Collection 29: Tweets containing keyword "humanities," c. 2014-2017

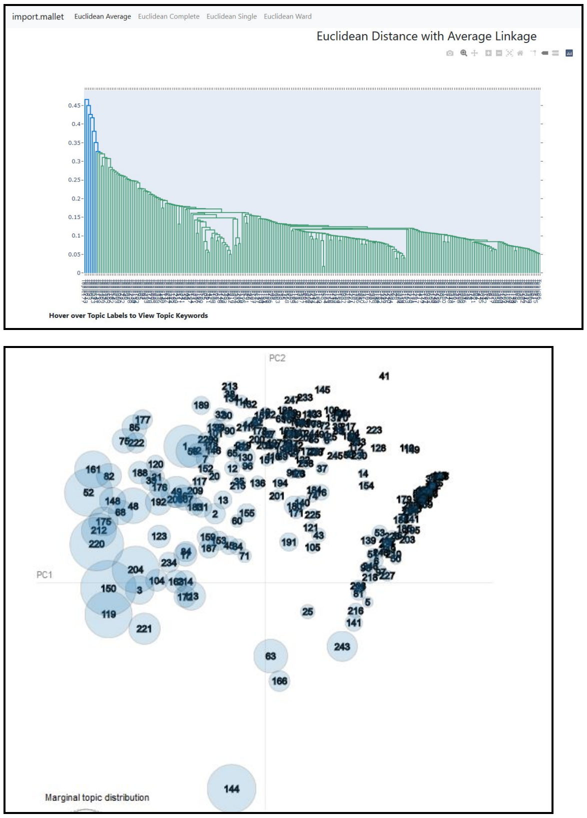
(tweets aggregated by author)

Collection 29 (C-29) 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. This version of our Twitter corpus compiles tweets by each author into single “documents” for topic-modeling analysis, resulting in 132,562 total documents. (See also C-28.)

Explanation: WE1S originally collected a corpus of 1,589,462 individual tweets containing the keyword “humanities,” posted between Jan. 1, 2014 and June 30, 2019. However, due to the unwieldy number of documents, we opted to reduce this collection to include only tweets posted between 2014-2017, which cut the individual tweets by almost half. This resulted in our Collection 28 (C-28).

Collection 29 is an off-shoot of C-28. But even after omitting one-time authors, C-28 still created a notable strain on our visualization tools due to the high number of individual documents (799,744). To reduce the processing load while retaining the full content, C-29 aggregates tweets by author, essentially treating everything tweeted by a given author as a single “document.” This cut the collection to 132,562 documents without sacrificing the content of C-28.

The advantage of aggregating by author is reduction in numbers of documents, which makes modeling the corpus with the WE1s visualization tools easier. However, this comes with the caveat that we lose the identity of tweets as individual posts and also generate more topics in our models that seem to lack coherence, possibly due to the concentration of Twitter markup (hashtags, mentions) in each aggregated document. (See also M-8 for an evaluation of trade-offs between C-28 and C29.)



Topic model of this collection (250 topics visualized in WE1S Topic Model Observatory: pyLDAvis, Dfr-browser, and DendrogramViewer)

Further Information

WE1S Collection Registry ID:
20190722_1758_tweets_by_author-aggregated
Data Source: Twitter (via Twint)

Collection dataset: [TBD]
Mallet topic model data files: [TBD]
Topic models available (# of topics): 50, 100, 150, 200, 250
Topic model visualizations: Start page

6 July 2020; rev. 30 Sept. 2020 (Joseph Jaffray)