

History Crawling with Warcbase

Ian Milligan
Assistant Professor
@ianmilligan1



UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History

Jimmy Lin
Professor and David R. Cheriton Chair
@lintool



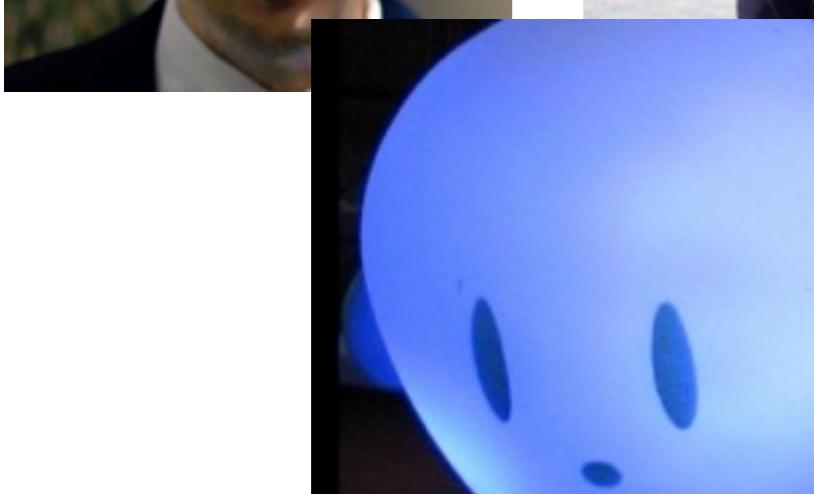
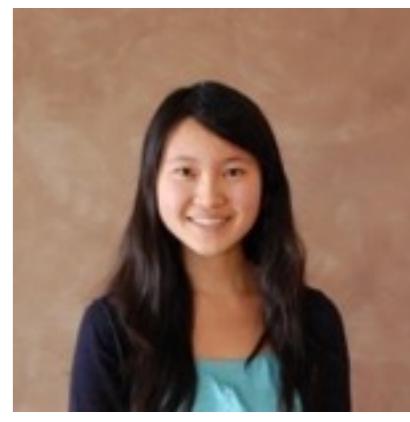
UNIVERSITY OF WATERLOO
FACULTY OF MATHEMATICS
David R. Cheriton School
of Computer Science

Team Warcbase

Historians



Computer Scientists



Librarians



Networks



Two Case Studies

- **Archive-It Research Services:** “Canadian Political Parties and Political Interest Groups”
- 2005 - 2015
- WARC files

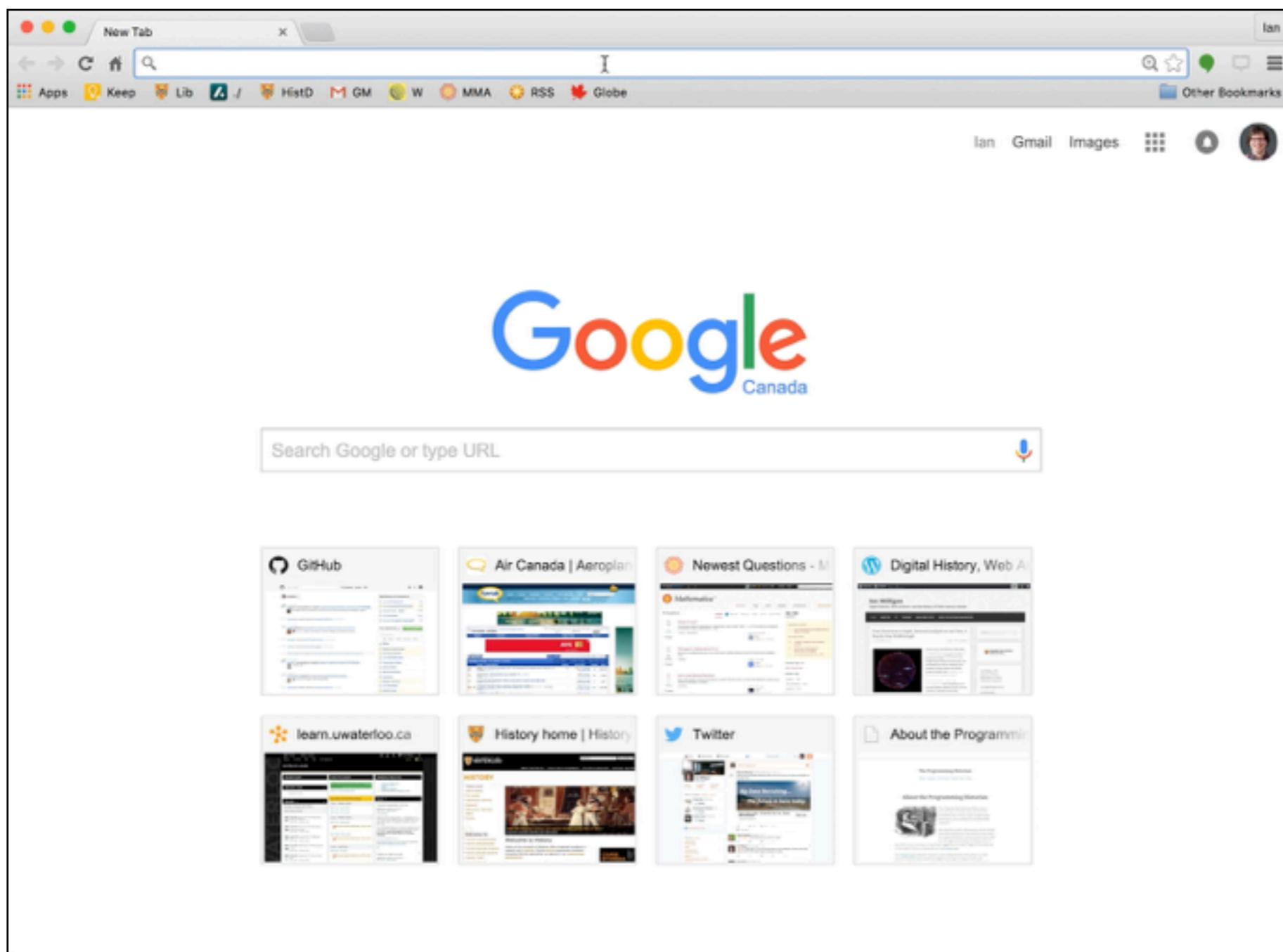
The screenshot shows a web browser window with the URL <https://archive-it.org/collections/227>. The page title is "Archive-It - Canadian Political Parties and Political Interest Groups". The header includes links for HOME, EXPLORE, LEARN MORE, and CONTACT US. A sidebar on the right says "The leading provider of digital preservation services for cultural institutions. Built by librarians for librarians." Below the header, it says "Explore >> University of Toronto >> Canadian Political Parties and Political Interest Groups". A large green box contains the "ARCHIVE-IT" logo and the title "Canadian Political Parties and Political Interest Groups" with "Collected by: University of Toronto". It also states "Archived since: Oct, 2005" and "Description: Canadian Political Parties and Political Interest Groups, national Canadian political parties, and a number of special interest groups". The subject is listed as "Politics & Elections" and the collector as "University of Toronto". Below this, there's a section titled "Narrow Your Results" with a search bar and a list of subjects: New Democratic Party of Canada (2), Assembly of First Nations (1), Bloc Québécois (1), Canada First (1), Canada West Foundation (1). There are buttons for "Sites" and "Search Page Text". At the bottom, it says "Page 1 of 1 (54 Total)" and "Sort By: Title (A-Z) | Title (Z-A) | URL (A-Z) | URL (Z-A)".

Two Case Studies



- **GeoCities**
- End-of-life crawl from 2009
- WARC files
- 4.1 TB, 186 million HTML documents

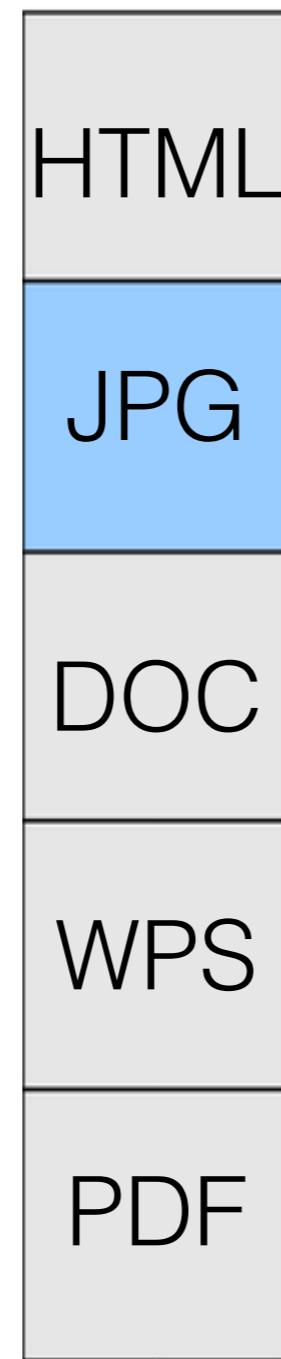
Moving beyond the Wayback Machine for scholarly access



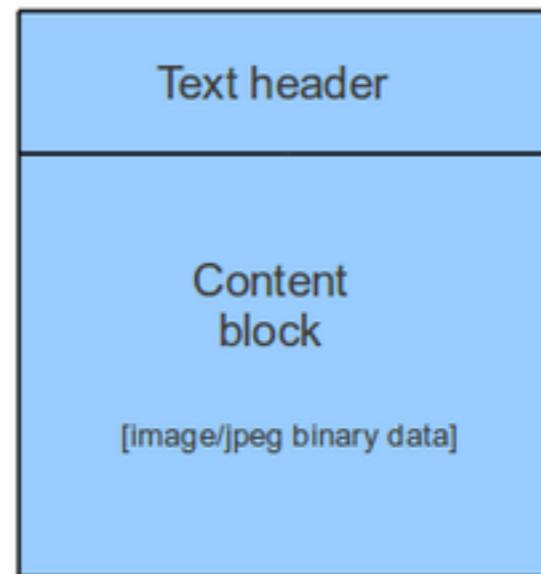
Why Warcbase?



WARC file



WARC record



WARC/1.0
WARC-Type: resource
WARC-Target-URI: file:/var/www/htdoc/images/logoc.jpg
WARC-Date: 2006-09-19T17:20:24Z
WARC-Record-ID: <urn:uuid:92283950-ef2f-4d72-b224-f54c6ec90bb0>
Content-Type: image/jpeg
WARC-Payload-Digest: sha1:CCHXETFVJD2MUZY6ND6SS7ZENMWF7KQ2
WARC-Block-Digest: sha1:CCHXETFVJD2MUZY6ND6SS7ZENMWF7KQ2
Content-Length: 1662

...etc.

Warcbase

An open-source platform for managing web archives

<http://warcbase.org>

Two main facets

- A flexible data store: your own Wayback Machine
- **Scriptable analytics and data processing**

Funded by Mellon, SSHRC, NSERC, and Government of Ontario.



Warcbase

- Scalable
 - From Raspberry Pi, to laptop, to powerful desktop, to single-node beefy server, to cluster
- Potentially very powerful
 - *Trantor cluster*: 1.2PB of disk, 25 compute nodes totalling 3.2TB memory and 300 current-generation Intel cores.



docs.warcbase.org

The screenshot shows a web browser window with the following details:

- Address Bar:** lintool.github.io/warcbase-docs/Spark-Extracting-Domain-Level-Plain-Text/
- Page Title:** Extracting Domain Level Plain Text
- Page Content:**
 - Left Sidebar (Extracting Domain Level Plain Text):** A sidebar menu with the following items:
 - All plain text
 - Plain text by domain
 - Plain text by URL pattern
 - Plain text minus boilerplate
 - Plain text filtered by date
 - Plain text filtered by language
 - Plain text filtered by keyword
 - Main Content Area:**

Extracting Domain Level Plain Text

All plain text

This script extracts the crawl date, domain, URL, and plain text from HTML files in the sample ARC data (and saves the output to out/).

```
import org.warcbase.spark.rdd.RecordRDD._  
import org.warcbase.spark.matchbox.{RemoveHTML, RecordLoader}  
  
RecordLoader.loadArchives("src/test/resources/arc/example.arc.gz", sc)  
  .keepValidPages()  
  .map(r => (r.getCrawlDate, r.getDomain, r.getUrl, RemoveHTML(r.getContent  
String)))  
  .saveAsTextFile("out/")
```

If you wanted to use it on your own collection, you would change "src/test/resources/arc/example.arc.gz" to the directory with your own ARC or WARC files, and change "out/" on the last line to where you want to save your output data.

Note that this will create a new directory to store the output, which cannot already exist.

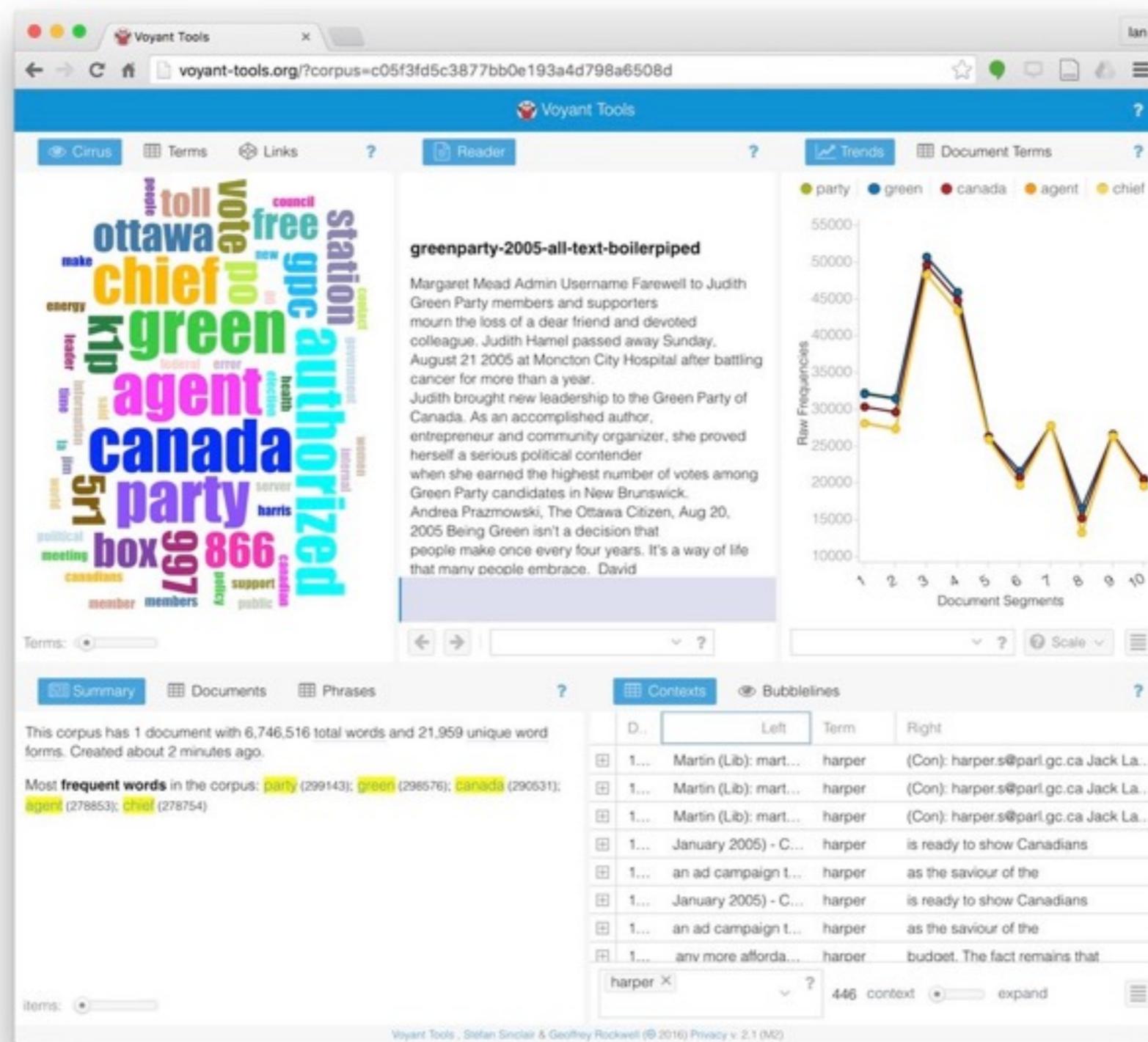
If you want to run it in your Spark Notebook, the following script will show in-notebook plain text:

```
val r = RecordLoader.loadArchives("/path/to/warcs", sc)  
.keepValidPages()  
.map(r => {
```

Extract all Text

```
1. i2millig@rho: ~/derivatives/cpp.all.plaintext (ssh)
Python      bash      bash      bash      i2millig@rho:... i2millig@rho:...      bash
(20060222,liberal.ca,http://liberal.ca/bio_e.aspx?id=35049,Liberal Party of Canada HOME THE TEAM THE P
ARTY MEDIA CENTRE COMMISSIONS YOUR RIDING      Omar Alghabra www.omaralghabra.com Home > Mississauga--E
rindale Riding Map (PDF) Omar Alghabra came to Canada at a very young age, and immediately knew Canada
was his home. He was first elected 2006 as the Member of Parliament for Mississauga-Erindale. Mr. Al
ghabra is an experienced entrepreneur. For the past six years, he has worked for a large multinational
corporation, carrying out different responsibilities including quality assurance, project management, s
ales, contract management and management of a complete department handling a global mandate. Mr. Alghab
ra is an active member of his community. He is the former National President of the Canadian Arab Feder
ation (2004-2005) and a former member of the Community Editorial Board for the Toronto Star. (2003-2004
). Mr. Alghabra is currently a member of the Diversity Council for General Electric Canada and is activ
e in Junior Achievement for the Toronto Region. He was a member of the Multicultural Inter-Agency of Pe
el from 2001 to 2002. Mr. Alghabra has a degree in Mechanical Engineering from Ryerson University and a
Masters in Business Administration (MBA) from York University.      Omar Alghabra 790 Burnamthorpe West,
Unit 10 905-276-2806      info@omaralghabra.ca Riding President Elias Hazineh Send an email
      Home | News | Your Riding | Contact Us | français This website is the property of the
Liberal Party of Canada and may not be reproduced in whole or in part without express written permission.
© Liberal Party of Canada 2006. All rights reserved. Authorized by the registered agent for the Libe
ral Party of Canada. Privacy Policy)
(20060222,liberal.ca,https://liberal.ca/news_e.aspx?id=11470,Liberal.ca HOME THE TEAM THE PARTY MEDIA C
ENTRE COMMISSIONS YOUR RIDING      Celebrating our National Flag February 15, 2006 February 15 is Nationa
l Flag Day in Canada, which marks the 41st anniversary of the first raising of the maple leaf flag on P
arliament Hill. Today is a celebration of our shared values, common citizenship and sense of pride in t
his great country we call home. The Canadian flag is one of the most recognizable symbols in the world
and flies proudly to remind us all of who we are and where we come from. The maple leaf's symbolic orig
ins date back to the beginning of our nation's history, while the red and white bars on the flag repres
ent strength and unity. Canada's flag was adopted in 1964 under the courageous leadership of Liberal Pr
ime Minister Lester B. Pearson. The idea of changing the Red Ensign which featured the Britain's Union
Jack, was very controversial at the time, with the Conservative Party strongly opposed to changing the
status quo. Facing strong Conservative resistance in the House of Commons, Pearson's minority governme
nt fought hard in the name of national unity and Canada's multicultural future to make the new flag a r
eality. In an impassioned speech to the House of Commons, Pearson said: "Mr. Speaker, it is for this g
eneration, for this Parliament, to give them and to give us all a common flag; a Canadian flag which, w
hile bringing together but rising above the landmarks and milestones of the past, will say proudly to t
he world and to the future: I stand for Canada." Thanks to Pearson's courageous leadership, Canadians a
cross this great nation celebrate our flag and what it stands for – a country and a citizenship that ar
e the envy of the world.
      Home | News | Your Riding | Contact Us | fran
cais This website is the property of the Liberal Party of Canada and may not be reproduced in whole or
```

Extract all Text



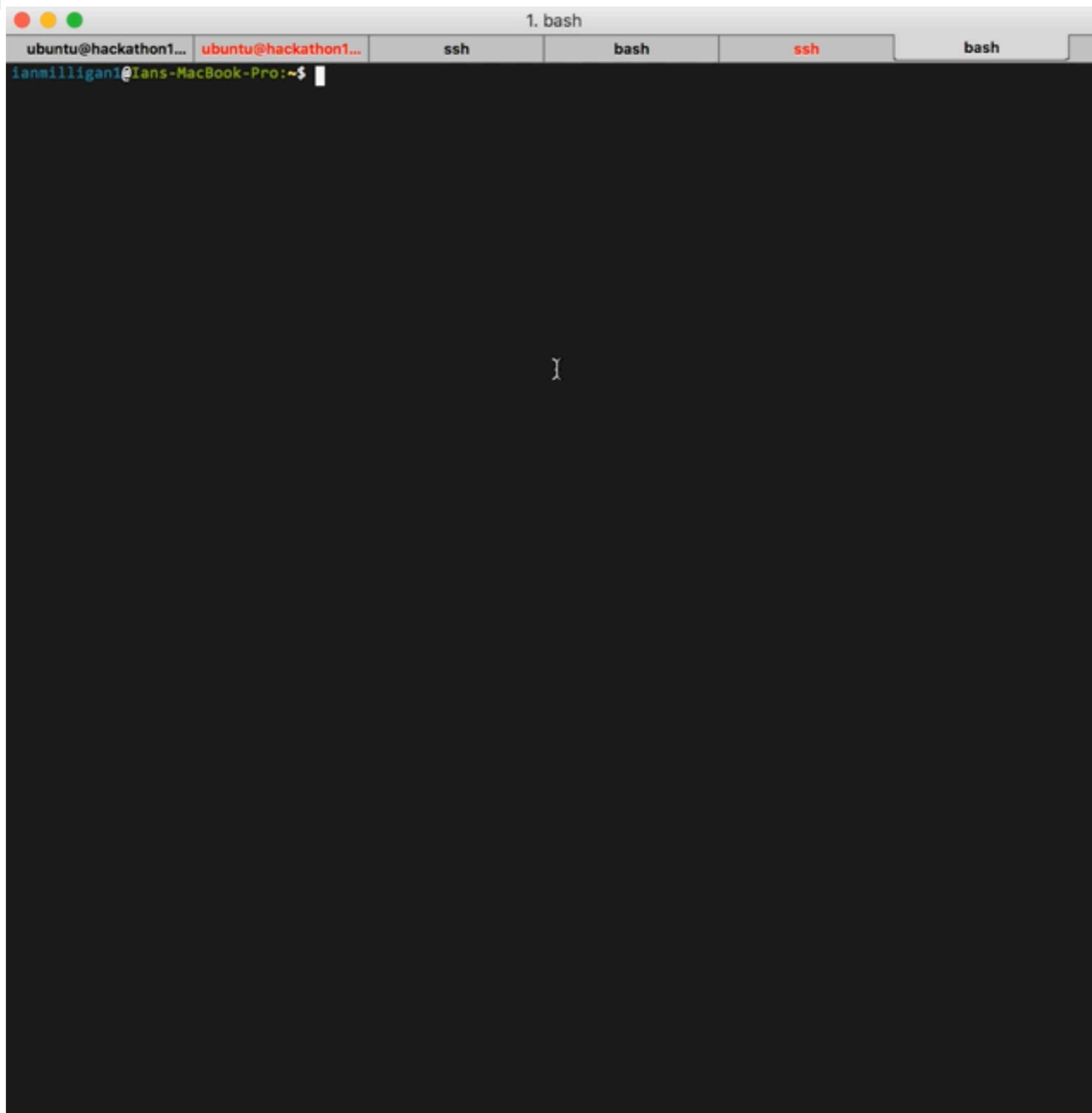
Using Warcbase to Learn
Cool Stuff about web
archives!

Step One: Grabbing WARCs



```
1. i2millig@rho: /mnt/vol1/data_sets/geocities/warc (ssh)
bash                                bash                                i2millig@rho: /mnt/vol1/data...
GEOCITIES-20091029114236-00191-ia400118.us.archive.org.warc.gz
GEOCITIES-20091029115416-00171-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029123834-00172-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029130439-00173-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029134536-00174-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029140344-00192-ia400118.us.archive.org.warc.gz
GEOCITIES-20091029141553-00193-ia400118.us.archive.org.warc.gz
GEOCITIES-20091029141726-00175-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029144445-00176-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029152151-00177-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029160824-00178-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029164941-00179-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029165037-00194-ia400118.us.archive.org.warc.gz
GEOCITIES-20091029170431-00195-ia400118.us.archive.org.warc.gz
GEOCITIES-20091029171605-00180-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029174154-00181-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029180018-00182-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029182725-00183-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029185058-00184-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029193728-00185-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029194541-00196-ia400118.us.archive.org.warc.gz
GEOCITIES-20091029195911-00197-ia400118.us.archive.org.warc.gz
GEOCITIES-20091029202841-00186-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029221340-00198-ia400118.us.archive.org.warc.gz
GEOCITIES-20091029222459-00199-ia400118.us.archive.org.warc.gz
GEOCITIES-20091030021147-00197-ia400103.us.archive.org.warc.gz
GEOCITIES-20091030021444-00198-ia400103.us.archive.org.warc.gz
GEOCITIES-20091030022413-00171-ia400104.us.archive.org.warc.gz
i2millig@rho:/mnt/vol1/data_sets/geocities/warc$ du -h
4.1T .
i2millig@rho:/mnt/vol1/data_sets/geocities/warc$
```

Step Two: Basic Shell Analysis

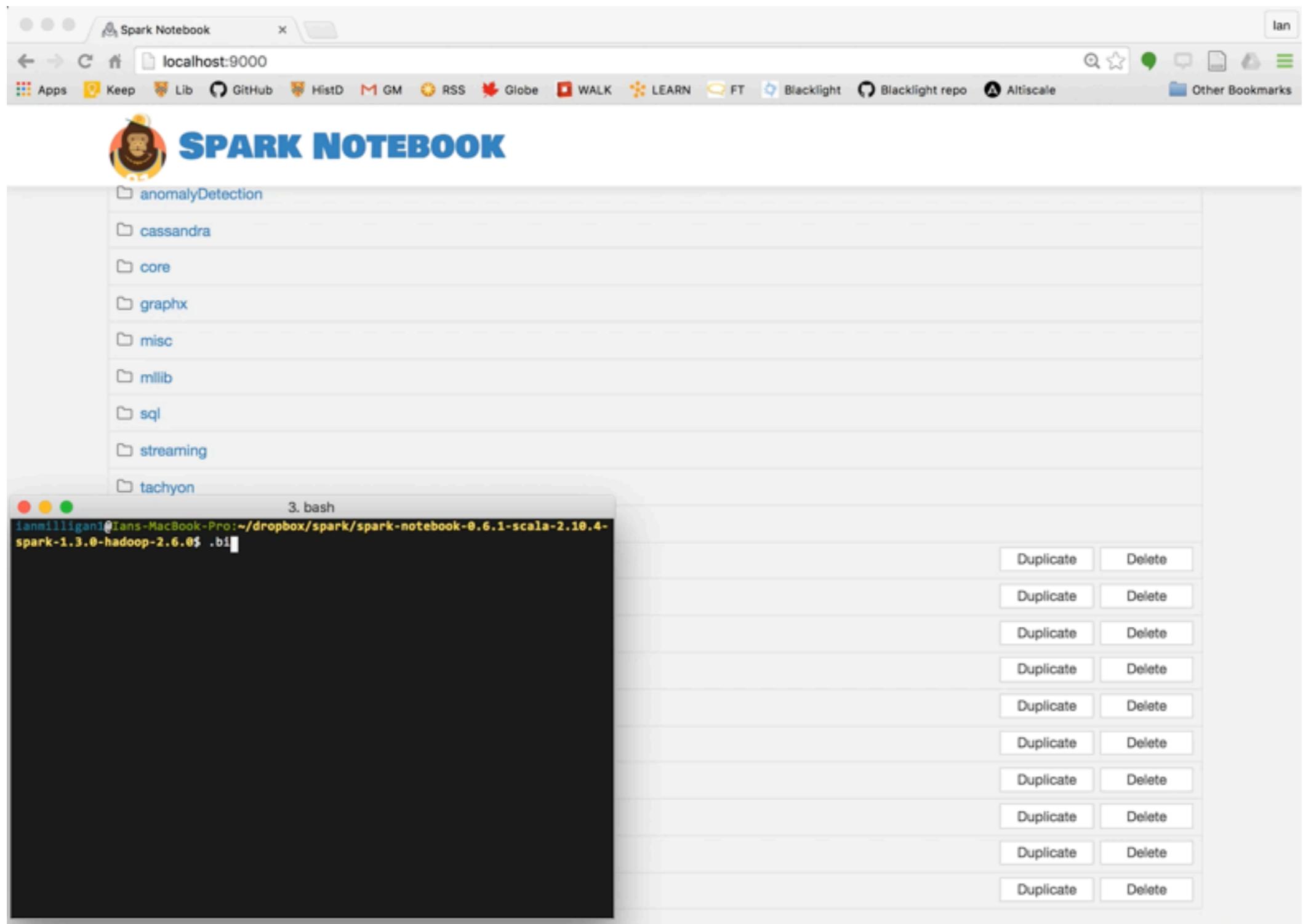


A screenshot of a terminal window titled "1. bash". The window has four tabs at the top: "ssh", "bash", "ssh", and "bash". The active tab is the third one, which is red. The terminal shows a user session:

```
ubuntu@hackathon1...:~$ ianmilligan1@Ians-MacBook-Pro:~$ }
```

The cursor is positioned at the end of the command line, indicated by a vertical bar.

Step Two: Basic Analytics



Step Three: Filtering a Corpus

```
1 import org.warcbase.spark.matchbox.{ExtractTopLevelDomain,  
    ExtractLinks, RecordLoader}  
2 import org.warcbase.spark.rdd.RecordRDD._  
3  
4 RecordLoader.loadArc("/mnt/vol1/data_sets/geocities/warcs/*", sc)  
5 .keepValidPages()  
6 .map(r => (r.getCrawldate, ExtractLinks(r.getUrl, r.  
    getContentString)))  
7 .flatMap(r => r._2.map(f => (r._1, ExtractTopLevelDomain(f._1).  
    replaceAll("^\\s*www\\\\.", ""), ExtractTopLevelDomain(f._2).  
    replaceAll("^\\s*www\\\\.", ""))))  
8 .filter(r => r._2 != "" && r._3 != "")  
9 .countItems()  
10 .filter(r => r._2 > 5)  
11 .saveAsTextFile("/mnt/vol1/data_sets/geocities/geocities.  
    sitelinks")
```

A Link Graph

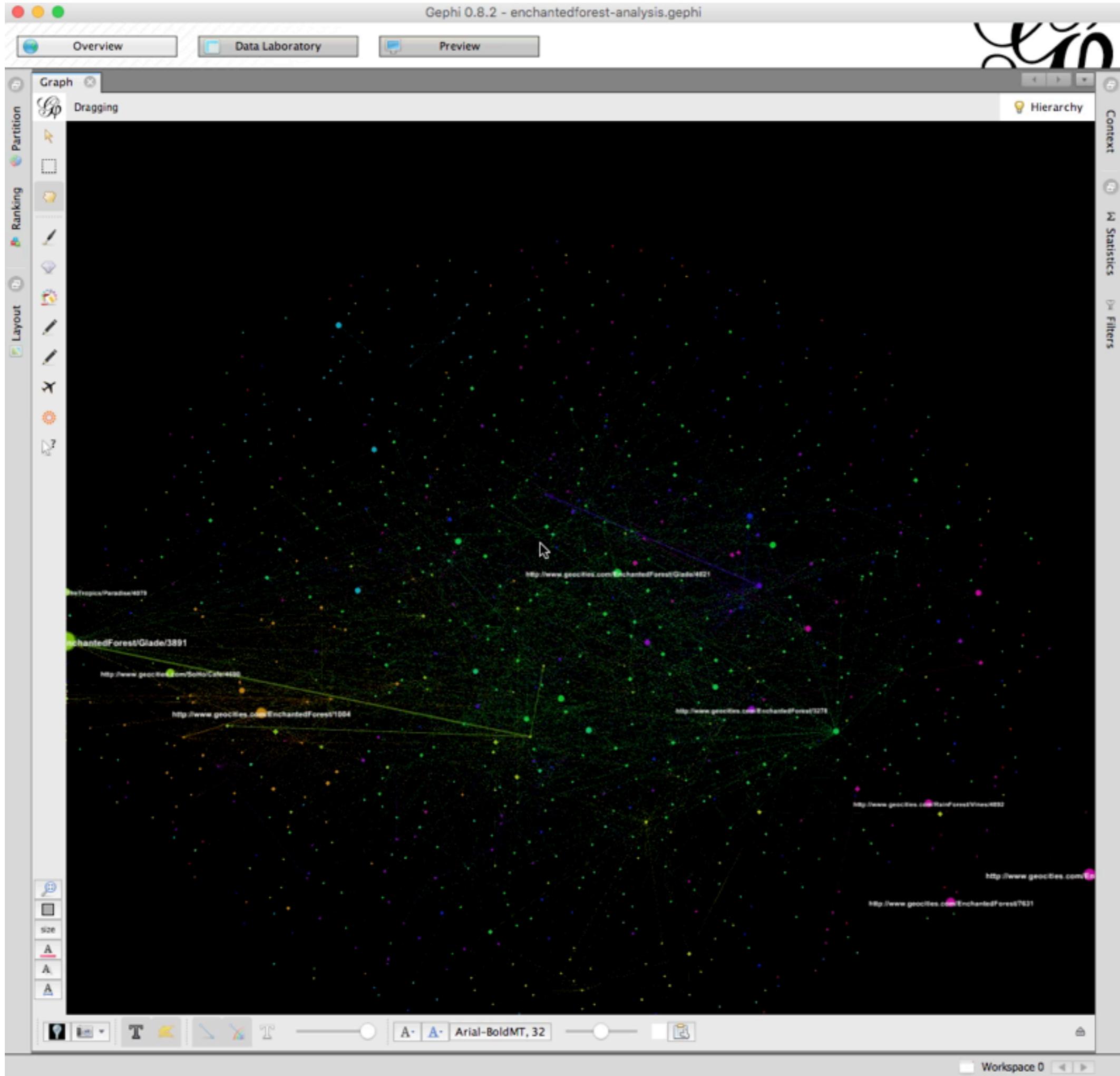
Step Three: Filtering a Corpus

```
1 ((20090903,http://geocities.com/saganaki2000/ADSLGR/adslgr.htm,  
http://www.adslgr.com),15337)  
2 ((20091026,http://geocities.com/saganaki2000/ADSLGR/adslgr.htm,  
http://www.adslgr.com),15337)  
3 ((20091027,http://geocities.com/spankbank69hard/,http://pg.photos  
.yahoo.com/ph/spankbank69hard/my_photos/),9807)  
4 ((20090903,http://geocities.com/spankbank69hard/index.html,http://  
/pg.photos.yahoo.com/ph/spankbank69hard/my_photos/),9807)  
5 ((20091027,http://geocities.com/CollegePark/Locker/8187/,http://  
www.comercialuruapan.com),8056)  
6 ((20090903,http://geocities.com/CollegePark/Locker/8187/,http://  
www.comercialuruapan.com),8056)
```

Results

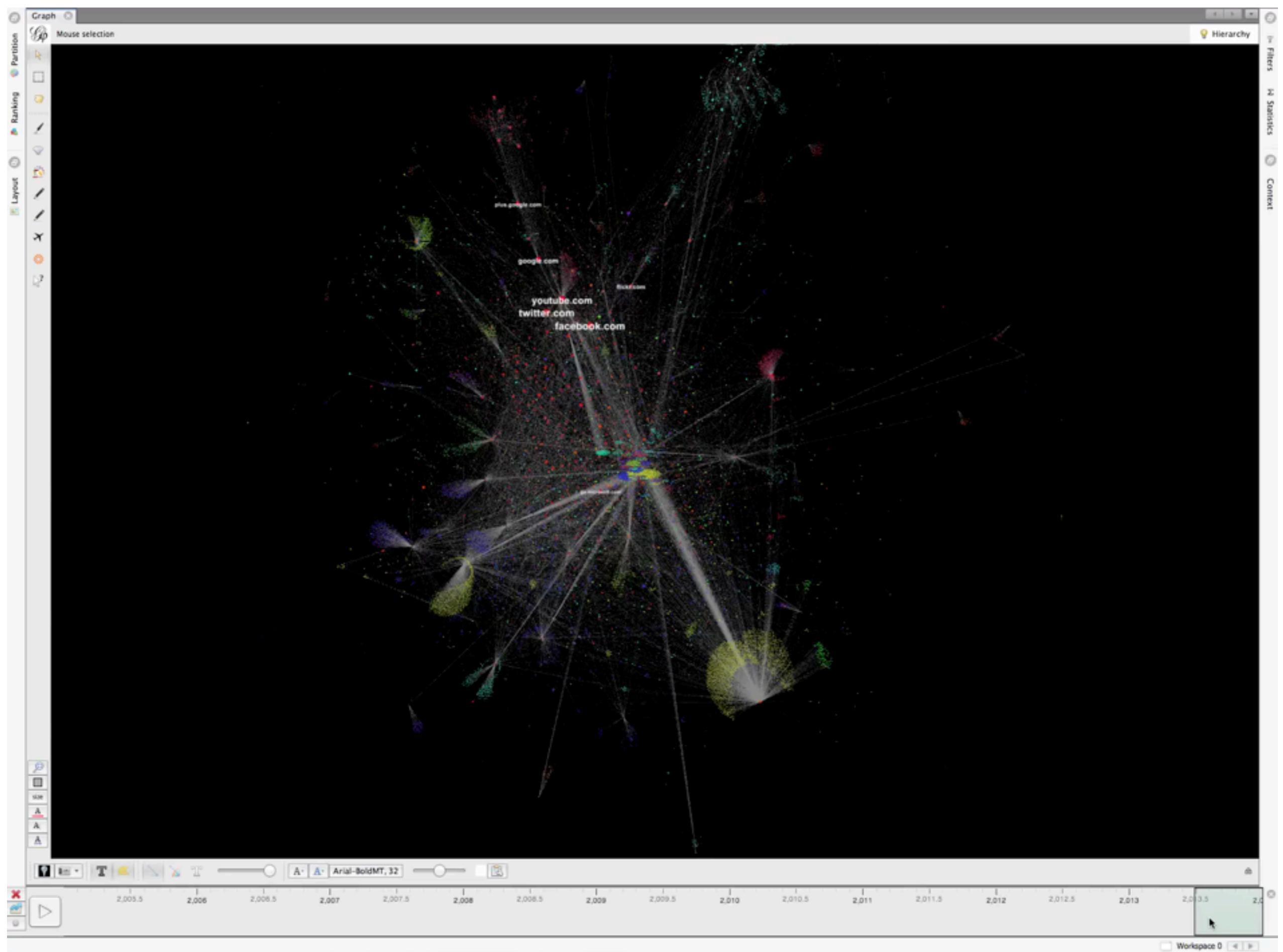
Filtering



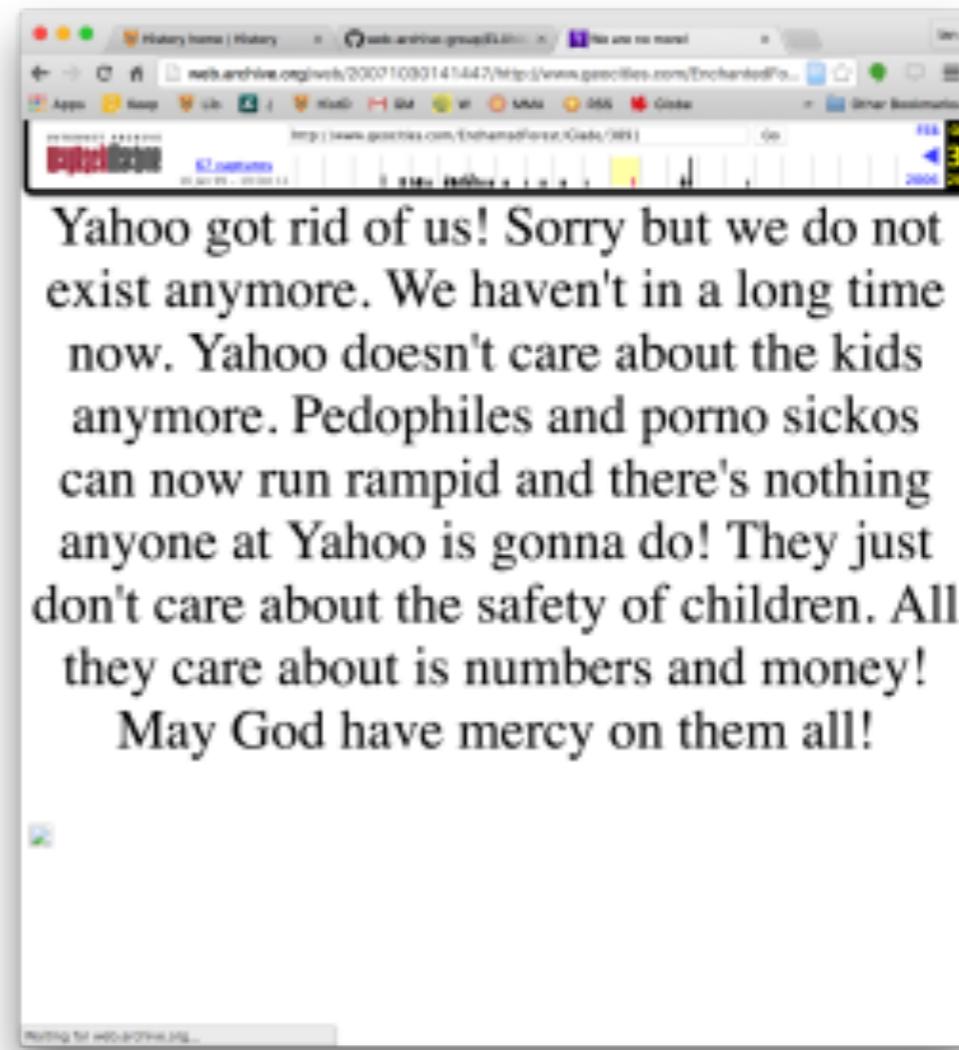


Finding cool sites!

Label	▼ PageRank	In-Degree	Out-Degree	Degree
http://www.geocities.com/EnchantedForest/Glade/3891	0.008	145	1	146
http://www.geocities.com/EnchantedForest/Glade/9378	0.005	6	13	19
http://www.geocities.com/EnchantedForest/1004	0.005	63	26	89
http://www.geocities.com/EnchantedForest/7631	0.004	6	3	9
http://www.geocities.com/SoHo/Cafe/4690	0.004	241	0	241
http://www.geocities.com/EnchantedForest/Glade/4021	0.004	151	0	151
http://www.geocities.com/TheTropics/Paradise/4079	0.003	248	0	248
http://www.geocities.com/RainForest/Vines/4892	0.003	5	6	11
http://www.geocities.com/EnchantedForest/3278	0.003	106	0	106
http://www.geocities.com/EnchantedForest/3696	0.003	70	0	70
http://www.geocities.com/EnchantedForest/Dell/5914	0.003	180	1	181
http://www.geocities.com/EnchantedForest/1469	0.003	16	49	65
http://www.geocities.com/EnchantedForest/Tower/9644	0.003	19	42	61
http://www.geocities.com/EnchantedForest/Dell/9501	0.003	79	362	441
http://www.geocities.com/EnchantedForest/Glade/8851	0.003	17	0	17
http://www.geocities.com/Heartland/Meadows/6263	0.003	9	0	9
http://www.geocities.com/Heartland/6188	0.003	56	0	56
http://www.geocities.com/EnchantedForest/4213	0.003	158	0	158
http://www.geocities.com/Athens/Acropolis/1465	0.003	20	0	20
http://www.geocities.com/EnchantedForest/8012	0.003	42	197	239
http://www.geocities.com/EnchantedForest/3810	0.003	98	147	245
http://www.geocities.com/EnchantedForest/Glade/3899	0.002	14	11	25
http://www.geocities.com/EnchantedForest/3015	0.002	64	0	64
http://www.geocities.com/EnchantedForest/Tower/8143	0.002	50	40	90
http://www.geocities.com/EnchantedForest/Meadow/1426	0.002	41	185	226



Step Four: Finding Significant Sites w/ PageRank



Step Five: Text Analysis

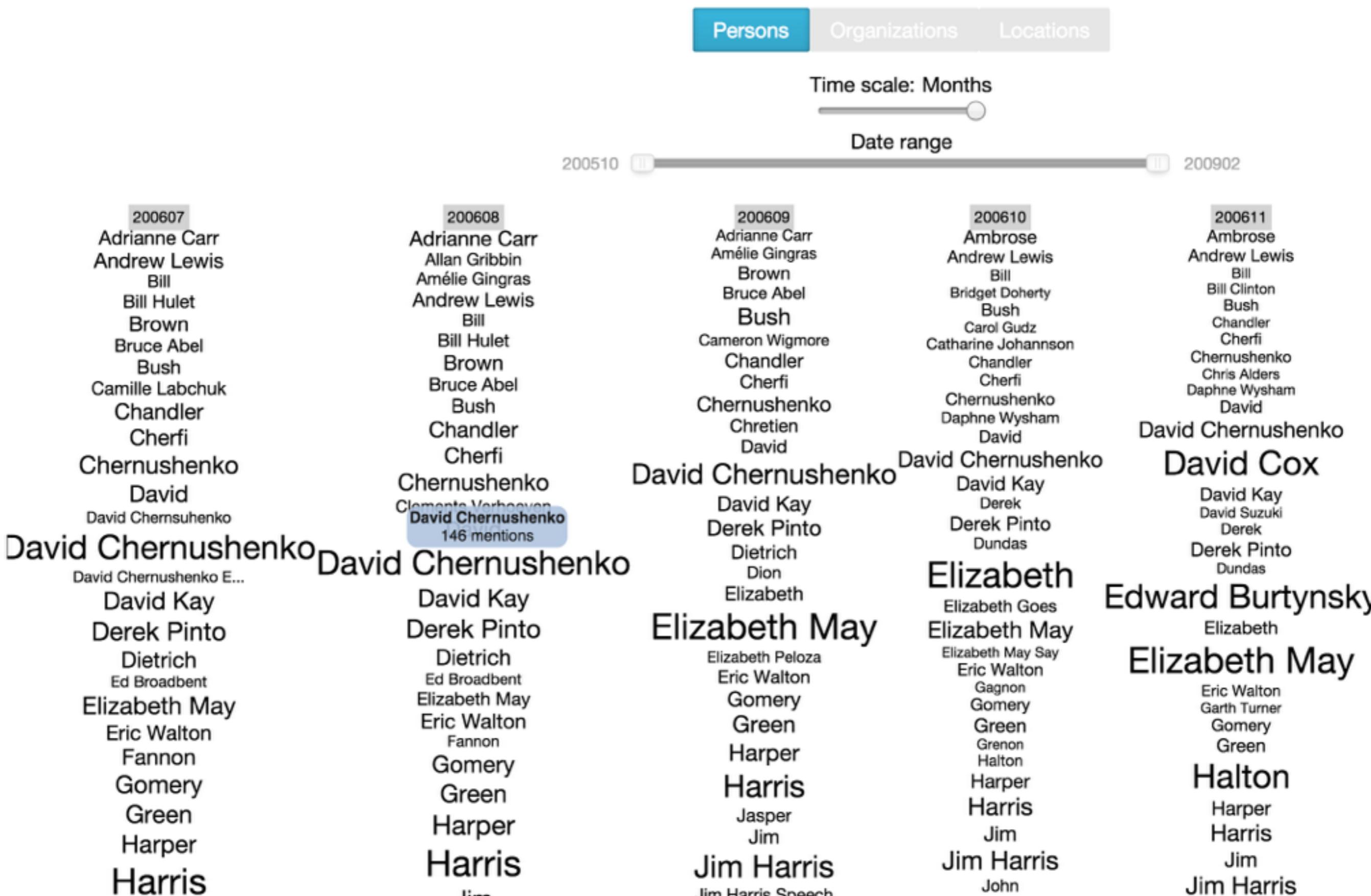
Different Ways to Filter

- Get everything
- Filter by domain (i.e. all pages in “[greenparty.ca](#)”)
- Filter by URL pattern (i.e. all pages in “[greenparty.ca/vegetables/*](#)”)
- Filter out boilerplate (i.e. advertisements, navigational elements, templates, etc.)
- Filter by date (i.e. all pages on July 4th, 2015)
- Filter by languages (i.e. only French language pages from [greenparty.ca](#))
- Or any of the above!

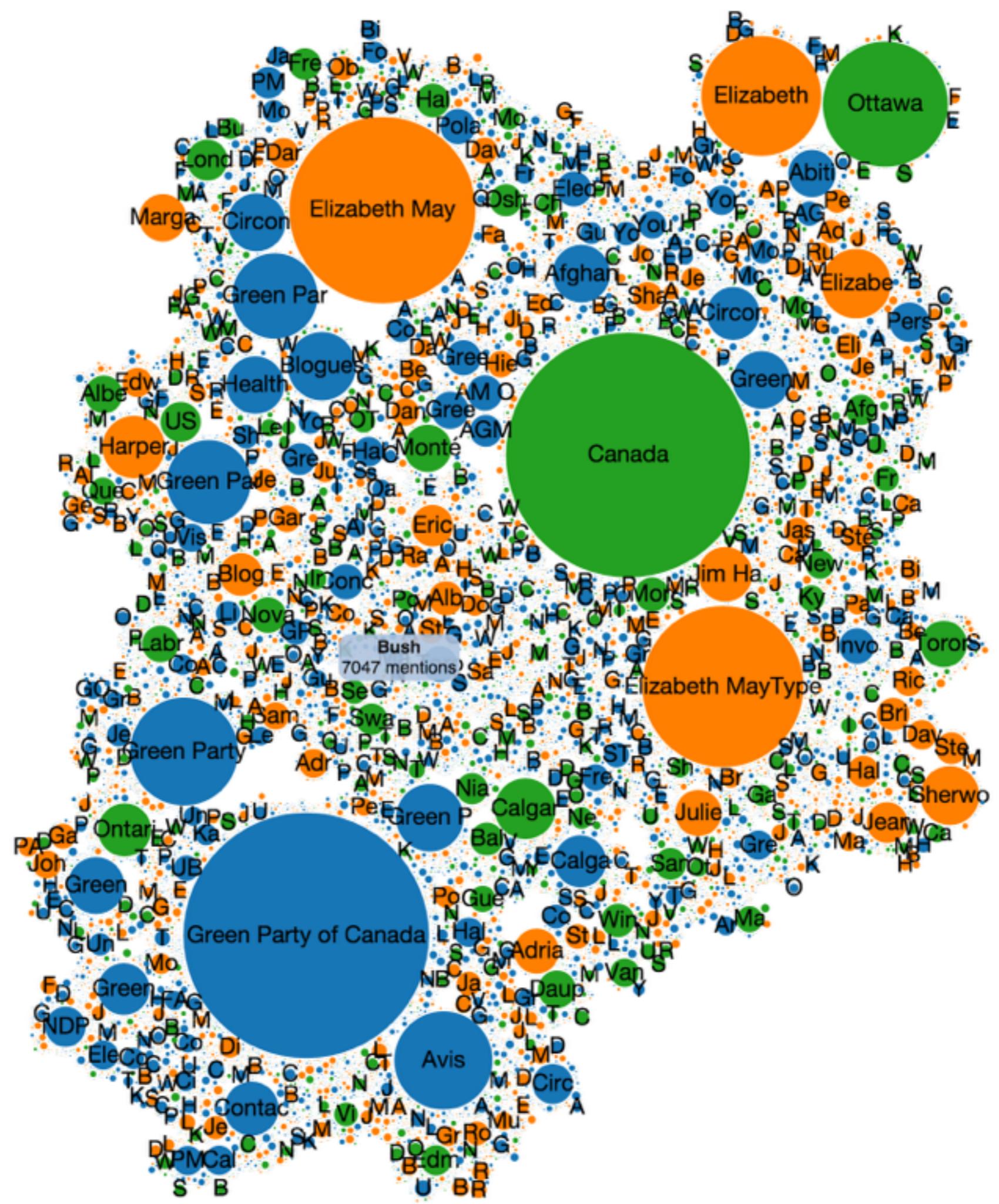


Named Entity Visualization

Data source: [greenparty.csv](#)







Or generate Solr indexes
using Warcbase too!

Welcome to the Web Archives for Historical Research political parties portal. Before diving in, we encourage you to visit our [about](#) page.

The Canadian Political Parties and Political Interest Groups Portal

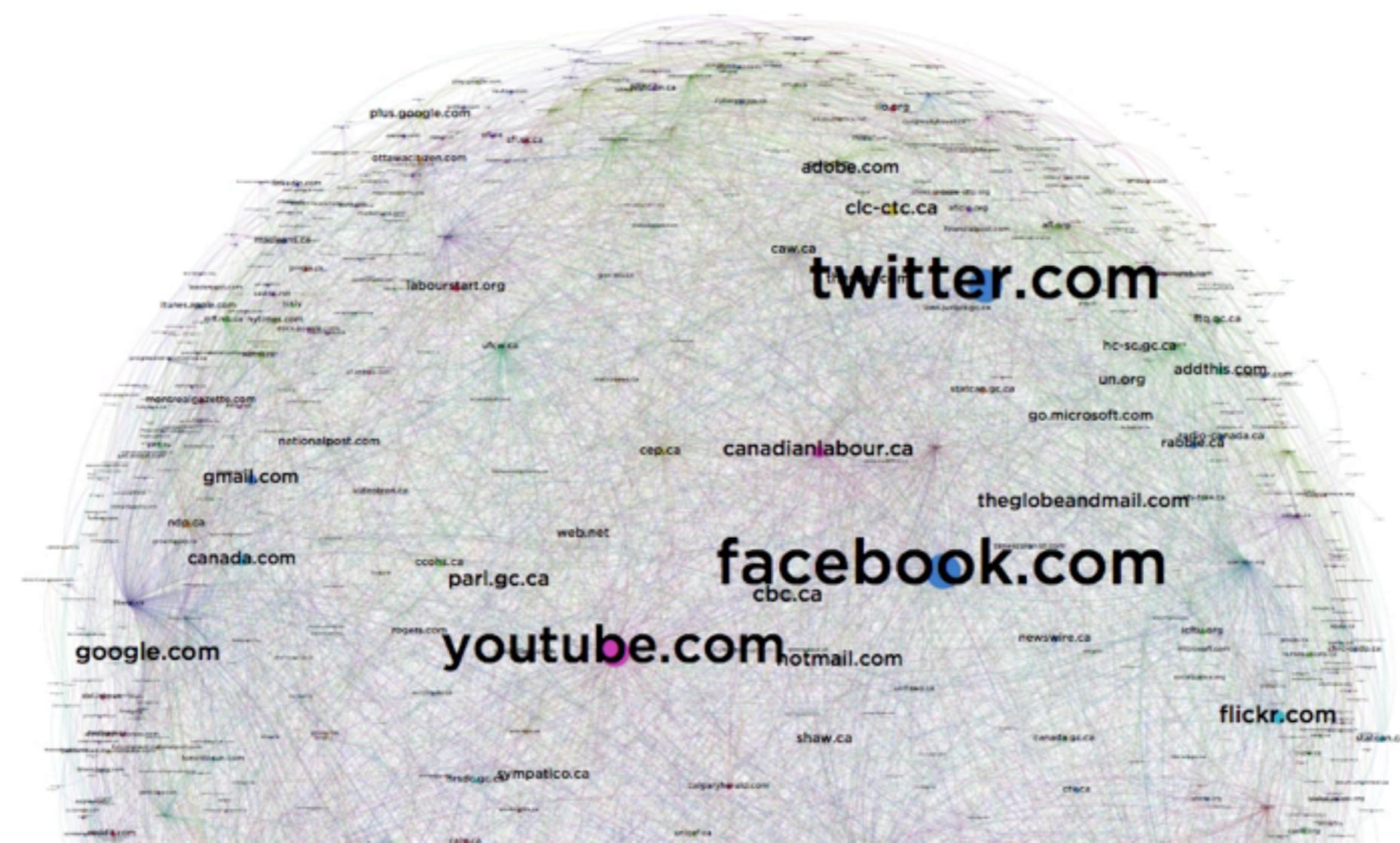
On this website, you can search web archived content from 50 political parties and political interest groups, from October 2005 to March 2015.

Curious how the Liberal Party of Canada responded to the 2008 financial crisis ([a search for "recession" in 2008, liberal.ca](#))? How the Canadian Centre for Policy Alternatives [reacted to Michael Ignatieff](#)? Now you can check it all out.

Options include

- Basic keyword searching [Example: "Rob Ford", only Liberal.ca]
 - Graphing trends over time [Example: Liberal Opposition Leaders, 2005-2015]
 - Advanced search, including words in proximity to each other [Example: environmental and tax within 25 words of each other]

Below, here are all of the links for the entire time period, visualized below.



docs.warchbase.org

or the living docs below:





Social Sciences and Humanities
Research Council of Canada

Conseil de recherches en
sciences humaines du Canada

Canada



compute | **calcul**
canada | canada

THE
ANDREW W.
MELLON
FOUNDATION



NSERC
CRSNG



UNIVERSITY OF
WATERLOO

Ian Milligan
Assistant Professor
@ianmilligan1



UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History

Jimmy Lin
Professor and David R. Cheriton Chair
@lintool



UNIVERSITY OF WATERLOO
FACULTY OF MATHEMATICS
David R. Cheriton School
of Computer Science