HathiTrust+Bookworm

HathiTrust+Bookworm (HT+BW) visualizes word trends in 13.7 million works held by HathiTrust. It enables scholars to discover new textual use patterns across the entire corpus, including in-copyright and public domain volumes.

The world's great research libraries have, over time, carefully assembled a rich body of metadata pertaining to the books in their collections. Since the HTRC has access to volume-level metadata as well as volume-level content, we have constructed a Bookworm of the HathiTrust corpus in order to provide scholarly researchers with the means of exploring trends. John Unsworth has noted that a fundamental goal of the humanities is appreciation: "by paying attention to an object of interest, we can explore it, find new dimensions within it, notice things about it that have never been noticed before, and increase its value" (2004). Shifting from traditional close reading to a large-scale view of text presents a profound discomfort for humanities scholars, due to the difficulty in retaining the same sensitivity to what is actually contained in the works being studied. HTRC-Bookworm will function as a link between quantitative analysis (distant reading) and close reading. According to Frederick Gibbs and Daniel Cohen, "any robust digital research methodology must allow the scholar to move easily between distant and close reading, between the bird's eye view and the ground level of the texts themselves" (2011). This is what HTRC-Bookworm intends to accomplish, within the limitations of applicable copyright laws.

Using HT+BW

HathiTrust+Bookworm is available from the Explore page on HTRC Analytics. From that page, you can also find links to experiment "advanced" Bookworm interfaces.

Use HathiTrust+Bookworm Follow a tutorial

The data

HT+BW runs on top of HTRC Extracted Features data and represents snapshot of the HathiTrust when the Digital Library was at 13.7 million volumes. It performs best with modern English or European languages, and because of the way the input data was parsed to create the Extracted Features, it does not perform as well on non-Latin characters. As with the HathiTrust Digital Library, there are duplicative volumes represented in the data. The OCR has not been corrected and search is less accurate for volumes published prior to the nineteenth century.

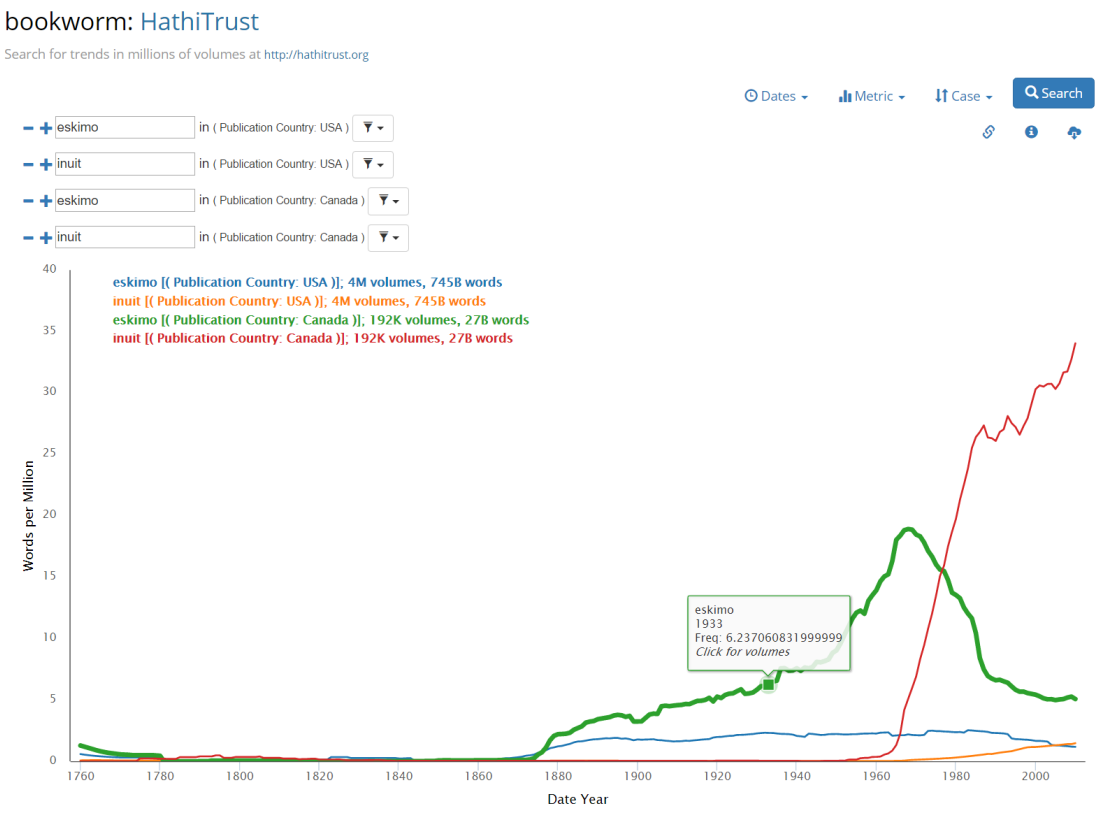
The metadata used to facet the search is pulled from the bibliographic metadata for the volumes. Some metadata fields are not universal for volumes in HathiTrust. For example, not every volume has a Library of Congress classification. Faceting based on those non-universal criteria will limit search to only volumes where that metadata field applies.

Examples

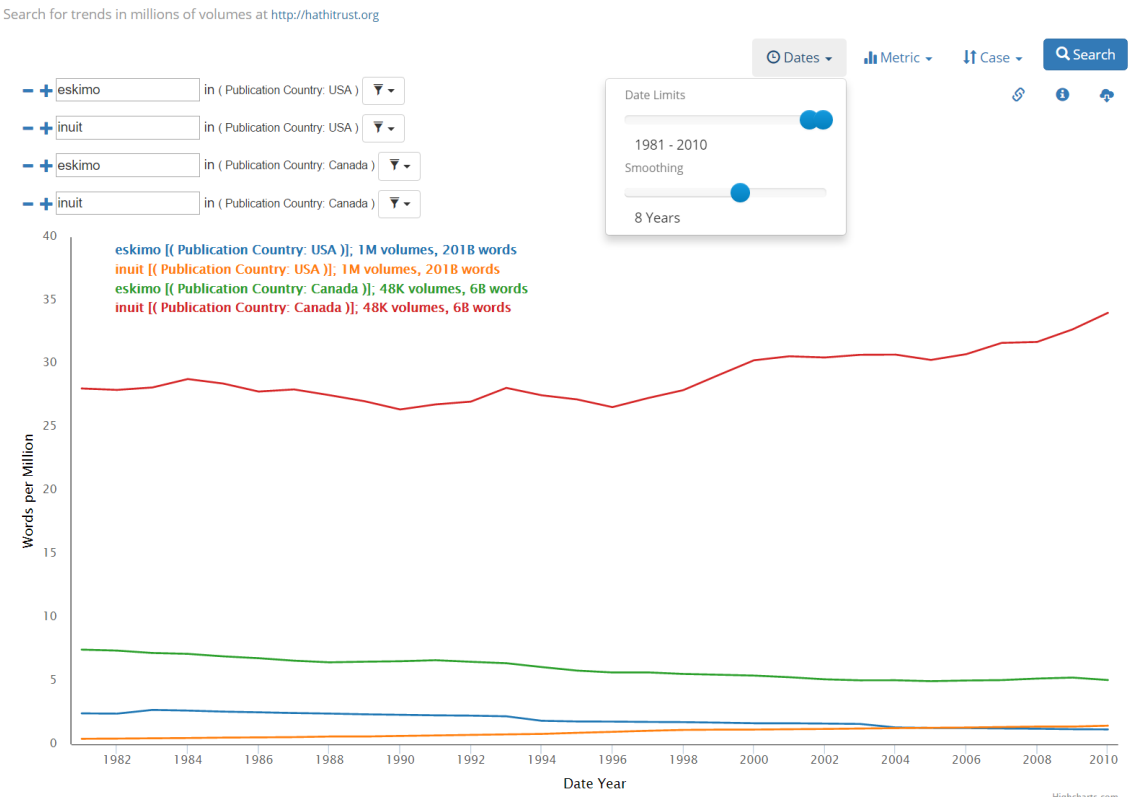
Inuit or Eskimo

Another typical word preference discussion is “Inuit or Eskimo”. The term "Eskimo" is commonly used in Alaska to refer to all Inuit people, however, this name given by non-Inuit people is considered offensive in many other places. Linda Lanz, a Ph.D. in linguistics from Rice University in Houston, claims that “In Canada, the term Inuit is preferred over Eskimo, which is considered offensive.” She sums it up with “Canada: Inuit; United States (i.e., Alaska): Eskimo.”

Doctor Lanz is not the only one who thinks so. Actually, her argument is in line with many linguistic studies. But “common opinions” are not necessarily the truth. To verify such theory, it is helpful to visualize word popularity of Inuit and Eskimo with Bookworm. According to the Bookworm output, for United States, Linda’s argument definitely makes sense in the period from 1880 to 1980. However, since 2000, the situation has changed. To further examine such change, we use the “Dates” settings on the panel to “zoom” in. Then we can see more details. For example, we can have a close look at the turning point when word counts of Inuit first surpassed that of Eskimo in the United States.

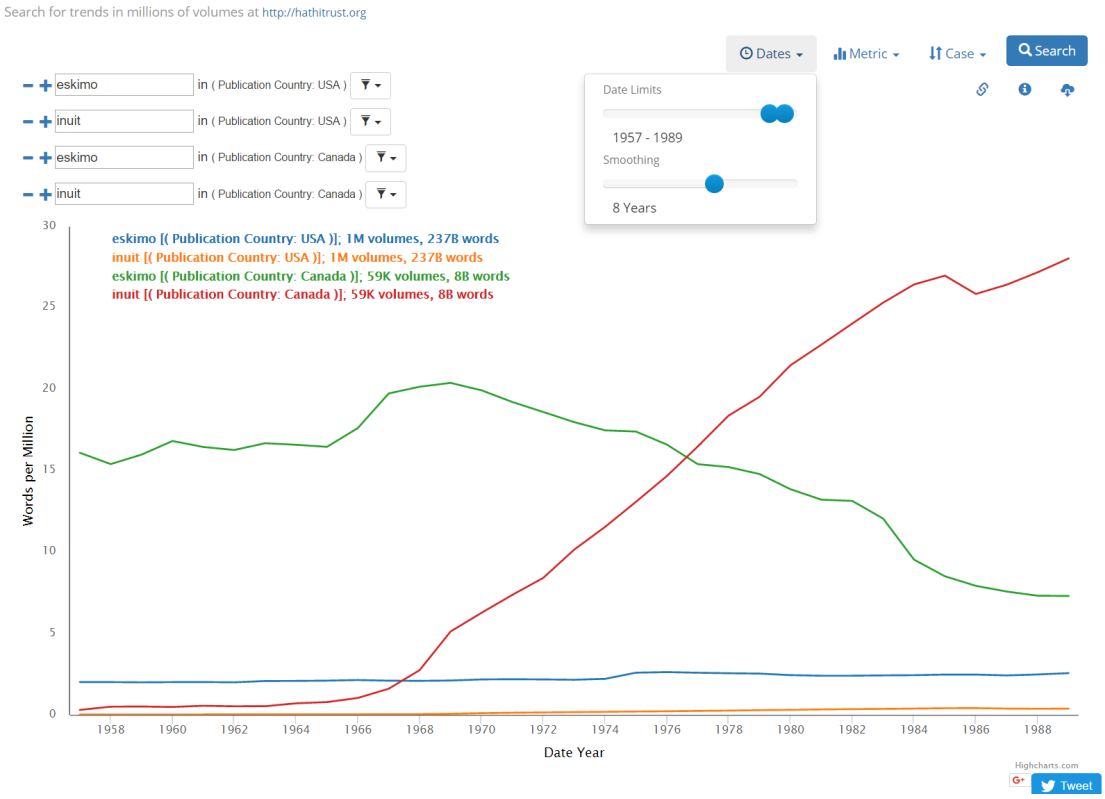


Word Popularity lines of Inuit VS Eskimo in United States and Canada from 1760 to 2000



Word Popularity Lines of Inuit VS Eskimo in United States and Canada from 1981 to 2010

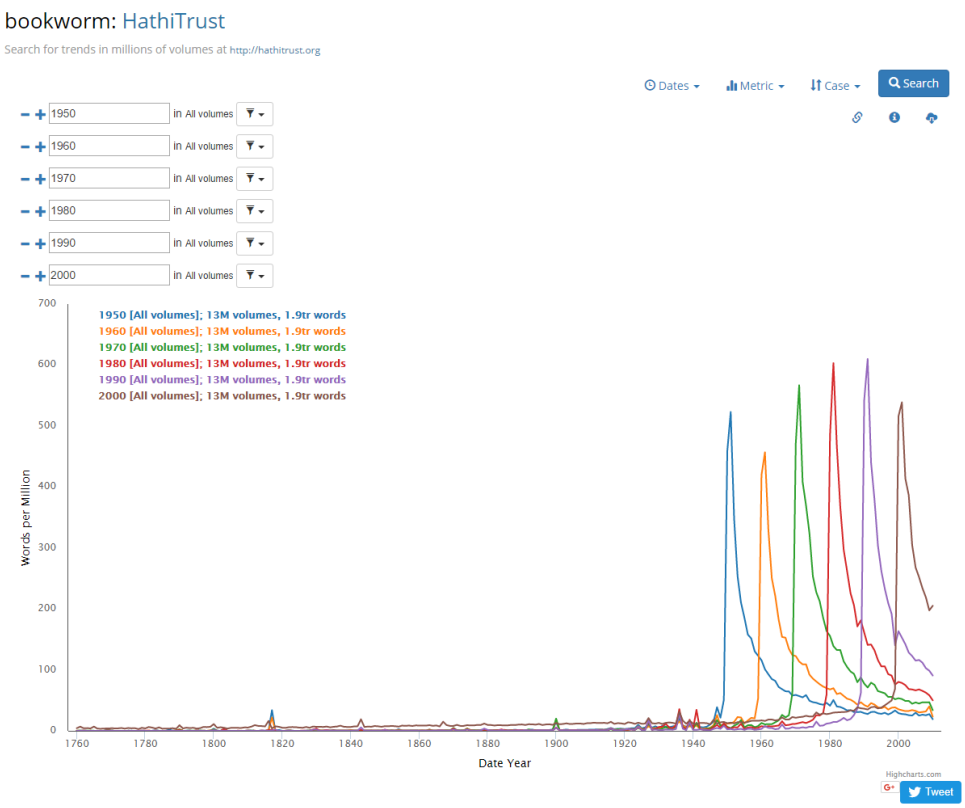
The word count lines showing Canadian publications is also dramatic, challenging the “common opinions” on the Inuit or Eskimo issue. As we zoom in, we can tell that for some reasons, the use of Inuit surged from 1968-1984 while the use of Eskimo dropped quickly at the same time. Why? What caused such changeover? Was it the same case in Canadian people‘s oral speaking? With such findings, we can raise more questions against the previous arguments and start our own research with new evidence.



Word Popularity Lines of Inuit VS Eskimo in United States and Canada from 1957 to 1989

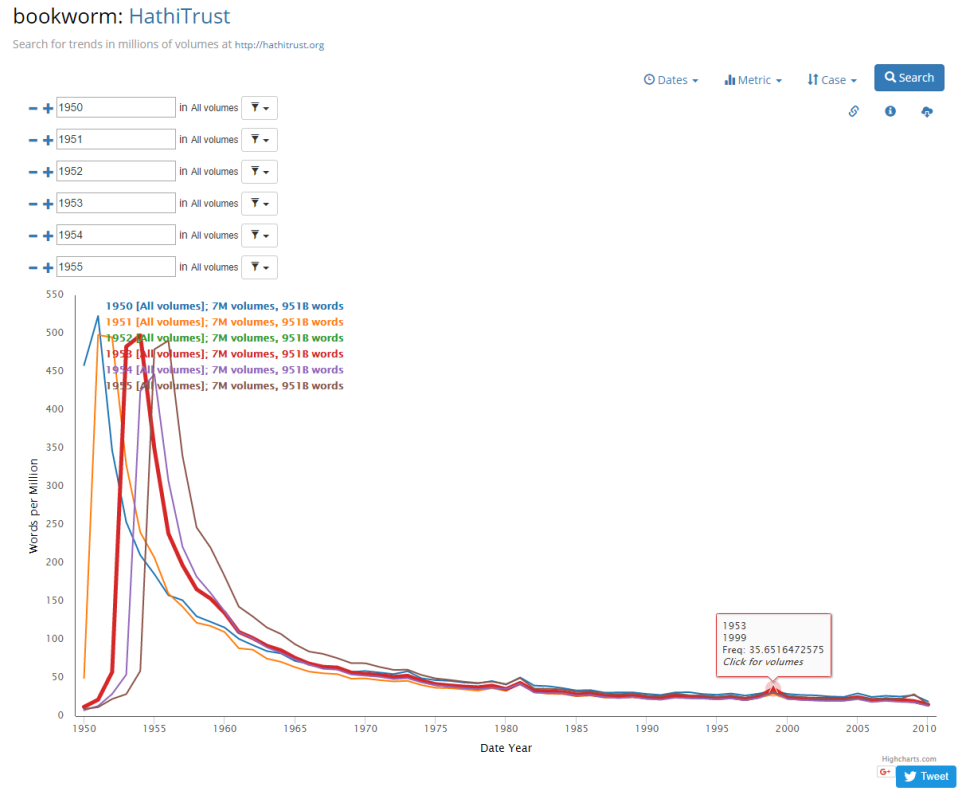
Popularity of years

Bookworm gives users a simple visualization platform that allow people to understand a large scale of data over many years in seconds. To make full use of this feature, we visualize the popularity of 6 single years: 1950, 1960, 1970, 1980, 1990, and 2000 with Bookworm to understand the varying popularity of these years. Learned from Bookworm, publications mentioned each of these years most frequently in 1 to 3 years after them. For example, the popularity of 1950 peaked in 1952. Once the popularity reached its peak, it kept shrinking in the following years.



Popularity of Years: 1950, 1960, 1970, 1980, 1990, 2000 (smoothing: 0 years

For a point of verification, we also plot another group of years: 1950, 1951, 1952, 1953, 1954, and 1955. Here is the visual.



Popularity of Years: 1950, 1951, 1952, 1953, 1954, 1955smoothing: 0 years

From this visual, we can see more clearly what the popularity of a single year is like. The popularity of year A always rockets up in 1-3 years after A itself. Then it decreases progressively. Decreasing rate is high at first and then tails off. Finally, the popularity maintains at a low level. Is this pattern universal? Does it reflect people's memory and oblivion of history?

Other interesting findings with Bookworm

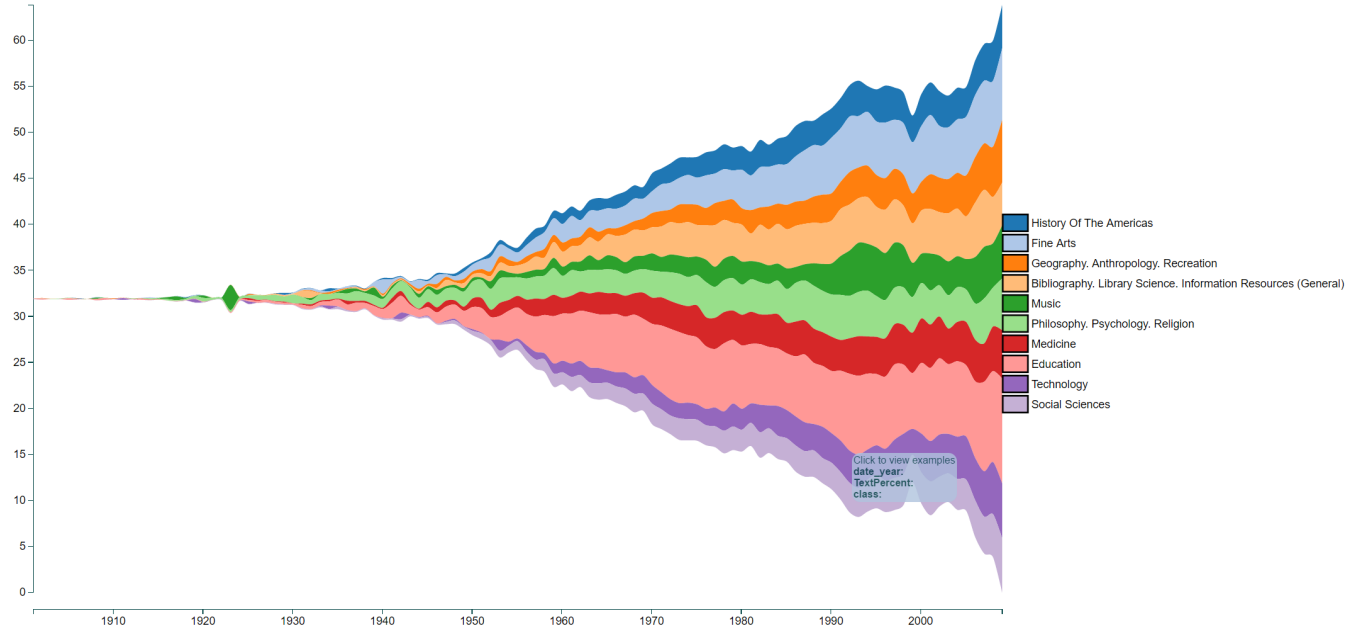
Each of them may shed light on several questions. For example:

What story does it tell for us? Is it attracting or persuasive?

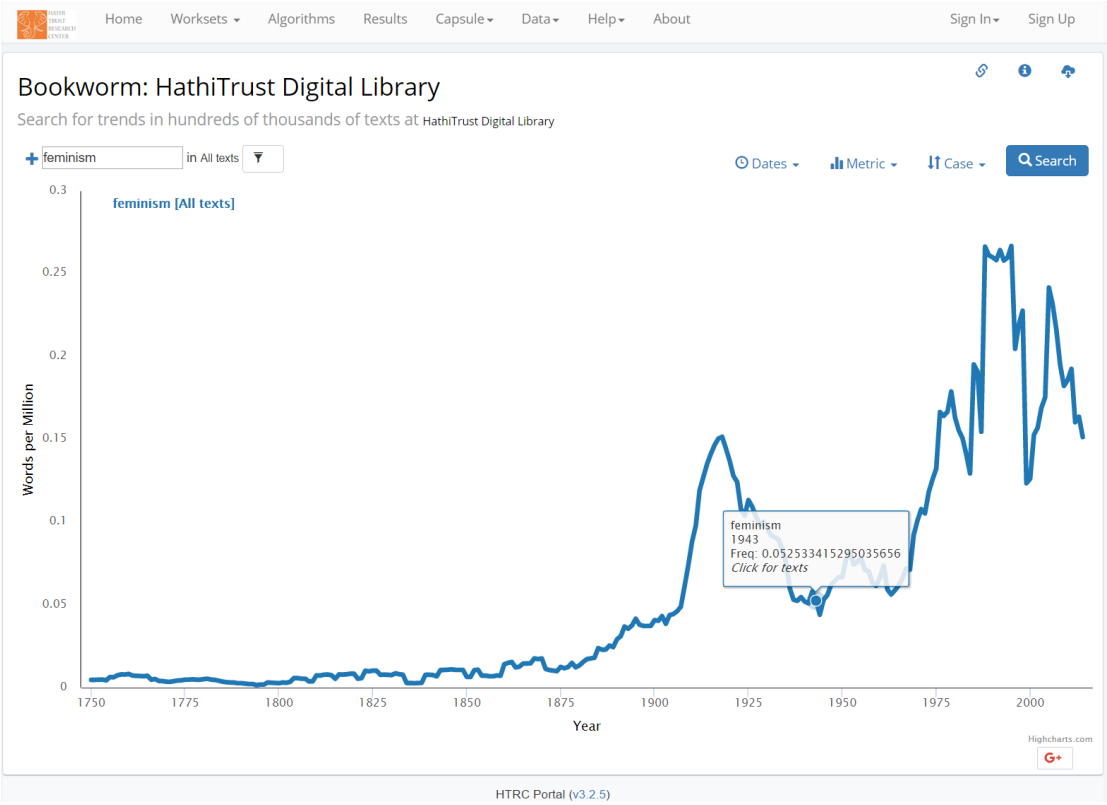
What are the reasons behind? How come the features of the visual?

What are the visible conclusions and what are the invisible bias?

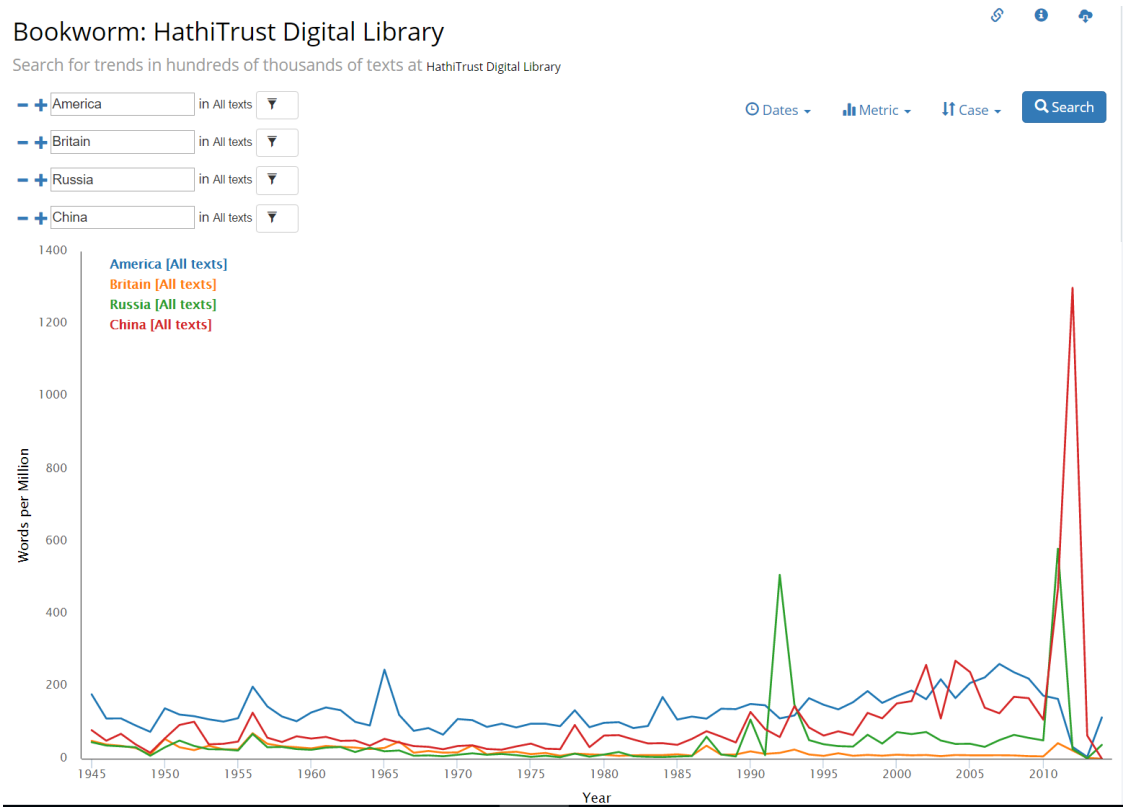
Streamgraph of "creativity" growth in the 20th century, by class



Word popularity of "Feminism" after 1750

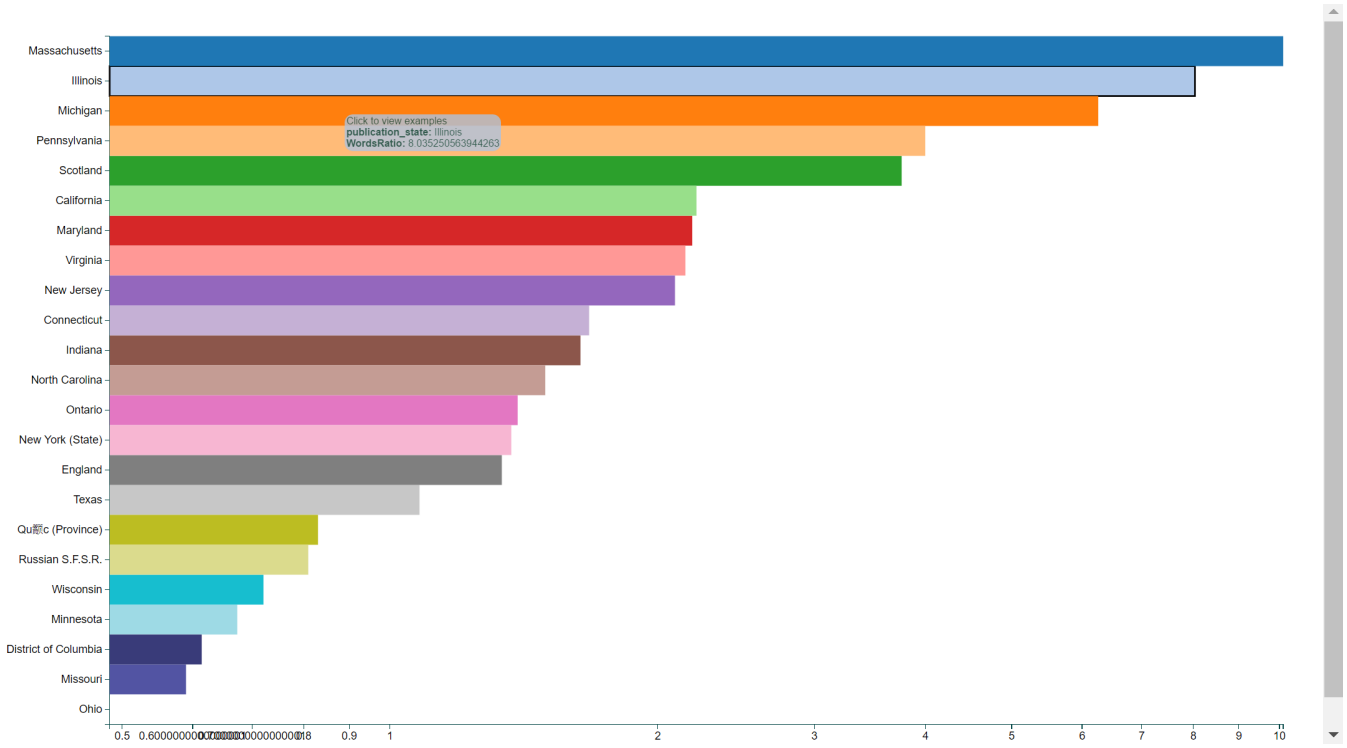


Word Popularity of 4 countries after 1945



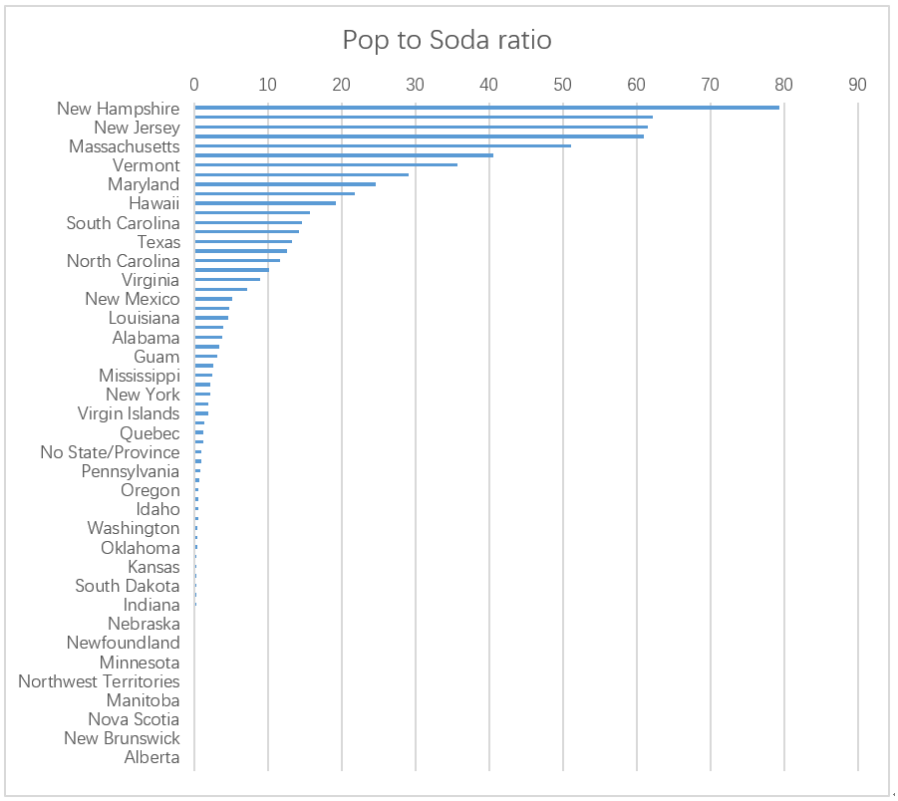
Soda vs Pop

Soda or *pop*? How do Americans in different places refer to their soft drinks? Besides a variety of scientific papers and journal articles arguing about it, the Pop vs Soda project plots the regional variations in the use of the terms "pop" and "soda" to describe soft drinks. Current statistics from the project are available at http://popvssoda.com/statistics/USA.html. Their statistics and mappings are interesting to read. However, what if we want to look back into the history and find the hidden statistics? Where can we get historical evidence for our question? What will the results look like if the statistics are based on authorized publications rather than people's voting online? Try Bookworm! With data extracted from millions of publications from 1940 to 2015, we visualized the" Soda to Pop Ratio" by state . The y-axis represents the publication states while the x-axis shows the word ratio of "soda" to "Pop". For example, we can see from the graph that publications in Massachusetts use *soda* for the soft drinks almost ten times the frequency of using *pop*.



Soda to Pop Ratio Graph

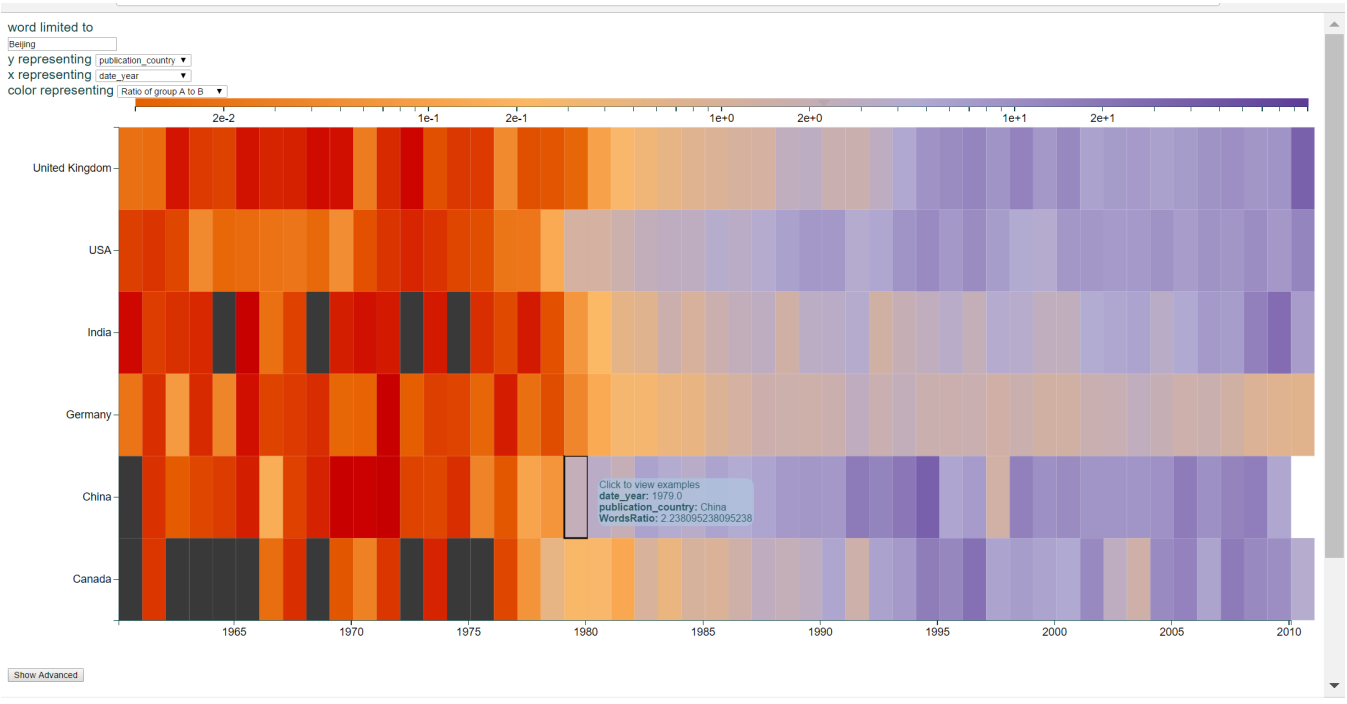
We can also visualize the Soda-to-Pop ratio with the statistics from the *Pop vs Soda* Page. Now you can look at a whole picture to find answers to more questions. What are the states where the word "Soda" always dominate in publications? Are publication language sharing the same words preferences for soft drinks as people's oral language do? Start your exploration with Bookworm.



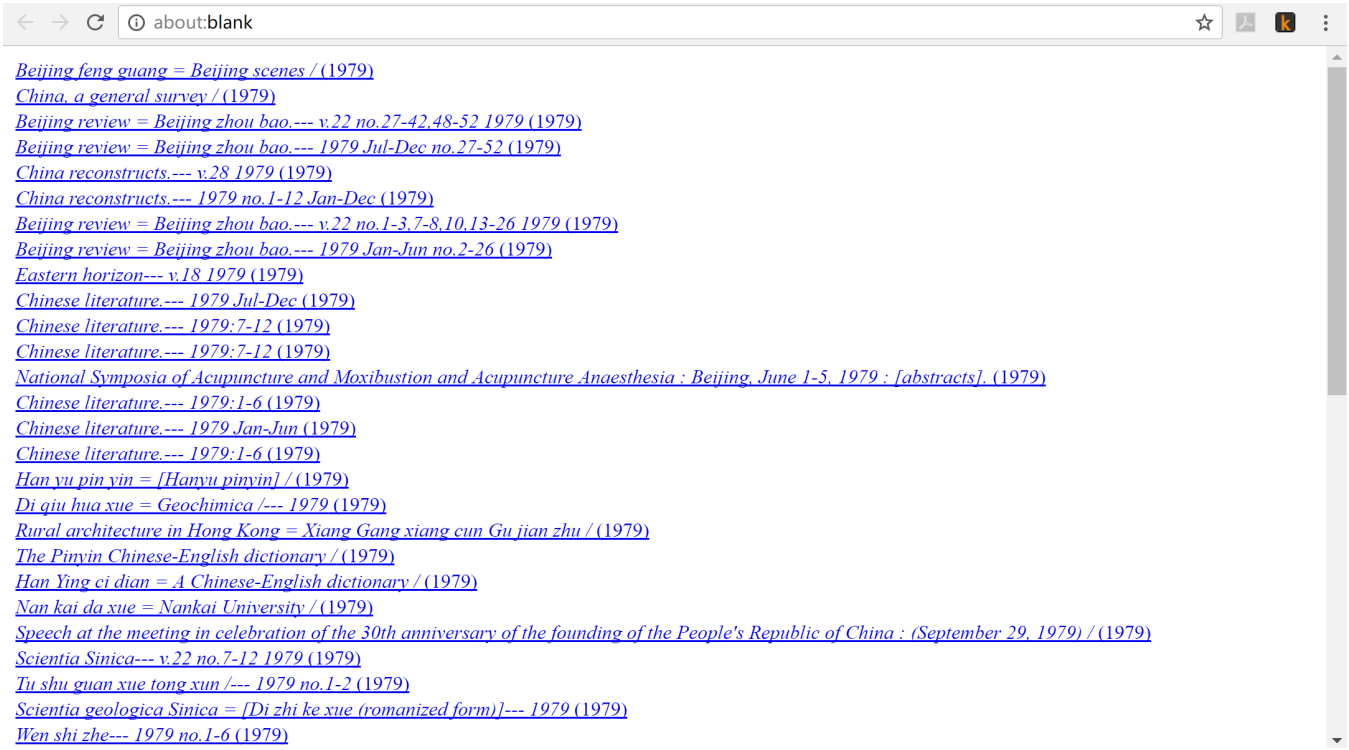
Soda to Pop Ratio Graph (Based on the Pop vs Soda Page statistics)

Peking to Beijing

What is the capital city of China? Interestingly, some would say Peking, while the others would say Beijing. Referring to the same city, Beijing is pretty close phonetically to the original Mandarin while Peking has been used for a longer time internationally. Some findings argue that the Chinese government is insistent on the more modern transliteration Beijing rather than Peking. What's more, they claim that with China's rapid development and increasing power, the trend of replacing Peking with Beijing grows. To further investigate this argument, we used Bookworm to find out the word usage in publications of six countries from the 1960s to 2010s. Then we generated a graph showing the log ratio of Peking to Beijing grouped by Country. The y-axis marks the publication country while x-axis shows the time of the publications. Blocks of different colors indicate different ranges of the ratio. Click on a block and you will find a list of related publications during a certain period in the country you pick. Try different settings and input various words, you will find more!



Log Ratio of Peking to Beijing (by country)



Part of the Lists of Related Publications

References

Michel, Jean-Baptiste, Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., ... & Aiden, E. L. "Quantitative analysis of culture using millions of digitized books." *Science*331.6014 (2011): pp. 176-182.

Unsworth, John. “Forms of Attention: Digital Humanities Beyond Representation," delivered at "*The Face of Text: Computer-Assisted Text Analysis in the Humanities,*" *The Third Conference of the Canadian Symposium on Text Analysis (CaSTA)*, McMaster University, November 19-21, 2004.

Gibbs, Frederick W., and Daniel J. Cohen. "A Conversation with data: prospecting Victorian words and ideas." *Victorian Studies*, Vol. 54, No. 1 (2011): pp. 69-77.