

# Вебсйтес оф тхе Формер Совиет Унион & Еастерн Еуропе

Тeam компромат

# Websites of the Former Soviet Union & Eastern Europe

Тeam компромат

# Research Question:

What countries are linking to each other in the collection?

# Methods

1. Reverse engineered the seed list with Seed ID in WARC filename + Archive-It API
2. Learned about WARCs and the AUT derivatives
3. Made derivatives with AUT, extracted all URLs from HTML
  - a. 93% parseable, 7 million URLs
4. Manual coding of Seed List, Top Level Domains and Crawled URLs (Google Sheets)
5. [Notebooks](#) to parse URLs with coded TLDs to edge list that we can't run
6. Visualised Data via Gephi, Google Maps, Charts
7. Visualisations helped inform further research questions
8. Took some Paracetamol ; ate Chocolate
9. Made our presentation



```
In [60]: fullurls = "/mnt/data/ivy/former-soviet-union/derivatives/all-domains/11360-fullurls.txt"

output_csv(get_counts('\.(\w+)$', fullurls), 'tlds.csv')
```

```
org,108119
hu,76011
ru,61108
pl,47130
com,26025
ua,25600
al,19623
ba,11268
rs,8297
mk,7930
me,4918
info,1800
net,723
ro,282
tv,135
us,13
fm,12
io,7
ca,6
de,6
uk,6
it,5
```



# <https://archive-it.org/collections/11360>

HOME EXPLORE LEARN MORE CONTACT US

Login

The leading web archiving service  
for collecting and accessing  
cultural heritage on the web  
*Built at the Internet Archive*



---

[Explore](#) >> [Ivy Plus Libraries Confederation](#) >> [Websites of the Former Soviet Union & Eastern Europe](#)



## Websites of the Former Soviet Union & Eastern Europe

Collected by: [Ivy Plus Libraries Confederation](#)

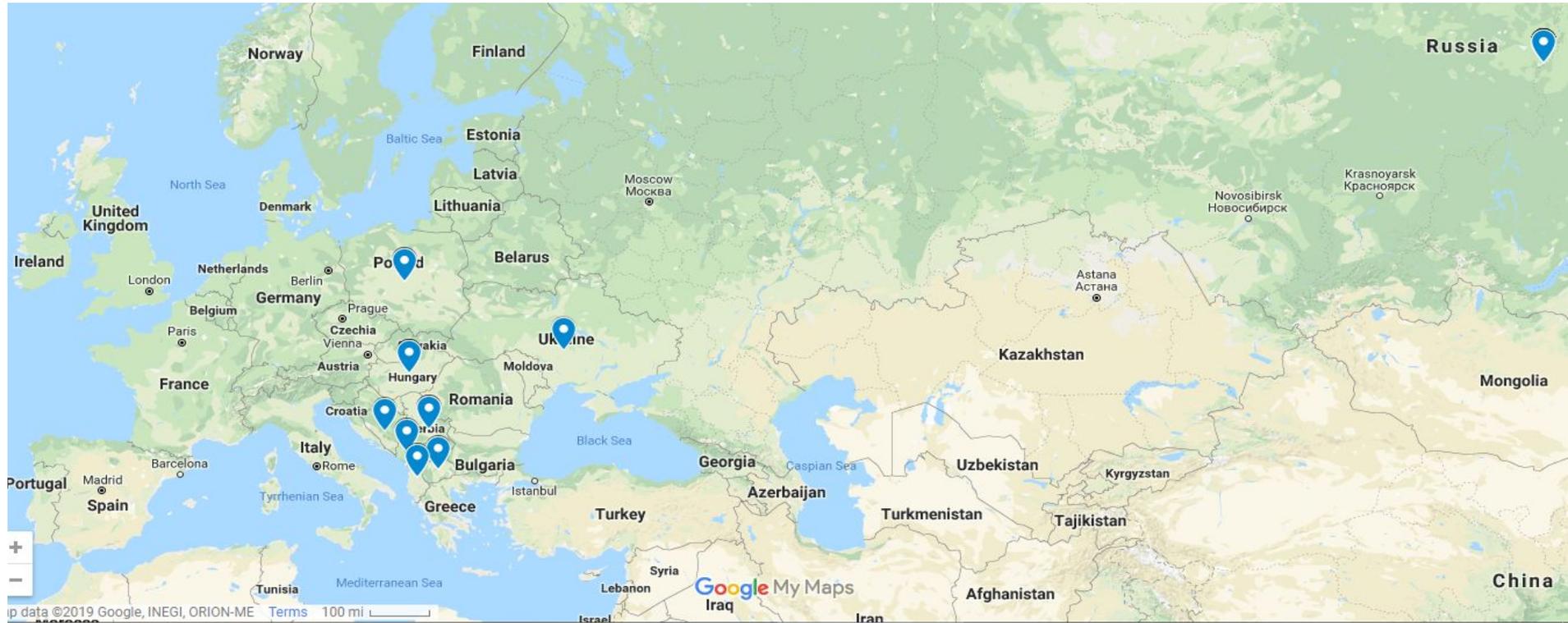
**Archived since:** Nov, 2018

**Description:** The Eastern Europe and Former Soviet Union Web Archive is an initiative developed by librarians at Columbia, Princeton, Yale, and New York Universities, and the New York Public Library, in partnership (as the Ivy Plus Libraries Confederation), with Brown University, the University of Chicago, Cornell University, Dartmouth University, Duke University, Harvard University, Johns Hopkins University, and the University of Pennsylvania. The collection is curated by Thomas Keenan (Princeton), Robert Davis (Columbia), Anna Arays (Yale), Alla Royslance (New York University), and Bogdan Horbal (New York Public Library). The Archive represents an effort to preserve research-valuable web content from Eastern Europe and the territories of the Former Soviet Union by a group of research librarians responsible for that part of the world. The countries of the region in recent years have been publishing a wide variety of websites likely to be of value to contemporary and future humanities, social science, and history projects, and this Archive has been established as an attempt to identify, capture, and preserve this material. The thematic and generic scope of the archive is deliberately broad, and includes websites published by political parties, non-governmental organizations and activist groups, artists and cultural collectives, and historians, philosophers, and other intellectuals.

**Subject:** [Arts & Humanities](#), [Society & Culture](#), [Government](#)

**Creator:** [Ivy Plus Libraries Confederation](#)

# Seed List Map

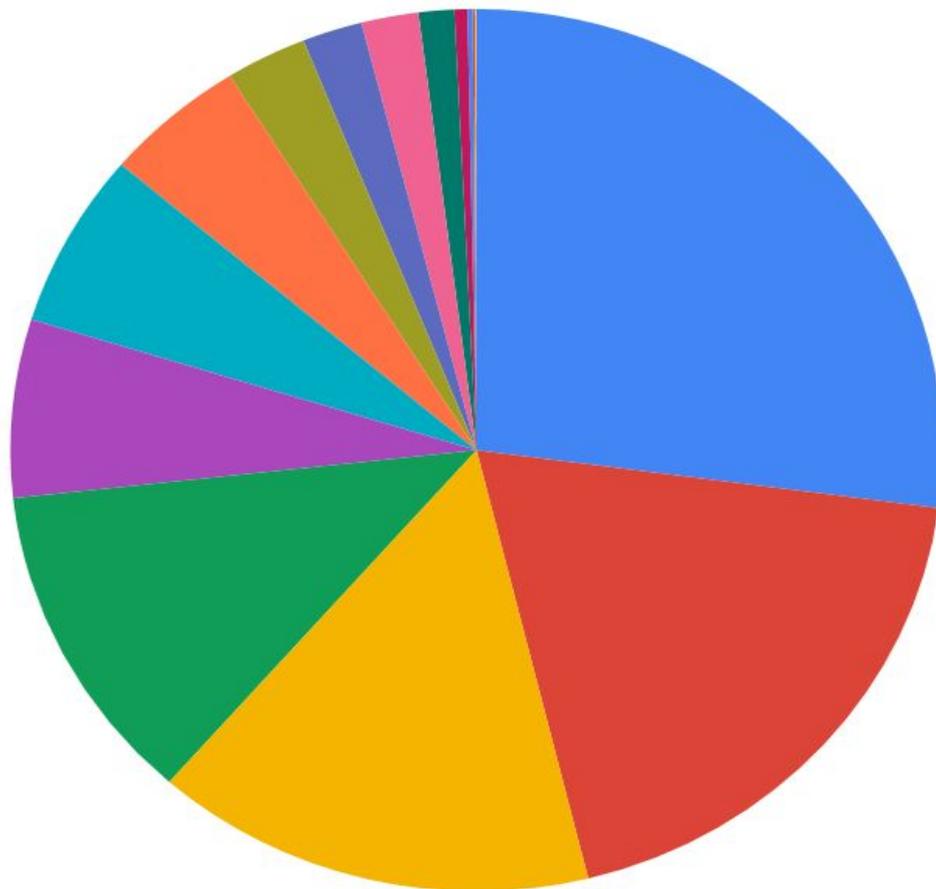


# Crawled URLs



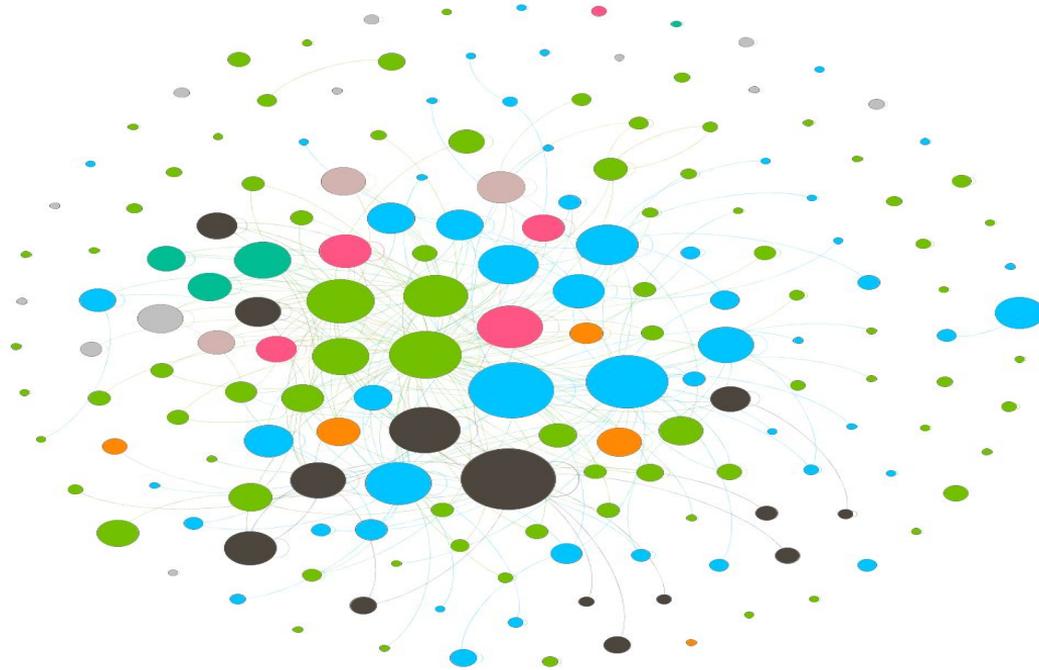
# Crawled URLs

TLD	Number
ru	15
org	8
hu	8
rs	4
al	3
mk	3
pl	3
ba	2
me	2
com	1
info	1
ua	1

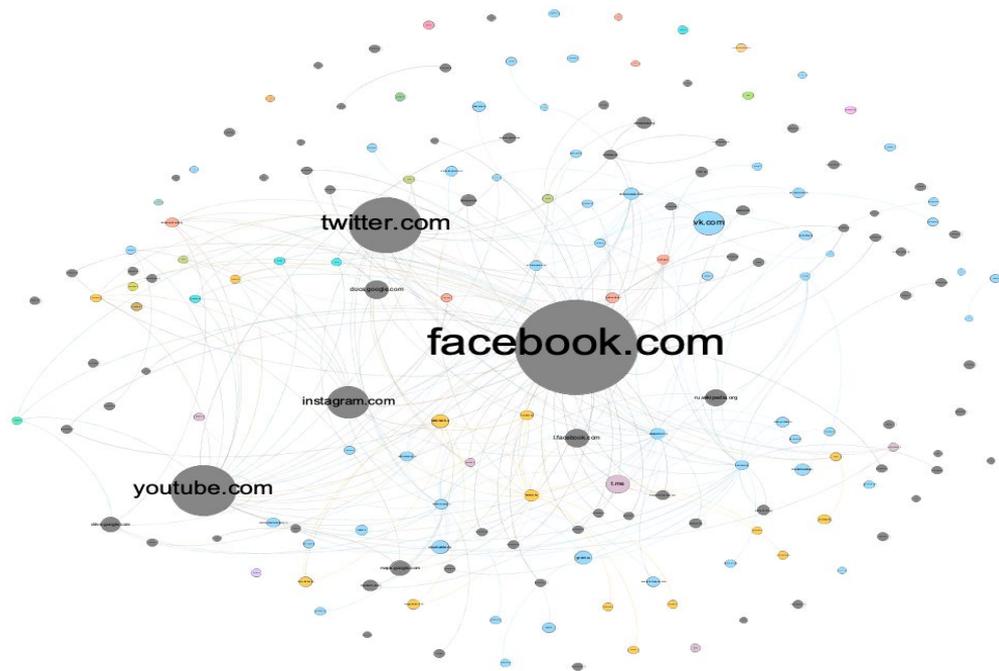


- org
- hu
- ru
- pl
- com
- ua
- al
- ba
- rs
- mk
- me
- 28 more

# Network Analyses of What is in the collection



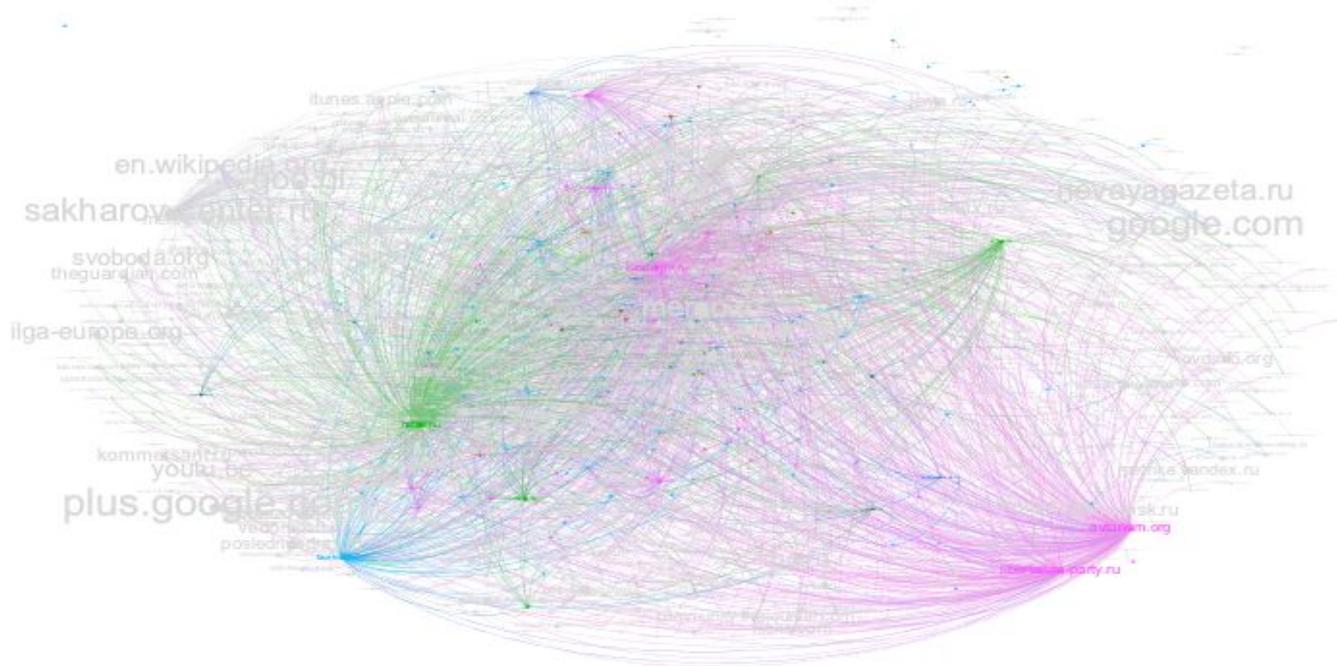
# Links out from the archived content



Uncrawled	(90.99%)
Transnational	(4.11%)
Russia	(2.65%)
Hungary	(0.73%)
Montenegro	(0.26%)
Poland	(0.26%)
Serbia	(0.21%)
Albania	(0.16%)
Canada	(0.05%)
Federated States of Micronesia	(0.05%)
Tuvalu	(0.05%)
Bosnia	(0.05%)
Macedonia	(0.05%)
Spain	(0.05%)
Belize	(0.05%)
Germany	(0.05%)
Ukraine	(0.05%)
British Indian Ocean Territory	(0.05%)
United Kingdom	(0.05%)
Romania	(0.05%)

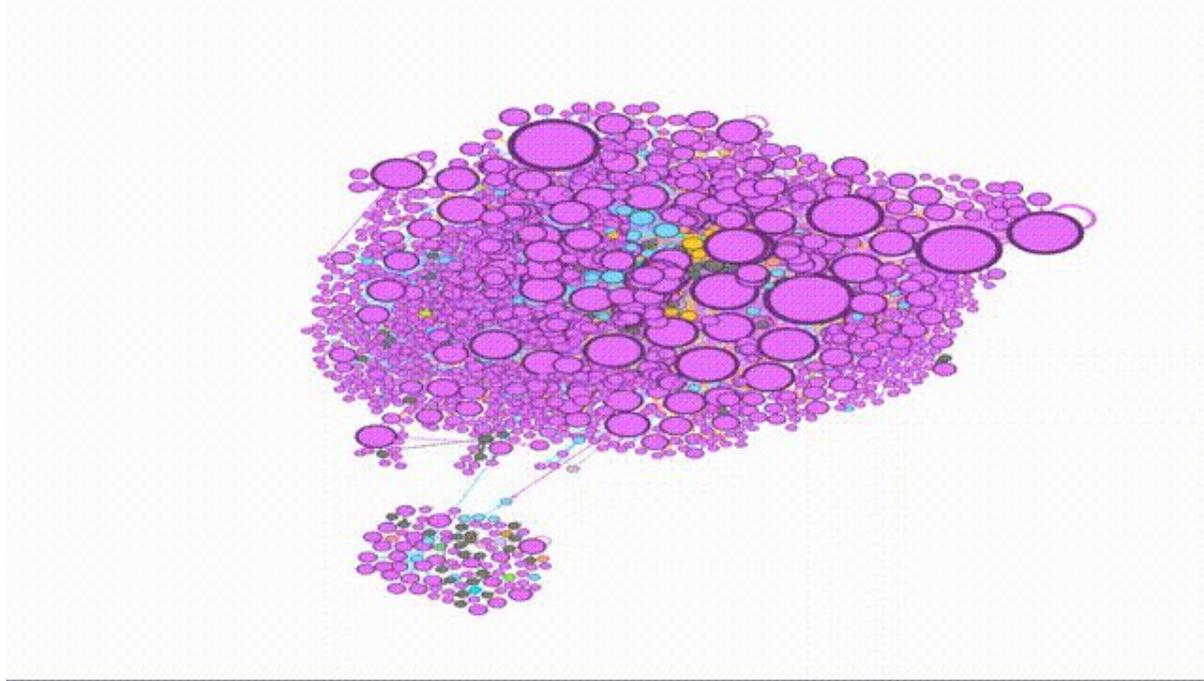


# Uncrawled Domains



Uncrawled	(90.99%)
Transnational	(4.11%)
Russia	(2.65%)
Hungary	(0.73%)
Montenegro	(0.26%)
Poland	(0.26%)
Serbia	(0.21%)
Albania	(0.16%)
Canada	(0.05%)
Federated States of Micr...	(0.05%)
Tuvalu	(0.05%)
Bosnia	(0.05%)
Macedonia	(0.05%)
Spain	(0.05%)
Belize	(0.05%)
Germany	(0.05%)
Ukraine	(0.05%)
British Indian Ocean Ter...	(0.05%)
United Kingdom	(0.05%)
Romania	(0.05%)

# Uncrawled Hot Mess



# Facebook Groups Not Crawled (35)

Прямужинская вольная артель  
Public group

About  
Discussion  
Members  
Events  
Photos  
Files

Search this group

Shortcuts  
Nogushi

RECENT ACTIVITY  
Mala Vida shared a post.  
March 14 at 8:57 AM

DESCRIPTION  
Прямужинская артель – это активистский проект по восстановлению... See More

GROUP TYPE  
Club  
Chat (7)

Похищение Европы: хроники гибридной войны  
Public group

About  
Discussion  
Members  
Events  
Videos  
Photos  
Files

Search this group

Shortcuts  
Nogushi

RECENT ACTIVITY  
Mihails Silovs shared a link.

DESCRIPTION  
Важные угрозы, провокации, кх погумн пропагандистские... See

ОСОБОЕ МНЕНИЕ  
Public group

About  
Discussion  
Members  
Events  
Videos  
Photos

Search this group

Shortcuts  
Nogushi

RECENT ACTIVITY  
Vladimir Polezyuk  
Yesterday at 12:16 PM

POPULAR TOPICS IN POSTS  
Саха Сэмин (1)

Оргкомитет Форума в Берлине  
Public group

About  
Discussion  
Members  
Events  
Videos  
Photos  
Files

Search this group

Shortcuts  
Nogushi

RECENT ACTIVITY  
Игорь Эйдман shared a post.  
4 hrs

Игорь Эйдман

GROUP BY  
Forum russischsprachiger Euro...  
181 like this

Chat (9)

Debreceni Színtér Egyesület  
Closed group

Shortcuts  
Nogushi

Join this group to see the discussion, post and comment.

About This Group

Description  
A Debreceni Színtér Egyesület azzal a céllal jött létre, hogy a régióban élő LGBT (leszbikus, meleg, biszexuális, transznemű) emberek és

HISTORY  
Group created on December 10, 2012  
See More

Chat (9)

Свободу Александру Кольченко! / Free Aleksandr Kolchenko!  
Public group

About  
Discussion  
Members  
Events  
Photos  
Files

Search this group

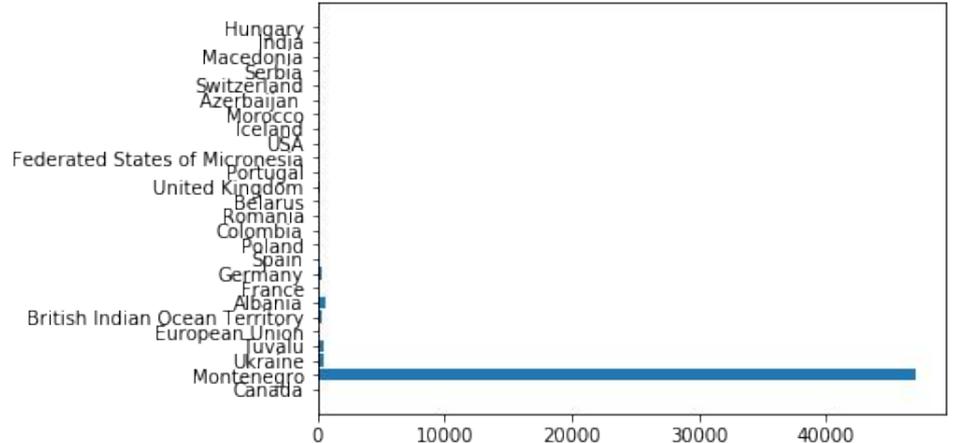
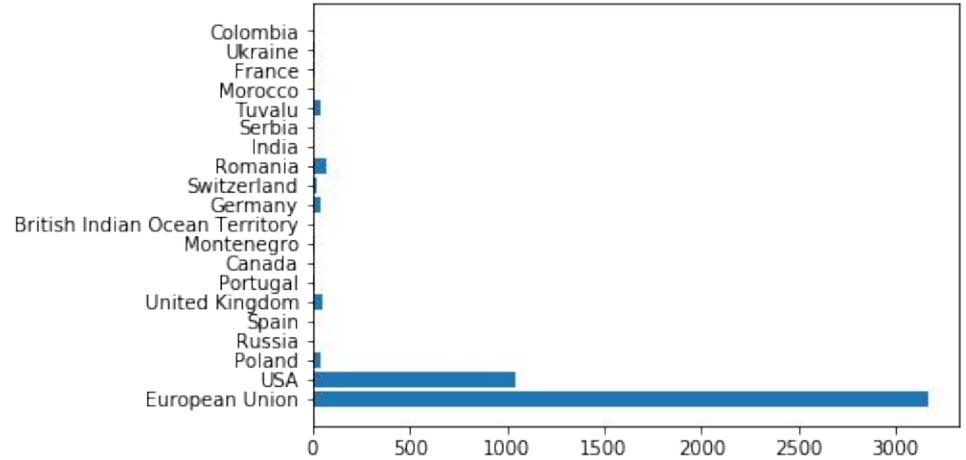
Shortcuts  
Nogushi

RECENT ACTIVITY  
Марина Попова shared a post.  
March 3 at 5:11 AM

DESCRIPTION  
В группе публикуются новости кампании поддержки крымского политического Александра Кольченко

Chat (9)

# Outlinks: Hungary vs. Russia TLDs



# Lessons Learned

Provenance of collections is important to inform your research questions

Defining geographic regions is challenging as many interpretations and politically charged

The web is not easy to define to one geographic region as there are lots of transnational web platforms

Still need to do a visual analysis of a sample of the content and manually code content

Supplied AUT domain visualization included uncrawled content that was not part of the collection.



# Q&A

Thank you for listening!