Operationalizing Historical Questions

Datafying the Library of Congress Vietnam Collections

Cindy A. Nguyen, History
University of California, Berkeley

Introduction

cindyanguyen.com

@cindyanguyen

Ph.D. Candidate in Vietnamese history

Fields in Colonial Knowledge, History of Science & Classification, Digital Humanities

…interdisciplinary journey of cross-pollination: from source criticism to experimental design, open-access and diversifying the language of the internet

Builders & Users: Constructing the Vietnamese Library (1887-1986)

What is the history of Vietnamese 'public' libraries during the political regime changes of the 20th century?

How do libraries demonstrate the relationship among:

- the state (colonial, post-colonial)
- the public (reading preferences)
- education, language, and literacy
- ideas about modernity, access to knowledge, civil society

TRIỂN LÃM TÀI LIỆU LƯU TRỮ
QUY HOẠCH ĐÔ THỊ VÀ ĐỊA GIỚI HÀNH CHÍNH HÀ NỘI
1873 - 1954
TRUNG TÂM LƯU TRỮ QUỐC GIA I

Fieldwork: Archives & Research collaborations in Vietnam

University of California, Berkeley

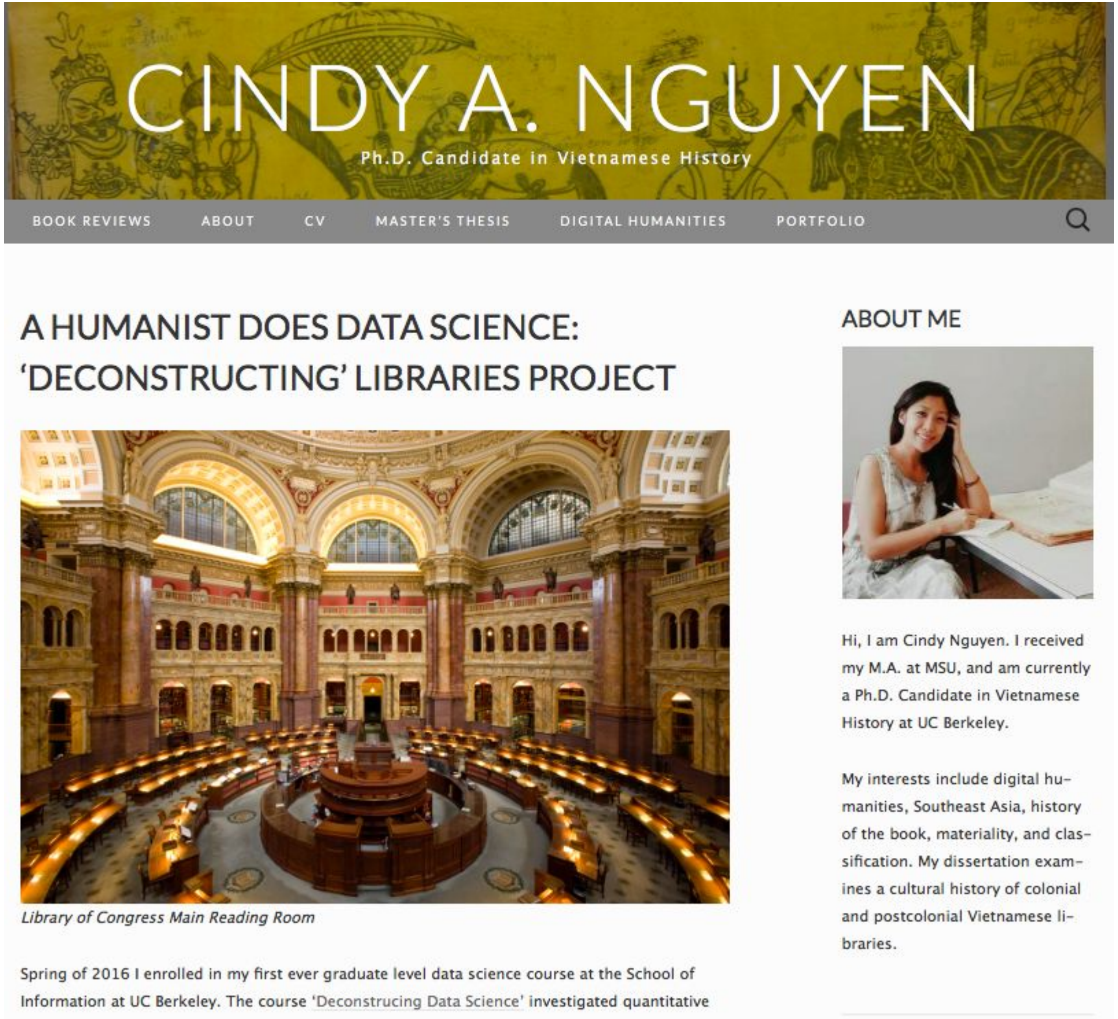
Spring 2016- Information Science 290

Final Project

A Humanist does Data Science: Deconstructing Libraries

Cindy A. Nguyen, History

Jordan Shedlock, School of Information



CINDY A. NGUYEN

Ph.D. Candidate in Vietnamese History

A HUMANIST DOES DATA SCIENCE: 'DECONSTRUCTING' LIBRARIES PROJECT

Library of Congress Main Reading Room

Spring of 2016 I enrolled in my first ever graduate level data science course at the School of Information at UC Berkeley. The course 'Deconstrucing Data Science' investigated quantitative

ABOUT ME

Hi, I am Cindy Nguyen. I received my M.A. at MSU, and am currently a Ph.D. Candidate in Vietnamese History at UC Berkeley.

My interests include digital humanities, Southeast Asia, history of the book, materiality, and classification. My dissertation examines a cultural history of colonial and postcolonial Vietnamese libraries.

Website

cindyanguyen.com

Github

https://github.com/cindyanguyen/deconstructing-libraries

Blog post

https://cindyanguyen.com/2016/12/02/a-humanist-does-data-science-deconstructing-libraries-project/

1. Operationalizing research questions
2. Tools to clean, read, analyze data
3. Concluding lessons

1. Operationalizing research questions

How to ask questions and how to answer them

- Translating research questions into operationalizable tasks
    - What evidence do I have to support my research claim? How do I argue this?
- Critical inquiry into data: its production and limitations
- What question does this actually answer?

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
1946-1954	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

Iterating between Research Question & Data

→ What type of data gestures at my research topic? (libraries, government, readership, libraries development, collections)

- Ideal data source?

→ What more refined research question can emerge from my data source? (curation, control, censorship, ideology, language, literacy)

- Example research question?

Finding Data

Humanistic data: difficult to obtain, uneven, unformatted, not easily machine readable

Not ‘data of convenience’ - well structured in the English language

→ justifiably ‘interesting’ and significant

VIETNAMESE HOLDINGS
in the Library of Congress

A Bibliography

Compiled by A. Kohar Rony
Southern Asia Section, Asian Division

LIBRARY OF CONGRESS WASHINGTON 1982

MONOGRAPHS

A

1
An Khê. Bông lúa sa-mo; truyện dài. [Saigon] Miên Nam, 1968.
PL4378.9.A45B6
510 p.

2
An Khê. Chân trời nào cho em; tiểu thuyết tình cảm xã hội. [Saigon] Đồng Nai, 1971.
PL4378.9.A45C5
276 p.

3
An Khê. Môi tình đâu; tiêu-thuyêt tình-cam xã-hôi. [Saigon] Miên Nam [1965]
PL4378.9.A45M6
479 p.

4
Anh Đào. Việt Nga hội thoại. [Hà-nội] Sông Lô, 1956.
PG2121.A54
118 p.
Errata slip inserted.

5
Anh Đức. Bức thư Cà-mau; truyện ngắn, bút ký. [In lần thứ 2] Hà-nội, Văn Học, 1966.
PL4378.9.A5B8 1966
135 p.
"Giai thương chinh thức văn học Nguyên Đình Chiêu."

6
Anh Đức. Đứa con của đất: tiểu thuyết. [s.1.] Văn Học Giải Phóng, 1975.
PL4378.9.A5D8
487 p.

7
Anh Đức. Một truyện chép ơ bệnh viện; truyện. In lân thú 2. [Hà-nội] Văn Học, 1963.
PL4378.9.A5M6 1963
151 p.

8
Anh Đức. Thằng My. Bìa và minh họa của Văn Đa. Hà-nội, Kim Đồng, 1966.
PL4378.9.A5T48
31 p. illus.
"Trích trong Bức thu Cà mâu."

9
Anh Thơ. Mùa xuân, màu xanh: thơ 1967–1973. Hà-nội, Văn Học, 1974.
PL4378.9.A55M8
109 p.

10
Anh Tho. Theo cánh chim câu; tho, 1945–1960. [Hà-nội] Văn Học, 1960.
PL4378.9.A55T5
105 p.

Anh-Tuấn Nguyễn Tuấn Phát:
see Nguyễn Tuấn Phát

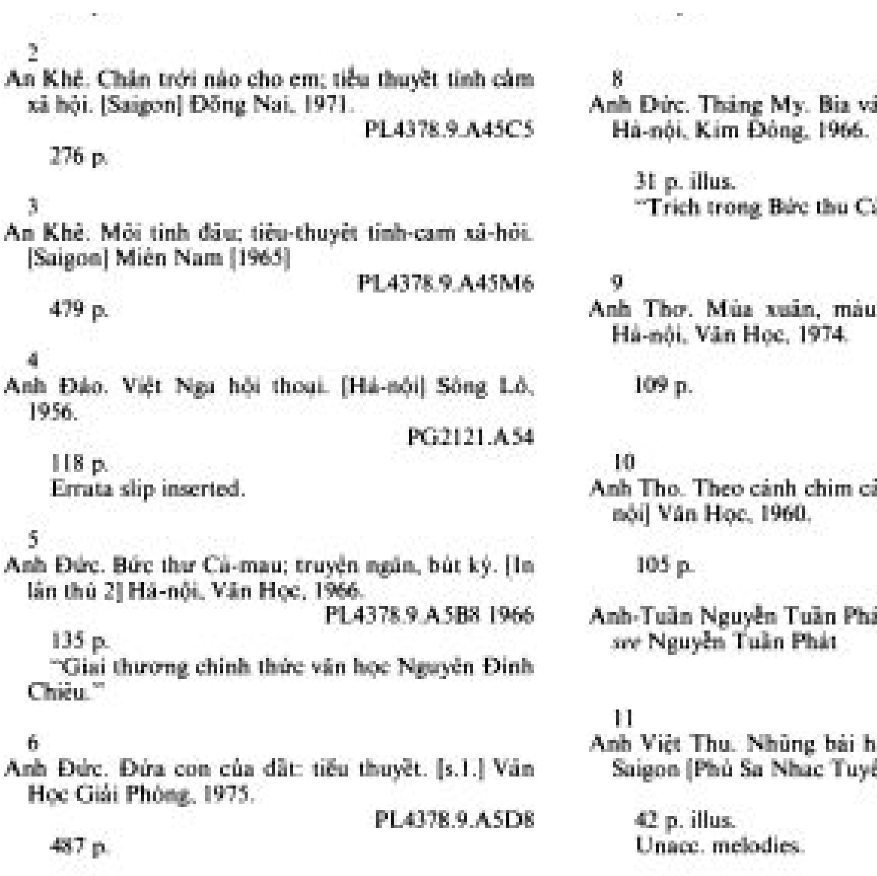
11
Anh Việt Thu. Những bài hát mới viết trên tóc me. Saigon [Phù Sa Nhạc Tuyển], 1967.
M1824.V5A5
42 p. illus.
Unacc. melodies.

B

12
Bá Dũng. Nắng sông Lam: truyện ký. [Hà-nội] Phụ Nữ, 1974.
PL4378.9.B2N3
117 p.

13
Ba Hồng và Phạm Hồng. Vòng quanh Sài-gòn: ký sự. [s.l.] Văn Nghệ Giải Phóng, 1975.
PL4378.9.B24V6
124 p.

- Author
- Title
- Publication City
- Publisher
- Publishing date
- Number of pages
- Library of Congress classification number



2
An Khê. Chân trời nào cho em; tiểu thuyết tình cảm xã hội. [Saigon] Đồng Nai, 1971.
PL4378.9.A45C5
276 p.

3
An Khê. Mối tình đầu; tiểu-thuyết tình-cam xã-hội. [Saigon] Miền Nam [1965]
PL4378.9.A45M6
479 p.

4
Anh Đào. Việt Nga hội thoại. [Hà-nội] Sông Lô, 1956.
PG2121.A54
118 p.
Errata slip inserted.

5
Anh Đức. Bức thư Cà-mau; truyện ngắn, bút ký. [In lần thứ 2] Hà-nội, Văn Học, 1966.
PL4378.9.A5B8 1966
135 p.
"Giải thưởng chính thức văn học Nguyễn Đình Chiểu."

6
Anh Đức. Đứa con của đất: tiểu thuyết. [s.1.] Văn Học Giải Phóng, 1975.
PL4378.9.A5D8
487 p.

8
Anh Đức. Thăng My. Bìa vi
Hà-nội, Kim Đồng, 1966.
31 p. illus.
"Trích trong Bức thư Cà

9
Anh Thơ. Mùa xuân, màu
Hà-nội, Văn Học, 1974.
109 p.

10
Anh Tho. Theo cánh chim cả
nội] Văn Học, 1960.
105 p.

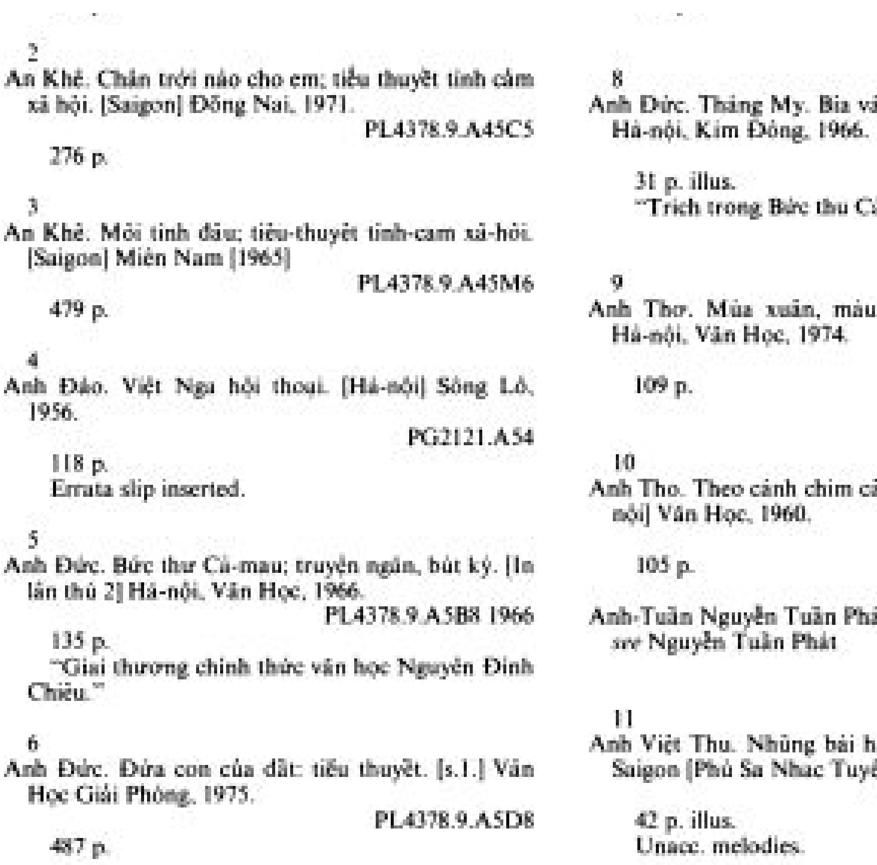
Anh-Tuấn Nguyễn Tuấn Phá
see Nguyễn Tuấn Phát

11
Anh Việt Thu. Những bài h
Saigon [Phù Sa Nhạc Tuyể
42 p. illus.
Unacc. melodies.

Why is this data source significant?
What story could I tell?

- Author
- Title
- Publication City
- Publisher
- Publishing date
- Number of pages
- Library of Congress classification number

- Library curation
- History of publishing in Vietnam
- Circulation of knowledge on Vietnam



2
An Khê. Chân trời nào cho em; tiểu thuyết tình cảm xã hội. [Saigon] Đồng Nai, 1971.
PL4378.9.A45C5

276 p.

3
An Khê. Mối tình đầu; tiểu-thuyết tình-cam xã-hội. [Saigon] Miền Nam [1965]
PL4378.9.A45M6

479 p.

4
Anh Đào. Việt Nga hội thoại. [Hà-nội] Sông Lô, 1956.
PG2121.A54

118 p.
Errata slip inserted.

5
Anh Đức. Bức thư Cà-mau; truyện ngắn, bút ký. [In lần thú 2] Hà-nội, Văn Học, 1966.
PL4378.9.A5B8 1966

135 p.
"Giai thương chinh thức văn học Nguyễn Đinh Chiều."

6
Anh Đức. Đứa con của đất: tiểu thuyết. [s.1.] Văn Học Giải Phóng, 1975.
PL4378.9.A5D8

487 p.

8
Anh Đức. Thăng My. Bia vi
Hà-nội, Kim Đồng, 1966.

31 p. illus.
"Trích trong Bức thu C

9
Anh Thơ. Mùa xuân, màu
Hà-nội, Văn Học, 1974.

109 p.

10
Anh Tho. Theo cánh chim c
nội] Văn Học, 1960.

105 p.

Anh-Tuấn Nguyễn Tuấn Ph
see Nguyễn Tuấn Phát

11
Anh Việt Thu. Những bài h
Saigon [Phù Sa Nhac Tuy

42 p. illus.
Unacc. melodies.

What it is and what it could be:

Data includes information on:

- 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.
- Both retrospective works and new items published in post-war Socialist Republic of Vietnam
- Author
- Title
- Publication City
- Publisher
- Publishing Date
- Number of Pages
- Library of Congress Classification
- Language
- Monograph or Serial

Data could shed light on:

- A part of publishing history
  - North and South Vietnam (during the War)
  - Post-war Socialist Republic of Vietnam
  - Diasporic Vietnamese populations
- Topics and types of works published
- Library curation - logic of collection
- Knowledge access to Americans on the subject of ‘Vietnam’

Assumptions - History of the book

‘TITLE’

- What *information* does a work’s title contain?

‘PUBLICATION CITY’

- What does the publication city suggest?

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. Probability of words in a title by city
   c. Topic Models

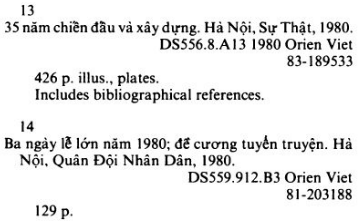
2. Tools to clean, read, analyze data

Data Cleaning

DATA

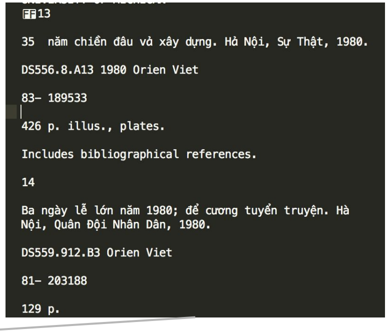
•Data cleaning and preparation:

- **Optical Character Recognition** (Abbyy Finereader): from PDF to text file
- **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]

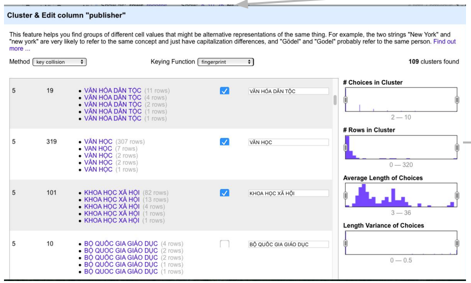


13
35 năm chiến đấu và xây dựng. Hà Nội, Sự Thật, 1980.
DS556.8.A13 1980 Orien Viet
83-189533
426 p. illus., plates.
Includes bibliographical references.

14
Ba ngày lễ lớn năm 1980; đề cương tuyển truyện. Hà Nội, Quân Đội Nhân Dân, 1980.
DS559.912.B3 Orien Viet
81-203188
129 p.



```
FF13

35  năm chiền đâu vả xây dựng. Hà Nội, Sự Thật, 1980.

DS556.8.A13 1980 Orien Viet

83- 189533

426 p. illus., plates.

Includes bibliographical references.

14

Ba ngày lễ lớn năm 1980; để cương tuyển truyện. Hà
Nội, Quân Đội Nhân Dân, 1980.

DS559.912.B3 Orien Viet

81- 203188

129 p.
```



Cluster & Edit column "publisher"

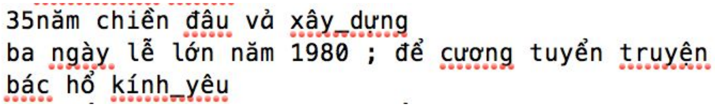
This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more ...

Method: key collision — Keying Function: fingerprint — **109** clusters found

5	19	VĂN HÓA DÂN TỘC (11 rows) VĂN HÓA DÂN TỘC (4 rows) VĂN HOÁ DÂN TỘC (2 rows) VĂN HOÁ DÂN TỘC (1 rows) VĂN HÓA DÂN TỘC (1 rows)	☑	VĂN HÓA DÂN TỘC
5	319	VĂN HỌC (307 rows) VAN HỌC (7 rows) VÀN HỌC (2 rows) VÂN HỌC (2 rows) VÁN HỌC (1 rows)	☑	VĂN HỌC
5	101	KHOA HỌC XÃ HỘI (82 rows) KHOA HỌC XÀ HỘI (13 rows) KHOA HỌC XÂ HỘI (4 rows) KHOA HỌC XÃ HỘI (1 rows) KHOA HỌC XA HỘI (1 rows)	☑	KHOA HỌC XÃ HỘI
5	10	BỘ QUỐC GIA GIÁO DỤC (4 rows) BỘ QUOC GIA GIÁO DỤC (2 rows) BỘ QUỐC GIA GIÁO DỤC (2 rows) BỘ QUỐC GIA GIÁO DỤC (1 rows) BỘ QUOC GIA GIÁO DỤC (1 rows)	□	BỘ QUỐC GIA GIÁO DỤC

35 years of war and building

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



```
35năm chiền đâu vả xây_dựng
ba ngày lễ lớn năm 1980 ; để cương tuyển truyện
bác hổ kính_yêu
```

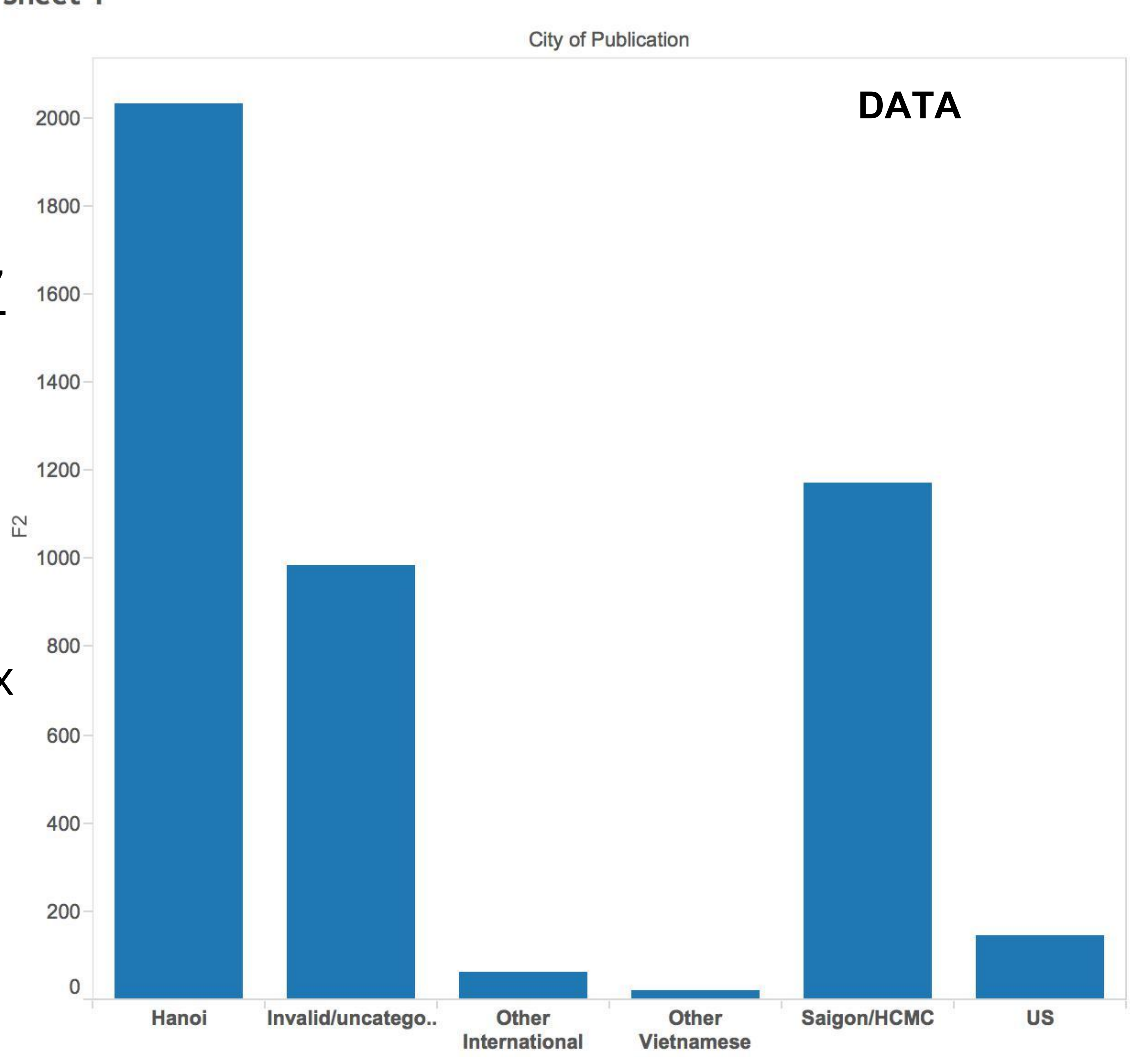
Frequency counts

4417 Total records LOC 82, 87

2016 Hanoi

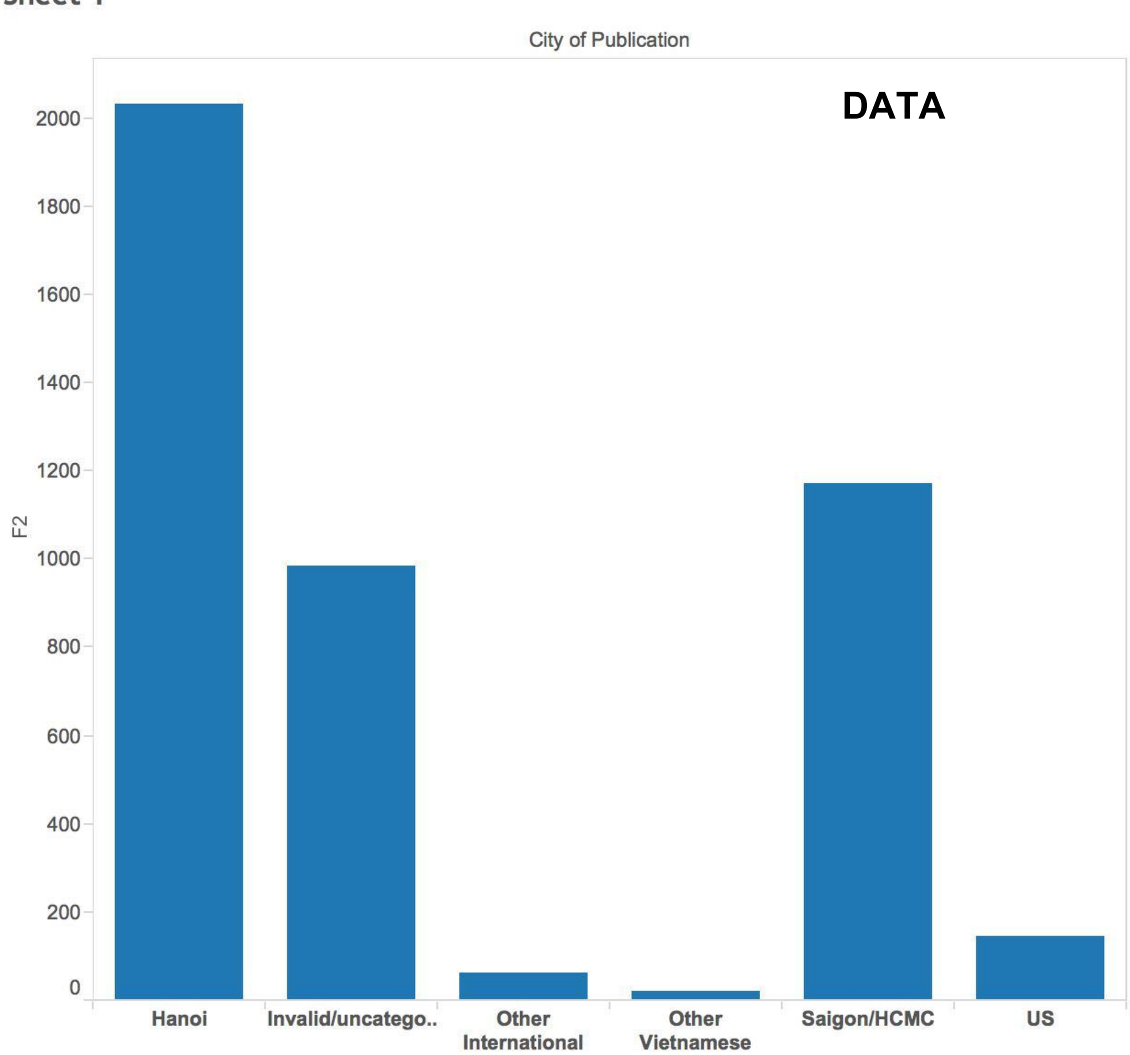
1139 Saigon

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*



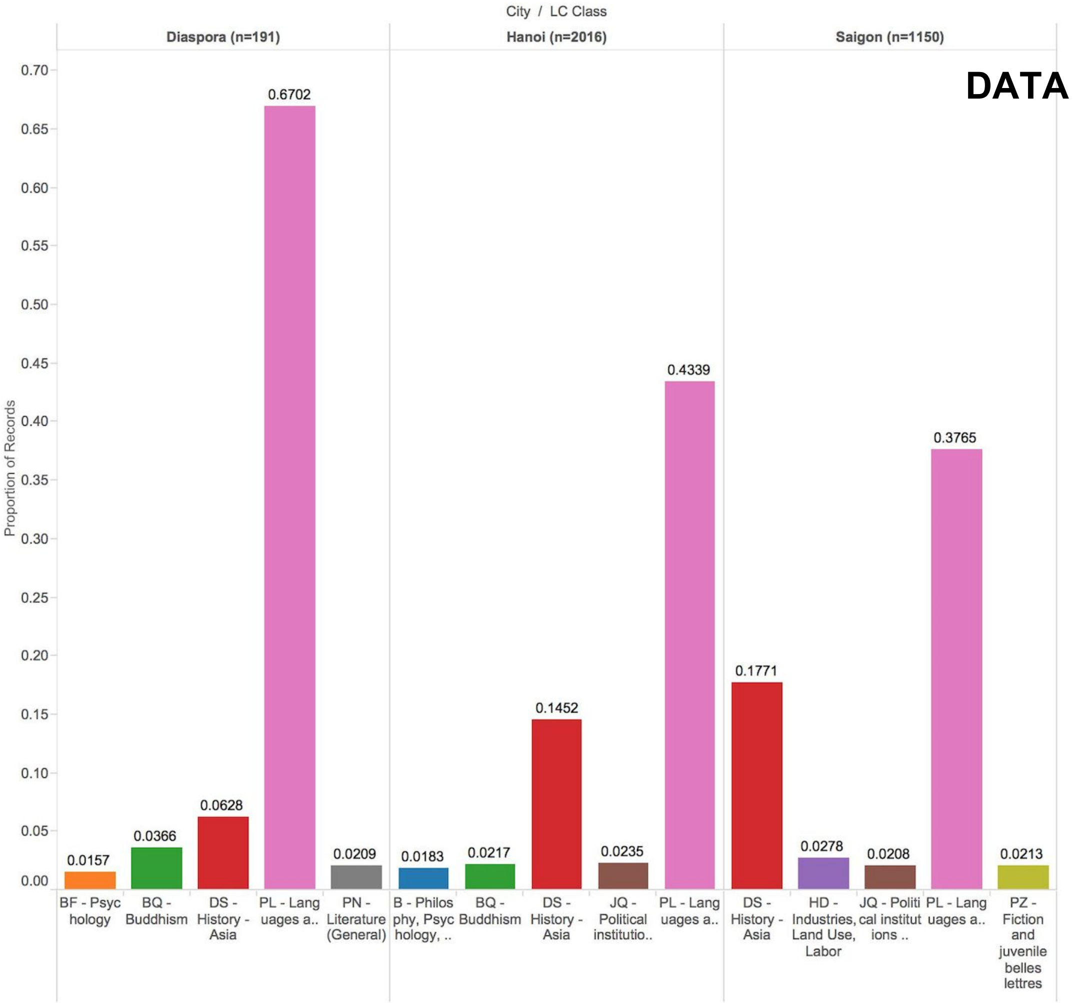
DATA

Library of Congress Classification (LCC) Breakdown by City

DATA

- **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

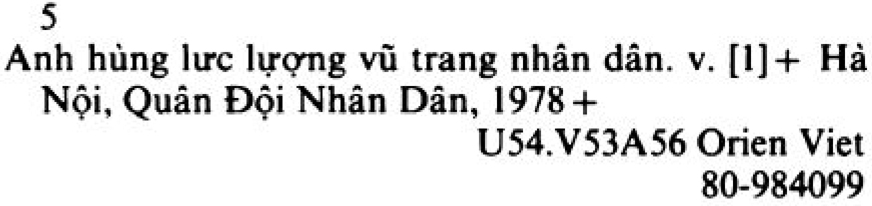
truyện_dài	0.0048038917604135	*long story*
truyện_ngắn	0.0007297050775311645	*short story*
giáo_dục	0.0006688963210702341	education
cách_mạng	0.0005472788081483734	***revolution***
công_dân	0.000486470051687443	***citizen***
văn_hóa	0.000486470051687443	culture
văn_học	0.000486470051687443	*literature*
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 lưc lượng vũ trang nhân dân”|Y=Hanoi) = 1.5648340863728223

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

True value =



5
Anh hùng lưc 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 ngấn 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 đẫt truyện_ngắn núi lửa đầt rửng lứa giữa nhiểu tác_giả thơ_văn đẫt 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 lẩn lần truyện_dài tái cúa xuẫt lãn bản đời xuẫt xuât khanh xuầt hoang lấn tựa phụ hậu Number, times, long story, published	**Textual format, republished, genre**
78	0.05733	nam việt miễn tại cộng_sản phong_trào bắc phụ_nữ công_đoàn câu nứa tiềng chương hiện_đại khoa_học_xã_hội tra chù thám lâp 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 bấc 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 vãn chú tường hữu lâm bùi ngữ biên Literature, introduction, language	**Literature, language**
65	0.05164	thuyềt tiểu tiếu thuyẽt thuyèt tiéu người gián_điệp thuyêt rừng lịch_sử thuybt tiẻu 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 điều 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

Concluding Remarks

1. Power of counting
2. Power of probability

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

Possible Ways to Featurize Data

Future Directions

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?

Reading by Machines: Open Refine on OCR Output Data

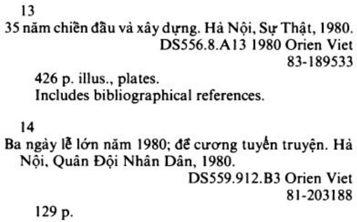
Cindy A. Nguyen, History
University of California, Berkeley

Data Cleaning

DATA

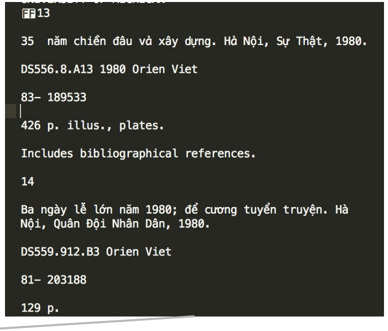
•Data cleaning and preparation:

- **Optical Character Recognition** (Abbyy Finereader): from PDF to text file
- **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]

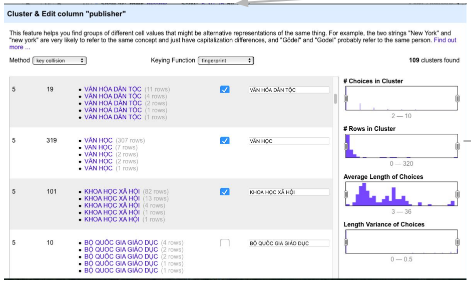


13
35 năm chiến đấu và xây dựng. Hà Nội, Sự Thật, 1980.
DS556.8.A13 1980 Orien Viet
83-189533
426 p. illus., plates.
Includes bibliographical references.

14
Ba ngày lễ lớn năm 1980; đề cương tuyển truyện. Hà Nội, Quân Đội Nhân Dân, 1980.
DS559.912.B3 Orien Viet
81-203188
129 p.



```
FF13

35  năm chiến đâu vả xây dựng. Hà Nội, Sự Thật, 1980.

DS556.8.A13 1980 Orien Viet

83- 189533

426 p. illus., plates.

Includes bibliographical references.

14

Ba ngày lễ lớn năm 1980; để cương tuyển truyện. Hà
Nội, Quân Đội Nhân Dân, 1980.

DS559.912.B3 Orien Viet

81- 203188

129 p.
```



Cluster & Edit column "publisher"

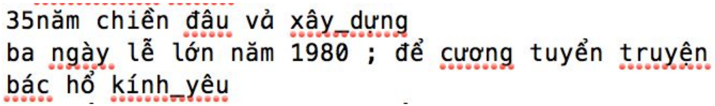
This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more ...

Method: key collision — Keying Function: fingerprint — **109** clusters found

5	19	• VĂN HÓA DÂN TỘC (11 rows) • VĂN HÓA DÂN TỘC (4 rows) • VĂN HOÁ DÂN TỘC (2 rows) • VĂN HOÁ DÂN TỘC (1 rows) • VĂN HÓA DÂN TỘC (1 rows)	☑	VĂN HÓA DÂN TỘC
5	319	• VĂN HỌC (307 rows) • VAN HỌC (7 rows) • VÀN HỌC (2 rows) • VÂN HỌC (2 rows) • VÁN HỌC (1 rows)	☑	VĂN HỌC
5	101	• KHOA HỌC XÃ HỘI (82 rows) • KHOA HỌC XÀ HỘI (13 rows) • KHOA HỌC XÂ HỘI (4 rows) • KHOA HỌC XÃ HỘI (1 rows) • KHOA HỌC XA HỘI (1 rows)	☑	KHOA HỌC XÃ HỘI
5	10	• BỘ QUỐC GIA GIÁO DỤC (4 rows) • BỘ QUOC GIA GIÁO DỤC (2 rows) • BỘ QUỐC GIA GIÁO DỤC (2 rows) • BỘ QUỐC GIA GIÁO DỤC (1 rows) • BỘ QUOC GIA GIÁO DỤC (1 rows)	□	BỘ QUỐC GIA GIÁO DỤC

35 years of war and building

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



```
35năm chiến đâu vả xây_dựng
ba ngày lễ lớn năm 1980 ; để cương tuyển truyện
bác hổ kính_yêu
```

Open Refine

OpenRefine (formerly Google Refine) is an open powerful tool for working with messy data: cleaning it; transforming it from one format into another; and extending it with web services and external data.

Why?

Exploring, finding, fixing inconsistencies due to typos and OCR output

Undo/Redo (track changes): freedom to make mistakes

Tutorial Resources

Programming Historian

http://programminghistorian.org/lessons/cleaning-data-with-openrefine

Videos & Community Board (Install Google Refine 2.5 - Stable)

http://openrefine.org

Digital Humanities & Open Refine (Today's walkthrough)

http://thomaspadilla.org/dataprep/

Data

https://github.com/cindyanguyen/tutorial-data

Favorite Operations

1. Text Facet/Numeric Facet: Locating inconsistencies in OCR, typos (sort by count/name)
2. Clustering: Helpful with diacritical errors in Vietnamese
3. Google Refine Expression Language(GREL): `value.replace('||', '|')`
   a. https://github.com/OpenRefine/OpenRefine/wiki/GREL-Functions
4. Common error: multiple facets/filters layered

First Steps:

1. Remove trailing whitespace
2. Capitalization: all lower case

Resources

Open Refine Lessons for Digital Humanities

https://data-lessons.github.io/dh-openrefine/

Tidy Data by Hadley Wickham - Datasets easy to manipulate, model, visualize, have specific structure: each variable is a column, each observation is a row, and each type of observational unit is a table.

https://vita.had.co.nz/papers/**tidy-data**.pdf

Digital Humanities x Data Science Short Reading List

https://cindyanguyen.com/2016/04/09/digital-humanities-data-science-list/