

4 TOPIC DISCOVERY IN WIKIPEDIA

We introduce a new software tool called SPOT, for **S**eeking the **P**rovenance **O**f **T**opics. Our tool SPOT supports several features:

- the collection (scraping) of a text dataset from Wikipedia;
- the learning of a topic model from a text dataset;
- the exploration of the discovered topics; and
- the determination of the provenance of those topics.

Our tool is implemented in python, and has been released open-source.[9] We discuss each of the above aspects in more detail next.

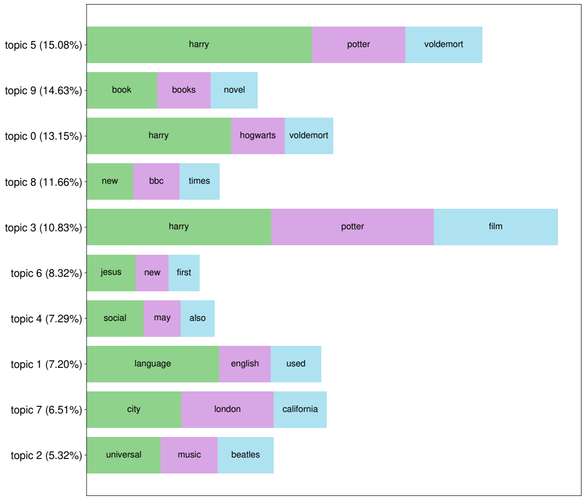


Figure 2: Ten Topics About `Harry Potter`

Collecting Datasets from Wikipedia. Our tool SPOT facilitates the collection of text datasets about any given subject. Suppose that one wants to collect a dataset of articles about Harry Potter. We propose to first download the Wikipedia article `Harry Potter`, and then download all articles *linked to* by the article `Harry Potter`. This comprises a small corpus consisting of articles related to Harry Potter. A dataset on any given subject can be collected in this way. In the case studies in the next section, we collect datasets across different subjects in this manner. Here, we remove all non-alphabetic characters, then filter out stop words, all short words, and all infrequently appearing words.

Learning a Topic Model. Our tool SPOT includes a simple implementation of topic model learning based on a numpy implementation of Expectation-Maximization (EM) using the BN structure of a topic model [2, 11]. In our case studies, we run EM for a fixed number of iterations, using Dirichlet priors with a pseudo-count of 2.

Exploring a Topic Model. Our tool SPOT provides a simple web interface to collect a dataset, and explore a topic model that was trained from it; cf. [5, 13, 16, 19]. Initially, a simple text field is displayed, allowing a user to specify an initial Wikipedia article by name. SPOT downloads the corresponding articles, and stores the result in a local database (so that articles do not need to be re-downloaded).[10] Next, a topic model is learned from the dataset, which provides a summary of the dataset in terms of the topics that appear in it. See, for example, the bar chart of Figure 2 which summarizes the ten topics of a small dataset of articles related to Harry Potter. Each bar represents a topic, with each bar labeled with the top three words of the topic. For example, the most prominent topic consists of the words `harry`, `potter`, and `voldemort`, which pertain to the plot and characters of the Harry Potter books. We will discuss this example in more depth in our case study of Section 5.

[6] Note that if we are computing $Pr(T \mid W, D)$ given word W and document D, which is a distribution over topics T, then $Pr(D)$ and $Pr(W, D)$ are independent of T, and are effectively normalization constants. That is, since $\sum_T Pr(T \mid W, D) = 1$, then $\sum_T Pr(W \mid T)Pr(T \mid D) = Pr(W \mid D)$.

[7] Similarly, [4] uses (a variational approximation of) $Pr(T \mid W, D)$ to label each word in a document by its most likely topic, to visualize the document as a mixture of topics.

[8] It is also common to find two topics in a topic model that should be merged together, or a single topic that should be split into two. Such questions may arise while debugging topic models, and have been addressed by others; see, e.g., [1, 21, 22].

[9] Available at `https://github.com/misha072/topicmodel`.

[10] More specifically, SPOT is a web-based application built on the Flask framework. The wikipedia python module is used to scrape articles from Wikipedia, which are subsequently stored in a SQLite database (using the sqlite3 module). The matplotlib module is used for plotting bar charts, and other visualizations.

Determining the Provenance of Topics. Suppose that a user is considering the summary of a dataset via its topics, and the user spots a topic T where they ask: *Why does this topic appear in this dataset?* To answer this question, our tool spot allows the user to click on topic T, and visualize the collection of the documents most relevant to the topic T, by their probability $Pr(D \mid T)$ (as in Section 3). Intuitively, a topic arises in a topic model if enough documents discuss a common topic, and hence, we want to visualize those documents.

Next, say that a user sees a document D in this list, where they ask: *Why does this topic appear in this document?* To answer this question, spot allows the user to click on the document D, and visualize its text. Here, we highlight each word W based on how relevant it is to the topic T, by the probability $Pr(T \mid W, D)$, as in Section 3. Intuitively, a topic appears in a document if enough words pertain to that topic. Hence, we want to quickly visualize those words, in addition to the context in which those words appear.

In the next two sections, we provide two case studies illustrating how spot provides insights about a dataset, both in terms of visualizing the topics that were learned from a dataset, but also through examining the provenance of those topics.

5 CASE STUDY: HARRY POTTER

Using our tool spot, we collected a small corpus of 573 documents based on the source article `Harry Potter`. Harry Potter is a popular fantasy series of seven novels, which were adapted into eight films. Consider first the ten most common words appearing in the dataset:

1. `harry` 2. `potter` 3. `also` 4. `first` 5. `one`
6. `book` 7. `film` 8. `new` 9. `series` 10. `world`

This list of common words does not provide any insights about the dataset, beyond what little we have said about Harry Potter so far.

In contrast, consider Figure 2 which depicts, using a bar chart, the 10 topics discovered by spot's topic model. This visualization can be viewed as a concise summary of the entire dataset. Each row corresponds to a topic, from most to least prevalent (top to bottom) based on the probability of the topic $Pr(T)$ occurring in the dataset. Each topic is summarized by its top-3 most prevalent words, where the width of the bar indicates the importance of the word, i.e., the probability of the word given the topic, $Pr(W \mid T)$. For example, we find a topic whose top-3 words are `city`, `london` and `california`. These words are indicative of a topic about cities and geography.

For a more in-depth example, consider the most common topic of the dataset, topic 5, whose top-3 words are `harry`, `potter`, and `voldemort`. Consider Table 1, where each selected topic has a table of its top-10 words, along with the probability of each word. We see that topic 5 also includes words like `hogwarts`, `dumbledore`, and `hermione`. These words suggest a topic about the plot and characters of the Harry Potter series. Going more deeply, we can enumerate the most relevant documents relating to this topic, in Table 2 (left), based on the probability $Pr(D \mid T)$, as in Section 3. We find that the ten most relevant documents are all articles relating to the plot and characters of the Harry Potter series. Table 2 provides more support for our interpretation of the topic.

As another example consider topic 8, whose top-3 words are `new`, `bbc`, and `times`. These words likely refer to the New York Times and the British Broadcasting Corporation. In Table 1, we see that topic 8 also includes words like `hbo` and `playstation`. Together, this indicates a topic about news and media. This interpretation is further supported by the list of most relevant documents, in Table 2 (center), which mentions different news organizations, video game publishers, etc. Arguably, this is a topic whose appearance may not be the first that one would predict. However, its presence makes sense when one considers the popularity of Harry Potter across multiple forms of media, which manifests in its appearance in best-seller lists (NYT), as well as in TV adaptations (HBO), and video game adaptations (Playstation, Electronic Arts).

Next, consider topic 6 whose top-3 words are `jesus`, `new`, and `first`. This topic also includes the word `bible` in the top-10. Words like `jesus` and `bible` suggest a topic on religion or Christianity, which may seem incongruous with Harry Potter, which is a fantasy series about witches, wizards and magic. One may ask: *Why does this topic appear in this dataset?* As in Section 3, we consider the documents from the dataset that were most relevant to this topic, given in Table 2 (right). We find multiple articles on religion appearing, including `Jesus Christ`, `Christian Bible`, `King James Version`, and `Resurrection of Jesus`. Articles such as `Anglo-Saxon` and `Beowulf` also have religious contexts. Each of these articles was linked to by the source `Harry Potter` article, which explains why religion appears as a topic in this dataset.

However, this now begs the question *Why does this topic appear in this document?* That is, why does this topic on Christianity appear in the original `Harry Potter` article? Consider Figure 3, which highlights a passage of raw text from the source `Harry Potter` article (after stop words were removed). In particular, words are highlighted based on how relevant they are to the religion and Christianity topic, as in Section 3; the more yellow a word is, the less relevant it is to the topic, and the more blue a word is, the more relevant it is. Based on this visualization, one can see that the source `Harry Potter` article includes a discussion on allusions that Harry Potter makes to Arthurian legend, but also to Christian allegories, including resurrection. This visualization helps explain why a topic on religion and Christianity appears in the `Harry Potter` article.

6 CASE STUDY: 1984

Again, using our tool spot, we collected a small corpus of 447 documents based on the source article `Nineteen Eighty-Four`. 1984 is a dystopian novel written by George Orwell, published in 1949. More specifically, 1984 imagines a future society that is subject to perpetual war and governed by a totalitarian state. The novel explores the conditions on a society that would allow totalitarian control of it to exist, as well as the consequences of those conditions.

Consider the bar chart of Figure 4, which summarizes the ten topics discovered by spot's topic model, from most prevalent (top) to least (bottom). Table 3 takes a closer look at three of the ten topics, and Table 4 considers the documents most relevant to them. Consider first topic 3, on the left of Table 3. This topic includes words like `government`, `political`, `party`, `world`, and `war`, which suggest a topic about governments, politics, and war. This interpretation is supported by Table 4, which lists the documents most relevant to this topic, which mostly relate to governments and war.

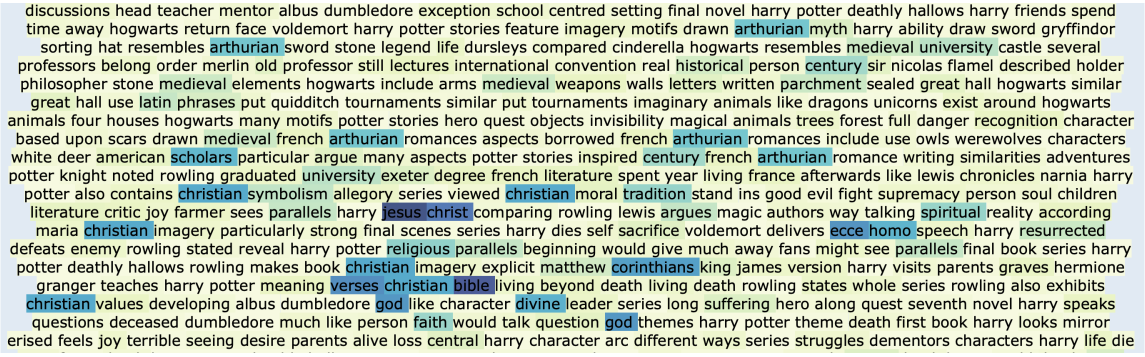
Consider next topic 9, which includes the words `soviet`, `union`, `communism`, `stalin`, and `lenin`. These words suggest a topic about the Soviet Union and communism. If one is not familiar with the

Table 2: Most Relevant Documents for Selected Topics About `Harry Potter`

topic 5	
Document d	$Pr(d\|t)$
Basilisk (Harry Potter)	4.4117
House-elves	4.4116
Dementors	4.4115
Magical creatures in Harry Potter	4.4114
Rose Granger-Weasley	3.5293
James Sirius Potter	3.5293
James Potter (character)	2.7000
Order of the Phoenix (organisation)	2.7000
Mad-Eye Moody	2.7000
Lily Potter	2.7000

topic 8	
Document d	$Pr(d\|t)$
BBC	4.9900
The New York Times	3.9625
HBO Max	3.7129
Max (streaming service)	3.7129
PlayStation	3.7001
The Guardian	3.5359
E-book	2.7277
Who Wants to Be a Millionaire?	2.6921
BBC News	2.6869
Electronic Arts	2.6112

topic 6	
Document d	$Pr(d\|t)$
Anglo-Saxon	7.7156
Jesus Christ	6.8127
Christian Bible	6.6076
King James Version	5.0929
Resurrection of Jesus	4.2745
British pound	4.2532
Rudyard Kipling	4.0399
Beowulf	3.4390
Abolitionist	3.3103
Dana Gioia	3.0089



discussions head teacher mentor albus dumbledore exception school centred setting final novel harry potter deathly hallows harry friends spend
time away hogwarts return face voldemort harry potter stories feature imagery motifs drawn arthurian myth harry ability draw sword gryffindor
sorting hat resembles arthurian sword stone legend life dursleys compared cinderella hogwarts resembles medieval university castle several
professors belong order merlin old professor still lectures international convention real historical person century sir nicolas flamel described holder
philosopher stone medieval elements hogwarts include arms medieval weapons walls letters written parchment sealed great hall hogwarts similar
great hall use latin phrases put quidditch tournaments similar put tournaments imaginary animals like dragons unicorns exist around hogwarts
animals four houses hogwarts many motifs potter stories hero quest objects invisibility magical animals trees forest full danger recognition character
based upon scars drawn medieval french arthurian romances aspects borrowed french arthurian romances include use owls werewolves characters
white deer american scholars particular argue many aspects potter stories inspired century french arthurian romance writing similarities adventures
potter knight noted rowling graduated university exeter degree french literature spent year living france afterwards like lewis chronicles narnia harry
potter also contains christian symbolism allegory series viewed christian moral tradition stand ins good evil fight supremacy person soul children
literature critic joy farmer sees parallels harry jesus christ comparing rowling lewis argues magic authors way talking spiritual reality according
maria christian imagery particularly strong final scenes series harry dies self sacrifice voldemort delivers ecce homo speech harry resurrected
defeats enemy rowling stated reveal harry potter religious parallels beginning would give much away fans might see parallels final book series harry
potter deathly hallows rowling makes book christian imagery explicit matthew corinthians king james version harry visits parents graves hermione
granger teaches harry potter meaning verses christian bible living beyond death living death rowling states whole series rowling also exhibits
christian values developing albus dumbledore god like character divine leader series long suffering hero along quest seventh novel harry speaks
questions deceased dumbledore much like person faith would talk question god themes harry potter theme death first book harry looks mirror
erised feels joy terrible seeing desire parents alive loss central harry character arc different ways series struggles dementors characters harry life die

Figure 3: Visualizing Why Religion Appears as a Topic in the `Harry Potter` **Article**

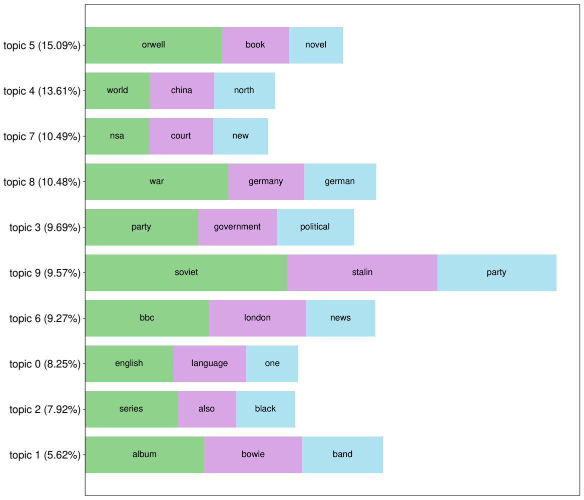


Figure 4: Ten Topics About `Nineteen Eighty-Four`

context in which the novel 1984 was written, then this might be considered an unexpected topic. In particular, the authoritarian state of 1984 is often claimed to be modeled off of the Stalinist government of the Soviet Union at that time (1984 was published in 1949). Considering the documents most relevant to this topic, in Table 4, we see many articles in the dataset relating to communism.

Table 3: Three Selected Topics About `Nineteen Eighty-Four`

topic 3	
w	$Pr(w)$
party	1.2492
government	0.8662
political	0.8503
war	0.7478
state	0.6643
world	0.6436
states	0.6206
democratic	0.6005
united	0.5991
new	0.5304

topic 9	
w	$Pr(w)$
soviet	2.2362
stalin	1.6588
party	1.3118
trotsky	0.9571
union	0.8600
russian	0.7580
communist	0.6468
lenin	0.6428
war	0.5880
revolution	0.5540

topic 7	
w	$Pr(w)$
nsa	0.7076
court	0.7062
new	0.6074
surveillance	0.6039
times	0.5261
states	0.4578
also	0.4535
library	0.4507
united	0.4452
public	0.4178

Finally, consider topic 7, which includes the word `surveillance`, which by itself is expected as it is a central theme of the novel. However, the topic also prominently features words like `united`, `states`, and `nsa` (as in the US National Security Agency). If one assumes that 1984 was based on the Soviet Union, then one may not necessarily expect a topic about the United States in this context. Again, we ask: *Why does this topic appear in this dataset?* Looking at the documents most relevant to this topic in Table 4, we see a number of articles dealing with surveillance and the US government, with the most prominent article dealing with the Snowden surveillance disclosure. Again, all of these articles were linked to by the source article `Nineteen Eighty-Four`. This leads to the question: *Why*

Table 4: Most Relevant Documents for Selected Topics About Nineteen Eighty-Four

topic 3	
Document d	$Pr(d \mid t)$
Stalinist Poland	6.8401
Politics of the United Kingdom	6.8243
Clement Attlee	5.5412
New World Order (conspiracy theory)	4.6025
Democratic socialism	3.8272
Democratic socialist	3.8272
British Commonwealth	3.7148
Authoritarian state	3.1714
Nuclear weapon	3.1050
Cold War	2.9694

topic 9	
Document d	$Pr(d \mid t)$
Leon Trotsky	12.2500
Joseph Stalin	11.0682
CPSU	7.2072
Soviet Union	7.1413
Great Purges	5.4070
Trotskyism	4.1028
Stalinism	3.6651
Lavrentiy Beria	3.3849
October revolution	3.3439
Stalinist Poland	3.2375

topic 7	
Document d	$Pr(d \mid t)$
2013 mass surveillance scandal	7.8376
Elevator	7.3736
US Supreme Court	7.2041
NSA	6.7517
Surveillance	4.5373
Mass surveillance	4.5003
The New York Times	4.1079
The Guardian	3.3904
Global surveillance	3.2680
Internet Archive	3.1969

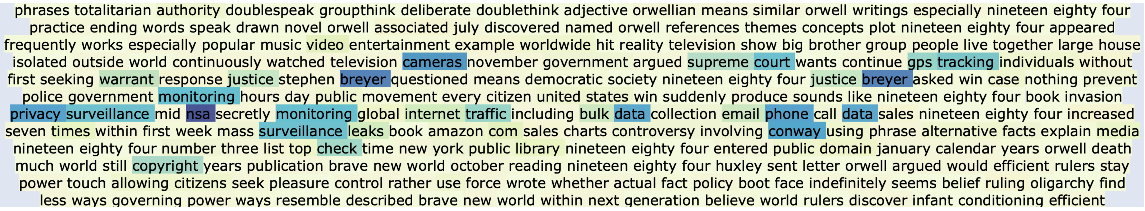

phrases totalitarian authority doublespeak groupthink deliberate doublethink adjective orwellian means similar orwell writings especially nineteen eighty four practice ending words speak drawn novel orwell associated july discovered named orwell references themes concepts plot nineteen eighty four appeared frequently works especially popular music video entertainment example worldwide hit reality television show big brother group people live together large house isolated outside world continuously watched television cameras november government argued supreme court wants continue gps tracking individuals without first seeking warrant response justice stephen breyer questioned means democratic society nineteen eighty four justice breyer asked win case nothing prevent police government monitoring hours day public movement every citizen united states win suddenly produce sounds like nineteen eighty four book invasion privacy surveillance mid nsa secretly monitoring global internet traffic including bulk data collection email phone call data sales nineteen eighty four increased seven times within first week mass surveillance leaks book amazon com sales charts controversy involving conway using phrase alternative facts explain media nineteen eighty four number three list top check time new york public library nineteen eighty four entered public domain january calendar years orwell death much world still copyright years publication brave new world october reading nineteen eighty four huxley sent letter orwell argued would efficient rulers stay power touch allowing citizens seek pleasure control rather use force wrote whether actual fact policy boot face indefinitely seems belief ruling oligarchy find less ways governing power ways resemble described brave new world within next generation believe world rulers discover infant conditioning efficient

Figure 5: Visualizing Why Surveillance Appears as a Topic in the Nineteen Eighty-Four Article

does this topic appear in this document? Figure 5 depicts a passage from the source article, where we have highlighted words that are most relevant to this topic. One of the cultural impacts of 1984 is its association with mass surveillance and authoritarian government (e.g., 1984 introduced the concept of Big Brother). This passage highlighting several instances where US government surveillance policy has been compared with the one of 1984.

7 CONCLUSION

In this paper, we sought the provenance of topics discovered by topic models. We considered two types of questions: *Why does this topic appear in this dataset?* and *Why does this topic appear in this document?* We proposed an answer to each question, based on the underlying probabilistic semantics of a topic model, which points us to the parts of the data that are most relevant to the emergence of a topic. We further introduced a new tool for visualizing topic models called SPOT. which we utilized to analyze two datasets collected from Wikipedia, one based on Harry Potter and another based on 1984. We showed how to explain the presence of unexpected topics, which provides insights about a dataset, and trust in the topics that were learned from it. For future work, we posed more challenging questions related to the provenance of topics.

REFERENCES

[1] David Andrzejewski, Xiaojin Zhu, and Mark Craven. 2009. Incorporating Domain Knowledge into Topic Modeling via Dirichlet Forest Priors. In *International Conference on Machine Learning (ICML)*. Quebec, Canada, 25–32.

[2] Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee-Whye Teh. 2009. On Smoothing and Inference for Topic Models. In *Conference on Uncertainty in Artificial Intelligence (UAI)*. Quebec, Canada, 27–34.

[3] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. 2010. How to Explain Individual Classification Decisions. *Journal of Machine Learning Research (JMLR)* 11 (2010), 1803–1831.

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research (JMLR)* 3 (2003), 993–1022.

[5] Allison J.B. Chaney and David M. Blei. 2012. Visualizing Topic Models. In *International Conference on Weblogs and Social Media (ICWSM)*. Dublin, Ireland.

[6] James Cheney, Laura Chiticariu, and Wang-Chiew Tan. 2009. Provenance in Databases: Why, How, and Where. *Foundations and Trends in Databases* 1, 4 (2009), 379–474.

[7] Adnan Darwiche. 2009. *Modeling and Reasoning with Bayesian Networks*. Cambridge University Press.

[8] Adnan Darwiche. 2010. Bayesian Networks. *Commun. ACM* 53, 12 (2010), 80–90.

[9] Thomas L. Griffiths and Mark Steyvers. 2004. Finding Scientific Topics. *Proceedings of the National Academy of Sciences (PNAS)* 101, suppl. 1 (2004), 5228–5235.

[10] Matthew D. Hoffman, David M. Blei, and Francis R. Bach. 2010. Online Learning for Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems 14 (NIPS)*. Vancouver, Canada, 856–864.

[11] Thomas Hofmann. 1999. Probabilistic Latent Semantic Analysis. In *Conference on Uncertainty in Artificial Intelligence (UAI)*. Stockholm, Sweden, 289–296.

[12] Zachary C. Lipton. 2018. The Mythos of Model Interpretability. *Commun. ACM* 61, 10 (2018), 36–43.

[13] Jaimie Murdock and Colin Allen. 2015. Visualization Techniques for Topic Model Checking. In *Conference on Artificial Intelligence (AAAI)*. Austin, USA, 4284–4285.

[14] Kevin Patrick Murphy. 2012. *Machine Learning: A Probabilistic Perspective*. MIT Press.

[15] Judea Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

[16] Steffen Pielström, Severin Simmler, Thorsten Vitt, and Fotis Jannidis. 2018. A Graphical User Interface for LDA Topic Modeling. In *Digital Humanities Conference (DH)*. Mexico City, Mexico, 651–652.

[17] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Conference on Knowledge Discovery and Data Mining (KDD)*. San Francisco, USA, 1135–1144.

[18] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-Precision Model-Agnostic Explanations. In *Conference on Artificial Intelligence (AAAI)*. New Orleans, USA, 1527–1535.

[19] Severin Simmler, Thorsten Vitt, and Steffen Pielström. 2019. Topic Modeling with Interactive Visualizations in a GUI Tool. In *Digital Humanities Conference (DH)*. Utrecht, Netherlands.

[20] Mark Steyvers and Tom Griffiths. 2007. Probabilistic Topic Models. *Handbook of Latent Semantic Analysis* 427, 7 (2007), 424–440.

[21] Kiri Wagstaff, Claire Cardie, Seth Rogers, and Stefan Schrödl. 2001. Constrained k-means Clustering with Background Knowledge. In *International Conference on Machine Learning (ICML)*. Williamstown, USA, 577–584.

[22] Tiansheng Yao, Arthur Choi, and Adnan Darwiche. 2017. Learning Bayesian Network Parameters under Equivalence Constraints. *Artificial Intelligence Journal (AIJ)* 244 (2017), 239–257.