Mining the Dispatch under Supervision: Using Casualty Counts to Guide Topics from the Richmond Daily Dispatch Corpus

Thomas Templeton
School of Information
University of Maryland
thomas.c.templeton@gmail.com

Travis Brown
Institute for Technology in the Humanities
University of Maryland
trbrown@umd.edu

Sayan Bhattacharyya
School of Information
University of Michigan
bhattach@umich.edu

Jordan Boyd-Graber
School of Information
University of Maryland
jbg@umiacs.umd.edu

Abstract

Large digitized text collections are of immense potential value to historians but are notoriously difficult to digest, given the near-impossibility of reading the entirety of their content within a reasonable amount of time. Making sense of such collections on the basis of searching with keywords is usually inadequate because it is often hard to know beforehand what the appropriate keywords ought to be. However, while large corpora present these challenges for close reading, digital techniques promise new avenues for engaging with text through "distant reading" (Moretti, 2005). On recent interest has been topic modeling, as exemplified by Latent Dirichlet Allocation (LDA) Blei et al. (2003), which has recently been applied to civil war texts Nelson (2010). Topic modeling discovers coherent topic threads that wind through large corpora, enabling readers to discover connections and patterns would otherwise be lost to the data's volume.

While LDA allows researchers to explore corpora in an undirected fashion, it does not enable *directed* research to focus on specific themes or issues. In this work, we employ supervised Latent Dirichlet Allocation (SLDA), which allows us to discover cross-cutting topics, like LDA, but with respect to a particular subject of interest. In this work, we explore using casualty figures allowing us to explore texts viewed through the lens of military sucess (or failure).

1 Introduction

Large digitized text collections are of immense potential value to historians but are notoriously difficult to digest, given the near-impossibility of reading the entirety of their content within a reasonable amount of time. Making sense of such collections on the basis of searching with keywords is usually inadequate because it is often hard to know beforehand what the appropriate keywords ought to be. However, while large corpora present these challenges for close reading, digital techniques promise new avenues for engaging with text through "distant reading" (Moretti, 2005).

Topic modeling furnishes researchers with a powerful approach for moving between close and distant reading. The technique generates a set of topics for a corpus, content-based descriptions of broad subjects that can be followed through time or compared between portions of the corpus. Projects such as the University of Richmond's *Mining the Dispatch* (Nelson, 2010) and the Woodchipper application developed at the Maryland Institute for Technology in the Humanities (Brown, 2011) use topic modeling to represent the structure of large text corpora in ways that are visual and intuitive, allowing the users to identify patterns that they might otherwise miss.

Text processing projects in the humanities that employ topic modeling prefer plain latent Dirichlet allocation (LDA) Blei et al. (2003) algorithm. LDA, however, does not allow a researcher to bias the algorithm towards discovering topics to specific kinds of content — nor does it allow relevant information from outside the corpus to inform the generation of topics. A modification of the basic LDA algorithm called supervised latent Dirichlet allocation (SLDA) addresses the above two problems. SLDA encourages the formation of topics that are good predictors

of a response variable (also sometimes called an observation variable). This ends up guiding the topics discovered by the algorithm towards the specific area of historical interest that the response-variable represents.

In this paper we use SLDA on a corpus composed of Confederate newspaper articles from the Richmond Daily Dispatch. To guide the model to discover interesting wartime topics, we use Confederate casualty counts as our response variable.

2 Background

Classic LDA topic modeling assumes that a corpus is produced through the following generative process:

1. For each of K topics
   (a) Choose a distribution over words $\phi_k \sim \text{Dir}(\lambda)$
2. For each document d in the corpus
3. Choose a distribution over topics $\theta_d \sim \text{Dir}(\alpha)$
4. For each word n in the document
   (a) choose a topic $z_{d,n}$ from $\text{Mult}(\theta_d)$
   (b) choose a word $w_{d,n} \sim \text{Mult}(\phi_{z_{d,n}})$

Based on this assumption, posterior inference seeks to discover the values of the latent variables that best explain how the data — under this assumed model — came to be. These latent variables represent the words associated with each topic ϕ_k, the topics associated with each document θ_d, and the per word topic assignments $z_{d,n}$. LDA's best guess at the underlying structure of the corpus lays the groundwork for the researcher's hermeneutic work. Distant reading happens at two levels: topics are interpreted and identified based on the lexical items that compose them, and patterns at the corpus level are identified based on changes or differences in topical composition. Additionally, topic modeling supports discovery for close reading - researchers can potentially pursue interesting topics by drilling down to documents that strongly evince them.

Applications of topic modeling to historical data sets have tended to emphasize changes in topical composition over time. At least three such projects have focused on historical newspapers. Newman and Block's work with the Pennsylvania Gazette pioneered time-based topical analysis of historical newspapers (Newman and Block, 2006). Nelson's Mining the Dispatch project continued in the same tradition, applying topic modeling to a prominent Civil War era newspaper (Nelson, 2010). Most recently, a team led by Tze-I Yang applied topic modeling to the content of Texas newspapers from 1829 to 2008 (Yang et al., 2011).

The results of humanities topic modeling projects often confirm established research findings (as in Newman and Block's work) or suggest interesting and counter-intuitive research directions, as in Cameron Blevins' research on Martha Ballard's diary (Blevins, 2010) and Yang et al's findings on the Spanish-American War (Yang et al., 2011). However, uncovering such findings using LDA is a bit like beachcombing — the hope is to find *something* cool, though exactly *what* is unspecified. This is frequently framed as an advantage, and LDA does indeed allow the researcher to escape from preconceived historical categories. As Newman and Block put it, "Because there is no a priori designation of topics — in fact there are very few "knobs to turn" in the method — historians do not need to rely on fallible human indexing or their own preconceived identification of topics" (Newman and Block, 2006, p. 766). But the flip side of this advantage is that LDA does not respond nimbly to specific research questions.

Instead, inference captures structure that best fill in the gaps of the LDA's assumptions. With LDA, we can tweak the number of topics as well as several parameters affecting the inner workings of the inference mechanism. While LDA allows researchers to discover patterns of usage in a corpus, however, it cannot explain how the words are affected by, explain, and interact with their broader historical context. A refinement to LDA called Supervised Latent Dirichlet Allocation (SLDA) provides a mechanism for capturing how word usage, as typified by LDA topics, can be connected with this context.

SLDA extends LDA to posit that a per-document observation y_d arises from a normal distribution with mean $\mu \cdot \bar{z_d}$, where $\bar{z_d}$ is the empirical topic distribution of document d. In other words, each document has a numerical value associated with it, and SLDA explains the value of that numerical value by forming a regression (parameterized by the vector μ). Note that this is not equivalent to simply running LDA and then using the resulting topic assignments in a

regression; SLDA's joint inference finds the topics and topic assignment that best explain both the words in a document and each document's associated value.

SLDA has been used to explain sentiment, popularity, and regional variation in dialect (Blei and McAuliffe, 2007; Eisenstein et al., 2010, inter alia). Here, we seek to use SLDA to focus our exploration of the Richmond Daily Dispatch corpus on specific aspects of Civil War history. SLDA focuses topics on a data set of historical observations, encouraging "terms with similar effects on the observation ... to be in the same topic" (Boyd-Graber and Resnik, 2010, p. 48). In our experiments with Confederate casualty counts, every document published in a given week was associated with the casualty count ydfor that week.

SLDA constrains the formation of topics so that the empirical topics assignment $\bar{z_d}$ in a document (the proportion of words in the document associated with each topic) are good predictors of the observation variable. For example, a positive parameter value μ_k in the casualty count experiments indicates a topic in the corpus that is associated with periods of high casualties.

3 Data processing

We used casualty data from from Greer's "Counting Civil War Casualties" (Greer, 2005) as the response variable for SLDA. After filtering a standard set of stopwords from the corpus and transforming all tokens to lower case, we found that modeling against the untransformed casualty counts yielded unreliable results, as SLDA assumes a normally distributed response variable. To remedy this problem, we log-transformed the casualty count data before running SLDA. We ran SLDA on the log-transformed casualty counts using 15, 25, 35, and 45 topics. The results reported here are from the 35 topic run of SLDA, which, in line with earlier work on this and other corpora, seemed to produce the most reasonable topics (Newman and Block, 2006) We ran SLDA for 2500 iterations.

Because we only had casualty information at a week's granularity, we associated the response variable y_d of a document based on the casualty figure associated with the appropriate *week* when the document was published.

4 Results

We ran sets of SLDA experiments on the Richmond Daily Dispatch corpus using casualty counts as the supervision variable. In the following discussion, a selection of results from these experiments illustrates the potential of SLDA for exploring large corpora. Following discussion of the results, future steps for fully taking advantage of the information available through SLDA are outlined.

Supervising LDA with casualty counts should lead terms to coalesce, based on common relationships to casualties, into topics that have some relevance to warfare. Of these topics, one might expect to see some that directly reflect military themes. Examining the parameters associated with these topics should yield insight into the nature of the topics, the corpus, and the war.

We discovered eleven topics (out of thirty five total) which, on inspection, included significant military elements. Of these, seven appeared to have notable explanatory power for casualty counts. For example, the three topics shown in Table 1 — one related to tactical accounts of battle, one containing terms consistent with military storytelling in a lofty register, and one related to reports of individual casualties — all show a relatively strong positive relationship with casualties.

On the other hand, a topic related to recruitment showed a negative correlation with casualties, and a topic consisting mostly of administrative units showed relatively insignificant association with casualties (Table 2).

We also found topics which appear to be related to each of three of the war's major theaters. Of these, only the topic related to the Eastern Theater is notably associated with casualties (Table 3). There are a number of possible explanations for this. First, the Eastern theater was by far the most contested and the biggest cause of loss of life; in constrast the other theaters, while strategically important, had smaller armies in the field. Second, other theaters often were more slowly reported (both because of relative importance and distance from Richmond), which might cause a lag in reporting vs. our casualty source.

Similarly, the naval topic and the Fort Sumter topic had relatively insignificant association with casualties (Table 4), as these had political, economic, and

Battle reports	Battle rhetoric	Casualty reports
$\mu = 0.49$	$\mu = 0.55$	$\mu = 0.38$
enemy	army	wounded
men	men	company
wounded	great	john
battle	enemy	capt
col	war	lieut
killed	battle	killed
left	time	regiment
gen	military	1st
regiment	richmond	slightly
artillery	country	privates
position	general	james
field	force	private
fire	field	col
fight	mcclellan	virginia
line	people	arm
brigade	armies	jas
time	troops	severely
cavalry	thousand	4th
back	make	captain
battery	long	smith
force	success	leg
loss	campaign	thomas
day	victory	ala
command	man	thos
general	made	miss

Table 1: These three military topics have relatively high response coefficients, suggesting that the correlate positively with casualties.

Recruitment	Administrative
$\mu = -1.12$	$\mu = 0.16$
company	general
men	gen
regiment	army
companies	major
capt	command
service	officers
volunteers	colonel
virginia	war
county	officer
camp	military
captain	col
col	commanding
richmond	lieutenant
state	brigade
troops	department
city	order
military	service
number	states
dispatch	brigadier
guard	chief
volunteer	lee
yesterday	staff
home	orders
good	commander
officers	president

Table 2: A topic related to recruitment has a negative correlation, and an administrative topic is not significantly correlated.

propaganda implications but not much of an effect on casualties. Intriguingly, the topic shown in Table 5 includes words prominent in reprinted stories from Northern papers (such as "rebel") and shows a very strong association with casualties, suggesting perhaps an increase use of alienaation techniques as casualties rose.

Log-odds analysis can highlight the contrast between thematically similar topics with divergent responses. For example, both topic 13 and topic 26 include terms denoting military units. They differ obviously in that topic 13 includes many words related to the violence of battle and is positively associated with casualties, while topic 26 is negatively associated with casualties. Log-odds is calculated by first taking the ratio of the likelihood of seeing the word in one topic over the likelihood of seeing it in the other. The log of the result is taken to orient the scores around 0, so that words at the positive and negative extremes are the most indicative of a word's strength of presence in one topic over the other. Log-odds analysis suggests that topic 26 is in fact a topic related strongly to recruitment.

So far we've only considered the topic level view of the data we obtained from SLDA. We are currently working with a topic model visualization suite to draw more comprehensive connections between levels of the corpus, to explore the changing influence of topics over time, and to better understand topics by examining documents in which they figure prominently, as Nelson does in the original Mining the Dispatch work. Additionally, in order to better connect this work to established historical knowledge and future historical research, we hope to involve subject domain experts in an interactive interpretation of the results and refinement of the model.

5 Future Work

Topic modeling is a promising tool because it allows a subject domain expert to quickly switch between

Trans-Mississippi	Western	Eastern
$\mu = -0.06$	0.13	$\mu = 0.86$
gen	tennessee	enemy
kentucky	army	yesterday
river	enemy	gen
federal	sherman	cavalry
mississippi	general	army
memphis	railroad	river
vicksburg	east	force
troops	miles	miles
missouri	atlanta	captured
nashville	river	lee
dispatch	chattanooga	morning
tennessee	georgia	night
morgan	north	prisoners
killed	line	general
men	hood	yankee
enemy	gen	yankees
mobile	knoxville	forces
col	south	petersburg
orleans	force	day
jackson	road	richmond
captured	bragg	loss
force	point	left
louisville	mountain	troops
miles	lines	received
texas	left	point

Table 3: Our model only associated the Eastern theater with a positive correlation with casualties.

Naval	Fort Sumter
$\mu = -0.04$	$\mu = -0.11$
iron	fort
river	enemy
navy	charleston
guns	island
water	fire
vessels	guns
feet	batteries
hundred	battery
work	city
made	sumter
great	firing
naval	night
ships	day
gun	morning
time	fired
ship	fleet
fleet	clock
land	point
yard	shot
war	shell
men	yesterday
clad	gunboats
miles	shells
twenty	attack
steam	flag

Table 4: These two topics show little association with casualties.

different levels of engagement with a corpus. Visualization tools support movement between document-level, topic-level, and corpus-level views that mutually support the meaning-making process. Using tools like Woodchipper and the Topical Guide (Gardner et al., 2010) and incorporating the insights of domain experts, we hope to continue our case study in relating historical corpora to pieces of the historical record by focusing topics with SLDA.

Other modifications to basic LDA have potential for use in humanities applications as well. Topics over Time uses date-stamps to encourage topics to cluster around a point in time, so that topics are more likely to conform to historical events (Wang and McCallum, 2006). Dynamic Topic Modeling (Blei and Lafferty, 2006) allows topics to evolve from year to year, capturing the intuition that scientific fields, for example, endure despite changing terminology. Finally, Dirichlet Forests allow the modeler to engender affinities or aversions between words based on prior knowledge of the content domain (Andrzejewski et al., 2009). As new tools emerge and become more widely available, topic modeling becomes a more flexible and precise tool for exploring large corpora.

Reprinted Northern Papers
rebel
york
rebels
general
army
gen
washington
men
union
herald
mcclellan
day
thousand
papers
hundred
war
made
united
received
rebellion
order
president
city
lincoln
baltimore

Table 5: A topic containing news reprinted from Northern papers shows a strong association with casualties (2.35477).

Low Casualty		High Casualty	
military	-7.04	prisoners	5.73
richmond	-6.98	division	5.68
city	-6.62	gen	5.65
call	-6.36	advance	5.64
services	-6.26	loss	5.34
militia	-5.97	rear	5.32
recruits	-5.97	captured	5.27
county	-5.94	woods	5.27
counties	-5.64	firing	5.21
drill	-5.61	road	5.14
governor	-5.58	front	4.95
equipped	-5.56	fell	4.93
mustered	-5.55	drove	4.69
state	-5.55	fire	4.63
enlist	-5.53	cut	4.59
enlisted	-5.51	attack	4.54
parade	-5.50	commenced	4.48
petersburg	-5.45	enemy	4.41
volunteer	-5.43	hill	4.39
citizens	-5.39	fall	4.38
raise	-5.32	began	4.37
months	-5.30	longstreet	4.25
esq	-5.24	missing	4.24
april	-5.23	continued	4.23
recruiting	-5.16	lines	4.23

Table 6: We took two military topics that had opposing response coefficients μ and compared probabilities of words within those two topics. This highlights how the topics differ.

Acknowledgments

This work was supported in part by the Maryland Institute for Technology in the Humanities and the Institute of Museum and Library Services.

References

Andrzejewski, David, Xiaojin Zhu, and Mark Craven. "Incorporating domain knowledge into topic modeling via Dirichlet Forest priors." In *Proceedings of International Conference of Machine Learning*. 2009.

Blei, David M., and John D. Lafferty. "Dynamic topic models." In *Proceedings of International Conference of Machine Learning*. New York, NY, USA, 2006.

Blei, David M., and Jon D. McAuliffe. "Supervised topic models." In *Proceedings of Advances in Neural Information Processing Systems*, MIT Press, 2007.

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

Blevins, Cameron. "Topic Modeling Martha Ballard's Diary." `http://historying.org/2010/04/01/topic-modeling-martha-ballards-diary/`, 2010.

Boyd-Graber, Jordan, and Philip Resnik. "Holistic Sentiment Analysis Across Languages: Multilingual Supervised Latent Dirichlet Allocation." In *Proceedings of Emperical Methods in Natural Language Processing*. 2010.

Brown, Travis. "About the Woodchipper: Byron and Austen." `http://mith.umd.edu/corporacamp/tool.php`, 2011.

Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. "A Latent Variable Model

for Geographic Lexical Variation." In *EMNLP'10*. 2010, 1277–1287.

Gardner, Matthew, Joshua Lutes, Jeff Lund, Josh Hansen, Dan Walker, Eric Ringger, and Kevin Seppi. "The Topic Browser: An Interactive Tool for Browsing Topic Models." In *Proceedings of the Workshop on Challenges of Data Visualization, held in conjunction with the 24th Annual Conference on Neural Information Processing Systems (NIPS 2010)*. Whistler, BC, Canada, 2010.

Greer, Darroch. "Counting Civil War Casualties, Week-by-Week, for the Abraham Lincoln Presidential Library and Museum." `http://www.brcweb.com/alplm/BRC_Counting_Casualties.pdf`, 2005.

Moretti, Franco. *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso, 2005.

Nelson, Robert K. "Mining the Dispatch." `http://dsl.richmond.edu/dispatch/`, 2010.

Newman, D., and S. Block. "Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper." *Journal of the American Society of Information Science and Technology* .

Wang, Xuerui, and Andrew McCallum. "Topics over time: a non-Markov continuous-time model of topical trends." In *Knowledge Discovery and Data Mining*. 2006, Knowledge Discovery and Data Mining. `http://doi.acm.org/10.1145/1150402.1150450`.

Yang, Tze-I, Andrew Torget, and Rada Mihalcea. "Topic Modeling on Historical Newspapers." In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Portland, OR, USA: Association for Computational Linguistics, 2011, 96–104. `http://www.aclweb.org/anthology/W11-1513`.