Hands-on with Warcbase
The Workshop

Ian Milligan
Assistant Professor
@ianmilligan1

Nick Ruest
Digital Assets Librarian
@ruebot

UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History

YORK UNIVERSITY
CATS: ALL YOUR BASE ARE BELONG TO US.
The Web as a Primary Source

- Web archives will fundamentally affect the way historians write history
  - We will have easier access to information on a previously-unknown scale, as well as improved capability to parse it;
  - Yet historians need to reflect on the shape that Web-based primary sources will take, and how we will be able to access them
Could one even study the 1990s and beyond without web archives?
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
Warcbase

WARC/ARC or Ingestion

Processing & Analytics

Applications and Services

Apache HBase

Hadoop
Warcbase

• Framework for distributed storage and distributed processing of very big data

• Scalable
  • From Raspberry Pi to Desktop Computer to Server to Cluster, all with the same scripts & commands

• Potentially very powerful
  • Trantor: 1.2PB of disk, 25 compute nodes (each w/ 128GB memory, 2×6-core Intel Xeon E5 v3 = 3.2TB memory and 300 current-generation Intel cores)
You can Warcbase Too (and will here!)

warcbase.org  docs.warcbase.org
Two Case Studies

- **Archive-It Research Services**: “Canadian Political Parties and Political Interest Groups”

- 2005 - 2015

- WARC files
Two Case Studies

- GeoCities
- End-of-life crawl from 2009
- WARC files
- 4.1 TB, 186 million HTML documents
Step One: Grabbing WARCs
Step Two: Basic Shell Analysis
Step Two: Basic Analytics
Step Three: Filtering a Corpus

```scala
import org.warcbase.spark.matchbox.{ExtractTopLevelDomain, ExtractLinks, RecordLoader}
import org.warcbase.spark.rdd.RecordRDD._

RecordLoader.loadArc("/mnt/vol1/data_sets/geocities/warcs/*", sc)
  .keepValidPages()
  .map(r => (r.getCrawldate, ExtractLinks(r.getUrl, r.getContentString())))
  .flatMap(r => r._2.map(f => (r._1, ExtractTopLevelDomain(f._1).replaceAll("^\s*www\./", ""), ExtractTopLevelDomain(f._2).replaceAll("^\s*www\./", ""))))
  .filter(r => r._2 != "" && r._3 != "")
  .countItems()
  .filter(r => r._2 > 5)
  .saveAsTextFile("/mnt/vol1/data_sets/geocities/geocities.sitelinks")
```

A Link Graph
Step Three: Filtering a Corpus

Results


Filtering
## Finding cool sites!

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

Welcome to the Enchanted Forest Awards Page!

Hip Hip Hurray!
Everyone likes to win awards right?
Well what are you waiting for?
The Enchanted Forest offers One main award and it is TOPS!
The New Enchanted Forest Award of Excellence!

Yahoo got rid of us! Sorry but we do not exist anymore. We haven't in a long time now. Yahoo doesn't care about the kids anymore. Pedophiles and porno sickos can now run rampant and there's nothing anyone at Yahoo is gonna do! They just don't care about the safety of children. All they care about is numbers and money! May God have mercy on them all!
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!
Or generate Solr indexes using Warcbase too!
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.
Step Five: $$$$$
https://github.com/web-archive-group/warcbase_workshop_vagrant