StormCrawler at Common Crawl and other use cases

IIPC 2019 Zagreb
Part 1

Julien Nioche - CameraForensics
Bristol, UK

Introduction to StormCrawler

Use cases
Part 2

Sebastian Nagel - CommonCrawl

News crawler
- Written in Java
- Open Source (Apache licence)
- Distributed / Scalable
- Robust
- Efficient yet polite
- Low latency
- Modular
- Lightweight
- Versatile
Ecosystem

- Rich set of features
  - Extraction
  - Filtering
  - Sitemaps
  - Robots.txt

- Modules
  - including WARC

- Monitoring

- Metrics
  - Kibana
  - Grafana
Storage agnostic

- Elasticsearch
- SOLR
- SQL
- Distributed Queues
- ...

...
Basic Dataflow

- Spout
- URL partitioner
- Fetcher
- Parser
- Indexer
- URL storage
- Status Updater
- Index
Sample of users

https://github.com/DigitalPebble/storm-crawler/wiki/Powered-By
Internet Advertisement

- Streaming URLs
- Distributed queue
- Simple architecture
- Low latency
Government of Northwest Territories

Search Engine

- Replace Google Search Appliance
- Limited number of sites
- Index to Elasticsearch
Top 5 Takeaways from the 2019/20 NWT Tourism Marketing ...
https://www.itl.gov.ca/fr/node/4651
Using Search Engine Optimization (SEO) and Search Engine Marketing (SEM), NWTT will pull more of the right people to read and interact with the NWT stories and content.

Top 5 Takeaways from the 2019/20 NWT Tourism Marketing ...
https://www.itl.gov.nt.ca/en/newsroom/top-5-takeaways-201920-nwt-tourism-ma...
Using Search Engine Optimization (SEO) and Search Engine Marketing (SEM), NWTT will pull more of the right people to read and interact with the NWT stories and content.

Mark this Page!
https://www.gov.nt.ca/en/newsroom/nwt-residents-get-new-direct-information...
RSS readers can be found using your favourite search engine. For more information contact: James MacKenzie Media Liaison Office of the Premier and Executive Council Government of the Northwest Territories Tel: (867) 669-2302 Fax: (867) 873-0111

Mark this Page!
https://www.gov.nt.ca/fr/newsroom/nwt-residents-get-new-direct-information...
RSS readers can be found using your favourite search engine. For more information contact: James MacKenzie Media Liaison Office of the Premier and Executive Council Government of the Northwest Territories Tel: (867) 669-2302 Fax: (867) 873-0111

Digital Literacy - Teacher Resource
https://www.ece.gov.nt.ca/en/services/resources-current-teachers principals/...
Its five modules focus on social networking and privacy; tricks of the trade of online advertisers; authentication skills and judging reliability through a mock search
Image tracking

- Image Rights Protection

- Different types of crawls
  - total 200M pages / day
  - up to 10B URLs
  - 10’s millions domains

- StormCrawler + Elasticsearch

Child protection

- Crawl photos to identify cameras
- Over 2 billion images
- Elasticsearch storage
- Heavily extended
  - Custom analysis
http://stormcrawler.net
StormCrawler at Common Crawl and other use cases

Part II – News Crawler

Sebastian Nagel
sebastian@commoncrawl.org

IIPC Web Archiving Conference, 6–7 June 2019, Zagreb, Croatia
we’re a non-profit that makes web data accessible to programmers and data scientists

for natural language processing, web science, ...

hosted as Open Data set on Amazon Web Services

main crawl data released monthly:
3 billion page captures, 50 TiB WARC files

4 Petabytes of data in total (2008 – 2019)
Why a news dataset?

- monthly crawl and release schedule (as for main crawls) is not optimal for the news genre
- continuously release freshly crawled data
- broad and global sample
- news feeds and sitemaps are used as seeds
  - well-maintained and clean (almost no spam, few ads)
  - only recent news articles crawled using the publication date announced in feeds and sitemaps
- small collection, for users easy to work with
- as usual: HTML pages (news articles) archived in WARC files
The News Data Set – Basic numbers

- started autumn 2016
- 5.5 TiB WARC files, 220 million pages in total (May 2019)
- currently: 8 – 10 million pages per month
- from 12,000 domains
- data from 200 top-level domains and over 100 languages
<table>
<thead>
<tr>
<th>%</th>
<th>% cumul.</th>
<th>TLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>44.4</td>
<td>44.4</td>
<td>com</td>
</tr>
<tr>
<td>6.9</td>
<td>51.3</td>
<td>de</td>
</tr>
<tr>
<td>4.0</td>
<td>55.3</td>
<td>it</td>
</tr>
<tr>
<td>3.3</td>
<td>58.6</td>
<td>ru</td>
</tr>
<tr>
<td>3.0</td>
<td>61.7</td>
<td>uk</td>
</tr>
<tr>
<td>2.8</td>
<td>64.4</td>
<td>es</td>
</tr>
<tr>
<td>2.4</td>
<td>66.9</td>
<td>co</td>
</tr>
<tr>
<td>2.0</td>
<td>68.8</td>
<td>net</td>
</tr>
<tr>
<td>1.8</td>
<td>70.6</td>
<td>fr</td>
</tr>
<tr>
<td>1.6</td>
<td>72.2</td>
<td>gr</td>
</tr>
<tr>
<td>1.4</td>
<td>73.6</td>
<td>tr</td>
</tr>
<tr>
<td>1.4</td>
<td>75.0</td>
<td>ua</td>
</tr>
<tr>
<td>1.3</td>
<td>76.3</td>
<td>ro</td>
</tr>
<tr>
<td>1.2</td>
<td>77.5</td>
<td>au</td>
</tr>
<tr>
<td>1.2</td>
<td>78.7</td>
<td>se</td>
</tr>
<tr>
<td>1.0</td>
<td>79.7</td>
<td>ch</td>
</tr>
<tr>
<td>1.0</td>
<td>80.7</td>
<td>nl</td>
</tr>
<tr>
<td>0.9</td>
<td>81.6</td>
<td>mx</td>
</tr>
<tr>
<td>0.9</td>
<td>82.5</td>
<td>org</td>
</tr>
<tr>
<td>0.7</td>
<td>83.3</td>
<td>cz</td>
</tr>
<tr>
<td>0.7</td>
<td>83.9</td>
<td>pl</td>
</tr>
<tr>
<td>0.7</td>
<td>84.6</td>
<td>ca</td>
</tr>
<tr>
<td>0.7</td>
<td>86.0</td>
<td>hu</td>
</tr>
<tr>
<td>0.7</td>
<td>86.6</td>
<td>br</td>
</tr>
<tr>
<td>0.7</td>
<td>87.3</td>
<td>at</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>%</th>
<th>% cumul.</th>
<th>language</th>
</tr>
</thead>
<tbody>
<tr>
<td>34.8</td>
<td>34.8</td>
<td>English</td>
</tr>
<tr>
<td>10.3</td>
<td>45.1</td>
<td>Spanish</td>
</tr>
<tr>
<td>8.9</td>
<td>54.0</td>
<td>German</td>
</tr>
<tr>
<td>5.7</td>
<td>59.6</td>
<td>Russian</td>
</tr>
<tr>
<td>5.5</td>
<td>65.2</td>
<td>Italian</td>
</tr>
<tr>
<td>4.6</td>
<td>69.7</td>
<td>Arabic</td>
</tr>
<tr>
<td>4.0</td>
<td>73.7</td>
<td>Turkish</td>
</tr>
<tr>
<td>3.7</td>
<td>77.4</td>
<td>French</td>
</tr>
<tr>
<td>2.3</td>
<td>79.7</td>
<td>Hindi</td>
</tr>
<tr>
<td>1.7</td>
<td>81.4</td>
<td>Greek</td>
</tr>
<tr>
<td>1.7</td>
<td>83.1</td>
<td>Portuguese</td>
</tr>
<tr>
<td>1.5</td>
<td>84.6</td>
<td>Romanian</td>
</tr>
<tr>
<td>1.3</td>
<td>85.9</td>
<td>Swedish</td>
</tr>
<tr>
<td>1.3</td>
<td>87.2</td>
<td>Dutch</td>
</tr>
<tr>
<td>0.8</td>
<td>88.0</td>
<td>Chinese</td>
</tr>
<tr>
<td>0.8</td>
<td>88.8</td>
<td>Korean</td>
</tr>
<tr>
<td>0.8</td>
<td>89.5</td>
<td>Hungarian</td>
</tr>
<tr>
<td>0.8</td>
<td>90.3</td>
<td>Polish</td>
</tr>
<tr>
<td>0.7</td>
<td>91.0</td>
<td>Czech</td>
</tr>
<tr>
<td>0.7</td>
<td>91.7</td>
<td>Ukrainian</td>
</tr>
<tr>
<td>0.6</td>
<td>92.3</td>
<td>Japanese</td>
</tr>
<tr>
<td>0.6</td>
<td>93.0</td>
<td>Tamil</td>
</tr>
<tr>
<td>0.6</td>
<td>93.6</td>
<td>Indonesian</td>
</tr>
<tr>
<td>0.6</td>
<td>94.1</td>
<td>Bulgarian</td>
</tr>
<tr>
<td>0.5</td>
<td>94.6</td>
<td>Vietnamese</td>
</tr>
</tbody>
</table>
Why StormCrawler?

- streaming approach allows to follow links very quickly
  (Note: for the main crawls we use Nutch, a batch-based crawler)
- flexible topology creation
- Elasticsearch index allows at any time to instantly
  - add news feeds or sitemaps
  - get metrics and statistics
- runs with little supervision over months
- on a single dual core machine with 32 MB RAM
Spout reads Elasticsearch status index and emits tuples `<URL,metadata> “ripe” for (re)fetch

- tuples are partitioned and passed to ...

- Fetcher bolts fetch URLs and
  - send fetch status to Status Updater bolt
  - pass content to
News crawler topology

- Sitemap Parser bolt
  - checks whether content is a news sitemap or sitemap index
  - if yes, parse sitemap and send links (subsitemaps or news articles) to Status Updater bolt
  - if not, pass forward to
- Feed Parser bolt
  - checks whether content is a RSS or Atom feed
  - if yes, parse and send links to Status Updater bolt
  - if not, pass forward to
- WARC writer and
- DummyIndexer which passes the tuple forward to
- Status Updater bolt
  - updates the fetch status in resp.
  - adds tuples to Elasticsearch status index
Detection of feeds and sitemaps

- the robots.txt is checked for announced sitemaps
- news sitemaps are detected by format (other sitemaps are ignored)
- bootstrap topology to detect feeds from alternate links in HTML pages
  (needs supervision to remove comment and multi-media feeds)
Re-fetch scheduling of feeds and sitemaps

- the crawler knows about 2 million feeds and sitemaps: 60k feeds, 800k news sitemaps, 20k sitemap indexes, 1.2 million other sitemaps
- need to focus on frequently changing feeds/sitemaps
  - “forgotten” feeds not maintained anymore
  - some sites provide a separate sitemaps every day
  - or provide hundreds of feeds/sitemaps (by topic or location), most of them infrequently updated
- scheduling is adapted to the change frequency (detected by signature comparison)
- and may range between 90 min. and 90 days
References:

- data prefix on Amazon S3: s3://commoncrawl/crawl-data/CC-NEWS/
- https://commoncrawl.org/2016/10/news-dataset-available/
- https://github.com/commoncrawl/news-crawl
- https://groups.google.com/forum/#!forum/common-crawl