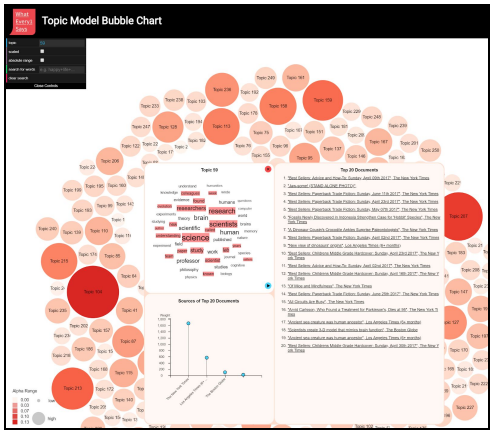
Collection 32: U.S. Top Newspapers (sample of all articles)

A collection of word-frequency and other data representing 204,617 unique articles (no duplicates or close variants) published during 2012-2018 in 15 top U.S. newspapers and their associated online blogs.[1]* WE1S and other researchers use this data to look for broad patterns and help guide closer study.

Included is data based on an approximately 1:40 proportional balance between articles mentioning “humanities” (about 5,000) and a sample of articles on everything else (about 200,000 more or less “random” documents found through searching on common English words). In essence, the collection is a sampled representation of “everything” in these sources for these years (limited by the fact that it is not feasible to know how many articles were actually published in these publications, to determine how completely they were collected in available database repositories, or to harvest everything from such databases.)

In this collection, the word “humanities” occurs 7,226 times in 4,976 documents. “Science” or “sciences” combined occur 25,693 times. (“science: 22,811 times in 12,319 documents; “sciences” 2,882 times in 2,277 documents). Mentions of the “humanities” are thus 28% the number of mentions of “science(s).”

News sources in this collection (in order of number of articles for each) are: *New York Times, Chicago Tribune, Los Angeles Times, News Day, New York Post, Houston Chronicle, Daily News, USA Today, Dallas Morning News, Denver Post, Washington Post, Boston Globe, Star Tribune* (Minneapolis)*, Seattle Times, Tampa Bay Times*.



Topic model of this collection (250 topics visualized in two of the visualization interfaces in the WE1S Topic Model Observatory: Dfr-browser & TopicBubbles)

Further Information

WE1S Collection Registry ID: 20191114_1518_us-humanities-comparison-top-newspapers-unsampled

Data source: LexisNexis (via LN Web Services Kit)

Collection dataset*: GitHub [TBD]

Mallet topic model data files: GitHub [TBD]

Topic models available (# of topics): 25, 50, 100, 150, 200, 250

Topic model visualizations: Start page

[1] 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.