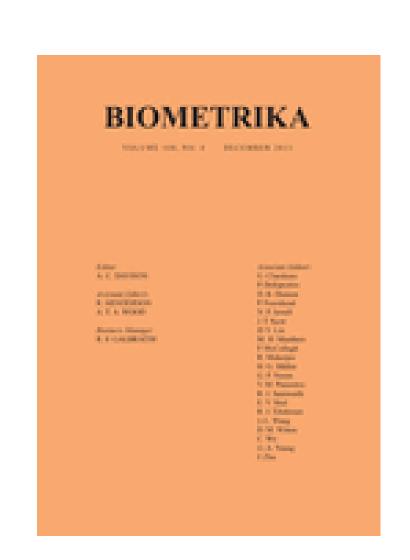
Text mining JSTOR

Quantitative approaches to histories of science

Ben Marwick & Ian Kretzler, UW Anthropology

Google's Ngram Viewer is a tool for visualizing the popularity of words over time in 5 million books digitized by Google. It has been described as the 'gateway drug' for data science in the humanities. While it is fun for casual searches, it has substantial limitations that prevent it from being a scholarly tool. The main problem is we simply don't know what is in the corpus. This makes validation through close reading of the texts impossible.

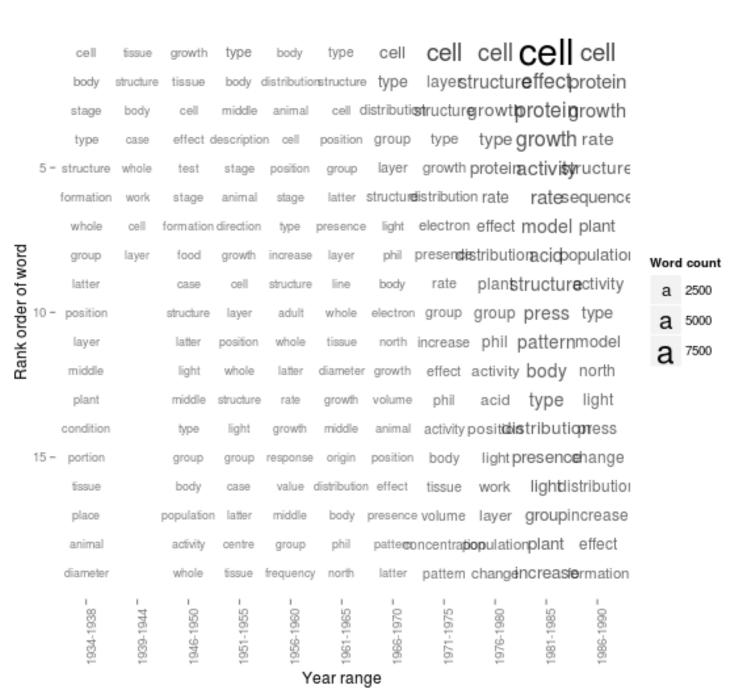
Inspired by Google's Ngram Viewer, we wrote JSTORr, an R package that visualizes ngrams for corpora held by JSTOR, a digital library of 2000 scholarly journals. JSTOR's Data for Research service allows users to download full text of journals. With JSTORr, the full text can be analysed for ngrams, word correlations, document clustering and topic modelling. This means that we can now carefully build sizable corpora of scholarly literature on specific topics and investigate them with text mining methods. Here we demonstrate some of the basic functions of the package to get insights into the histories of science.

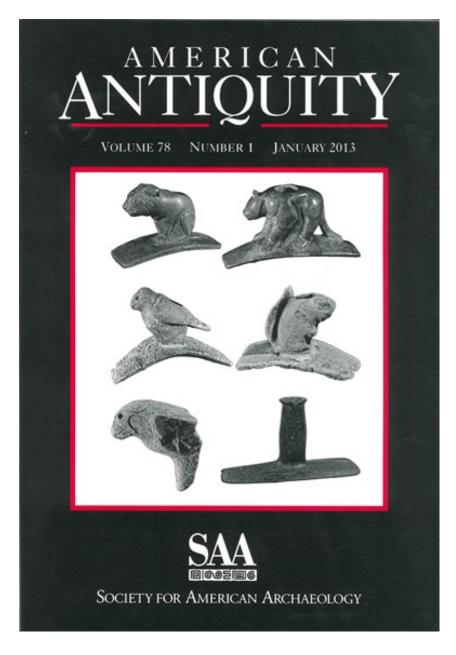


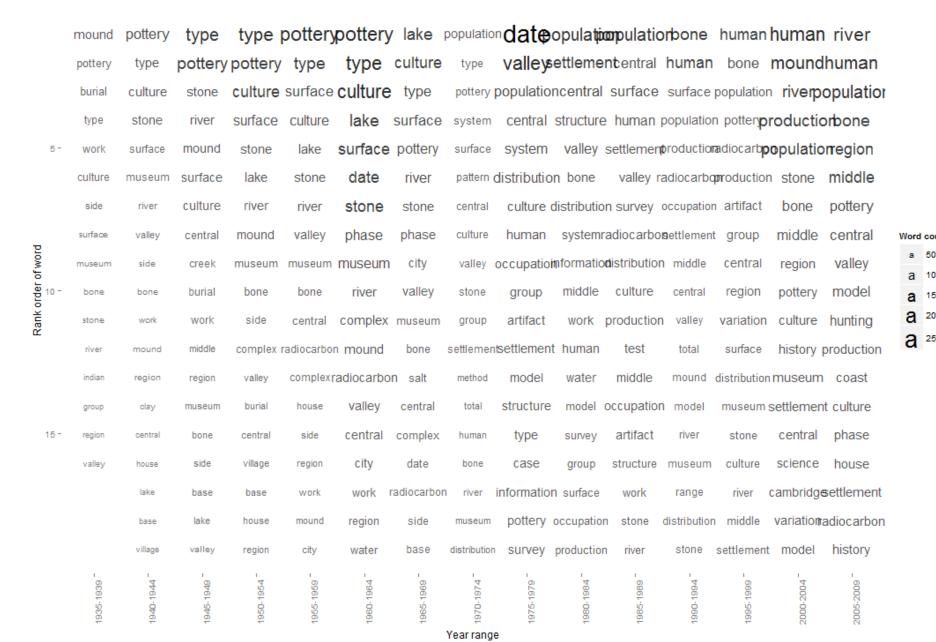
5-	difference standard result	population distribution standard	correlation method population difference	distribution correlation standard		population standard type	population standard type	method standard probability			population variance probability	probability function	probability function	function probability variance	model function likelihood	likelihood model function	mod ed ilikelihook	stributli kelihoti function	stribudio delihoo function	stribdti function	stribudi function		nunction distribution	
	error deviation	difference	error		equation difference	order method	probability difference	error		method	method order	order	method equation	model method	variance matrix		matrix maximum					egressior estimation		Word count a 2500 a 5000
Rank order of word	method individual average	method work individual	group order average		order population deviation	error deviation	method equation	work difference		significance equation standard	type correlation	approximatio hypothesisa population	approximatio	order on linear population	linear maximum method	maximum statist method	statist linear probability	maximum	matrix	variance		matrix method theorem	method regression	
Rank		a verage group	type result		regression	group	order	order function		power	significance	matrix	hypothesis		procedure	order	order	parameter	order	theorem	probability	orobability		a 7500
15 -			fact regression		error result	probability theory	formula hypothesis	hypothesis theory		hypothesis type		zero error n problem	statist result	likelihood error statist	statist estimations	linear estimation approximatio	regression		probability	linear	parameter	variance parameter standard	probability	
					fact	result average	function	variance problem		correlation	hypothesis error	model likelihood	parameter type	estimation maximum	error			method theorem a			standard	error	standard	
	_ 90	_ 0_		20 _	_ 52	formula line	problem	equation 4	- 4 ₅ -	theory	result	correlation	linear	theorem	parameter		ratio	standard	standard	example	error	maximum approach	approach	
	1901-1905	1906-1910	1911-1915	1916-1920	1921-1925	1926-1930	1931-1935	1936-1940	1941-1945	1946-1950	1951-1955	Year range	1961-1965	1966-1970	1971-1975	1976-1980	1981-1985	1986-1990	1991-1995	1996-2000	2001-2005	2006-2010	2011-2012	

For example, we can quickly see intellectual shifts in a high-impact statistics journal (above) from correlation to model-based analyses. Similarly, a prominent biology journal (below) reveals a turn from macrostructures to a focus on cells and proteins. We might hypothesize that these intellectual trends are related to technological changes. This hypothesis could be tested by close reading of a small sample of articles.

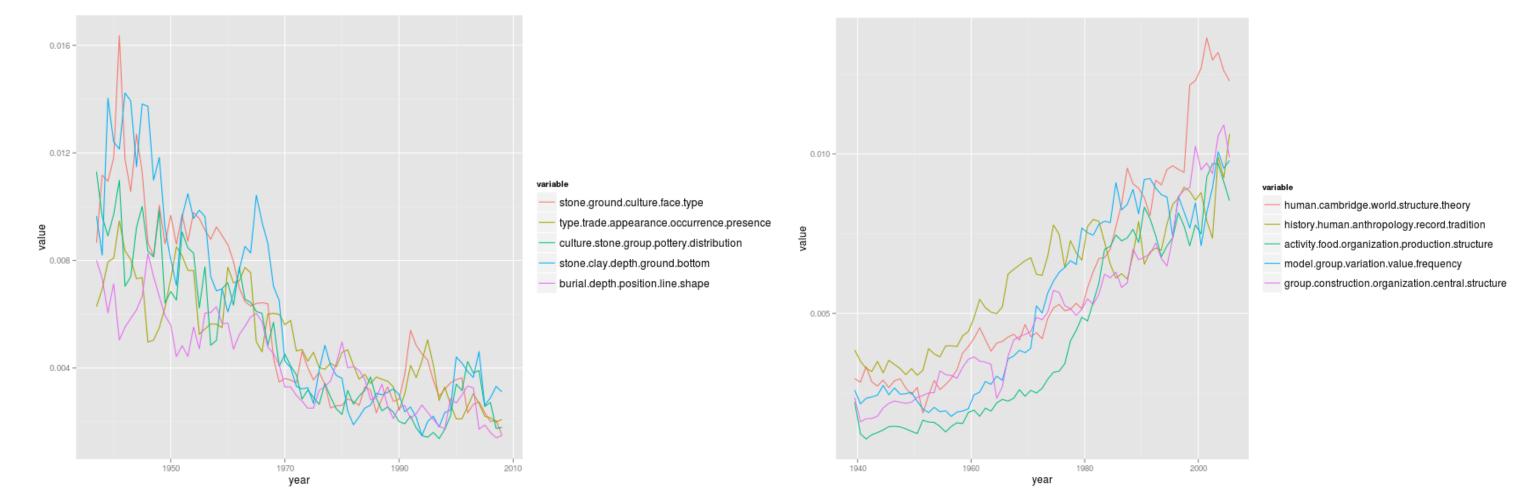




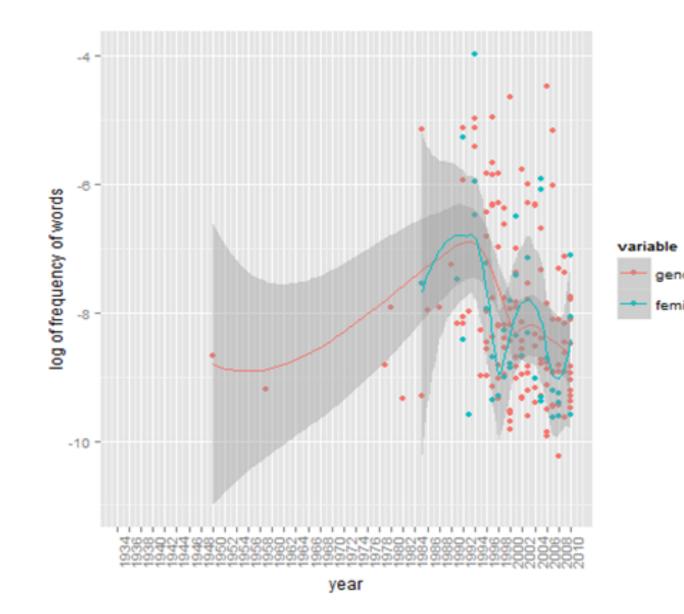




In archaeology we see a turn from typological analyses of pottery to work on population-level behaviors and settlement patterns. Dating becomes frequently discussed after radiocarbon methods become widely available.



Topic models generated by latent Dirichlet allocation (above) show hot and cold topics over time in American archaeology. These validate and add context to the most frequent words over time in the top figure.



Ngrams over time (left) show the complex relationship between gender and feminism in American archaeology. The two terms appear together then diverge in their frequencies.

K-means clustering of journal articles using word frequencies (right) reveals distinct approaches to gender in the American archaeological literature. In the lower cluster we have a cluster of articles on the archaeological record of the US Southwest. The upper right clusters are articles about gender theory and philosophy.

