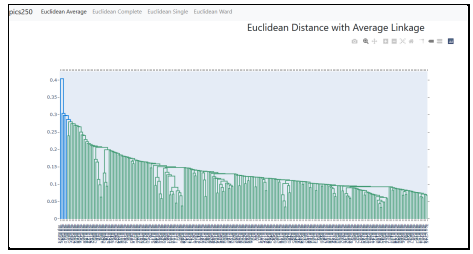
Collection 28: Tweets containing keyword "humanities," c. 2014-2017.

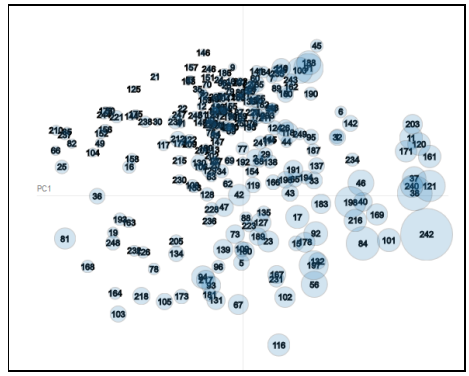
Collection 28 (C-28) of the WE1S Twitter corpus consists of 799,744 tweets containing the keyword “humanities” from authors who tweeted the term "humanities" more than once between Jan. 1, 2014, and Dec. 31, 2017. (See also C-29, which aggregates tweets by author.)

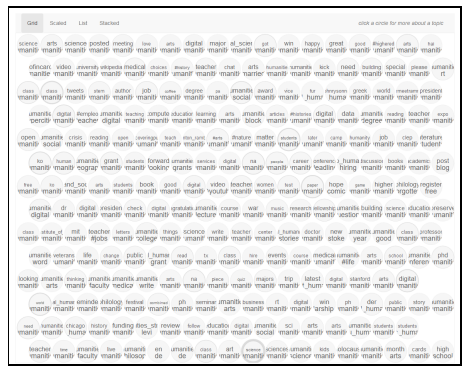
Explanation: Our WE1S Twitter Corpus originally consisted of 1,589,462 individual tweets containing the keyword “humanities,” posted between January 1st, 2014 and June 30th, 2019. However, due to the astronomical number of individual documents, we decided that reduction was necessary to have a workable model. We opted to include only tweets from 2014-2017 because Twitter switched to a larger character limit in Nov. 2017, and we wanted to work with the older format.

Even with these parameters, further reduction was necessary. Thus, in this collection we decided to omit tweets of authors who only tweeted once about the humanities. This step limited the corpus to 799,744 individual tweets, each of which is treated as an individual document.

The advantage of this method is that it retains tweets as individual documents, maintaining the ability to link directly to the tweets in the model. However, due to the large number of documents, visualizations can still become unstable and/or unresponsive. (But see C-29 for a variant corpus that aggregates authors’ tweets. See also M-8 for an evaluation of trade-offs between C-28 and C-29.)







Topic model of this collection (250 topics visualized in three of the visualization interfaces in WE1S Topic Model Observatory: Dendrograms, pyLDAvis, and Dfr-browser)

WE1S Collection Registry ID:

20190717_1646_tweets-by-author

Data Source: Twitter (via Twint)

Collection deposit (term frequencies, Mallet outputs, visualization files, and other non-consumptive data)**:**

Zenodo (DOI: TBD)

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

Topic model visualizations: Start page