Carlin MacKenzie

Namespace Database

carlinmack / NamespaceDatabase

Unwatch 2 | Star 1 | Fork 1

Code | Issues | Pull requests | Security | Insights | Settings

master | 3 branches | 1 tag | Go to file | Add file | Code

carlinmack Add 4 new treemaps 7680d8d on 24 Oct 281 commits

nsdb	Add 4 new treemaps	2 months ago
plots	Add 4 new treemaps	2 months ago
slurm	remove unnecessary delays as .get() is blocking	8 months ago
sql	add importing blocked IPs	6 months ago
.gitignore	start interfacing with database	10 months ago
DOCUMENTATION.md	fix namespaces parameter	7 months ago
LICENSE	Add license	10 months ago
README.md	Update READMEs	2 months ago
programflow.png	refactor mirrors.py into nsdb	7 months ago
requirements.txt	add new plot and ip version of last five	5 months ago
schema.md	update documentation and schema	8 months ago
xmlformat.md	Add links to the edits shown in the XML	10 months ago

About

A tool to create a database of Wikipedia edits

meta.wikimedia.org/wiki/research:clas...

wikipedia | python3 | mysql

Readme

EPL-2.0 License

Releases

1 tags

Create a new release

Languages

Python 99.2% Shell 0.8%

enwiki dump progress on 20201120

This is the Wikimedia dump service. Please read the copyrights information. See Meta:Data dumps for documentation on the provided data formats.

Older versions of the 7zip decoder on Windows are known to have problems with some bz2-format files for larger wikis; we recommend the use of bzip2 for Windows for these cases. 7zip 20.00 alpha (x64) on Win10 LTSC 2019 and later versions should work properly.

Please report problems with these dumps on Phabricator and add the Dumps-generation tag.

See all databases list.

Last dumped on 2020-11-01

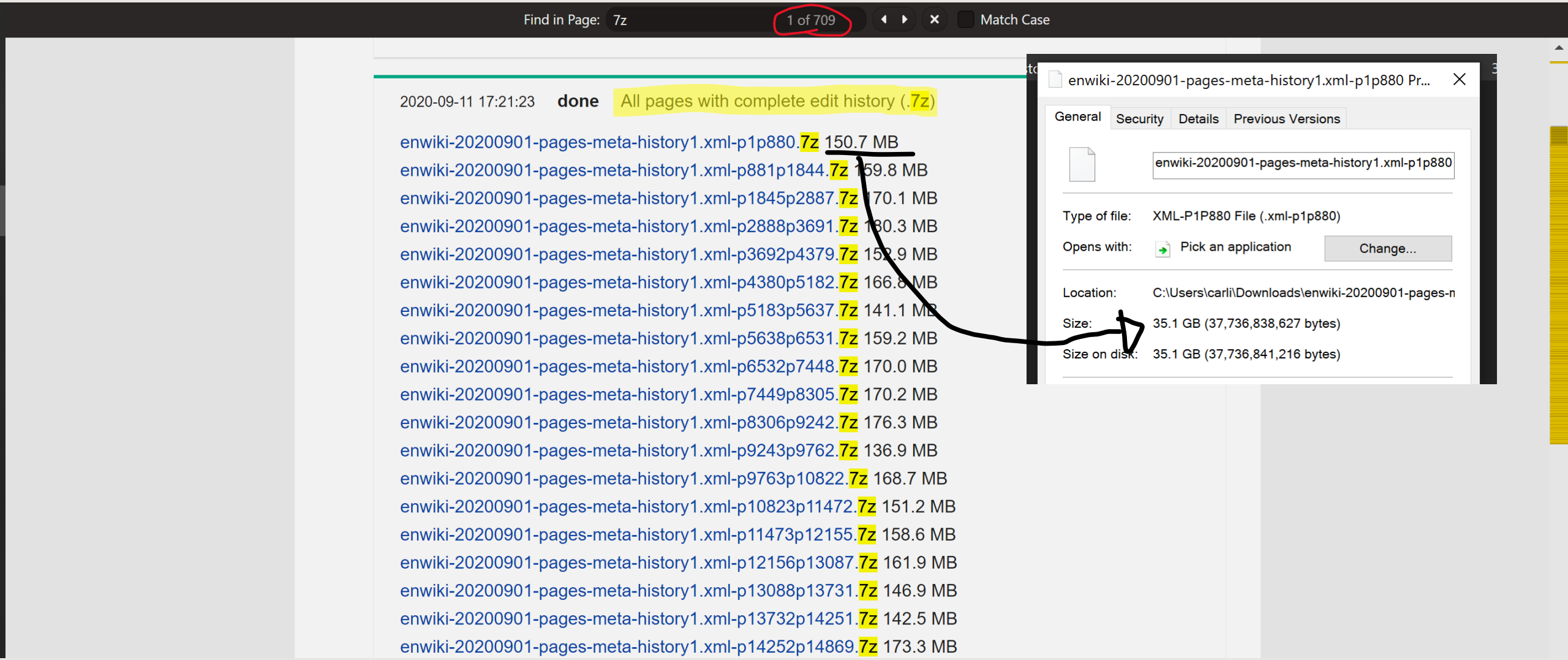
For a machine-readable version of the information on this page, see the json status file.

Dump complete

Verify downloaded files against the (md5), (sha1) checksums to check for corrupted files.

2020-11-22 12:32:41 **done** Recombine multiple bz2 streams

enwiki-20201120-pages-articles-multistream.xml.bz2 17.7 GB
enwiki-20201120-pages-articles-multistream-index.txt.bz2 217.0 MB


Find in Page: 7z
1 of 709
Match Case
2020-09-11 17:21:23 done All pages with complete edit history (.7z)
enwiki-20200901-pages-meta-history1.xml-p1p880.7z 150.7 MB
enwiki-20200901-pages-meta-history1.xml-p881p1844.7z 159.8 MB
enwiki-20200901-pages-meta-history1.xml-p1845p2887.7z 170.1 MB
enwiki-20200901-pages-meta-history1.xml-p2888p3691.7z 130.3 MB
enwiki-20200901-pages-meta-history1.xml-p3692p4379.7z 152.9 MB
enwiki-20200901-pages-meta-history1.xml-p4380p5182.7z 166.8 MB
enwiki-20200901-pages-meta-history1.xml-p5183p5637.7z 141.1 MB
enwiki-20200901-pages-meta-history1.xml-p5638p6531.7z 159.2 MB
enwiki-20200901-pages-meta-history1.xml-p6532p7448.7z 170.0 MB
enwiki-20200901-pages-meta-history1.xml-p7449p8305.7z 170.2 MB
enwiki-20200901-pages-meta-history1.xml-p8306p9242.7z 176.3 MB
enwiki-20200901-pages-meta-history1.xml-p9243p9762.7z 136.9 MB
enwiki-20200901-pages-meta-history1.xml-p9763p10822.7z 168.7 MB
enwiki-20200901-pages-meta-history1.xml-p10823p11472.7z 151.2 MB
enwiki-20200901-pages-meta-history1.xml-p11473p12155.7z 158.6 MB
enwiki-20200901-pages-meta-history1.xml-p12156p13087.7z 161.9 MB
enwiki-20200901-pages-meta-history1.xml-p13088p13731.7z 146.9 MB
enwiki-20200901-pages-meta-history1.xml-p13732p14251.7z 142.5 MB
enwiki-20200901-pages-meta-history1.xml-p14252p14869.7z 173.3 MB
enwiki-20200901-pages-meta-history1.xml-p1p880 Pr...
General
Security
Details
Previous Versions
enwiki-20200901-pages-meta-history1.xml-p1p880
Type of file: XML-P1P880 File (.xml-p1p880)
Opens with: Pick an application
Change...
Location: C:\Users\carli\Downloads\enwiki-20200901-pages-n
Size: 35.1 GB (37,736,838,627 bytes)
Size on disk: 35.1 GB (37,736,841,216 bytes)

Join WikiConference North America live online: Friday Dec 11- Sunday Dec 13!

Main page
Wikimedia News
Translations
Recent changes
Random page
Help
Babel

Community
Wikimedia Resource Center
Wikimedia Forum
Mailing lists
Requests
Babylon
Reports
Research
Planet Wikimedia

Beyond the Web
Meet Wikimedians
Events
Movement affiliates
Donate

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page

Datasets

Various places that ha

Also, you can now sto

List [edit | edit sour

Official Wikipedia dat
Parsoid exposes sem dewikibooks, ... The p modified) HTML bac
Taxobox - Wikipedia
Wikipedia³ is a conv
DBpedia Facts extrac
Multiple data sets (En
This is an alphabeti
Using the Wikiped
Wikipedia: Lists of co
Apache Hadoop is a
Wikipedia XML Data
Wikipedia Page Traffi
Complete Wikipedia e

- StatMediaWiki
- Wikimedia group on The Data
- mw:Backing_up_a_wiki How to

Category: Data dumps

WikiPrep Perl s

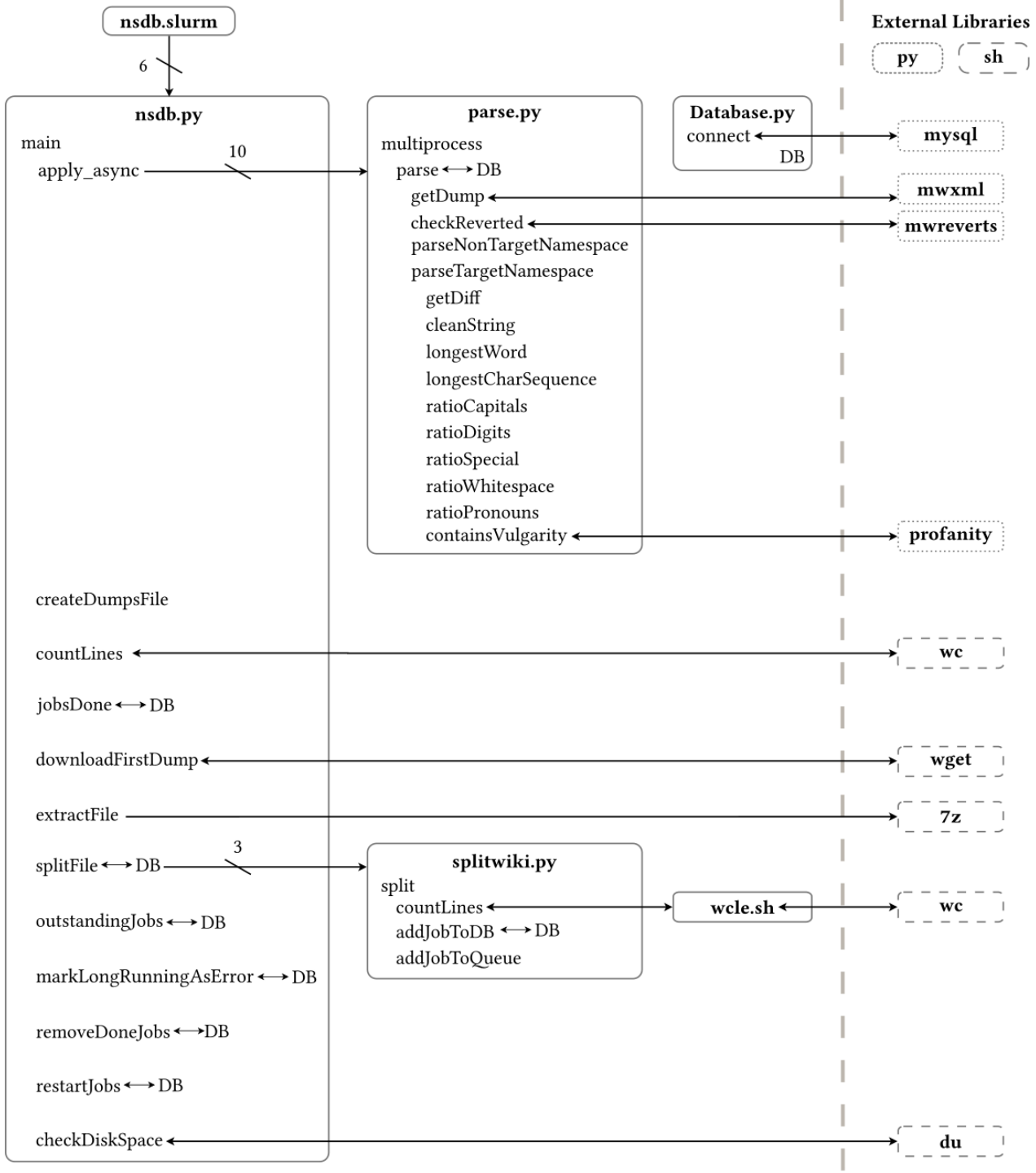
Tools to extract data from Wikipedia: [edit | edit source]

This table might be migrated to the Knowledge Extraction Wikipedia Article

Tool	Description	URL	Last Updated
Wikilytics	Extracting the dumps into a NoSQL database	[24]	2017
Wikipedia2text	Extracting Text from Wikipedia	[25]	2008
Traffic Statistics	Wikipedia article traffic statistics	[26]	Dead
Wikipedia to Plain Text	Generating a Plain Text Corpus from Wikipedia	[27]	2009
DBpedia Extraction Framework	The DBpedia software that processes Wikipedia pages using templates and stores the data obtained as structured data. Highly configurable for other MediaWikis also).	[28] [29] github	2019
[illegible]	[illegible] for archiving [illegible] Wikipedia	github	2019
History Flow	History flow is a tool for visualizing dynamic, evolving documents and the interactions of multiple collaborating authors	[30]	Dead
WikiXRay	This tool includes a set of Python and GNU R scripts to obtain statistics, graphics and quantitative results for any Wikipedia language version	[31]	2012
StatMediaWiki	StatMediaWiki is a project that aims to create a tool to collect and aggregate information available in a MediaWiki installation.Results are static HTML pages including tables and graphics that can help to analyze the wiki status and development, or a CSV file for custom processing.	[32]	Dead
Java Wikipedia Library (JWPL)	This is a open-source, Java-based application programming interface that allows to access all information contained in Wikipedia	[33]	2016
Wikokit	Wiktionary parser and visual interface	github	2019
wiki-network	Python scripts for parsing Wikipedia dumps with different goals	github	2012
Pywikipediabot	Python Wikipedia robot framework	[34]	2019
WikiRelate	API for computing semantic relatedness using Wikipedia (Strube and Ponzetto,2006)	[35]	2006

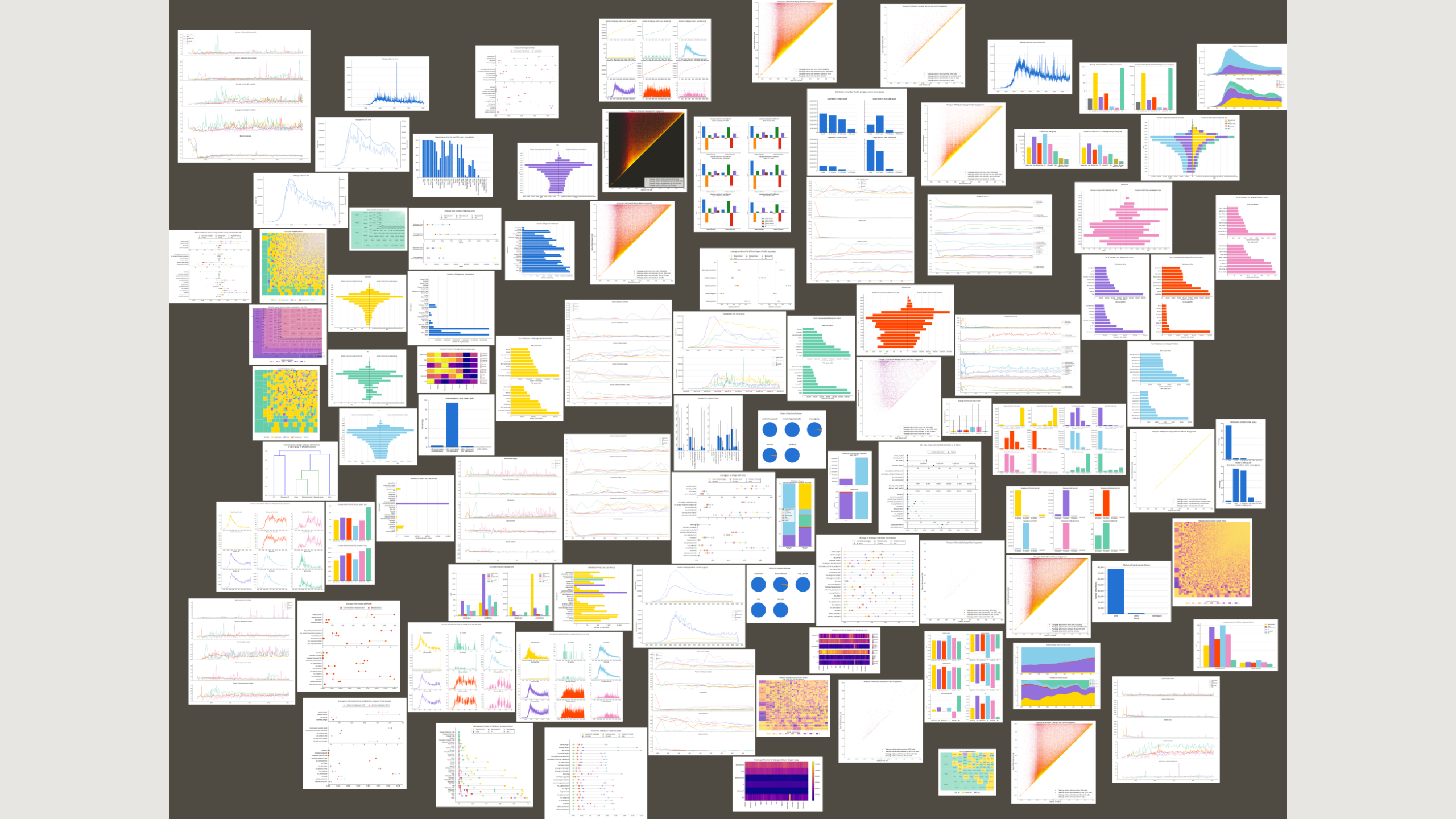
No actively updated tool to create a database of edits ☹

1. Finds the fastest mirror
2. Downloads and extracts
3. Splits into partitions
4. Extracts features in parallel
5. Inserts into a database


nsdb.slurm
6
External Libraries
py
sh
nsdb.py
main
apply_async
10
parse.py
multiprocess
parse ⟷ DB
getDump
checkReverted
parseNonTargetNamespace
parseTargetNamespace
getDiff
cleanString
longestWord
longestCharSequence
ratioCapitals
ratioDigits
ratioSpecial
ratioWhitespace
ratioPronouns
containsVulgarity
Database.py
connect
DB
mysql
mwxml
mwreverts
profanity
createDumpsFile
countLines
wc
jobsDone ⟷ DB
downloadFirstDump
wget
extractFile
7z
splitFile ⟷ DB
3
splitwiki.py
split
countLines
addJobToDB ⟷ DB
addJobToQueue
wcle.sh
wc
outstandingJobs ⟷ DB
markLongRunningAsError ⟷ DB
removeDoneJobs ⟷ DB
restartJobs ⟷ DB
checkDiskSpace
du

I have all this data, but what does it mean?

Plots!



To note

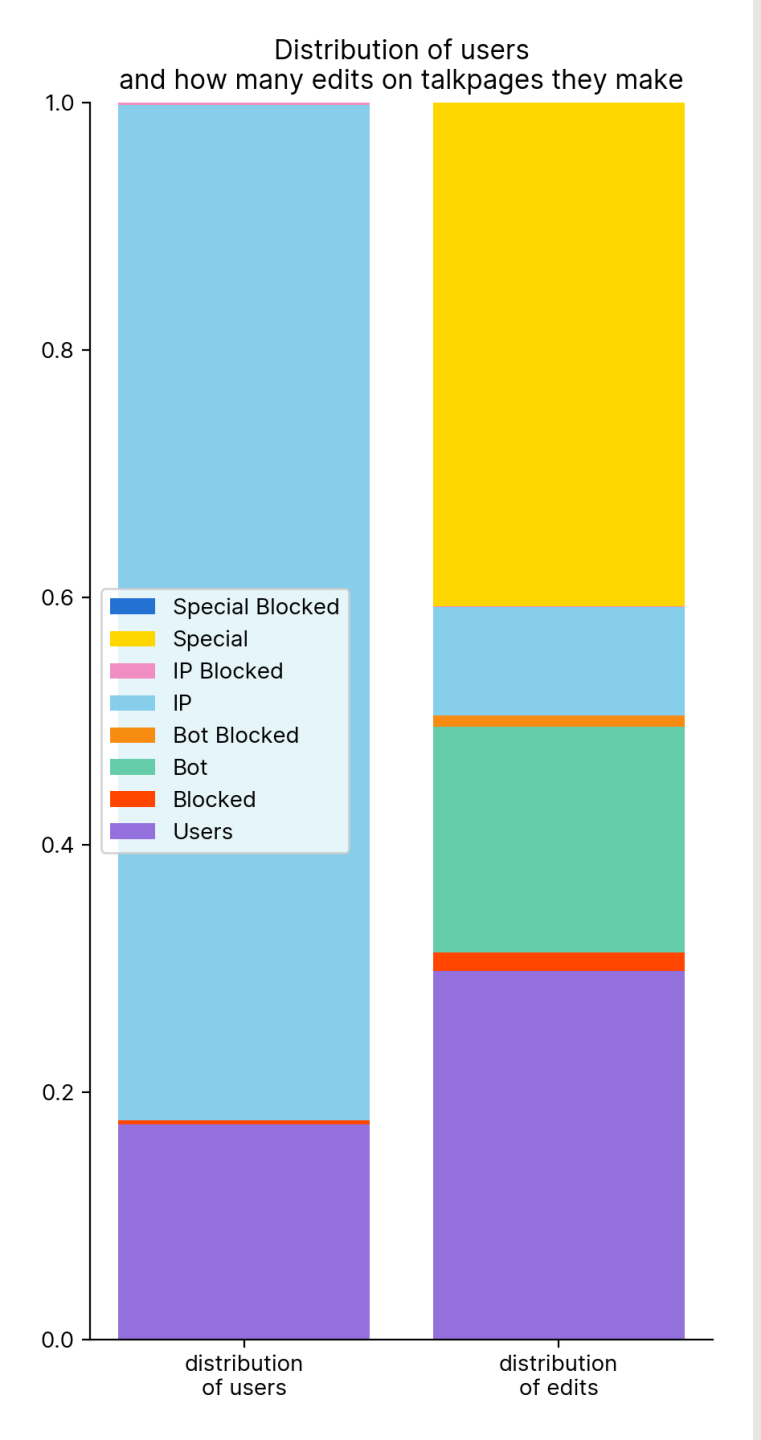
1. Users
2. Blocked Users
3. Special Users
4. IP Users
5. Blocked IP
6. Bot

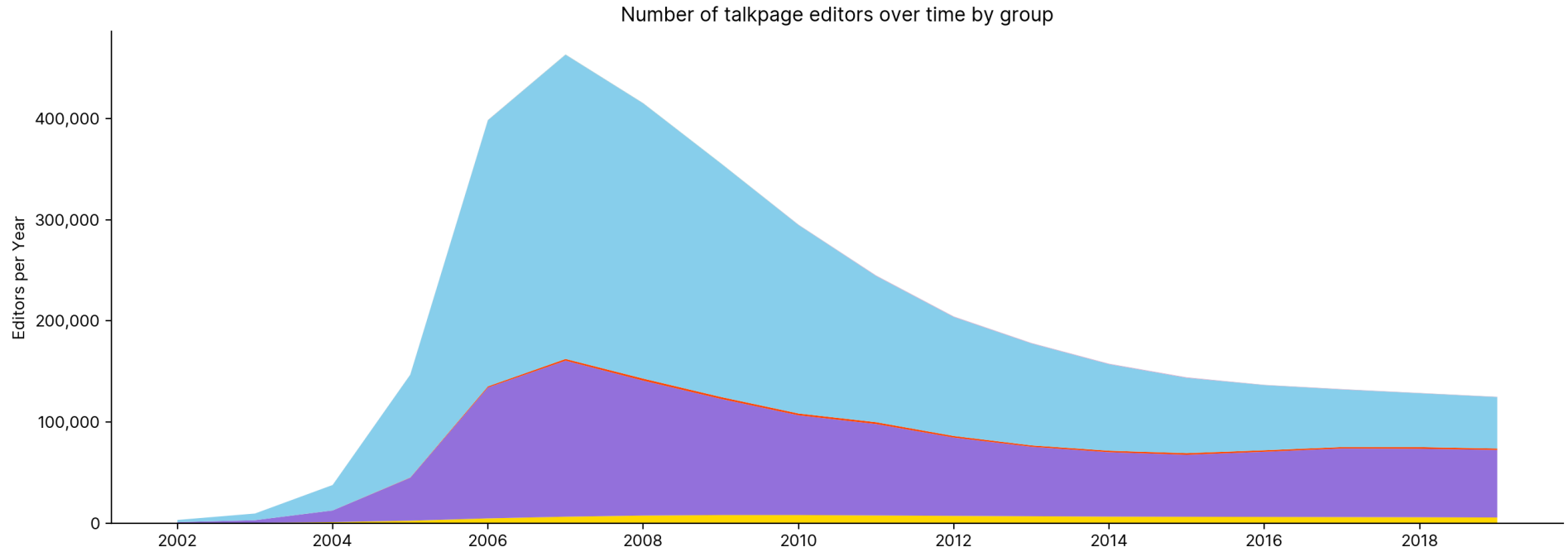
Also

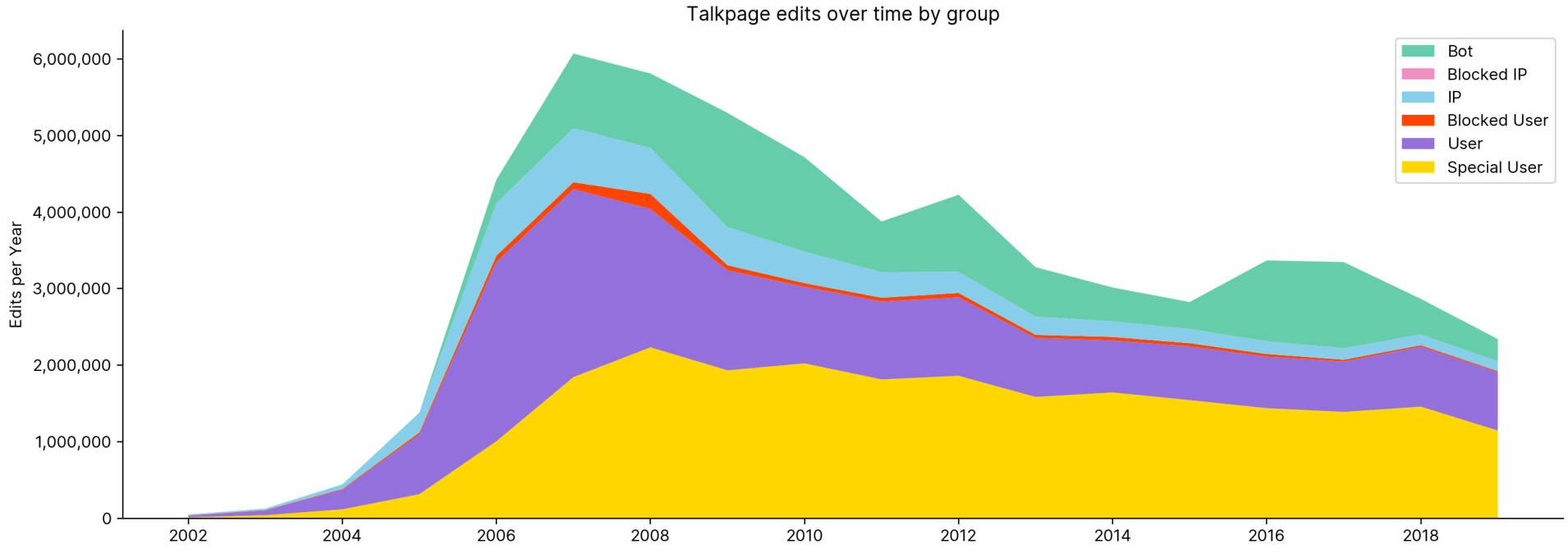
All these plots focus on article talkpages

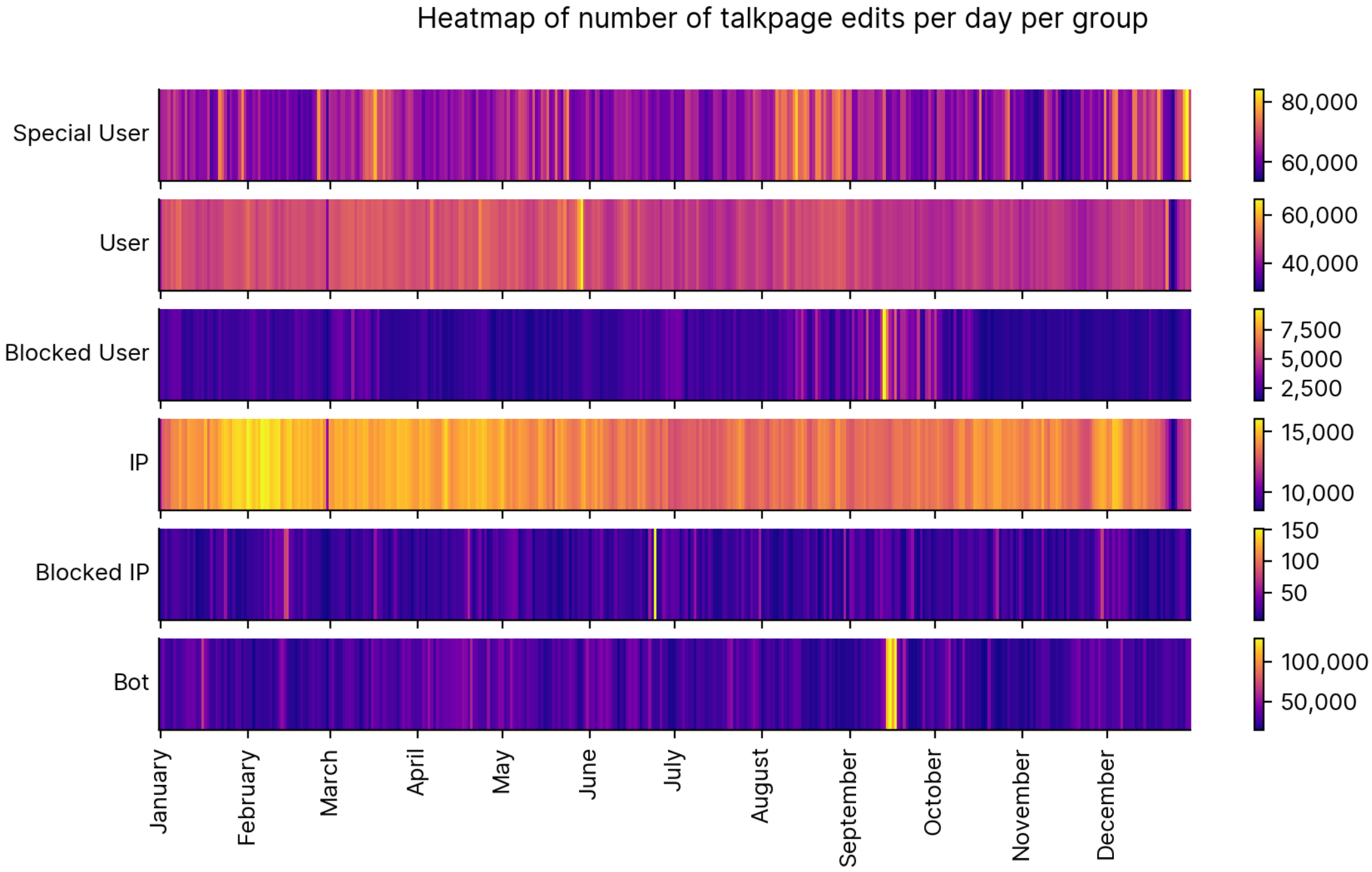
How many users per group?

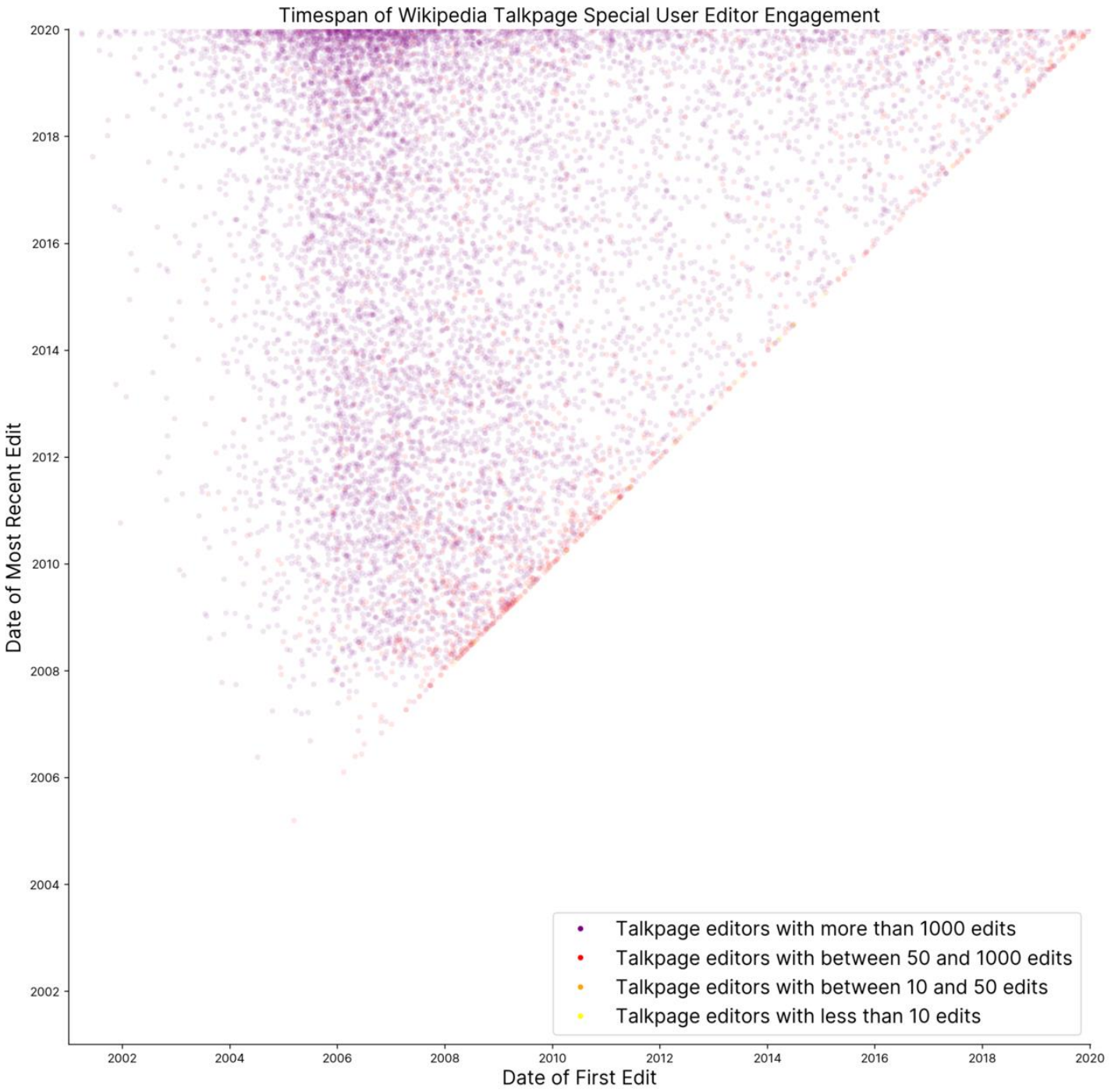
How many (talkpage) edits do they make?




Number of talkpage editors over time by group
Editors per Year
400,000
300,000
200,000
100,000
0
2002
2004
2006
2008
2010
2012
2014
2016
2018


Talkpage edits over time by group
Edits per Year
6,000,000
5,000,000
4,000,000
3,000,000
2,000,000
1,000,000
0
2002
2004
2006
2008
2010
2012
2014
2016
2018
Bot
Blocked IP
IP
Blocked User
User
Special User


Heatmap of number of talkpage edits per day per group
Special User
User
Blocked User
IP
Blocked IP
Bot
80,000
60,000
60,000
40,000
7,500
5,000
2,500
15,000
10,000
150
100
50
100,000
50,000
January
February
March
April
May
June
July
August
September
October
November
December


Timespan of Wikipedia Talkpage Special User Editor Engagement
Date of Most Recent Edit
2020
2018
2016
2014
2012
2010
2008
2006
2004
2002
Date of First Edit
2002
2004
2006
2008
2010
2012
2014
2016
2018
2020
Talkpage editors with more than 1000 edits
Talkpage editors with between 50 and 1000 edits
Talkpage editors with between 10 and 50 edits
Talkpage editors with less than 10 edits

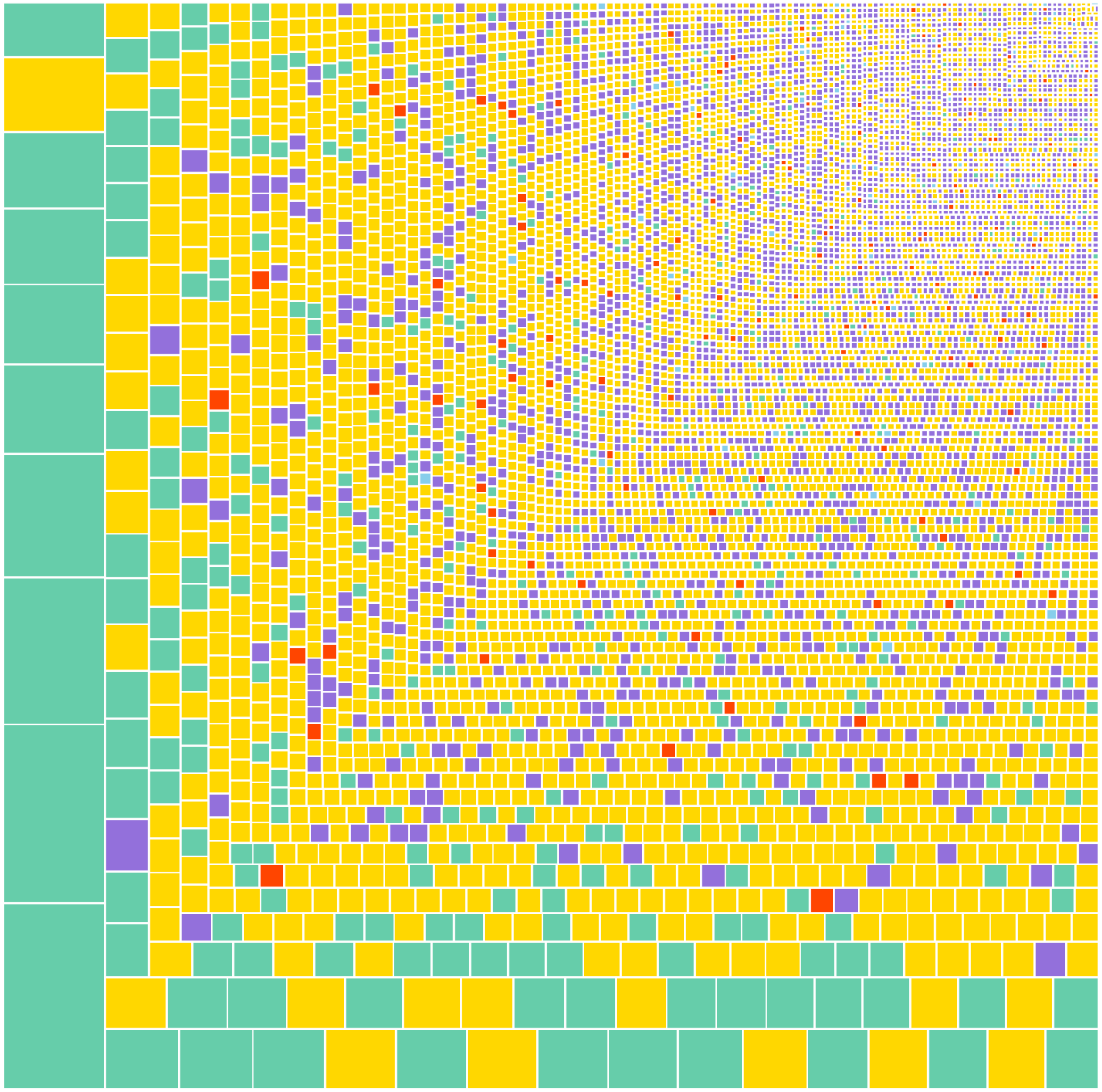
Next steps

Paper?

Toolforge?

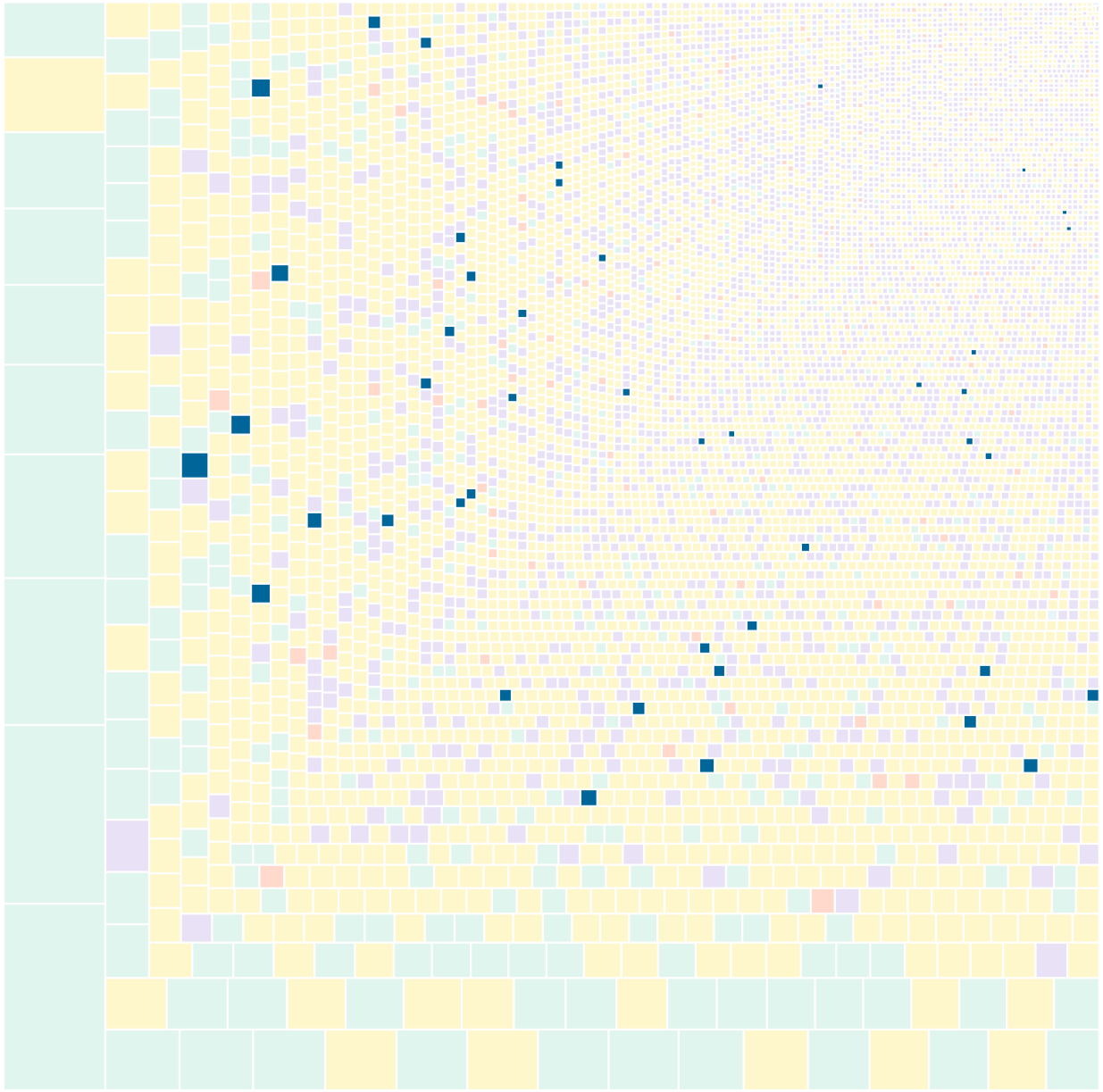
Run this on mainspace?

Top 10,000 Wikipedia editors



Bot Special User User Blocked User IP

Top 10,000 Wikipedia editors



Bot
Special User
User
Blocked User
IP
WikiCon NA 2020 Attendees