Distant Reading Methods for Historical Chinese Newspaper

...

The Case of Reading *Tianyi* with Topic Modeling

XIE Jia
26.01.2023

What is Distant Reading

Distant reading is an approach from digital literary studies in which **computational methods** are applied to often large amounts of text data without reading the texts themselves. The focus here is usually on **quantitative analyses**, but qualitative metadata can also be compared quantitatively. The term was coined by Franco Moretti (2000) in particular as a counter term to close reading.

(translated from fortext.net)

Distant Reading is **a supplement to the close reading** and allows us to interpret our extensive material as comprehensively as possible.

(XIE Jia, ongoing dissertation)

How to do Distant Reading

preparation: data (full text, word tokenization, user dictionary, ...)

processing: DH methods (topic modeling, text mining, ...)

presentation: visual outcomes (graphs), interpretation

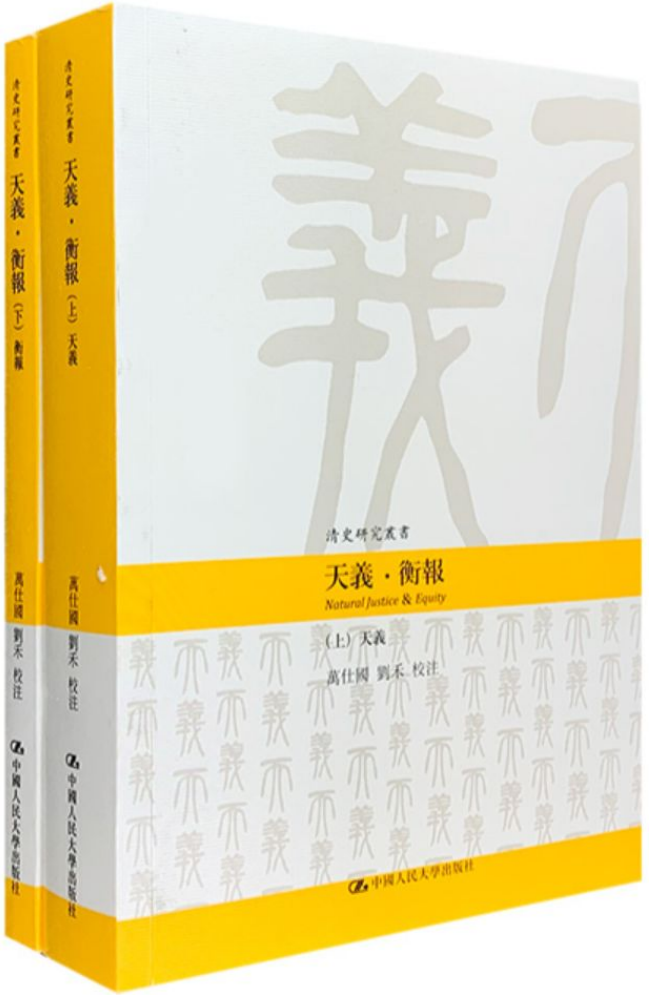
A case study on *Tianyi* 天義

Tianyi 天義

1907 - 1908 Tokyo and Shanghai

an anarchist publication

Tianyi supported anarchism and political, economic, and racial revolution, calling for women's emancipation and utopian socialism.



Tianyi 天義

full text data (in notepad)

total length: 571,661

total characters: 359,350

average length of sentence: 5.44

(Corpus Grams Tool from DocuSky)

社說
女子宣布書
載《天義》第一號‵‵社說〞欄，一九〇七年六月十日， 頁
一至七，題下注‵論著一〞，署‵震达〞；‵目錄〞標題同，署
‵彳可十〞故象。

嗚呼！世界之男女，其不平等也久矣。印度之女， 自焚以殉男。曰本
之女，卑屈以事男。歐美各國，雖行一夫一妻之制，號爲平等，然議政之
權，選擧之權，女子均鮮得干預。所謂平權者，果安在邪？更反觀之吾中
國，貝IJ男子之視女子也，幾不以人類相待。上古之民，戰勝他族，則係縶其
女，械繫其身，以爲妃妾。由是，男爲主而女爲奴，是爲剽掠婦女之時代。
繼因剽劫易起爭端，乃創爲儷皮之禮，故古禮所言納采、納徵，均沿財昏之
俗，蓋視女子爲財產之一也。由是，男爲人而女爲物，是爲買賣婦女之時代。
積此二因，由是男女之間遂不平等。今卽古制可玫者言之，厥有四事。
一曰嫁娶上之不平等。古代之時，位愈尊者妻愈衆。如殷代之制，天
子娶十二女，諸侯娶九女，大夫三女，士二女。至於周代，則爲天子者有
一后、三夫人、九嬪、二十七世婦、八十一御妻，豈非以百餘之女匹一男
子邪？而後世之嬪妃，則更無限制。貴顯之家，蓄妾尤衆。其不平者，一也。
二曰名分上之不平等。男權旣伸，其防範女子亦日嚴，創‵一與之
齊，終身不改〞《禮記》。之說，使女子終事一夫。又謂夫尊妻卑，夫猶天
而妻猶地，妻不去夫，猶地不得去天。《白虎通》說。由是，爵則從夫，姓
則從夫，而謚亦從夫，以女子爲男子附屬物。宋人因之，遂有扶陽鋤陰
之論。其不平者，二也。
三曰職務上之不平等。中國‵婦〞訓爲‵服〞，象持帚之形。而《禮
記•曲禮篇》亦言：‵納女于諸蔍曰備酒漿，於大夫曰備洒掃。〞是古代
之婦人，僅以服從爲義務。又創爲女子不逾閾之說，以禁其自由。後世
以降，爲女子者，舍治家而外無職務，以有才爲大戒，以卑屈爲當然。
其不平者，三也。
四曰禮制上之不平等。夫之於妻，僅服期喪；而妻之於夫，則服喪
三年。非惟爲夫服重喪也，卽夫之父母，亦爲之服斬衰；於己之父母，
轉降爲齊衰，非所謂厚於所薄、薄于所厚者邪？且古代之時，父存母歿，
爲母服齊衰，尤爲失理之尤。其不平者，四也。
略擧四端，則男子之壓制女子，昭昭明矣。夫以男陵女，猶可言也。
女子而甘於自屈，抑獨何心？豈非社會之習慣、腐儒之學術有以箝制之
邪？吾今以一語告女界同胞：男子者，女子之大敵也。女子一日不與男
子平等，則此恨終不磨。試將女界所應爭者，分列如左。
一曰實行一夫一妻之制。如男子不僅一妻，或私蓄妾御，性好冶游

Tianyi 天義

(Corpus Grams Tool from DocuSky)

No.	Gram	DocFreq	TermFreq
1	社會	1	1285
2	政府	1	997
3	主義	1	905
4	女子	1	882
5	天義	1	757
6	中國	1	572
7	革命	1	544
8	無政	1	495
9	人民	1	402
10	自由	1	349
11	會主	1	346
12	日本	1	343
13	衡報	1	336

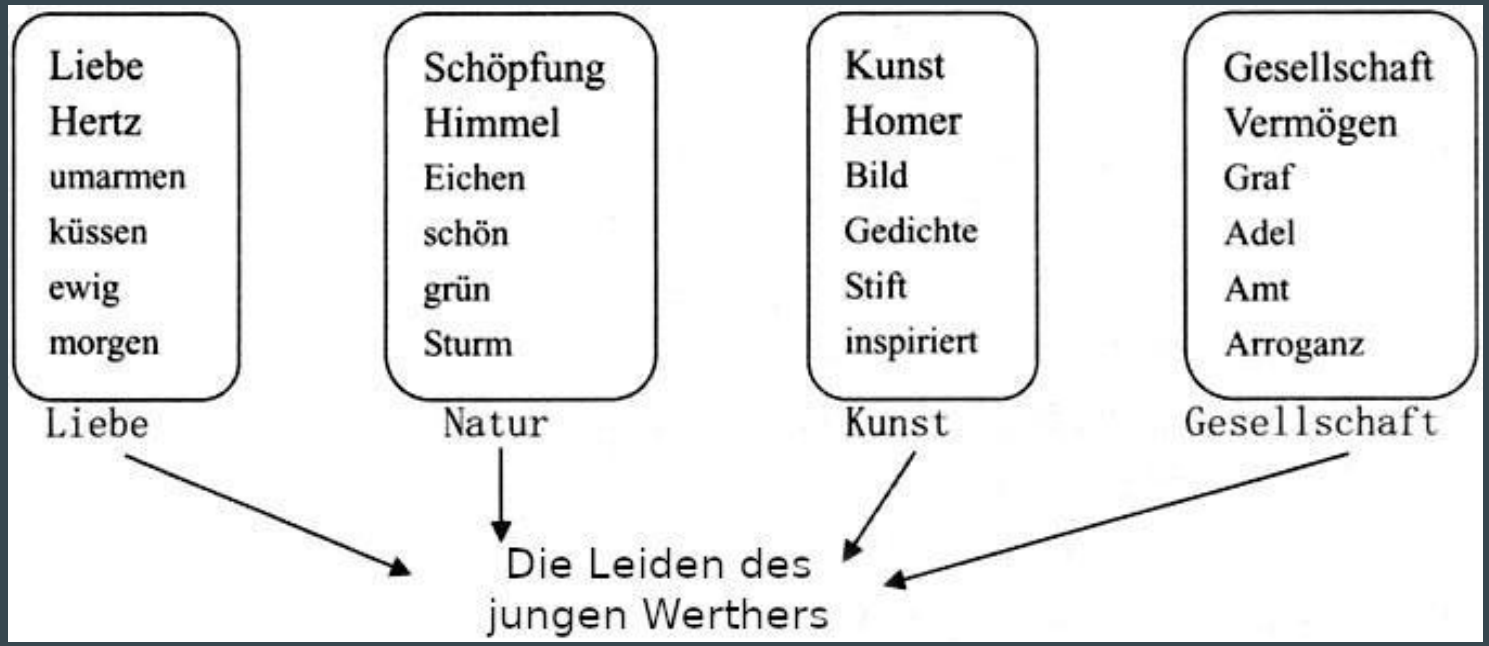
How to do Distant Reading

Topic Modeling :

Put in simpler terms, a topic model is a computer aided program that from a text generates ‘topics’ or ‘themes’: strings of words that are supposed to be indicative of themes addresses within the text. The basic idea is that words that cluster ‘closely’ share a meaningful connection, i.e. a ‘topic’, ‘theme’ or ‘motif’ of a text, which in lay terms could be understood as the ‘important’ or ‘signicant’ key words of sharedtheme.

Fridlund, Mats & Brauer, Rene. (2013). ”Historizing topic models: A distant reading of topic modeling texts within historical studies”.

Topic Modeling



Matt Erlin, ed., *Distant Readings. Topologies of German Culture in the Long Nineteenth Century*, New York: Camden House, 2014, p.59.

Topic Modeling

Latent Dirichlet Allocation (LDA) 2003

- a machine learning algorithm that uses the principles of statistics and probability,.
- application: Marketing, Accounting, Management, Multimedia studies,

  Cultural analytics, Historical studies, ...

Blei, David M., Andrew Y. Ng und Michael I. Jordan (2003): „Latent Dirichlet Allocation“. In: *Journal of Machine Learning Research*. 3, 993–1022.

Topic Modeling

Latent Dirichlet Allocation (LDA)

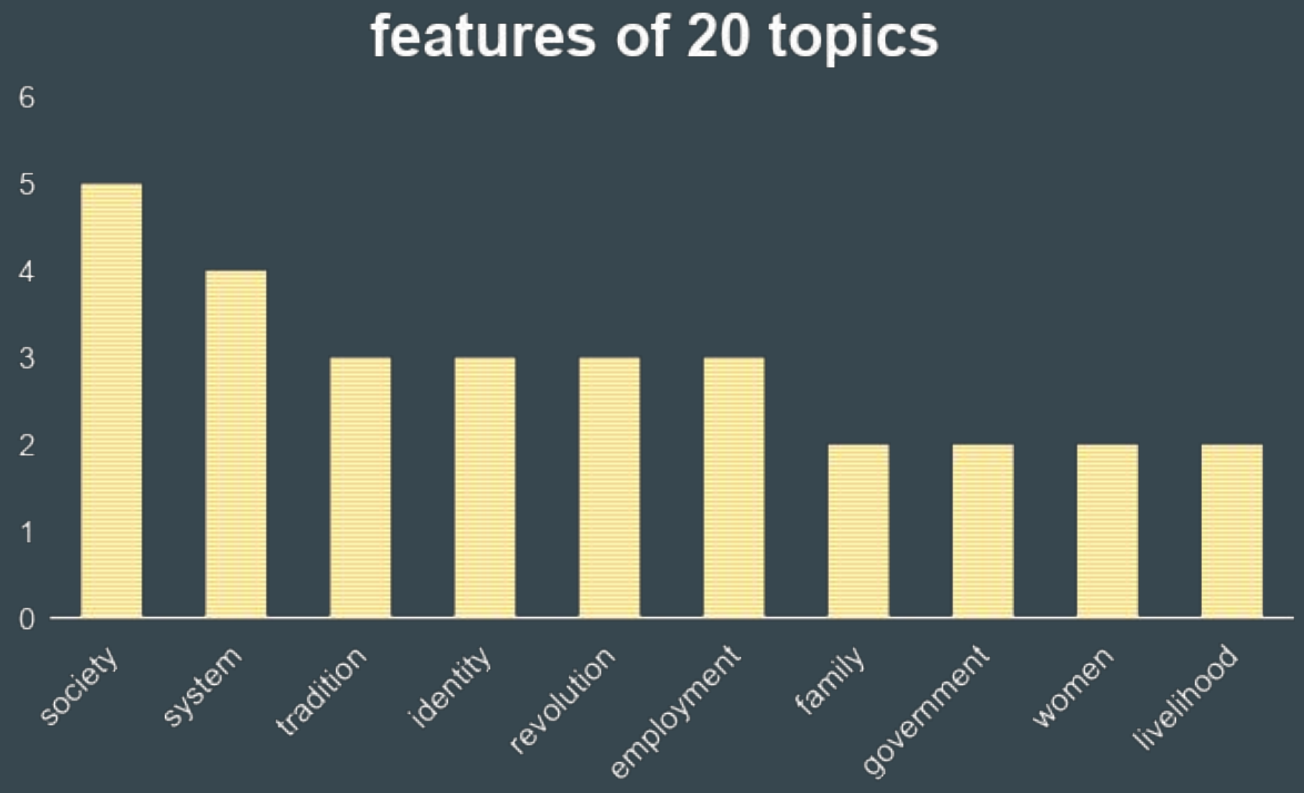
- Source: full text of *Tianyi*
- Topic number = 20
- Word number = 50

Blei, David M., Andrew Y. Ng und Michael I. Jordan (2003): „Latent Dirichlet Allocation“. In: *Journal of Machine Learning Research*. 3, 993–1022.

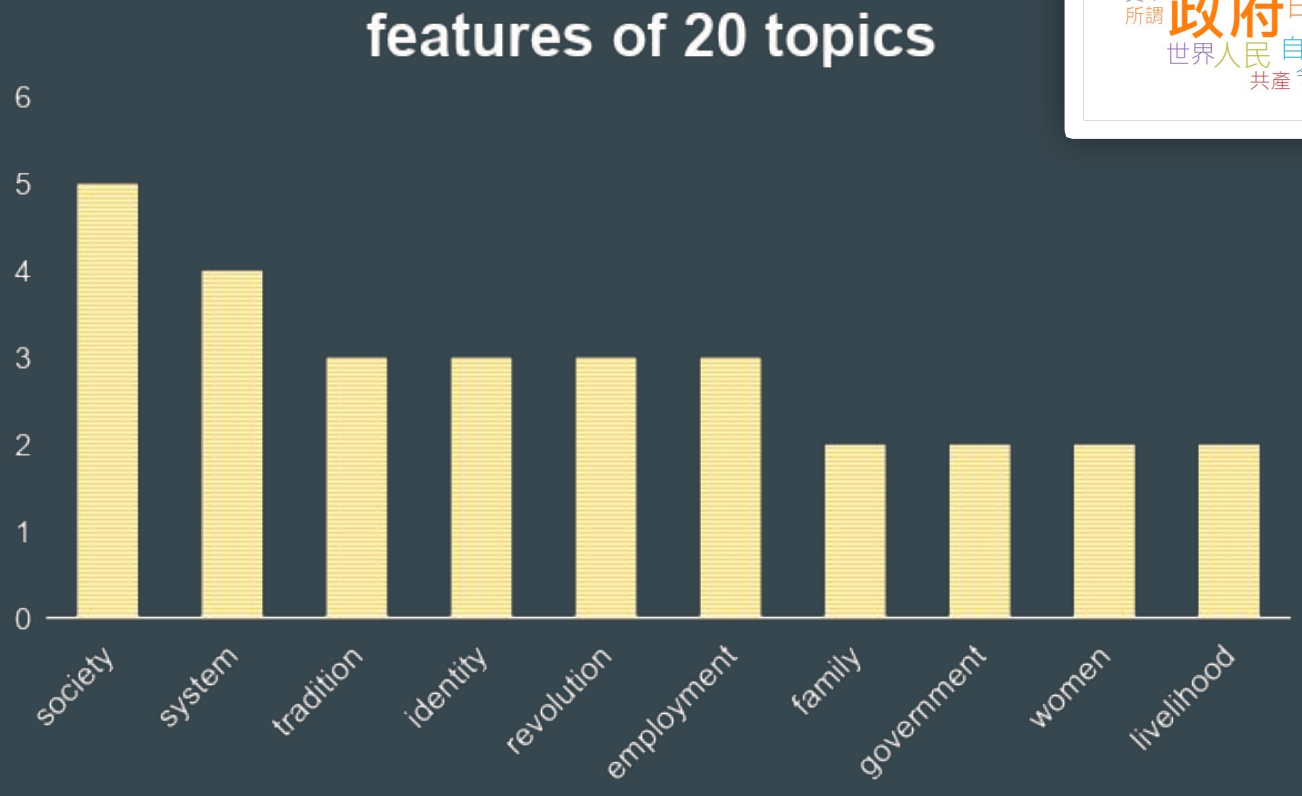
topic1	topic2	topic3	topic4	topic5	topic6	…
印度	女子	民族	富民	政府	社會	
獨立	男子	特權	貧民	人民	自由	
亞洲	男女	革命	田主	社會	吾人	
日本	婦人	滿人	貧富	政治	制度	
美洲	家庭	異族	富者	中國	人類	
白人	中國	中土	農民	自由	組織	
奴婢	平等	大同	貧者	世界	私有	
波斯	自由	吾人	富人	平民	進化	
安南	解放	政府	公司	日本	平等	
法人	結婚	中國	制度	階級	生活	
世界	父母	受制	佃民	各國	中心	
强權	婦女	少數	吳下	革命	財產	
排斥	婚姻	種族	疾苦	多數	感情	
英人	世界	滿洲	人民	文明	自然	
朝鮮	奴隸	梵文	中國	主義者	文明	
主人	女界	未有	美國	歐美	苦氏	
人民	夫婦	革命者	不均	共和	近世	
弱種	之權	自利	慈善	目的	互助	
中國	子女	以求	女史	國家	扶助	

Colony	Gender	Nation	Economics	Government	Society	…
印度	女子	民族	富民	政府	社會	
獨立	男子	特權	貧民	人民	自由	
亞洲	男女	革命	田主	社會	吾人	
日本	婦人	滿人	貧富	政治	制度	
美洲	家庭	異族	富者	中國	人類	
白人	中國	中土	農民	自由	組織	
奴婢	平等	大同	貧者	世界	私有	
波斯	自由	吾人	富人	平民	進化	
安南	解放	政府	公司	日本	平等	
法人	結婚	中國	制度	階級	生活	
世界	父母	受制	佃民	各國	中心	
强權	婦女	少數	吳下	革命	財產	
排斥	婚姻	種族	疾苦	多數	感情	
英人	世界	滿洲	人民	文明	自然	
朝鮮	奴隸	梵文	中國	主義者	文明	
主人	女界	未有	美國	歐美	苦氏	
人民	夫婦	革命者	不均	共和	近世	
弱種	之權	自利	慈善	目的	互助	
中國	子女	以求	女史	國家	扶助	

Distant Reading



Distant Reading

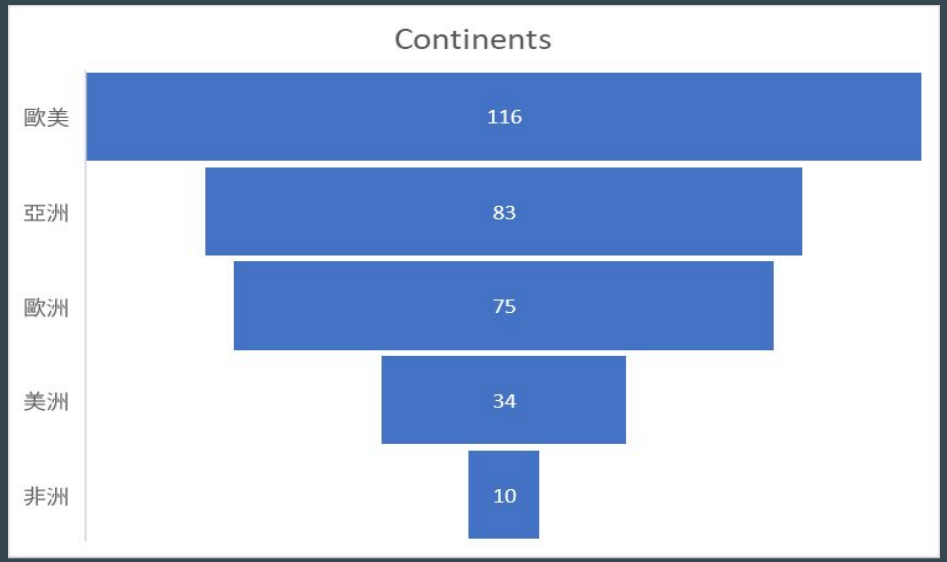


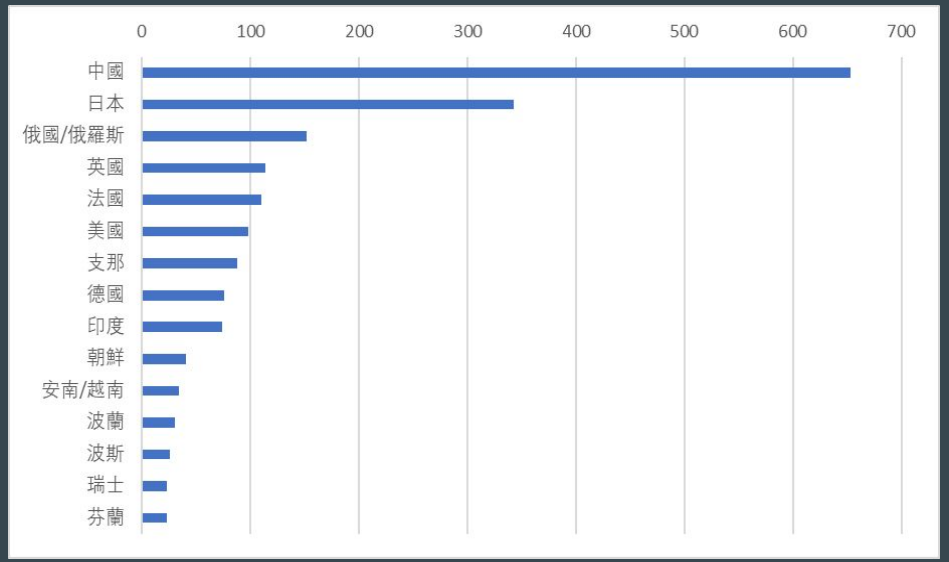
Distant Reading

After further organizing, we can get

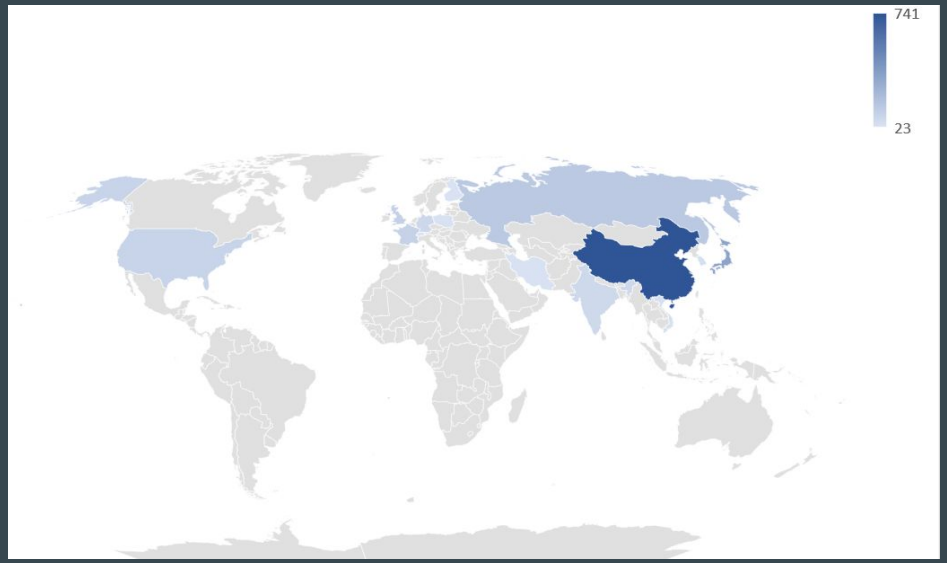
- **People:** Marx, Kōtoku Shūsui (幸徳秋水), Qiu Jin (秋瑾), Kropotkin, Confucius, Ba Jin (巴金), Bakunin, ...
- **Nations:** China, Japan, UK, US, France, Germany, Swiss, Russia, Joseon, India, Persia, ...
- **Cities:** Tokyo, Shanghai, New York, Paris, Siberia, ...
- **Social identities:** servant, prostitute, worker, emperor, merchant, soldier, common, ...
- **Thoughts:** Communism, Socialism, Anarchism, Da Tong (大同), Evolution, Mutual Aid, ...

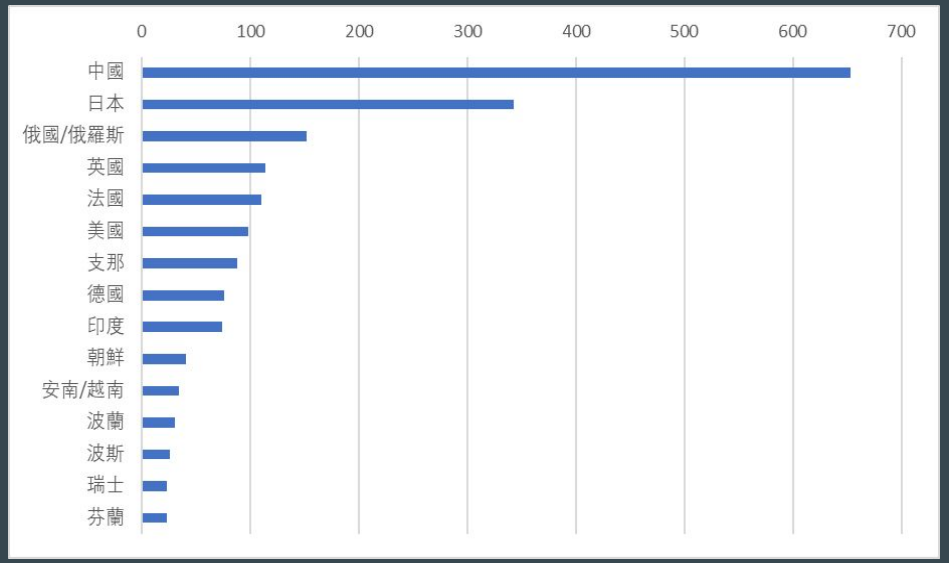
Distant Reading





Distant Reading





Distant Reading

Statistics imply:

Anarchists pay close attention to Asian countries like India, Joseon, Vietnam, Persia, ...

and lack of the attention to Africa.

Distant Reading & Close Reading

波斯、中國、朝鮮之暗殺，亦隱與無政府黨暗符。是則數年之內，社會主義、無政府主義必爲亞洲所通行。此誠亞洲之大幸，而吾輩所可預測者也。

波斯雖設議會，然國都騷亂日甚。首相既遭暗殺，而國王亦以被戕著聞。中國人民，有倡民族主義者，有倡共和主義者，暗殺、暴動之事亦層見叠出。蓋現今亞洲諸弱種，受制强族之下者，既曰思脫離其羈絆；卽受制君主、官吏之下者，亦思脫離虐政，以伸民氣。此均亞洲諸弱種不甘受壓抑之證也。

“Asian Perspective”

Li-ching Chou 周麗卿, “The Encounter of Revolutionary Movements and the Asiatic Perspective of Late-Qing China: A Case Study of the ‘Asiatic Humanitarian Brotherhood’” (清末革命運動與亞洲觀的遇合：以「亞洲和親會」為個案的討論), *21 Century* 二十一世紀 146 (2014): 17-27

Conclusion

Advantages of Topic Modeling:

1. fast and efficient
2. statistics supporting
3. likely to discover hidden clues

“Note that the statistical models are meant to help interpret and understand texts; it is still the scholar’s job to do the actual interpreting and understanding.”

David M. Blei, "Topic Modeling and Digital Humanities", *Journal of Digital Humanities,* vol 2, no. 1 (2012)

Reflection

Requirements:

1. well-produced and well-orginized full-text data
2. some linguistics and informatics knowledge

Usability:

3. contextless, logocentrist
4. textual content

Debate

Ying Bai, Ruixue Jia, Jiaojiao Yang, Web of Power: How Elite Networks Shaped War and Politics in China, *The Quarterly Journal of Economics*, 2022;, qjac041, https://doi.org/10.1093/qje/qjac041

Critic (Taisu Zhang): the conclusion of the quantitative method is not different from the consensus of traditional research…

Defender (Ruixue Jia): do some issues need to be proven by data when there is a historical narrative? Is data proof more credible than a historical narrative?