Deconstructing Libraries: A Dual Approach to Historical Research and Machine Learning

May 7, 2016

Jordan Shedlock
UC Berkeley, School of Information
jshedlock@ischool.berkeley.edu

Cindy A. Nguyen
UC Berkeley, History Department
cindyanguyen@berkeley.edu

Abstract

This project focuses on a complex non-English language historical data source-- bibliographies of the United States Library of Congress collections of Vietnamese language materials from 1979-1985. We employ a dual approach to this project: a contextualized historical reading and machine learning methods to understand the language, political leanings, and content of the collections and how they change over time.

1 Introduction

Within recent years, a body of important scholarship has developed which has integrated statistical and machine learning methods to analyze complex humanistic questions. While there have been significant work advanced in the fields of computational journalism (Sarah Cohen et. al., Sylvain Parasie) and literary studies (Ted Underwood, Andrew Goldstone, Franco Moretti and the Stanford Literary Lab), few historians have integrated statistical methods into their work. Furthermore, most of this scholarship has drawn from data of convenience-- well-structured data often in the English language.

This project is a proof of concept for computational methods in Cindy Nguyen's Ph.D. dissertation in history at UC Berkeley titled "Creating the Vietnamese Library: Builders and Users of Libraries in Modern Vietnam." Nguyen examines the cultural and political history of libraries in Vietnam from 1887 to 1986.[1] For the dissertation, computational methods and digital humanities are imperative to understand the intellectual and cultural history of book holdings, circulation, and readership. Furthermore, studies of non-Vietnamese library collections reveals the epistemology of 'Vietnam' by other curatorial regimes, such as the United States Library of Congress in the immediate post-Vietnam War period.

2 Hypotheses

Focusing on a small slice of the research questions, we examine the relationship between the titles and their city of publication within the Library of Congress collections of Vietnamese language materials retrospectively collected up to 1979 and 1979-1985. We hypothesize that different publication locations will have different distribution of topics. Subsequent hypotheses include the following:

H1: Hanoi will have more topics on Communism, war, revolution, and army than Saigon.
H2: Saigon will have more topics on US ideas (modernity, democracy, anti-Communism) than Hanoi.
H3: LOC collection will prefer Saigon (United States ally) materials over Hanoi.

To operationalize our hypothesis into a probabilistic problem, we ask the following question: What are the words in a title most characteristic of a publication city? Our hypotheses rest upon the important underlying assumption that a publication title can reveal important information about a work's content, potential audience, and literary style. Although not a substitute for reading the entirety of the work, a work's title reveals negotiated information between the author, the perceived audience, the publisher, and in this case, the library collection. Some of the titles were created by political or military entities such as the training pamphlets for North Vietnamese

[1] Nguyen's project consists of three parts: (1) a theoretical investigation into the library as an institution of knowledge and power; (2) a cultural history of reading practices and communities in Vietnamese libraries and (3) a focused social history of five libraries—the two national libraries in Hanoi and Saigon, the publishing project Bibliothèque de Vulgarisation of F.H. Schneider and Nguyễn Văn Vĩnh, and the association libraries of Société d'enseignement mutuel de Cochinchine, and Parti révolutionnaire du jeune Annam.

army and were intended to circulate within a specific community. Other titles were language textbooks or volumes of collected literary works and were intended for reference and wider circulation. Other titles might include English translations due to its ascension into the Library of Congress collection.

3 Related Work

This project has tremendous contributions to the important fields of history and digital humanities.[2] Furthermore this project applies machine learning, data science, and natural language processing for the "long tail" of complex, uncommon, and difficult Vietnamese historical data.[3] This topic also offers insight into contemporary Vietnamese library institutions, reading practices, and preservation approaches. We will work with and share findings with the ÉFEO, National Library in Hanoi, and General Sciences Library in HCMC in hopes of bringing attention to the long, important, but understudied history and heritage of libraries in Vietnam.

4 Data

The data consists of the official bibliography of the non-legal Vietnamese-language collection in the Asian Division at the Library of Congress (LOC) in 1982 and 1987 digitized by HathiTrust.[4] With the end of the Vietnam War in 1975, there was a rising demand from Vietnamese refugees and American scholars (barred from field research in Vietnam) to access Vietnamese-language sources. This demand led to the creation of the 1982 and 1987 bibliographic guides (compiled by A. Kohar Rony, Area Specialist of the Southern Asia Section, Asian Division) analyzed in our project. The Vietnamese-language collections includes both retrospective materials and newer items published in the Socialist Republic of Vietnam (post-1975 unified Vietnam). However, the Library of Congress has not disclosed their collection decisions for this collection.

The 1982 bibliography includes Vietnamese language materials the LOC collected up to June 1979. The 1987 bibliography includes Vietnamese language materials collected by the LOC from 1979 to 1985. The bibliographies consisted of metadata such as Author, Title, Publisher, Publishing Location, Year of Publication, Library of Congress Classification (LCC). For the purpose of our study, we only examine non-periodical works.

Data clean up included the following stages:

- Optical Character Recognition (OCR) with Abbyy Finereader
- Regular Expressions to extract author, title, publisher, publisher location, year of publication, and Library of Congress Classification (LCC)
- Google Refine to clean up data fields, resolve OCR errors on Vietnamese language diacritics
- vnTokenizer[5] to tokenize Vietnamese words, for example cách mạng = cách-mạng [revolution]

Final output: Tab separated text files formatted in UTF-16, lower-case words.

This stage of data cleanup involved a significant amount of time and manual labor. Of particular difficulty was the removal of noise from the OCR of the HathiTrust digitization. Despite their relatively standardized format, the records were closer to natural language than to structured data, and it was not always possible to differentiate between, for example, authors and titles. Even with manual clean up via Google Refine, the accuracy of OCR of Vietnamese diacritics could result in multiple spellings/misspellings for the same word. Furthermore, OCR accuracy could limit the tokenization of Vietnamese words. Given the limited time and resources, we proceeded with

[2] For a list of related work in the fields of history and digital humanities, see Appendix A "Related Work".

[3] David Bamman, "NLP for the Long Tail." Keynote at the Berkeley Digital Humanities Summer Institute. 21 August 2015. http://digitalhumanities.berkeley.edu/summer-institute-2015/david-bamman

[4] Library of Congress., and A. Kohar Rony. *Vietnamese Holdings in the Library of Congress: A Bibliography*. Washington: The Library : For sale by the Supt. of Docs., U.S. G.P.O., 1982. //catalog.hathitrust.org/Record/000109970. *Vietnamese Holdings in the Library of Congress. Supplement, 1979-1985*. Washington: The Library : For sale by the Supt. of Docs., U.S. G.P.O., 1987. //catalog.hathitrust.org/Record/002577324.

[5] "vnTokenizer -- Vietnamese Word Segmentation | Lê Hồng Phương." Accessed May 4, 2016. http://mim.hus.vnu.edu.vn/phuonglh/softwares/vnTokenizer. vntokenizer segments Vietnamese into lexical units with an accuracy of 98% on a test set extracted from the Vietnamese treebank.

the best possible version of cleaned and structured data.

From our extracted and structured data we decided to focus on Titles, Location of Publication, Year of Publication, and Library of Congress Classification (LCC).

4.1 Validity - Data and Methods

Sampling Validity: One of the most important aspects that contribute to the validity of this project is the question of Library of Congress collection subjectivity. We recognize that the collections of the LOC are subject to financial limitations, political preferences, and logistical constraints that make it impossible for the LOC to collect *all* that is published in Vietnamese between the years 1979 and 1985. Thus, in order to conform with sampling validity, we focused our research questions to consider patterns within the content of material collected, rather than general publishing patterns in Vietnamese globally and historically.

Another challenge is the quality of our data. The accuracy of the OCR to capture Vietnamese diacritics and correct spelling influences the Regex and tokenization of Vietnamese words. We used vnTokenizer, which has a high accuracy of 98% on a test set extracted from the Vietnamese treebank. However, our data comprised of lower-case Vietnamese titles with possible OCR diacritics errors. This led to a significantly lower accuracy for tokenization.

Semantic Validity: Furthermore, in order to maintain semantic validity, we conduct all of our research and modelling in the original language of Vietnamese. We provide English language translations for the purpose of communicating to a larger audience.

5 Method

5.1 Frequency Counts

The first steps we took with our newly-structured data were basic frequency counts to get a general sense for the numbers and proportions involved. In total, we had 4417 records from the two bibliographies. Of these, 2016 were from Hanoi, 1150 from Saigon (we included those labeled "Ho Chi Minh City" as part of the Saigon label), and 1252 from other origins. The "others" included other cities within and outside of Vietnam, records with no city specified (omitted or labeled 's.l.'), and records where our regular expressions failed to capture the city (null values, or incorrect ones such as years or LCC numbers. Of these, 624 had no value for "city" whatsoever (we used these later for validation). Frequency counts challenge our first hypothesis, that LOC collection policy would be biased towards collecting more material from Saigon, its political ally. In fact, even after the end of the Vietnam War, the LOC vigorously collected overtly Communist political and ideological texts from Hanoi, including such titles as *Understanding American Studies of Marx-Lenin* (1979) and *Beloved Uncle Ho* (1970).

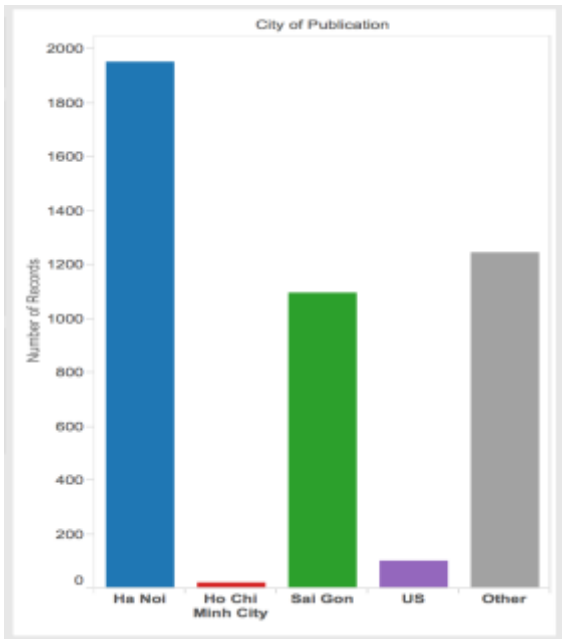

Figure 1. Distribution of records by city of publication.

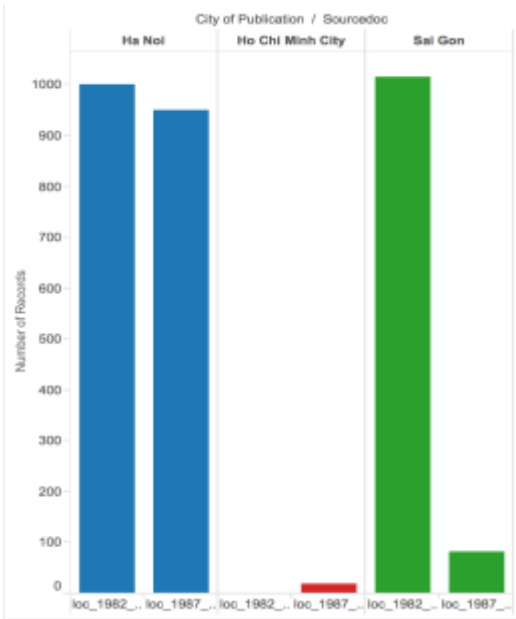

Figure 2. Broken down by year
by city of publication. (left column - 1982, right - 1987)

This observation of a higher collection of materials from Hanoi is seen even more clearly if we disaggregate our data by the 1982 and 1987 bibliographies. In 1982, 1028 records were from Hanoi, 1039 were from Saigon and Ho Chi Minh City. In 1987, 988 records were from Hanoi, while only 111 records were from Saigon and Ho Chi Minh City (figure 2).

Next, we looked at where the titles collected fell in the Library of Congress Classification, which sorts books by topic. This was related to our first two hypotheses: that materials collected from Hanoi would be slanted towards typically North Vietnamese Communist topics such as revolution, war, and Communist ideology, while those from Saigon would be more focused on US- and Western-friendly ideas such as modernity, democracy, and anti-Communism. Due to time and resource constraints, we limited our examination to the 21 top-level categories of the LCC and their immediate children (represented by the initial one or two letters in the classification number). We looked at three groups of records: those from Hanoi, those from Saigon, and diaspora publications (identified with a city outside of Vietnam).

Again, the numbers did not bear out our expectations. Across all 3 groups, language and literature (PL) was the most numerous category by a wide margin (43% of Hanoi records, 38% from Saigon, and 67% of diaspora records) with Asian history a strong second (15% from Hanoi and 17% from Saigon; with 6% of diaspora materials, it was less dominant in that group, but still twice as numerous as the next category). Only in the less numerous categories is any difference apparent; however, these differences also defy easy explanations of political bias. For example, both works published in diaspora communities and in Hanoi have a significant number of works on Buddhism and psychology, while political science and political institutions are one of the highest-ranking topics in both Hanoi and Saigon. Literature and juvenile fiction (Saigon, diaspora), and industries, land use, and labor (Saigon) round out the top five categories. However, it is worth emphasizing that all topics besides language and literature and Asian history occupy single-digit percentages of these groups.

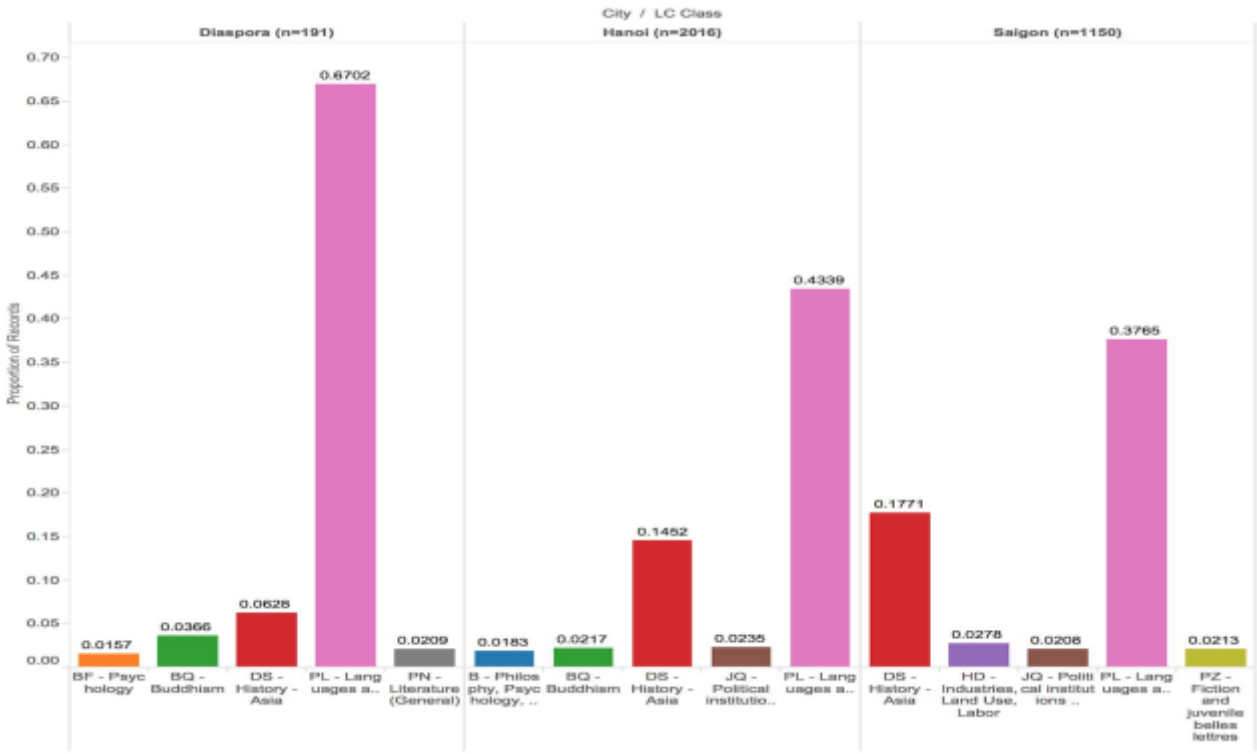

Figure 3. Records from Hanoi (center), Saigon (right) and diaspora cities (left) broken down by LCC. Pink is language and literature (PL), red is Asian history (DS)

5.2 Frequency Counts - Analysis

During the Vietnam War and in the aftermath, thousands of Vietnamese intellectuals from South Vietnam fled as political refugees or were imprisoned. Furthermore, in the after the war, the Socialist Republic of Vietnam enforced strict intellectual censorship of publications throughout the country. Historians characterize the Vietnamese intellectual culture in Hanoi and Saigon in the aftermath of war as an extension of socialist state building and control.

The low collection of materials from Saigon in the 1987 (101 records compared to 988 from Hanoi) could be due in part to the post-war intellectual, political, and social contexts such as the exile and arrests of many Southern Vietnamese intellectuals, the post-war publishing regime, or the impact of the border conflicts with Cambodia and China in 1979.[6] The sheer difference in number of works collected from Saigon confirm face validity based on the explained historical context above.

We had hypothesized (H3) that US collections would prioritize materials from Saigon (ally) over Hanoi (enemy). This was disproved based on the relatively low number of works collected in the 1987 collection from Saigon as well as the consistent number of total collections between Hanoi and Saigon. The frequency counts of materials by city, and of the distribution of LCC categories for each city, are lacking in face validity, as they do not reveal any meaningful political or cultural divisions between North, South, and diaspora Vietnamese publishing, or any political/regional bias on the part of the LOC. This suggests some limitations in the semantic validity of this approach. First, it appears that LOC collection policy is driven by subject, rather than political affinity, since there is no hesitation to collect books from Hanoi or on Communism. At the same time a historian can argue that US collections would prioritize materials from both Saigon and Hanoi, for issues of national security. US alliance with South Vietnam was a historically uneasy one, with low limits of trust in the administrative capacities of the South Vietnamese state, as well as the perceived Communist threat in the south from the Southern Insurgency (National Liberation Front, or Việt Cộng).

Overall, frequency counts allowed us to refine our initial assumptions regarding the collection strategies of the Library of Congress. Furthermore, this close look at our data also calls into question the ability to compare across the 1982 and 1987 collections, since the small number of Saigon materials implies a dramatically different collecting strategy and possibly publishing pattern. Reading frequency counts offered a useful benchmark entry into the data, but our questions regarding political and linguistic distinctions of collections are not captured.

5.3 Topic Modeling

Returning to our probabilistic and operationable question "What are the words most characteristic of a publication city?" we used Topic Modeling to compare the 'topics' between works published in Hanoi and Saigon. The purpose of topic models here was to provide coarse semantic structure of the texts and to compare between the topics of works published in Hanoi to Saigon. Topic models can allow us to learn patterns of content, language, and themes of works published in each city.

We examined only texts published in Hanoi (2016 records) and Saigon (1150 records including Ho Chi Minh City) from the 1982 and 1987 bibliographies combined. For topic modelling, the documents comprised the titles of texts and we retained the metadata of publication city. We used Mallet (latent Dirichlet allocation), and designated our corpora to be distributed over 100 topics.

From our topic distribution by document output, we calculated the average distribution of a certain topic over all the documents published in Hanoi and the same for Saigon. (This was done by taking the sum of all the probabilities of a topic in a document published in Hanoi, then dividing it by the number of documents published in Hanoi.) For example, what is the average distribution of topic 74 (cách mạng tháng hồi đạo tám lãnh trước mang nghĩa mười thắng vằn lọi hoành chạp lê-nin há-nội khẫu tât - Revolution, month party eight labor ideology Lenin Hanoi) for Hanoi? For Saigon? How does the ranking of the topic for Hanoi or Saigon compare to the rank in the total corpus?

[6] Between 1975 and 1978, the Socialist Republic of Vietnam was involved in isolated border conflicts with Democratic Kampuchea. In 1979, the Socialist Republic of Vietnam fought in a brief border war (also known as the Third Indochina War) with the People's Republic of China. These political military events were significant because they were wars between socialist states and extended the already war-exhausted Socialist Republic of Vietnam into over 30 years of consecutive warfare.

Rank in Total Corpus	Topic Number	Average Topic Distribution	Topic Vietnamese	Topic English
1	96	0.07636	truyện tập ngắn hoa làng lửa lòng ngấn vui bạn rừng kịch đất biển tuyển đèn vàng dâu địch ngoặt	Story volume short village fire spirit joy friend play
2	51	0.06723	văn hóa nghệ văn học thuật trong nội mới thuật vân gia nẽn tích quoc đoán lĩnh kiển rừng tải	Culture trade literature art new nation jungle
3	81	0.0661	nguyễn văn biên phạm ngọc hữu đức mai soạn thị xuân hồng huy thái trần chủ cúa thạch đặng châu	Nguyen write Pham Ngoc Huu (Vietnamese names)
4	7	0.06467	nam việt của miên miển viêt bắc qua phân các thức đường trí quan câu bạc hay dân-tộc quòc vản	South Viet region North nation/people
5	44	0.05612	thơ tập thi văn mới đổng nội phẩm nhà xuân đất đền bằng tày xuôi cầu đơn chân mưa viêt	Poem volume text new house temple spring bridge rain Viet

Table 1. Top five Average topic distribution of total titles published in entire corpus.

Topic Number	Rank of topic in total corpus	Rank of topic in Hanoi	Average Topic Distribution	Topic Vietnamese	Topic English
96	1	1	0.032328068	truyện tập ngắn hoa làng lửa lòng ngấn vui bạn rừng kịch đất biển tuyển đèn vàng dâu địch ngoặt	Story volume short village fire spirit joy friend play
81	3	2	0.025674659	nguyễn văn biên phạm ngọc hữu đức mai soạn thị xuân hồng huy thái trần chủ cúa thạch đặng châu	Nguyen write Pham Ngoc Huu (Vietnamese names)
55	10	3	0.023222841	nghĩa chủ hội dân chú quôc pháp thực hiện lập nghiã độc sản nhân nền lớn quý nin báo một	Meaning democracy nation France realize independence
7	4	4	0.022383684	nam việt của miên miển viêt bắc qua phân các thức đường trí quan câu bạc hay dân-tộc quòc vản	South Viet region North nation/people
51	2	5	0.022079638	văn hóa nghệ văn học thuật trong nội mới thuật vân gia nẽn tích quoc đoán lĩnh kiển rừng tải	Culture trade literature art new nation jungle

Table 2. Top five Average topic distribution of total titles published in Hanoi.

Topic Number	Rank of topic in total corpus	Rank of topic in Saigon	Average Topic Distribution	Topic Vietnamese	Topic English
72	7	1	0.029372025	thứ lần bán lần bản tái xuất lãn xuất xuất xuất lấn khôi vản-học trăng xuất độc lớp văn chiêm-vân-thi	Number published literature
90	22	2	0.026174056	lân thú ban tái xuất van tưa thủ vói tuyên sung điên làn luân hát khi xuất-ban thòi cua sưa	Number published time old/former
66	25	3	0.022560966	truyện dài tré bóng lưng tồi thù tóc huyển sương khuya nhó trách hoang tướng hạt mộng việt mắt ngân	Story long youth shining dew dream Viet eyes
50	20	4	0.022543548	truyện dài trời thiên mặt trường chân bên tím đêm cứa chúa tim phía tháp hòn phiá thu cho thẫy	Long story sky heavens school night gods spirit
7	4	5	0.021051846	nam việt của miên miển viêt bắc qua phân các thức đường trí quan câu bạc hay dân-tộc quòc vản	South Viet region North nation/people

Table 3. Number Top five Average topic distribution of total titles published in Saigon (and Ho Chi Minh City).

To test whether there were any statistically significant differences between the average topics, we ran a Monte Carlo permutation test over 10,000 iterations using the city as the label. For all topics but one, the permuted difference in means exceeded the true difference in every trial. The remaining topic was topic 61 (see table below), in which the permuted difference in means exceeded the true mean in only 36.15% of the trials; while this is a great difference from the other topics, it is still not statistically significant at the 5% significance level. Thus it appears that the titles from the Hanoi and Saigon records reflect the same or similar distribution of topics, based on our model.

việt-nam quôc-gia quốc-gia viêt-nam bibliography vietnamese hành-chánh chính-trị tại du-lich giáo-dục thu-viện thư-viện giáo-duc saigon thông-kê thu-tịch political retrospective
vietnam nation nation vietnam bibliography vietnamese administrative politics because travel education library library education saigon statistic bibliography political retrospective

Table 4. Topic 61 in Vietnamese and English. Note that some of the tokens had OCR errors and thus variations in spelling were counted as different tokens.

5.4 Topic Modeling - Analysis

For our topic model, we designated the title for each publication record as the 'document.' There has been recent research that demonstrates LDA to work well on short and sparse texts such as tweets.[7] At the same time, we realize that the short and variable document size of the titles (such as "Ho Chi Minh" or "Building solidarity among the people and protecting the Socialist country of Vietnam") can influence the topic model.

For this project we wanted to use topic models first to explore the topics of titles and second to analyze if the topics are related to the city of publication. From the average

[7] Jianhua Yin and Jianyong Wang. 2014. A Dirichlet Multinomial Mixture Model-based Approach for Short Text Clustering. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 233–242. Naveed, N., Gottron, T., Kunegis, J., Che Alhadi, A.: Searching microblogs: Coping with sparsity and document quality. In: CIKM'11: Proceedings of 20th ACM Conference on Information and Knowledge Management. pp. 183–188 (2011).

distribution of a certain topic over all the all the documents published in Hanoi and the same for Saigon, we observed several significant aspects. (A reminder that an important assumption for our project is that a title can reveal important information about a work's content, potential audience, and literary style.) First, many of the topics including topics 81 and 7 related to the nature of collected works on Vietnam--that is, information explicitly about 'Vietnam' or certain 'Vietnamese' figures. This was expected, and some of the biases for a topic of Vietnamese names could have been due to OCR/Regex errors that pulled in author names together with the title. Second, we expected many topics to cluster together words about editions, volumes, and translations given the semantic structure of a title. However, what the topic model output shows are certain other words that co-occur with these edition-words such as the spiritual emotions: "spirit, joy, fire, friend" (topic 96) and space/time: "house, temple, spring, rain" (topic 44). This suggests something about the style and content of works that might be multi-volumed and literary. Third, the association of certain topics around democratic politics: "Meaning democracy nation France realize independence" (topic 55) ranked third highest average probability for Hanoi while ranked 29 for Saigon hints at the popularity of ideologically charged 'democratic' language in socialist Hanoi versus anti-Communist Saigon. Fourth, based on the ranking of topics by city, Hanoi texts and topics more characteristic of the topics in the entire corpus than Saigon.

For our second question regarding the relationship between topics of titles and city of publication, we used a permutation test to see if the difference of topic distribution between Hanoi and Saigon was due to chance or was due to the condition of publication city. Our permutation test reveals that the city of publication Hanoi or Saigon is not statistically significant to document topics.

Thus, topic models was not as useful to understand the condition of publication city as it relates to topics but offered a way to explore distribution of topics and thus insight into what type of works the Library of Congress collected at a finer grain detail than the LCC subject classification.

5.5 Probability distributions (Naive Bayes)

Another approach we took to assessing the difference in word distributions among Saigon and Hanoi records was to calculate the probability of the words' appearance, conditional upon city. To calculate these, we used Bayes' rule (adding 1 to each number of instances to prevent zero values). We then examined the features, looking at the most probable tokens for Hanoi and Saigon to see if any interesting patterns emerged. Before doing so, we eliminated those values that would not be interpretable, namely punctuation and single-syllable tokens (single syllables in Vietnamese are often meaningless out of context, and not all were caught by the tokenizer). The initial results were tantalizing: among the most likely tokens for Hanoi were words associated with Communist rhetoric, such as cách mạng (revolution), nhân dân (people), xây dựng (build), and anh hùng (hero), while the Saigon tokens included more words that could be seen as democratic or nationalist, e.g. công dân (citizen), phật giáo (Buddhism), quê hương (homeland), and hiện đại (modern). When we manually calculated the probability of one title, "Anh hùng lực lượng vũ trang nhân dân" *(The Hero of the People's Armed Forces)*, we found that its conditional probability for Hanoi outweighed that for Saigon by a wide margin, as would be expected, given the words. This matched the true label of "Hanoi."

Hanoi			Saigon		
cách_mạng	0.002979032	**revolution**	truyện_dài	0.004803892	long story
truyện_ngắn	0.001986021	short story	truyện_ngắn	0.000729705	short story
nhân_dân	0.001871443	**people**	giáo_dục	0.000668896	education
minh_họa	0.001795058	illustrate	**cách_mạng**	0.000547279	**revolution**
dân_tộc	0.001527709	nation	**công_dân**	0.00048647	**citizen**
truyện_ký	0.001413131	memoir	văn_hóa	0.00048647	culture
xây_dựng	0.001413131	**build**	văn_học	0.00048647	literature
anh_hùng	0.001298552	**hero**	chú_thích	0.00048647	note
công_tác	0.00126036	activity	chúng_ta	0.000425661	we
nghiên_cứu	0.001183974	research	**phật_giáo**	0.000425661	**Buddhism**
giới_thiệu	0.001145782	introduction	giới_thiệu	0.000364853	introduction
nhiệm_vụ	0.001069396	duty	cộng_hòa	0.000364853	republic
biên_soạn	0.000993011	compile	xã_hội	0.000364853	society
khoa_học	0.000993011	science	phiên_dịch	0.000364853	translate
văn_học	0.000993011	literature	**quê_hương**	0.000364853	**homeland**
xã_hội	0.000916625	society	con_người	0.000304044	person
nông_nghiệp	0.000878433	agriculture	**hiện_đại**	0.000304044	**modern**
lịch_sử	0.000802047	history	giáo_sư	0.000304044	professor
nghệ_thuật	0.000802047	art	khuôn_mặt	0.000304044	face
chú_thích	0.000802047	note	cuộc_đời	0.000304044	lifetime
bổ_sung	0.000725662	supplement	lịch_sử	0.000304044	history

Table 4. Tokens with strongest Hanoi and Saigon probabilities. "Ideologically significant" words are in bold.

5.6 Probability Distribution - Analysis

We evaluated this approach on some unlabeled data. Our data collection and cleaning process had produced a natural "test set" of 624 records for which we had not been able to extract city of origin; although the regular expressions had not identified the city, a human reader would be able to. We calculated the conditional probabilities for three labels: Hanoi, Saigon, and "other" (records with a valid city label other than Hanoi and Saigon), and ran a Python script to produce the probabilities. We then manually compared each result to the original scanned text, discarding 37 records with a clearly invalid title (author's name, very short fragments, or other text inappropriately pulled in).

Overall, our calculated probabilities predicted the city of origin with 46.6% accuracy. Broken down by city, Hanoi records were correctly labeled in 56.9% of cases, 'other' in 62.4%, and Saigon records in only 20.5%. In general, there seems to be bias against Saigon: very few cities were incorrectly labeled as Saigon, while a large number of Saigon records are labeled as Hanoi and 'other.' It is possible that grouping all other cities together 'diluted' the results by including words that might represent a wide range of political stances, including those of diaspora communities those that might be similar to Saigon. The fact that the 'other' label is the most accurately assigned one, and the number of Saigon records mislabeled as 'other,' seem to support this (a smaller but very significant proportion of Hanoi records were also labeled 'other'). We cannot claim that our Bayesian model has discriminant validity.

True v. Predicted	Hanoi (pred.)	Saigon (pred.)	Other (pred.)
Hanoi (true)	**70**	5	48
Saigon (true)	40	**42**	123
Other (true)	93	4	**161**

Table 5. Confusion Matrix for Held-Out Records with Cities Predicted by Naive Bayes Model

6 CONCLUSION

Our project analyzed the bibliographies (1982, 1987) of Library of Congress Vietnamese-language collections in order to understand library collecting patterns and the relationship between topics and publication location of non-serial works. We approached these questions using three methods: frequency counts, topic models, and Naive Bayes. Frequency counts allowed us to critically think through the variation in library collecting patterns and thus develop more well-informed hypotheses. Topic models offered a closer look beyond the LCC classification and into the topics of the titles. These topics revealed some insight into the types of books which were collected from Hanoi and Saigon. Furthermore, topic models on the 'title' of a work suggested certain semantic and literary structures of titles around themes like spiritual emotions, space/time, and democratic politics.

Naive Bayes offered a more structured and convincing method of analyzing the difference in word distributions among Saigon and Hanoi records. We calculated the probability of the words' appearance conditioned upon its publication city. Among the most likely tokens for Hanoi were words associated with Communist rhetoric, such as cách mạng (revolution), nhân dân (people), xây dựng (build), and anh hùng (hero). In comparison, the Saigon tokens included more words that could be seen as democratic or nationalist, e.g. công dân (citizen), phật giáo (Buddhism), quê hương (homeland), and hiện đại (modern). For validation, we 'predicted' unknown cities (due to OCR/Regex) and cross-validated that with human-reading of the original bibliography. Although our cross-validation method yielded a 46.6% predictive accuracy, Naive Bayes revealed the most interesting and convincing results to answer our research question: What words in titles most characterize a publication city?

The results of this project have strong convergent validity on the following points: the recognition of subjective political leanings in library collections, the differences in topics of published materials from Saigon and Hanoi, and the semantic and linguistic structure of titles. Our project did not directly consider publication cities outside of Hanoi and Saigon, and thus continues the standard (and arguably limiting) analytical frame within the field of Vietnamese history. Overall, this project contributes new ways of analyzing the relationship between library collections, topics of works, and their publication location.

This project demonstrates the strengths of a dual "close" and "distant" approach to experimental design: historically situated and argument driven as well as scientifically justified in terms of method, validation, and transparency on limitations.

REFERENCES

Blei, David M. "Probabilistic Topic Models." *Communications of the ACM* 55 (April 2012): 4.

Monroe et al. Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. Political Analysis (2008) 16 (4): 372-403.

Moretti, F. (2007) Graphs, Maps, Trees: Abstract Models for a Literary History, Verso, London.

Underwood, T. "The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us" (*New Literary History* 45, no. 3 [Summer 2014].

APPENDIX

A. Related Work

Histories of Libraries, Archives, State Building, Vietnam

a. Beredo, Cheryl. Import of the Archive: U.S. Colonial Rule of the Philippines and the Making of American Archival History. Series on Archives, Archivists and Society, number. 5. Sacramento, CA: Litwin Books, 2013
b. Fitzpatrick, Elizabeth B. "The Public Library as Instrument of Colonialism: The Case of the Netherlands East Indies." Libraries & the Cultural Record 43, no. 3 (2008): 270–85.
c. Jarvis, Helen. "The National Library of Cambodia: Surviving for Seventy Years." Libraries & Culture 30, no. 4 (October 1, 1995): 391–408.
d. Krajewski, Markus. Paper Machines: About Cards & Catalogs, 1548-1929. History and Foundations of Information Science. Cambridge, Mass: MIT Press, 2011.
e. Maack, Mary Niles. Libraries in Senegal: Continuity and Change in an Emerging Nation. Chicago: American Library Association, 1981.
f. Marco Beretta, Bibliotheca Lavoisieriana: The Catalogue of the Library of Antoine Laurent Lavoisier (Florence, 1995), 13– 58.
g. Wijasuriya, D. E. K., Huck Tee Lim, and Radha Nadarajah. The Barefoot Librarian: Library Developments in Southeast Asia with Special Reference to Malaysia. Hamden, Conn: Linnet Books, 1975.
h. Yu, Priscilla C. "Leaning to One Side: The Impact of the Cold War on Chinese Library Collections." Libraries & Culture 36, no. 1 (January 1, 2001): 253–66.

Digital Humanities, Computational Studies of Literature, Quantitative Methods

i. Ted Underwood's work- understanding genre in a collection of a million volumes
    i. https://figshare.com/articles/Understanding_Genre_in_a_Collection_of_a_Million_Volumes_Interim_Report/1281251
j. Lincoln, Matthew D. "Foreign and Domestic Interaction in the Early Modern Printmaking Network." *Matthew Lincoln* (blog), 17 Oct 2014, http://matthewlincoln.net/2014/10/17/foreign-and-domestic-interaction-in-the-early-modern-printmaking-network.html

B. Other Data and Corpora

Future Directions: We can use the output of word lists from Naive Bayes to train a classifier for uncoded documents, such as a bibliography of titles without a publication city. We can also apply the same process of topic models and permutation test with other corpora to compare most prevalent topics (highest topic distribution across documents).

Databases

1. Vietnamese Intellectual Networks Database (VIND)[8]: databases of Vietnamese intellectuals, publishers, and their textual output

Official bibliographies, catalogs, deposits (things published in Indochina)

1. 1859-1954: bibliography of official indochina publications (hardcopy)
2. 1887-1919: bibliography of indochinese studies society
3. 1897: catalog of indochinese studies society
4. 1916: EFEO catalog of library
5. 1922-1954: catalog of Indochinese works in the national library in France (microfiche)
6. 1923-1931: semesterly legal deposit slips of books published in Indochina, deposited in the national library of France
7. 1931: bibliography of official indochina publications edited by the Governor General of Indochina presented at French colonial exposition
8. 1958: bibliography of periodicals published in Vietnam by MSU

[8] "Vietnamese Intellectual Networks Database | Digital Humanities." Accessed May 4, 2016. http://digitalhumanities.berkeley.edu/projects/vietnamese-intellectual-networks-database.

9. 1960-1969: bibliography on Vietnamese Official publications (Thư tịch về ấn phẩm công việt nâm) Directorate of National Archives and LIbraries, Saigon) - to request
10. 1962: bibliography of periodicals published in Vietnam by MSU
11. 1982, 1987: Vietnamese holdings in the LOC
12. 2009-2015: monthly or yearly bibliography of works published in Vietnam in the National Library of Vietnam (contemporary Hanoi)
    a. 2009: yearly
    b. 2012: monthly bibliography of works published in Vietnam in the National Library of Vietnam (contemporary Hanoi) ^$ (problem with diacritics encoding for 2-2012 and 3-2012)
    c. 2015: monthly
    d. 2016: monthly, only have 01-2016 so far

Thematic/Analytical Bibliographies (things curated on subject of 'Indochina')

1. 1915: Indosinica bibliography by Cordier
2. 1922: catalog of scientific institute of Indochina
3. 1932: "To Better Know Indochina" by Central Library of Hanoi, Paul Boudet
4. 1940: Indochina books by Harvard University
5. 1950: "Bibliography of a land and its people" by LOC
6. 1952: recent articles on Vietnam by MSU
7. 1954: Vietnamese legal materials by Phuong
8. 1955: National Institute of Administration holdings by MSU
9. 1958: listing of journals on public administration by MSU
10. 1958: recent articles on Vietnam by MSU
11. 1959: MSU Vietnam project bibliography by MSU
12. 1959: "What to read on Vietnam" by MSU
13. 1962: "Bibliography on the political history of Vietnam 1802-1962" by MSU, Roy Jumper
14. 1963: South Vietnam Intelligence Bibliography by US Bureau of Intelligence and Research
15. 1963: North Vietnam Intelligence Bibliography
16. 1966: Recommended books form and about Vietnam by Charlotte Polin; US Committee to Aid the National Liberation Front of Vietnam
17. 2000s: Thematic catalog of materials on Hanoi in the National Library of Vietnam (contemporary)