Inside the Creativity Boom

HTRC Final Report

PI: Samuel Franklin (Brown University)
Support: Peter Organisciak (HTRC), Ryan Dubnicek (HTRC), Eleanor Dickson (HTRC)

Intent

The project set out to make use of the HathiTrust corpus to map the career of the words "creative," "creativity," and their less common variants over the last several hundred years, with an emphasis on the twentieth century. It was already known that the word "creative" emerged gradually over the course of the modern era, increasing in use rapidly in the twentieth century, and that "creativity" only barely appeared around the turn of the twentieth century and exploded into the regular English lexicon in the post-WWII era. In order to discern the relationship between these two patterns, and to figure out if the recent rise of "creativity" signifies simply the popularization of a pre-existing concept or a new conceptual formation, a more granular analysis of these trends was necessary. Have the increases in the use of the word types "creative" and "creativity" been distributed evenly throughout the printed corpus, or have they clustered around certain fields, genres, or communities of discourse? To what topics, activities, and types of people have those words pertained? Is it possible to discern variation and change in the meanings of those words across and between genres, fields, and eras? To answer these questions, I proposed utilizing a number of different analytical tools such as collocates, topic modelling, and faceted queries, using the HathiTrust corpus and subsets therein.

Process

After an initial period of weekly calls in July 2016. The team met via phone every other week from August-December 2016, with periodic follow-ups in early 2017.

Setting the Agenda

Early on, we decided to first tackle the parts of the project that could be done using the new Extracted Features Dataset for in-copyright works, which was soon to be released. The Bookworm research would wait until the next iteration of Bookworm was ready. Because the Extracted Features Dataset was bag-of-words on the page level, collocates and concordances were not possible, but topic modelling and co-occurrences (on the page level) would be. Since the research goal was to discover "what we talk about when we talk about creativity," we decided it would be acceptable to work on a custom corpus of pages only containing the relevant keywords, which we called the "creativity corpus."

Building the Creativity Corpus

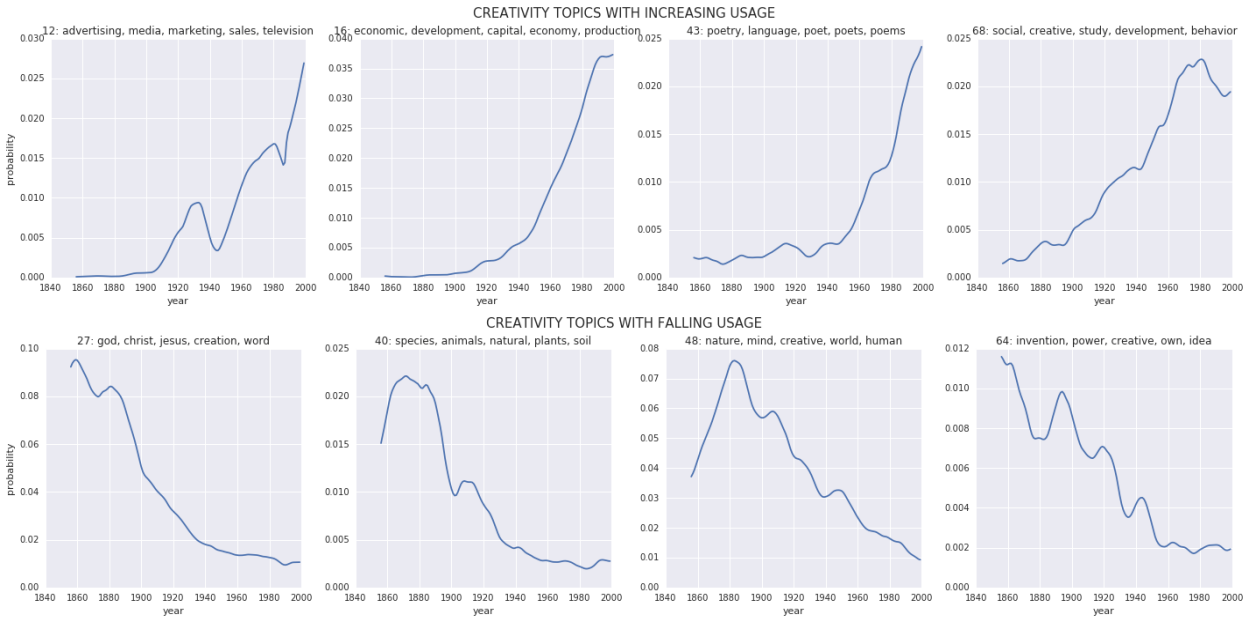
We first decided on criteria and process for generating the creativity corpus. We decided to cast the net widely, including all pages with at least one occurrence of *creativ** words in English books, the idea being the set would be small enough to work with and large enough to discern broad patterns of the uses of those words over time. We decided to de-duplicate identical volumes or editions, preserving multiple editions of a single title, the idea being that more widely published works are also more widely circulated and therefore more influential and/or paradigmatic of the way language is used in general. (This is an arguable claim that I would be interested in testing more in the future). We also introduced date field criteria to weed out obviously erroneous dates (e.g. 0000 or 9999) that appeared during manual inspection. HTRC then obtained from HT a CSV file of 2.7 volume IDs.

Topic Modelling

We began to experiment with topic modelling using Latent Dirichlet Allocation (LDA). We quickly identified a potential problem with the existing method for topic modelling: Normally, the order in which texts are sampled is either chronological (starting from the beginning of the list) or randomized to control for time. Because we wanted the model to be able to identify topics specific to their particular eras, but did not want older topics to be drowned out by the massive number of works in the last twenty years or so, a compromise was needed. Peter Organisciak developed code for temporally weighting the training sample (chronologically, by decade, randomizing within each decade) in order to soften the temporal bias without entirely removing it.

Additionally, we applied asymmetric document-topic priors in order to allow more specific topics to rise to the top. By assuming the top three topics would be very prevalent, we assigned them most of the probability mass, dividing the remainder between the rest. This passed the sanity test, with topics such as [0: creative, own, god, world, human, art, does, power, social, mind] and [1: world, creative, christian, modern, way, own, human, religious, social, power] taking top positions, and others with clear specificity (in other words, looking like "topics" in the conventional sense of the word) such as [12: advertising, media, marketing, sales, television, business, market, agency, service, creative] and [13: art, artist, artists, painting, creative, artistic, arts, form, world, architecture] following somewhat farther down the list.

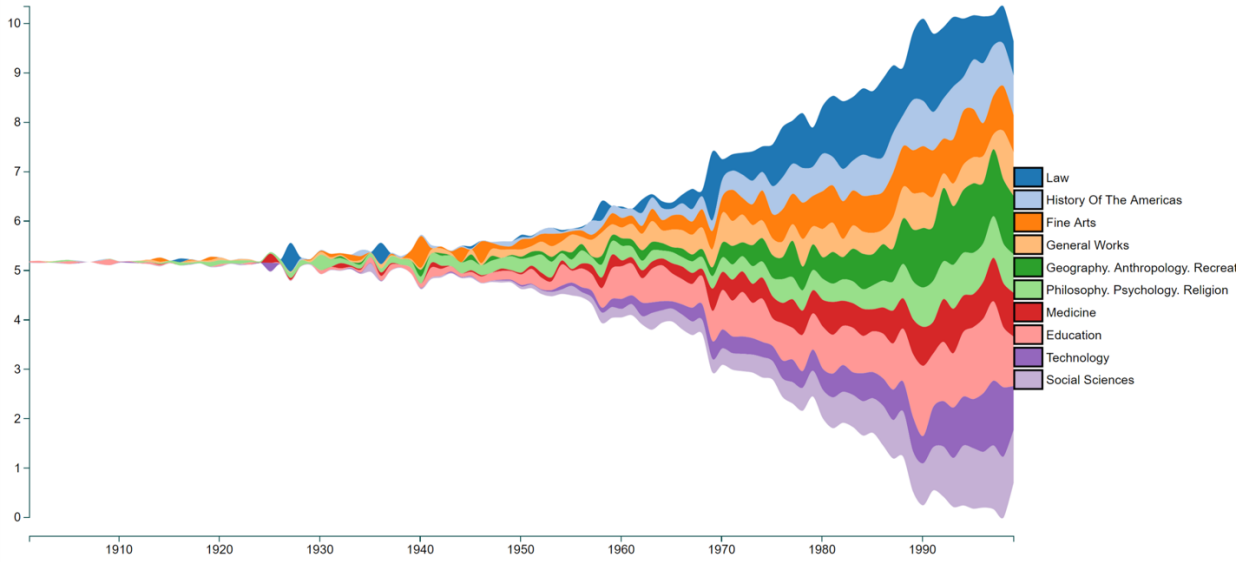
We were then able to plot these over time (see charts below). Some of these patterns comport with a general narrative about the word "creative" (namely that its originally divine connotations were gradually shed as it became more closely associated with the human arts), and others pose questions for further research. In all, the exercise revealed that topic modelling using a deliberately narrow sample (the creativity corpus) and trained in a temporally weighted manner could yield comprehensible topics whose rise and fall can be charted through time. With further research we can compare and compile these topics to get a sense of what proportion of creativity talk each one represents over time, and where overlaps exist.



(credit: Peter Organisciak)

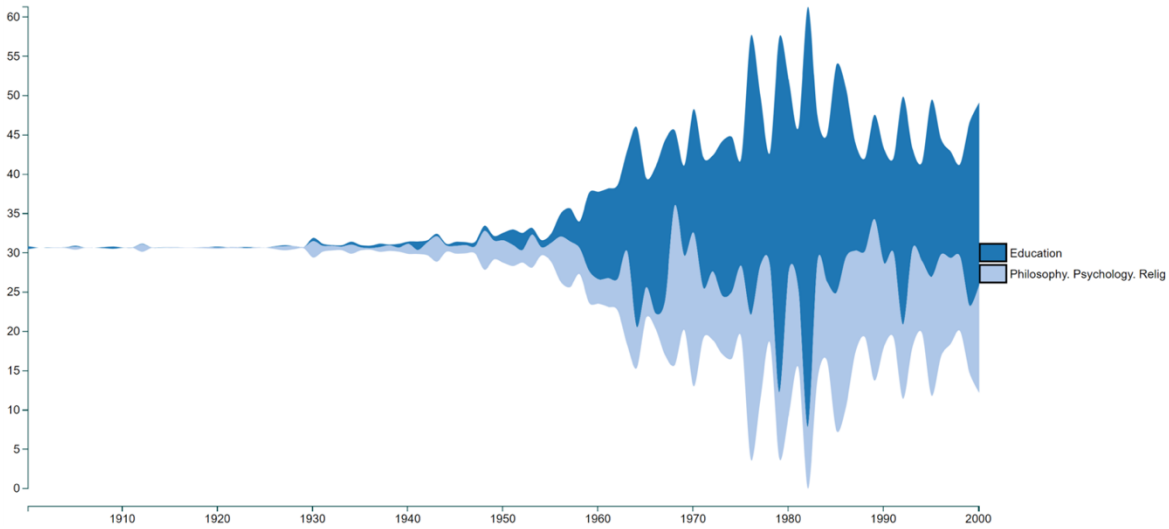
Bookworm

Though the rollout of the new Bookworm on the Extracted Features dataset was delayed until after the ACS grant expired, we were later able to work through a set of queries and experiments to run, which will be incorporated into future work. Using Library of Congress classification metadata we can see in what genres "creative" and "creativity," respectively, got their starts, and whether the increases we see in the overall curve of the Google n-gram viewer are evenly spread across fields and genres, or if some are overrepresented over the entire period or in certain historical moments. A visualization of this appears to show a fairly even spread across the LOC classifications:



(credit: Peter Organisciak)

From the above graph, it seems as though *Philosophy, Psychology, Religion* and *Education* were the strongest veins in which "creativity" was used in the middle 50 years of the 20th Century. Zooming in on those (see below), we can see that occurrences of "creativity" really exploded in Education in the mid-1960s. This is an important insight for me: While I treated the field of education in my dissertation, I may have underemphasized its centrality to the discourse of creativity. As I revise the dissertation into a book, I may dig deeper into these numbers and consider emphasizing education, which, as a major site of enculturation, would have a major impact on our general public discourse.



(credit: Peter Organisciak)

N-grams

Because n-grams are not available through the Extracted Features Dataset, Peter aided me in downloading a dataset from the Google books API with unigrams and 2-5-grams. I will use this data to replicate the graphs from Chapter 1 of the finished dissertation (which are currently screengrabs from the Google n-gram viewer) as I revise for publication.

Outcomes

The major outcomes of the ACS research so far have been a conference paper, the creativity corpus itself, and the foundation for an article in progress. Peter and Sam collaborated on a paper on the novel workflow for topic modelling on temporally-biased corpora, which Peter presented at DH2017. The Creativity Corpus is currently being used by the HTRC for demonstrations and should be made available for other researchers.

Continuing work

Much of the work begun during the ACS term continues. Over the next academic year I will revise Chapter 1 of my dissertation (which showcases large corpus text analysis) for publication in a journal of history, intellectual history/history of ideas, or American studies. I will utilize the Creativity Corpus data to dive deeper into co-occurrences, which we only barely got started on before our time ran out, and I will use the new Bookworm running on part of the Extracted Features Dataset to extend my research on and generate more visualizations of the facets of creativity discourse.