Collection 39 (and Corpus-A): Reddit -- Students on the Humanities

Collection 39 (C-39) is a subset of WE1S’s C-38 Reddit collection tailored to focus on student discourse about the humanities. Where C-38 includes Reddit comments longer than 225 words from 2006 to 2019 containing the terms humanities, liberal arts, or the arts, C-39 consists of 66,290 comments from that larger collection (about half the original number) that also contain at least one of the terms student, major, or college (including plurals and other forms). (Similar to C-38 is WE1S’s Corpus-A, an earlier version of the same collection, but including only the years 2006-2018.)

Explanation: WE1S’s rationale and methodology for collecting Reddit comments to study public discourse about the humanities (especially by students) is explained in the blog post by WE1S’s lead Reddit researcher Raymond Steding: “A Digital Humanities Study of Reddit Student Discourse about the Humanities” (2019).

As Steding notes, he initially collected 3.3 terabytes of Reddit data from 2006 to 2018 (approximately five billion comments) by downloading it from pushshift.io in JSON format. Data for 2019 was later added.

This data was filtered to retain only comments containing at least one of the terms humanities, liberal arts, or the arts. Then to improve the coherence of topic models of the collection and to make the number of documents more tractable, WE1S subtracted comments under 225 words. (We did not weed out duplicates.)

Further Information
WE1S Collection Registry ID: 20200714_2157_reddit-students-and-the-humanities
Data Source: Reddit (via data harvested at pushshift.io)
Collection dataset: [TBD]
Mallet topic model data files: [TBD]
Topic models available for C-39 (# of topics): 100, 200
Topic model visualizations: C-39 start page (See also Corpus-A: Dfr-browser 200; pylDAvis 200)

---

1 Pushshift is a “big-data storage and analytics project” providing APIs for searching and downloading Reddit and also downloadable raw data (comments) from Reddit. (See Pushshift FAQ.)

2 Corpus-A, the earlier version of C-39 missing Reddits in 2019, also subtracted comments by users with a Reddit “karma” score lower than 2. (“When posts or comments get upvoted, that user gains some karma” -- Reddit, “What is Karma?”)