
#metoo Digital Media Collection

Jane Kelly

Web Archiving Assistant

Schlesinger Library on the History of Women in America
Radcliffe Institute for Advanced Study, Harvard University

Collection Scope

- 2007-present
 - Bulk of collection will date from October 2017-present
- English and other languages
 - Primarily English; very small amount of Spanish language content
- All regions of the United States
 - Not just big coastal cities
- Across industries
 - Not just entertainment and media
- Variety of political ideologies
 - Supporters and critics

What do we have so far?

- 21.5 GB of data in Archive-It
 - 899 seeds
 - 854 one-time crawls (mostly single page)
 - 45 semi-annual crawls
- Twitter data
 - 2.4 million tweets collected using Social Feed Manager since December 2018
 - 19 million tweets licensed directly from Twitter (coming soon!)

The case for crude and quick scoping

- Breadth instead of depth
- Text and media content prioritized over preserving look and feel of website
- Staff time
- Large data budget

Example: BuzzFeed article

Crawl history for one seed

Settings	Metadata	Crawling History	Notes	Seed Scope			
Crawl List (2 Crawls)							
<input type="text" value="Type to Filter Crawls"/>				Download Seed Crawls List			
Crawl ID	Date & Time Completed ▾	Status	Seed Status	Frequency	Data	Documents	Report Link
933425	Jun 12, 2019	Finished	Crawled	Saved Test	10.9 MB	246	View Report >
912833	May 15, 2019	Finished	Crawled	Deleted Test	101.7 MB	316	View Report >

Seed-Level Scoping

Host report

Host List (1 to 100 of 338 Hosts)

Type to Filter Hosts

[Download Host List](#) < 1 2 3 ... >


Edit Rules Run Patch Crawl

Host	Docs	New Docs	Data	New Data	Blocked	Queued	Out of Scope
<input type="checkbox"/> r5---sn-n4v7snee.googlevideo.com	3	1	79 MB	79 MB	0	0	0
<input type="checkbox"/> www.buzzfeed.com	110	83	18.5 MB	16.2 MB	4	0	2,081
<input type="checkbox"/> abs.twimg.com	13	1	2.2 MB	556.3 KB	0	0	65
<input type="checkbox"/> twitter.com	9	5	1.5 MB	1.5 MB	0	0	424
<input type="checkbox"/> www.pinterest.com	5	5	1.1 MB	1.1 MB	0	0	459
<input type="checkbox"/> tasty.co	2	2	853.2 KB	853.2 KB	0	0	177
<input type="checkbox"/> www.youtube.com	2	2	798.8 KB	798.8 KB	0	0	367
<input type="checkbox"/> www.buzzfeednews.com	4	3	781.9 KB	780.4 KB	0	0	152

Seed-Level Scoping

Documents from a specific host, sorted by amount of data.

★ Pro tip: look at this information in your crawl reports!

Document URL	New 	Data
https://www.buzzfeed.com/static-assets/js/core.c96c2948bd27e2ff9c52.js	Yes	776.9 KB
https://www.buzzfeed.com/static-assets/js/core.f23660feb1bf01393147.js	Yes	776.5 KB
https://www.buzzfeed.com/archive	Yes	559 KB
https://www.buzzfeed.com/buzz	Yes	436.6 KB
https://www.buzzfeed.com/sarahemcbride/why-its-so-hard-for-trans-women-to-talk-about-sexual-assault?utm_term=.xn3n	Yes	411.7 KB
https://www.buzzfeed.com/sarahemcbride/why-its-so-hard-for-trans-women-to-talk-about-sexual-assault	No	411.7 KB
https://www.buzzfeed.com/sarahemcbride/why-its-so-hard-for-trans-women-to-talk-about-sexual-assault?utm_term=.xn3n	No	411.7 KB
https://www.buzzfeed.com/entertainment	Yes	411.5 KB
https://www.buzzfeed.com/celebrity	Yes	409.7 KB
https://www.buzzfeed.com/rewind	Yes	408.2 KB
https://www.buzzfeed.com/food	Yes	405.5 KB
https://www.buzzfeed.com/music	Yes	405.4 KB
https://www.buzzfeed.com/parents	Yes	405 KB

Seed-Level Scoping

LOTS of blocked hosts!

Applies To	Rule	Url Match	Value
For seed https://www.buzzfeed.com/sarahemcbride/why-its-so-hard-for-tra	block URL if it	contains	"googlevideo"
For seed https://www.buzzfeed.com/sarahemcbride/why-its-so-hard-for-tra	block URL if it	contains	"buzzfeed.com/archive"
For seed https://www.buzzfeed.com/sarahemcbride/why-its-so-hard-for-tra	block URL if it	contains	"buzzfeed.com/buzz"
For seed https://www.buzzfeed.com/sarahemcbride/why-its-so-hard-for-tra	block URL if it	contains	"buzzfeed.com/entertainment"
For seed https://www.buzzfeed.com/sarahemcbride/why-its-so-hard-for-tra	block URL if it	contains	"buzzfeed.com/celebrity"
For seed https://www.buzzfeed.com/sarahemcbride/why-its-so-hard-for-tra	block URL if it	contains	"buzzfeed.com/rewind"
For seed https://www.buzzfeed.com/sarahemcbride/why-its-so-hard-for-tra	block URL if it	contains	"buzzfeed.com/food"
For seed https://www.buzzfeed.com/sarahemcbride/why-its-so-hard-for-tra	block URL if it	contains	"buzzfeed.com/music"
For seed https://www.buzzfeed.com/sarahemcbride/why-its-so-hard-for-tra	block URL if it	contains	"buzzfeed.com/parents"
For seed https://www.buzzfeed.com/sarahemcbride/why-its-so-hard-for-tra	block URL if it	contains	"buzzfeed.com/win"
For seed https://www.buzzfeed.com/sarahemcbride/why-its-so-hard-for-tra	block URL if it	contains	"buzzfeed.com/lol"
For seed https://www.buzzfeed.com/sarahemcbride/why-its-so-hard-for-tra	block URL if it	contains	"buzzfeed.com/weddings"
For seed https://www.buzzfeed.com/sarahemcbride/why-its-so-hard-for-tra	block URL if it	contains	"buzzfeed.com/bringme"
For seed https://www.buzzfeed.com/sarahemcbride/why-its-so-hard-for-tra	block URL if it	contains	"buzzfeed.com/tvandmovies"
For seed https://www.buzzfeed.com/sarahemcbride/why-its-so-hard-for-tra	block URL if it	contains	"buzzfeed.com/life"
For seed https://www.buzzfeed.com/sarahemcbride/why-its-so-hard-for-tra	block URL if it	contains	"buzzfeed.com/asis"
For seed https://www.buzzfeed.com/sarahemcbride/why-its-so-hard-for-tra	block URL if it	contains	"buzzfeed.com/goodful"
For seed https://www.buzzfeed.com/sarahemcbride/why-its-so-hard-for-tra	block URL if it	contains	"buzzfeed.com/nifty"
For seed https://www.buzzfeed.com/sarahemcbride/why-its-so-hard-for-tra	block URL if it	contains	"buzzfeed.com/shopping"
For seed https://www.buzzfeed.com/sarahemcbride/why-its-so-hard-for-tra	block URL if it	contains	"buzzfeed.com/quizzes"
For seed https://www.buzzfeed.com/sarahemcbride/why-its-so-hard-for-tra	block URL if it	contains	"buzzfeed.com/animals"
For seed https://www.buzzfeed.com/sarahemcbride/why-its-so-hard-for-tra	block URL if it	contains	"buzzfeed.com/community"
For seed https://www.buzzfeed.com/sarahemcbride/why-its-so-hard-for-tra	block URL if it	contains	"buzzfeed.com/trending"
For seed https://www.buzzfeed.com/sarahemcbride/why-its-so-hard-for-tra	block URL if it	contains	"buzzfeed.com/giftguide"
For seed https://www.buzzfeed.com/sarahemcbride/why-its-so-hard-for-tra	block URL if it	contains	"buzzfeed.com/videos"
For seed https://www.buzzfeed.com/sarahemcbride/why-its-so-hard-for-tra	block URL if it	contains	"buzzfeed.com/buzzfeed/settings"
For seed https://www.buzzfeed.com/sarahemcbride/why-its-so-hard-for-tra	block URL if it	contains	"buzzfeed.com/help"
For seed https://www.buzzfeed.com/sarahemcbride/why-its-so-hard-for-tra	block URL if it	contains	"buzzfeed.com/rss"
For seed https://www.buzzfeed.com/sarahemcbride/why-its-so-hard-for-tra	block URL if it	contains	"buzzfeed.com/consent-preferen

Want to hear more about our project?

Web Archiving & Metadata and Digital Object (MDOS) Joint Section Meeting

Saturday, August 3 • 11:30am - 12:45pm

Facilitated discussion on web archives and descriptive metadata!

Moderators: Carolyn Runyon & Julia Corrin

Panelists: Sumitra Duncan, Jane Kelly, & Greg Wiedeman

Session 702: Documenting Current Events and Controversial Topics

Monday, August 5 • 1:30pm - 2:30pm

Jennifer Weintraub: Documenting Current Events and
Controversial Topics

Jane Kelly: Collecting Material about the #metoo Movement

Samantha Abrams: Ivy Plus Libraries Partnership

Framework for Collection of Web Archives

Thank you!

Contribute!

<https://www.schlesinger-metoo-project-radcliffe.org/>

Contact me:

jane_kelly@radcliffe.harvard.edu

Contact the project team:

schlesinger_metoo@radcliffe.harvard.edu
