Collection 27: Full Twitter Corpus, c. 2014-2019

Collection 27 (C-27) is the full, raw set of Twitter posts that WE1S gathered, consisting of 4,909,180 tweets between January 2014 and July 2019 that contain one or more of the terms *humanities, science(s), STEM,* and *liberal arts*. From C-27, WE1S later carved out two more limited and curated subsets for topic modeling. (See C-28 and C-29.)
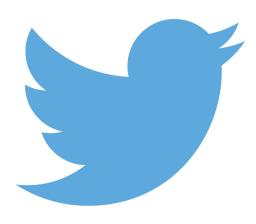
In C-27, tweets mentioning *humanities* number 1,589,462.[1] Tweets mentioning *science* or *sciences* number 2,446,899. (However, we downsampled the latter *science(s)* tweets by collecting only for one month per year: Jan. 2014, Jan. 2015, Feb. 2016, and Jan. 2017. This is due to the astronomical number of tweets mentioning *science(s)* by comparison with *humanities*. (Similarly, mainstream U.S. newspapers mention *science(s)* at a 40:1 ratio to *humanities*; see C-32.) C-27 also includes 865,156 tweets mentioning *STEM.*that were posted in Jan. 2016 or Feb. 2017. Like *science(s)* the term *STEM* also greatly outnumbered tweets with *humanities*. We elected to collect only two months' worth for comparison purposes.

Finally, C-27 includes 7,663 tweets mentioning *liberal arts* from Jan. 2017. After discovering how sparse such tweets are for this single month, we chose to abandon further collection based on this term.

Although the topic models we created for analysis ultimately only made use of our *humanities* tweets (in C-28 and C-29), we make publicly available the entirety of our raw C-27 corpus as a dataset deposited in

the Zenodo open-science repository. in case other researchers find it useful to carve out other subsets for study. WE1S itself made use of C-27 only to conduct a comparative analysis of the hashtags *#Humanitie*s, *#arts*, *#DigitalHumanities*, *#LiberalArts/ #Liberal_Arts*, *#HumanitiesMatter*, *STEM*,

and *science[s]* (see our KF-6-5).

Twitter logo as shown in Melani Rehak, "Who Made that Twitter Bird?" (2014)

---

[1] A portion of these, posted between Jan. 2014 and Dec. 2017, are included in C-28 and C-29.

WE1S Collection Registry ID: 27

Data source: Twitter (via Twint)

Collection dataset deposit
   Zenodo (DOI: TBD)

Topic models available (# of topics): None (WE1S did not model C-27, but only subsets C-28 and C-29)