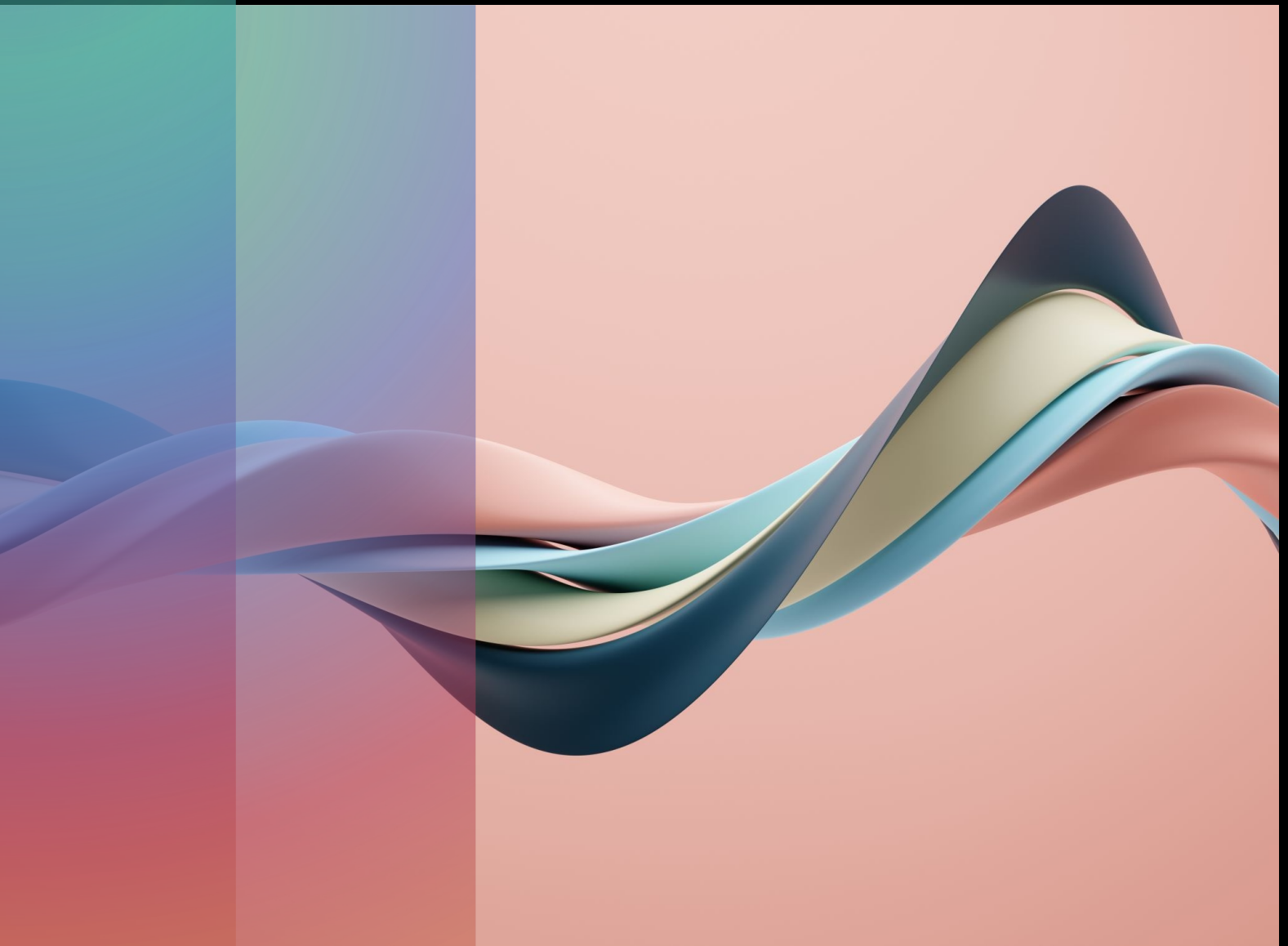


Analyzing web archives

Using Python and Solrwayback to explore Covid-19 content on Niagara Region municipal webpages

Introduction

- Fletcher Johnson - Research Assistant @ Brock
- B.Sc Computer Science, UofT (2010), MLIS Western University (2021)
- Previous industry experience, transition into library work

“Crisis Communication in the Niagara Region during the COVID-19 Pandemic”

- Analysis of 13 municipal webpages from April, 2020 – December, 2021
- Examination of how municipalities talked about Covid during the crisis
- Team members: Tim Ribaric, David Sharron, Cal Murgu, Dr. Karen Louise Smith, Dr. Duncan Koerber, and myself.
- My role: Technical support (Solrwayback + Notebooks)

The Data

- 300~ GB, captured using Archive-it.org service (IA commercial arm)
- WARC format
- Full datasets with API
- Derivatives accessible using Archives Research Compute Hub (ARCH)

The Tools: SolrWayback

- Project by Royal Danish Library for viewing Danish Internet (2005-onwards)
- “Search engine for WARCs”
- Like Wayback machine, but better

SolarWayback – Why?

- Much faster than Wayback machine –
  - We host ours on an AWS t2.small (1 cpu, 2 GB ram)

- Full text search (Solr)
- Complex queries with Booleans, wildcards, and facets
- No API support, though CSV data export available.

SolrWayback Facets

content
content_encoding
content_language
content_type
crawl_date
domain
links
links_domains
links_images
source_file
type

content
content_encoding
url
url_path
title
status_code
public_siffix

... and more

SolrWayback

Not secure | 3.136.85.220:8080/solrwayback/search?query=%2a%3A%2a&grouping=false&imgSearch=false&offset=0&urlSearch=false&facet...

TD Canada Trust - P... | Western Libraries -... | Johnson, Fletcher -... | JupyterLite | Mail - Fletcher John... | Home : Toronto Pu... | JavaScript | MDN | Exploring ES6 | Other bookmarks

SEARCH WITH UPLOADED FILE | GPS IMAGE SEARCH | TOOLBOX

FACETS

domain

- brocku.ca (291,511)
- gncc.ca (115,151)
- youtube.com (90,754)
- twitter.com (78,826)
- thestar.com (76,021)
- stcatharinesstand... (71,701)
- portcolborne.ca (38,605)
- googlevideo.com (34,399)
- google.com (29,478)
- ytimg.com (25,674)

more domains

content_type_norm

- html (841,776)
- other (385,462)
- image (266,563)
- text (100,611)
- video (2,066)
- pdf (1,681)
- audio (47)

RESULTS

Showing **0 - 20** of **1,598,206** entries matching query.

See available export options

Previous 20 | Next 20

LORD MAYOR AND COUNCIL | NIAGARA ON THE LAKE

#1
score: 21

Type: html, web page @ notl.com

Date: 09/07-2021

Url: https://www.notl.com/content/lord-mayor-and-council

Highlighted content:

"
Skip to main content Niagara-on-the-Lake is in Step 2 of the Provincial Roadmap to Reopen. Click
"

Images: showing 4 out of 10 | See all images




View data fields

DET KGL. BIBLIOTEK

SOLRWAYBACK

SEARCH HINTS

Two colons without quote signs. When a qualified search is performed, consider quoting the value - (domain:portcolborne.ca OR domain:welland.ca) AND (content_type_norm:html AND links_domains:niagararegion.ca AND crawl_date:"[2021-01-01T00:00:00Z" TO 2022-02-01T00:00:00Z]) covid

(domain:portcolborne.ca OR domain:welland.ca) AND (content_type_norm:html AND links_domains:niagararegion.ca AND crawl_date:[2021-01-01T00:00:00Z TO 2022-02-01T00:00:00Z]) covid

GROUPED SEARCH [?] IMAGE SEARCH [?] URL SEARCH [?]

SEARCH WITH UPLOADED FILE GPS IMAGE SEARCH TOOLBOX

FACETS

domain
- portcolborne.ca (658)
- welland.ca (21)

content_type_norm
- html (679)

type
- web page (679)

crawl_year
- 2021 (679)

status_code
- 200 (679)

public_suffix
- ca (679)

RESULTS

Showing **0** - **20** of **679** entries matching query.

See available export options

Previous 20 Next 20

COVID-19

#1
score: **39.759163**

Type: html, web page @ welland.ca

Date: 27/08-2021

Url: https://www.welland.ca/hottopics/COVID-19.asp

Highlighted content:

*"Skip to Main Content Search for: Menu Listen **COVID**-19 information details Welland Response"*

Images: showing 4 out of 4

View data fields

News / Events

- News Releases
- COVID-19 Media Briefings
- Regional Budget Info
- Councillor Profiles
- Upcoming Events
- Public Notices
- Media Contacts

Preparation for COVID-19 Immunizations for Residents 80+

Niagara Region Public Health is busy preparing for the launch of our COVID-19 mass immunization clinics in the coming weeks.

Last week the 11 clinic locations were announced. and Public Health continues to work with these community partners to be

se in frequency
a's 80+ residents as

- Drive time for large portions of our overall population
- Accessible by public transit
- Connectivity to cellular internet services
- Public safety implications, with a focus on possible impacts to major roadways

Once vaccine supply is increased, the clinics will be supplemented by local pharmacies and family doctors, so that by the time the majority of the population can receive their vaccine, they can do so even closer to home.

Clinics for residents 80 years of age and older will be starting within the next couple weeks, and residents will be able to book their appointments through the Provincial online and phone registration system that is scheduled to launch on Monday, March 15.

As soon as we have more information about it, including the website link, phone number, and any other pertinent information, we will share it with the public.

We are grateful for the large number of people who responded to our call for volunteers, and we are still looking to fill a small number of paid staff positions to help run our clinics. More information about these postings can be found on our career portal.

The eagerness for vaccines rollout to ramp up over the coming weeks and months is appreciated and echoed by everyone at Public Health. We ask for continued patience as this next phase begins, as there is a lot of interest in vaccine appointments across the province, and wait times in the registration portals, both online and by phone, are possible.

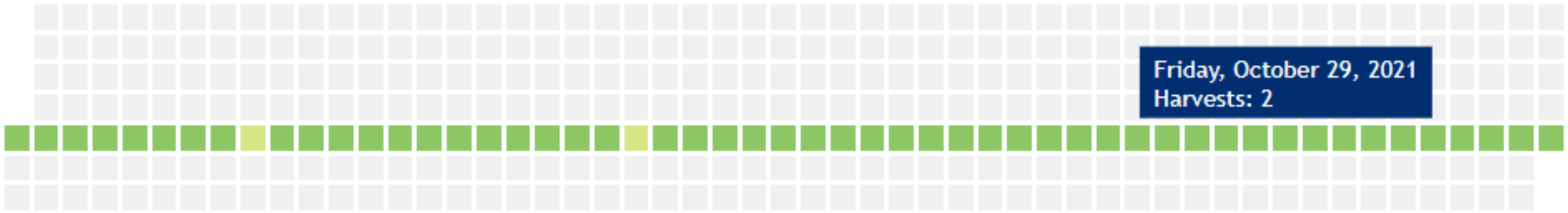
SOLRWAYBACK

Harvests for: http://welland.ca/hottopics/covid-19.asp

First harvest: **April 17, 2020**
Latest harvest: **December 31, 2021**
Total harvests: **179**

Show: Months - Days

2021 - Hide details



Less More

Harvests for October 29, 2021

1. October 29, 2021 15:54
2. October 29, 2021 15:55

The Tools: Python notebooks

- Notebooks
  - Mixed media: (Python) code, rich-text, results (pictures,videos)
  - Easy to distribute and use
  - Results can be viewed without running code
  - Browser based, often run in the cloud

- Why use notebooks at all?

  Answer: Sharable integrated environment with repeatable results.

Google Colab

Google Colab is an environment for running notebooks

- Available (free) to everyone, literally no config necessary
- No issues with dealing with infrastructure (GPUs offered!)
- Complete environment within browser – very portable

The Tools: Python + Google Colab

- Why is it useful for our project?
  - Distant reading (quantitative results)
  - Sharable graphs + and results
  - Drastically lowers barrier to entry for modifications
  - Reproducible results (post-publication)

Our notebooks use ...

- Text analysis libraries used extensively. (Scikit-learn, nltk, SpaCy)
- Pandas
- Matlibplot
- Github for notebooks, S3 for data (ARCH full text derivation)

Notebook examples

- Sentiment analysis
- Text similarity (of URLs)
- Update frequency + size of updates
- First mention of vaccine by municipality

Notebook listing

ARCH Data Exploration	Notebook	Open in Colab
COMM 4P35 Tutorial	Notebook	Open in Colab
Hackfest notebook	Notebook	Open in Colab
Muni Data Export	Notebook	Open in Colab
Prep Domain Data	Notebook	Open in Colab
Twitter Data Export	Notebook	Open in Colab
Municipal Data Similarity	Notebook	Open in Colab
Another example of Municipal Data Similarity using SpaCy	Notebook	Open in Colab
Municipal Data Similarity using TF-IDF	Notebook	Open in Colab
Content size of pages over time	Notebook	Open in Colab
Frequency of page updates over time	Notebook	Open in Colab
Crawl frequency visualized	Notebook	Open in Colab
Sentiment scores of Municipal Data	Notebook	Open in Colab
Vaccine keyword frequency	Notebook	Open in Colab
Vaccine keyword frequency normalized	Notebook	Open in Colab
First mention of vaccine keywords	Notebook	Open in Colab

Muni_Updates.ipynb

File Edit View Insert Runtime Tools Help

+ Code + Text Copy to Drive

Update frequency

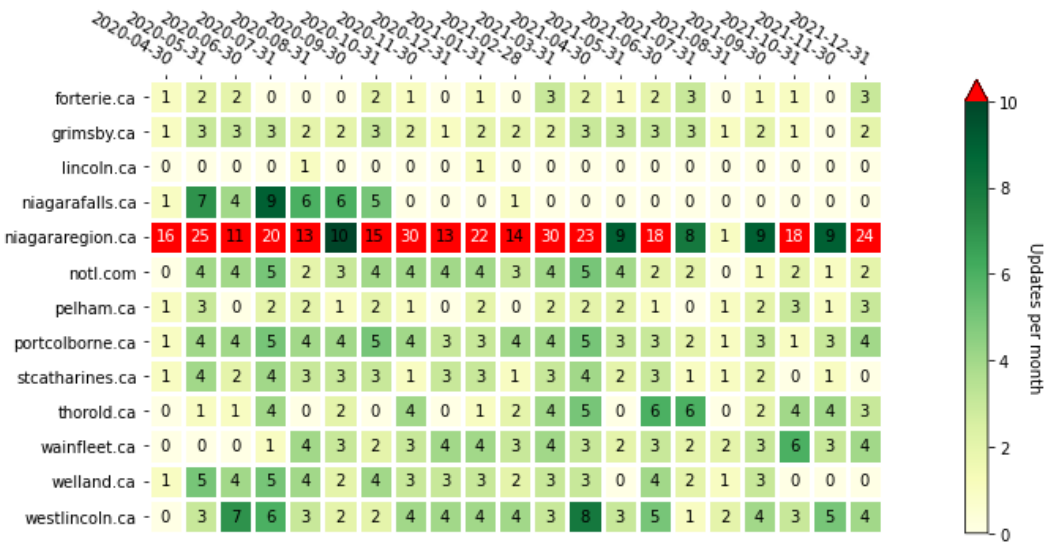
```
# 0.0 values in the delta_table mean that no updates occurred
# set these values to NaNs so the groupby count() aggregator ignores them
dt_cp = delta_table.copy()
dt_cp[dt_cp == 0] = np.nan

update_table = dt_cp.groupby(pd.Grouper(freq='1M')).count().transpose().groupby(level=0).sum()
update_table
```

```
fig, ax = plt.subplots(figsize=(15,5))

column_labels = list(map(lambda ts: ts.strftime('%Y-%m-%d'), update_table.columns))
im, cbar = heatmap(update_table, update_table.index, column_labels, ax=ax, cmap='YlGn',
                   cbar_kw={'extend':'max'}, cbarlabel="Updates per month", vmin=0, vmax=10)
texts = annotate_heatmap(im, valfmt="{x}", threshold=15)

fig.tight_layout()
plt.show()
```



Muni_Updates.ipynb - Colaborat

colab.research.google.com/github/BrockDSL/ARCH_Data_Explore/blob/main/Muni_Updates.ipynb

TD Canada Trust - P... | Western Libraries -... | Johnson, Fletcher -... | JupyterLite | Mail - Fletcher John... | Home : Toronto Pu... | JavaScript | MDN | Exploring ES6 | Other bookmarks

Muni_Updates.ipynb

File Edit View Insert Runtime Tools Help

Share

+ Code + Text Copy to Drive

Connect Editing

Brock University

Brock University Library

Digital Scholarship Lab

About

This notebook creates two heatmaps based on the number of updates and size of updates released by each municipality.

The methodology for determining update size is as follows:

1. For each municipality, determine the content length - the character count - of all crawls.
2. For each crawl calculate the difference in content length compared to the previous time it was crawled.
3. For each url calculate the average of the differences over a month timeframe.
4. For each municipality calculate the average of the monthly url averages.

The way the *number of updates* is determined is very similar. The number of crawls for each municipality are determined. The difference in content length compared to a previous crawl is calculated. If there is a difference it is considered an update. The updates are then summed over a period of a month.

Limitations

- The update size heat map is quite large. Making it smaller is difficult because it becomes hard to read the numbers.
- The assumption is that the character count of each crawl is indicative of information related to covid but this is not necessarily true

Take away

- Notebooks place code, data, results all in one place
- Google Colab makes running notebooks incredibly easy
- SolrWayback is a powerful WARC search engine: useful for anyone with large collection of WARCs
- Both make analyzing web archives much easier

Links

- BrockDSL "Covid-19 in Niagara" Notebooks

  https://github.com/BrockDSL/ARCH_Data_Explore

- SolrWayback

  https://github.com/netarchivesuite/solrwayback

- Our Archive-it.org web archive

  https://archive-it.org/collections/13781

- More about Google Colab

  https://research.google.com/colaboratory/faq.html