Collection 20: U.S. Top Newspapers, 2000-2018
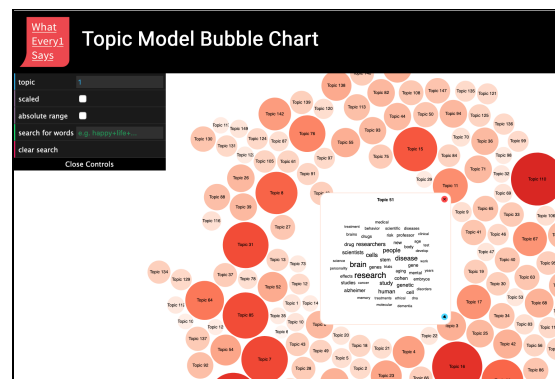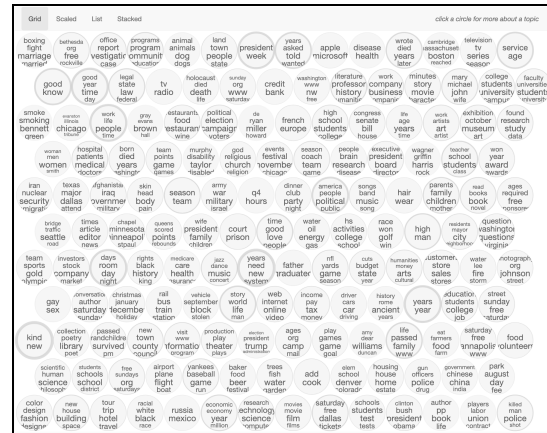## (sample of all articles

A collection of word-frequency and other data representing 29,183 unique articles (no duplicates or close variants) published during 2000-2018 in 15 top U.S. newspapers and their associated online blogs.* WE1S and other researchers use this data to look for broad patterns and help guide closer study.

Collection 20 contains data representing all 15,692 articles from its set of sources in these years mentioning "humanities" and a sample of the 641,617 articles on everything else from those same sources and years ("random" documents found through searching on common English words). It downsamples these "random" articles (while maintaining the proportions of articles from particular sources and years) to achieve a 50/50 balance of articles from each category. The purpose is to allow media discourse on the humanities to be studied alongside "everything else" and not be buried so far down in the statistical pile that it cannot easily be seen in detail. Collection 20 is thus not a representation of the relative weight of discussion of the humanities in media discourse in general but instead an aid to studying the fine features and structures of each.

News sources in Collection 20 include: *Boston Globe, Chicago Tribune, Daily News* (New York)*, Dallas Morning News, Denver Post, Houston Chronicle, Los Angeles Times, New York Post, New York Times*

(and its blogs)*, Newsday* (New York)*, Seattle Times, Star Tribune* (Minneapolis, MN)*, Tampa Bay Times, USA Today*, *Washington Post*.

Topic model of this collection (150 topics visualized in two of the visualization interfaces in the WE1S Topic Model Observatory: Dfr-browser & TopicBubbles)

* WE1S makes available only "non-consumptive use" word frequency, topic model, and other datasets along with their visualizations. Datasets cannot be used to access, read, or reconstruct the original texts.