Deconstructing Libraries

Spring 2016
Info 290

Cindy A. Nguyen, History
Jordan Shedlock, School of Information

MOTIVATIONS & RESEARCH QUESTION

Relationship between politics and culture through the lens of libraries

Nguyen's dissertation examines the cultural and political history of libraries in Vietnam from 1885 to 1986. How does the library develop as an institution of knowledge in Vietnam? In what ways do Vietnamese libraries transform through political regime changes? How do libraries express colonial and post-colonial state power as well as subvert it?

RESEARCH QUESTION

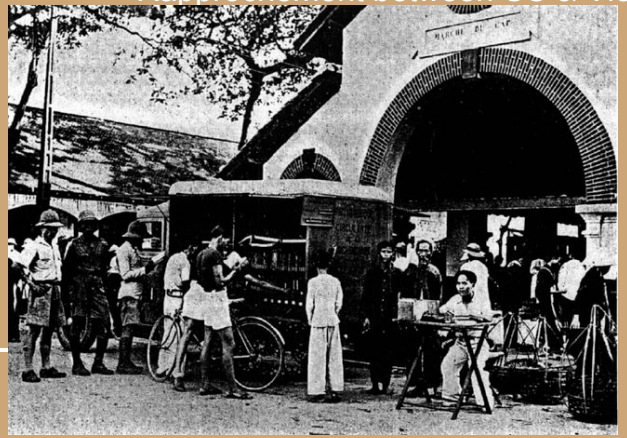
1885-1945 French Colonialism

1945-1954 First Indochina War

1954-1975 Second Indochina War/ Vietnam War

1975-1979 Vietnam-Cambodia War, Sino-Vietnamese Conflict

1986 Renovation Neo-Liberal Reforms & Rapprochement between US & Vietnam



Mobile library, 1936 Saigon

Testable Hypotheses

1. Different publication locations will have different distribution of topics. What are the topics/words most characteristic of a city?
    a. H1: Hanoi will have more topics on Communism, war, revolution, army than Saigon.
    b. H2: Saigon will have more topics on US ideas (modernity, democracy, anti-Communism)
    c. H3: LOC collection will prefer Saigon (ally) materials over Hanoi.

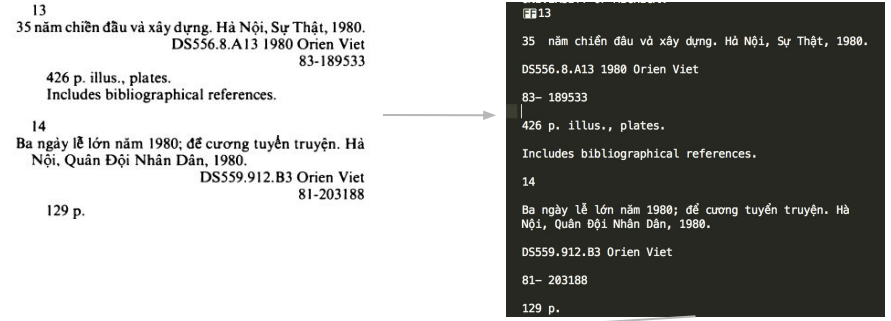
2. Operationalized through
    a. Distribution of collected works by Library of Congress Classification
    b. Frequency Distribution of words by city
    c. Topic Models

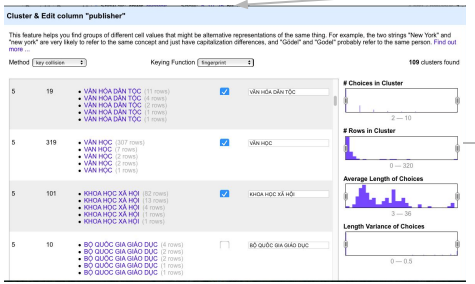
Data Collection and Cleaning

•Sources: bibliographies (metadata for books published in & about Vietnam) from Library of Congress

•Data cleaning and preparation:

- OCR (Abbyy Finereader)
- Regex: extract author, title, year, publisher, location
- Google Refine: clean up data fields, resolve OCR errors and misspellings
- vnTokenizer: tokenize Vietnamese words
  cách mạng = cách-mạng [revolution]





35 years of war and building

3 major holidays in 1980; an outline of the stories of beloved uncle Hồ

35năm chiến_đấu và xây_dựng
ba ngày lễ lớn năm 1980 ; đề cương tuyên truyện
bác hồ kính_yêu

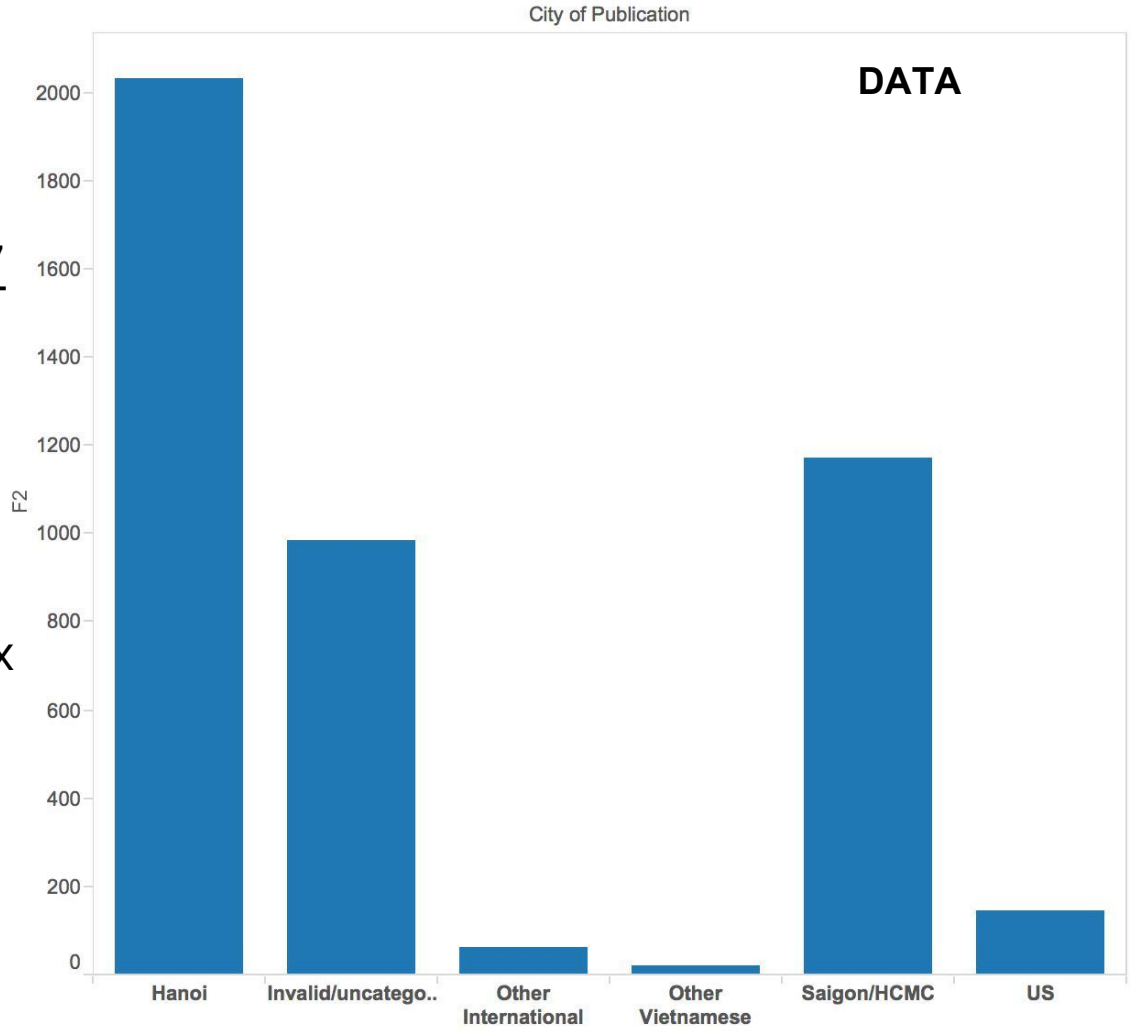
Frequency counts

4417 Total records LOC 82, 87

2016 Hanoi

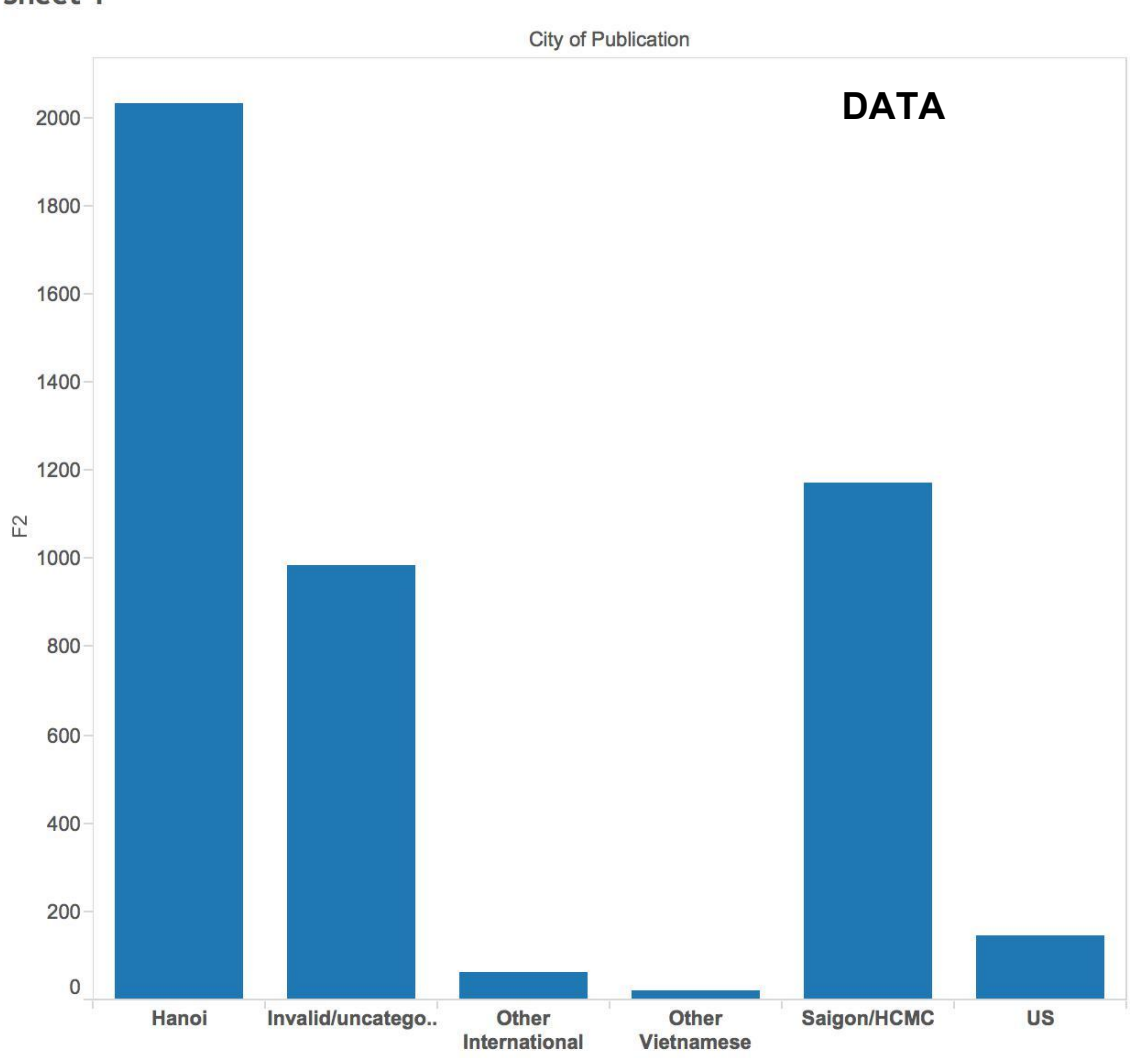
1139 Saigon

1252 Non-Hanoi or
Non-Saigon, Messy OCR/RegEx



Frequency counts

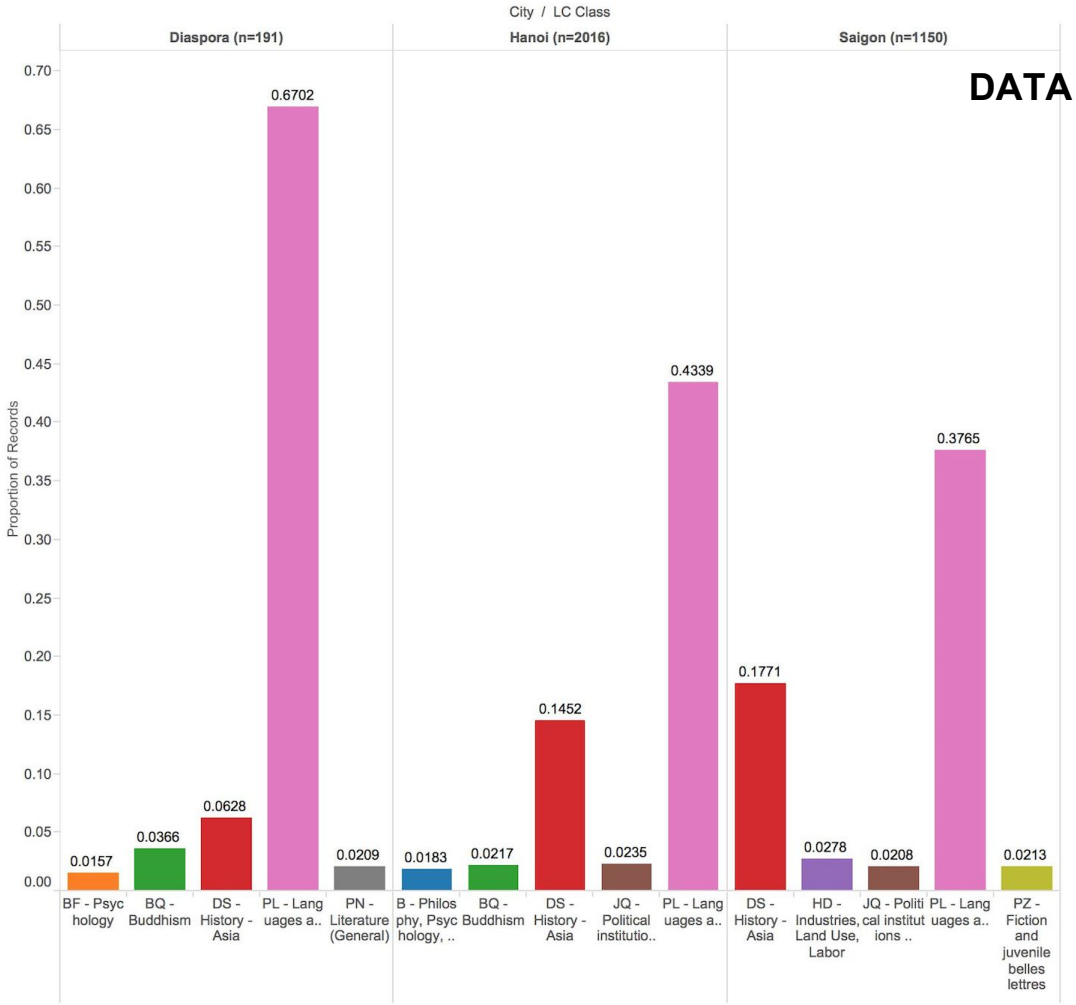
- LOC collected more books from Hanoi than Saigon.
- Even after the Vietnam War, LOC collecting regime interested in Hanoi politics and Communist ideology. *“Understanding American studies of Marx-Lenin” 1979, “Beloved Uncle Ho” 1970*



Library of Congress Classification (LCC) Breakdown by City

- **Topics consistent across cities:**
- **PL:** Languages and Literature
- **DS:** Asian History

This suggests that the data reflect the LOC's **collection policies** first and foremost; **ideological distinctions/content are not captured**



What are the words most characteristic of a city?

1. Exploratory
    a. Naive Bayes (Probability of word in title conditioned on city)
    b. Topic models (learn patterns that suggest types of books from a city)
2. Prediction
    a. Naive Bayes (Can we predict an unknown city based on the words in a title?)

Naive Bayes

Probability of word in title conditioned on city; bigrams only, no stopwords

Red text: words that co-occur in top 21 results of Hanoi and Saigon

Φ, Y=Hanoi

cách_mạng	0.002979032196463354	*revolution*
truyện_ngắn	0.0019860214643089027	*short story*
nhân_dân	0.001871443302906466	*people*
minh_họa	0.0017950578619715081	*illustrate*
dân_tộc	0.001527708818699156	*nation*
truyện_ký	0.0014131306572967193	*memoir*
xây_dựng	0.0014131306572967193	*build*
anh_hùng	0.0012985524958942826	*hero*
công_tác	0.0012603597754268037	*activity*
nghiên_cứu	0.0011839743344918459	*research*
giới_thiệu	0.001145781614024367	*introduction*
nhiệm_vụ	0.0010693961730894091	*duty*
biên_soạn	0.0009930107321544513	*compile*
khoa_học	0.0009930107321544513	science
văn_học	0.0009930107321544513	*literature*
xã_hội	0.0009166252912194935	*society*
nông_nghiệp	0.0008784325707520146	agriculture
lịch_sử	0.0008020471298170569	*history*
nghệ_thuật	0.0008020471298170569	*art*
chú_thích	0.0008020471298170569	*note*
bổ_sung	0.0007256616888820991	*supplement*

Φ, Y=Saigon

truyện_dài	0.0048038917604135	*long story*
truyện_ngắn	0.0007297050775311645	*short story*
giáo_dục	0.0006688963210702341	*education*
cách_mạng	0.0005472788081483734	*revolution*
công_dân	0.000486470051687443	*citizen*
văn_hóa	0.000486470051687443	*culture*
văn_học	0.000486470051687443	*literature*
chú_thích	0.000486470051687443	*note*
chúng_ta	0.0004256612952265126	*we*
phật_giáo	0.0004256612952265126	*Buddhism*
giới_thiệu	0.00036485253876558226	*introduction*
cộng_hòa	0.00036485253876558226	*republic*
xã_hội	0.00036485253876558226	*society*
phiên_dịch	0.00036485253876558226	*translate*
quê_hương	0.00036485253876558226	*homeland*
con_người	0.0003040437823046519	*person*
hiện_đại	0.0003040437823046519	*modern*
giáo_sư	0.0003040437823046519	*professor*
khuôn_mặt	0.0003040437823046519	*face*
cuộc_đời	0.0003040437823046519	*lifetime*
lịch_sử	0.0003040437823046519	*history*

Naive Bayes

Probability of word in title conditioned on city; bigrams only, no stopwords

Highlighted blue: words of interest for comparison of results between Hanoi and Saigon

Φ, Y=Hanoi

cách_mạng	0.002979032196463354	***revolution***
truyện_ngắn	0.0019860214643089027	*short story*
nhân_dân	0.001871443302906466	***people***
minh_họa	0.0017950578619715081	*illustrate*
dân_tộc	0.001527708818699156	*nation*
truyện_ký	0.0014131306572967193	*memoir*
xây_dựng	0.0014131306572967193	***build***
anh_hùng	0.0012985524958942826	***hero***
công_tác	0.0012603597754268037	*activity*
nghiên_cứu	0.0011839743344918459	*research*
giới_thiệu	0.001145781614024367	*introduction*
nhiệm_vụ	0.0010693961730894091	*duty*
biên_soạn	0.0009930107321544513	*compile*
khoa_học	0.0009930107321544513	science
văn_học	0.0009930107321544513	*literature*
xã_hội	0.0009166252912194935	*society*
nông_nghiệp	0.0008784325707520146	agriculture
lịch_sử	0.0008020471298170569	*history*
nghệ_thuật	0.0008020471298170569	*art*
chú_thích	0.0008020471298170569	*note*
bổ_sung	0.0007256616888820991	*supplement*

Φ, Y=Saigon

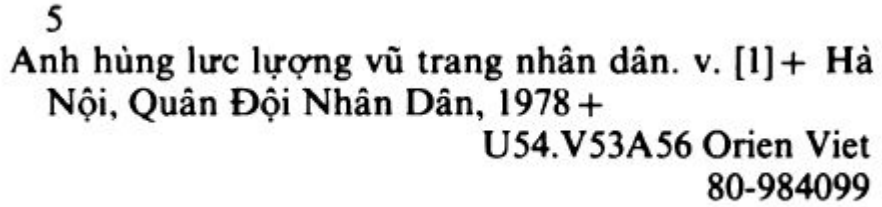
truyện_dài	0.0048038917604135	*long story*
truyện_ngắn	0.0007297050775311645	*short story*
giáo_dục	0.0006688963210702341	*education*
cách_mạng	0.0005472788081483734	***revolution***
công_dân	0.000486470051687443	***citizen***
văn_hóa	0.000486470051687443	*culture*
văn_học	0.000486470051687443	*literature*
chú_thích	0.000486470051687443	*note*
chúng_ta	0.0004256612952265126	*we*
phật_giáo	0.0004256612952265126	***Buddhism***
giới_thiệu	0.00036485253876558226	*introduction*
cộng_hòa	0.00036485253876558226	*republic*
xã_hội	0.00036485253876558226	*society*
phiên_dịch	0.00036485253876558226	*translate*
quê_hương	0.00036485253876558226	***homeland***
con_người	0.0003040437823046519	*person*
hiện_đại	0.0003040437823046519	***modern***
giáo_sư	0.0003040437823046519	*professor*
khuôn_mặt	0.0003040437823046519	*face*
cuộc_đời	0.0003040437823046519	*lifetime*
lịch_sử	0.0003040437823046519	*history*

Naive Bayes to Predict Unknown City

P(X = ”Anh hùng lực lượng vũ trang nhân dân”|Y=Hanoi) = 1.5648340863728223

P(X=”Anh hùng lực lượng vũ trang nhân dân”|Y=Saigon) = 0.0002585442251010566

True value =


5
Anh hùng lực lượng vũ trang nhân dân. v. [1]+ Hà Nội, Quân Đội Nhân Dân, 1978+
U54.V53A56 Orien Viet
80-984099

The Hero of the People’s Armed Forces. V. 1, (Hanoi: The People’s Army, 1978)

Library of Congress Subject Heading : U [Military Science]

Topics 1982 and 1987

#	Weight	Topic tokens	General Topic
20	0.06321	truyện tập ngắn ngày người vui dái bạn biển cơn gọi nơi sâu ngán thơ_ca chân_trời cứa thư non càmau Story, volume, short, poem, song, horizon	**Textual format, story, genre**
5	0.06214	tập thơ nước truyện_ký sáng trong hoa đất truyện_ngắn núi lửa đất rừng lứa giữa nhiều tác_giả thơ_văn đất mặt Volume, poem, country, memoir, short story, mountain, jungle, author, verse	**Poetry, nature, memoir, author**
96	0.06051	trong của mới một những cách_mạng năm văn_hóa đèn nội văn_nghệ giai_đoạn dân_tộc nhiệm_vụ hoạch vần đầu liên đền dòng_sông Revolution, literature, performance, stage, nation, action, function, channel	**Revolution, action**
22	0.05975	bán thứ lần lần truyện_dài tái cúa xuất lãn bản đời xuất xuất khanh xuất hoang lấn tựa phụ hậu Number, times, long story, published	**Textual format, republished, genre**
78	0.05733	nam việt miển tại cộng_sản phong_trào bắc phụ_nữ công_đoàn câu nứa tiềng chương hiện_đại khoa_học_xã_hội tra chù thám lập tiêng South, Viet, region, Communism, movement, north, women, union, modern, social sciences	**Communism, region, women, social sciences**
31	0.0571	thứ lần lẩn bổ_sung sửa_chữa hành tái_bản đại_hội tên một huân bắc bôi đẹp trang vàng pic quảng lửa thức Number, supplement, repair, reprint, meeting, north	**Number, reprint**
7	0.05663	văn nguyễn của phạm ngọc giới_thiệu dương hổng nguyên đinh xuân đức văn chú tường hữu lâm bùi ngữ biên Literature, introduction, language	**Literature, language**
65	0.05164	thuyết tiểu tiếu thuyẽt thuyèt tiéu người gián_điệp thuyêt rừng lịch_sử thuybt tiểu tiêu xanh bão vàng trè phò hận Novel, fiction, spy, forest, history	**Novel, spy, history**
97	0.04697	minh chí chi thành đổng phô truyện_ký trường_ca thông phồ lăng ca_dao thảnh đạo tiển lôc tiền cao triểu phường Memoir, epic, poem	**Memoir, epics**
24	0.04633	của người những bác anh làm thanh_niên tré điểu phẩn giông nghĩ tuổi súng tám cám trẻ như giờ bât Youth, young, think, age, time	**Youth**

Hanoi v. Saigon

1982

#	General Topic	Hanoi	Saigon
5	Poetry, nature, memoir, author	0.009808929	0.008638799
96	Revolution, action	0.030214885	0.021528166
78	Communism, region, women, social sciences	0.008836326	0.008266742

1987

#	General Topic	Hanoi	Saigon
5	Poetry, nature, memoir, author	0.011779472	0.009184172
96	Revolution, action	0.034526805	0.013463745
78	Communism, region, women, social sciences	0.004967198	0.007170078

- From 1982 to 1987 decrease in topics Communism 96 and 78 for Saigon.

- In 1987 topic 96 difference in Hanoi and Saigon

Results

1. Different publication locations will have different distribution of topics. What are the topics/words most characteristic of a city?
    a. H1: Hanoi will have more topics on Communism, war, revolution, army than Saigon.
        i. → Naive bayes suggest this is true.
        ii. → Topic models suggests* cN
    b. H2: Saigon will have more topics on US ideas (modernity, democracy, anti-Communism)
        i. → Naive bayes suggests other topics such as Buddhism, homeland
        ii. → Topic models suggests * CN
    c. H3: LOC collection will prefer Saigon (ally) materials over Hanoi.
        i. Frequency counts demonstrate preference over Hanoi than Saigon.

Concluding Remarks

- Challenges of data cleaning/OCR
- Confounding factor of **collection policy of LOC v. universe of published materials in Vietnam**
- Value of quantitative analysis for history discipline

Feature	**β**
Published in Hanoi	
Published in Saigon	
Topic 44 “Revolution, worker”	
Topic 51 “Saigon, home”	
Topic 5 “War poetry”	
$P(X = \text{"Hero"} \mid Y=\text{Hanoi}) = .0005298$	

Classification:

- Was this published in Hanoi? Saigon?
- Was this text pro-Communist or anti-Communist?

Research:

- What types of works were published in Saigon? Hanoi?
- What types of works were collected by the LOC?

Possible Ways to Featurize Data