Modeling Creativity: Tracking Long-term Lexical Change

Peter Organisciak
organis2@illinois.edu
University of Illinois, United States of America

Samuel Franklin
samuel_franklin@brown.edu
Brown University, United States of America

The concept of creativity underwent a period of shifting meaning and rapid adoption in the twentieth century. Following from a narrow early scope of usage, in which it carried largely religious connotations, the word 'creative' grew broader and adopted the more subjective meanings we are familiar with today. Though many contemporary observers point out the vagueness of the term, creativity's power comes from a particular mix of meanings and connotations accrued over time. Still, there is no clear inventory of the higher-level concepts around discussion of creativity and how they evolved. Additionally, because of the rapid increase in usage, early uses of 'creativity' may be overlooked as they are overshadowed by much more common later uses.

In this paper, we present a method for tracking the different styles of discourse around a concept over time, developed for following the evolution of 'creativity' but applicable to other domains. Our approach is an application of Latent Dirichlet Allocation (LDA) -trained topic models, with three novel steps in their preparation:

- a highly-selective keyword sampling of pages from a large text corpus,
- temporally weighted training sample ordering, and
- purposively-assigned asymmetric document-topic priors.

Motivation

This research supports a larger project on the discourse of 'creativity' in post-WWII America. The anecdotal observation that creativity has become a buzzword in recent years is supported by graphs of word frequency available through platforms such as the Google Ngram viewer and JSTOR Data for Research, which show creativity only entered the American lexicon in the twentieth century, diffusing rapidly after about 1950. 'Creative' appears to have enjoyed a similar growth spurt over the same period, but it preceded creativity by about three hundred years.

Unfortunately, these graphs do not reveal the long-term changes in meaning nor the distinct contexts in which the language of creativity accrued its contemporary salience. It is obvious from contemporary usage that the word 'creative' has a tangle of interrelated but distinct meanings, ranging from *generative* or *constructive* to *artistic* to *nonconformist*. These meanings are distributed unevenly over time and across communities of discourse. To understand why and through what routes creativity arose when it did, it will be essential to tease apart these various meanings of creative, and the contexts in which they have been strongest over the long term.

We believe topic modeling can help. First, it can help us identify and distinguish between the several discourses in which creative has been a keyword—for example in theology versus education versus psychology—whilst still reflecting the historically shifting connections and overlaps between those. Second, we can then apply those topics to only those texts containing the token 'creativity,' to reveal which of the discourses and meanings of 'creative' seem to be at work. By this process we can achieve a more granular picture of the creativity boom, helping us answer the basic question 'what do we talk about when we talk about creativity?'

Approach

Topic modeling enables us to observe more higher-level concepts than keyword searching and collocations would allow. Topic modeling depends on a certain class of mixed model clustering, but we believe that the two should not be conflated. The connotation of 'topic modeling' implies a qualitative interpretability. Surfacing what would be recognized as concepts is not solely a case of running a modeling algorithm on words from a text. Instead, it needs to be paired with a series of preparatory and parameterization steps tailored to the particular problem.

We developed a workflow for training better topic models to track a specific concept in a temporally-biased corpus. This involves standard pre-processing such as stoplisting words, but also contributes three novel steps: selective page-level sampling, weighted

training, and explicitly imbalanced prior assumptions on how likely a document is to be reflected by each topic. The sampling helps focus the models on creativity, the weighted training counteracts temporal biases to retain older topics to surface, and the asymmetric priors help find more granular topics.

For a dataset cross-cutting published work broadly, we used a recent release of the HTRC Extracted Features Dataset (Capitanu 2016). The Extracted Features Dataset includes term counts for every page of 13.7m volumes in the HathiTrust Digital Library and benefits from a mostly indiscriminate digitization policy, allowing us to observe a term's usage in a wide spectrum of texts.

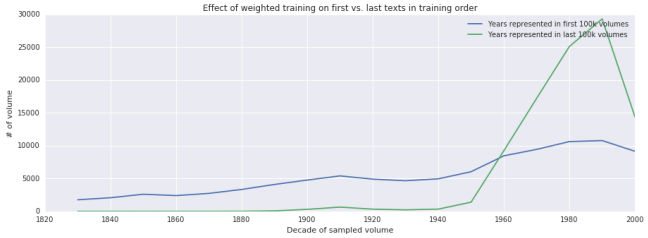
Topic Modeling Preparation

In topic modeling, the goal is surfacing patterns that represent qualitatively intuitive concepts. However, to the *methods* used for topic modeling, the mark of success is being able to represent documents in the desired number of topics with as little error as possible. This divergence between our needs and the machine's makes the text preparation important. One such preparation is to remove words that are not interesting to a human reader. An algorithm may find a meaning in a word like 'however' or 'whereas', but as a proxy for topicality, such words are usually not desired.

For tracking trends in creativity discourse, we used Latent Dirichlet Allocation (LDA) combined with standard preprocessing: removing the most common words in the English language, less interesting parts-of-speech (e.g. adverbs, determiners, numbers), and cutting off the sparser end of the vocabulary. In addition, we developed three less common preparations in the service of issues arising from tracking concept diffusion.

Sampling. One possible approach to finding the most common topics for a keyword is to look at the underlying term-topic probabilities for the keyword, post-training, and identifying the topics where the word is most common. This approach scales well to multiple keywords but provides low specificity for tracking them. Instead, we sampled only pages that use the word 'creativity' or variants of 'creative'. The size of the HTRC EF Dataset affords the small contextual window and selective sampling, as there were slightly more than 2 million volumes found that have at least a single mention of the keyword list.

Weighted training. When training topic models, earlier texts have an outsize influence on the topics that emerge. This is a problem for our use case, where we expected a topical shift alongside a steep increase in usage. A randomized training order would reflex later texts very strongly, at the risk of missing topics which are prominent in older texts. To counteract this, we applied weighting to the randomized training order, to soften the temporal bias without entirely removing is. When deciding on the next text to send to the training algorithm, texts are weighted for sampling with weight(decade) = 1/ n(decade). The following figure shows this weighting in action: at the important start of training, newer texts are only slightly more common. Since a disproportionate number of older texts are used early on, there are few left by the end of training.



Honeypot topics. As part of the estimation process for LDA topics, we have to formalize our best guess for how likely any given topic is to be assigned to a document. Past work has found value in allowing for these prior assumptions to be uneven - e.g. one topic can be considered more likely than another (Wallach, Mimno, and McCallum 2009). We found initial success with a heuristic intended to find many smaller trends in the collection by provided the first three topics the majority of the probability mass and dividing the remainder between the remaining topics. In qualitative comparisons with evenly distributed probabilities, we found that setting asymmetric priors in this way set traps to catch broadly common documents in predictable topics, while allowing other topics to surface more highly-specific topical hotspots.

0: creative, own, god, world, human, art, does, power, social, mind 1: world, creative, christian, modern, way, own, human, religious, social, power
12: advertising, media, marketing, sales, television, business, market, agency, service, creative 13: art, artist, artists, painting, creative, artistic, arts, form, world, architecture

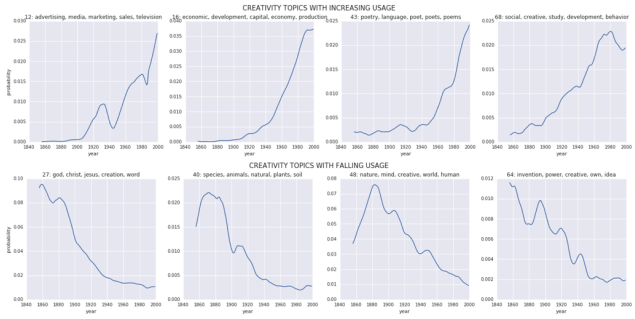
Two general topics and two niche topics

Results

The training yielded several topics which confirm where we would expect to find the language of creativity. Some of these reflect specialized uses, such as in advertising and evolutionary biology, while others reflect the broad humanistic discussions of the nature of thought, art, and religious creation. By graphing these

topics over time we can see that our temporally weighted sampling appears to have been successful in revealing archaic topics that are nonetheless essential to understanding the connotative textures of the language of creativity in our own time.

The following figures show a small selection of topics where the usage has grown in the past 150 years, and topics where it has fallen. Generally, we see that the language of creativity has transitioned from religious and natural notions of creation toward the language of economic and human capital.



Future work

This work has a number of future directions. We have thus far focused on a number of words (creative, creativity, creativeness); moving forward, we intend to map how the verb and noun uses compare. Also, while much of the development has been qualitatively development against our particular problem, we hope to compare variants of our workflow in more contexts.

Conclusion

In the proposed paper, we will present our method for tracking longitudinal trends in a diffuse and shifting context. Motivated by work on the language of creativity and particularly the noun 'creativity', our contributions are in text processing and parameterization for topic modeling, allowing clear and specific concepts to be revealed.

Bibliography

Capitanu, B., Underwood, T., Organisciak, P., Cole, T. J., Sarol, M. J., Downie, J. S. (2016). *The HathiTrust Research Center Extracted Features Dataset*. 1.0. HathiTrust Research Center. Dataset. http://dx.doi.org/10.13012/J8X63JT3

de Bolla, P. (2013). *The Architecture of Concepts: The Historical Formation of Human Rights*. Fordham University Press.

Wallach, H.M., Mimno, D.M., and McCallum, A. (2009). "Rethinking LDA: Why priors matter." A*dvances in neural information processing systems*.

Williams, R. (1976). *Keywords: A Vocabulary of Culture and Society*. New York: Oxford University Press.