CS6604 Digital Libraries
Global Events Team Final Presentation

Presenters:
Liuqing Li, Islam Harb, Andrej Galad
{liuqing, iharb, agalad}@vt.edu

Instructor:
Dr. Edward A. Fox

Virginia Polytechnic Institute and State University
Blacksburg, VA, 24061
April 27, 2017
Outline

• Background

• Implementation
  • Data Collection
  • Data Processing
  • Data Visualization

• Future Work

• Acknowledgement
Background

• GETAR*
  • Global Event and Trend Archive Research
  • Architecture

Implementation – Architecture

Event Focused Crawler (EFC)

WARC Files → CDX Writer → CDX Files

ArchiveSpark

Apache Spark → Stanford NER → Regular Expression → Score Function

Entity-based Results

Standalone HBase → Web Application

Data Collection

Data Processing

Data Visualization
# Events of Interest

<table>
<thead>
<tr>
<th>School Shooting Events</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Virginia Tech Shooting</td>
<td>2007</td>
</tr>
<tr>
<td>Northern Illinois University Shooting</td>
<td>2008</td>
</tr>
<tr>
<td>Dunbar High School Shooting</td>
<td>2009</td>
</tr>
<tr>
<td>University of Alabama Shooting</td>
<td>2010</td>
</tr>
<tr>
<td>Worthing High School Shooting</td>
<td>2011</td>
</tr>
<tr>
<td>Sandy Hook Elementary School Shooting</td>
<td>2012</td>
</tr>
<tr>
<td>Sparks Middle School Shooting</td>
<td>2013</td>
</tr>
<tr>
<td>Reynolds High School Shooting</td>
<td>2014</td>
</tr>
<tr>
<td>Umpqua Community College Shooting</td>
<td>2015</td>
</tr>
<tr>
<td>Townville Elementary School Shooting</td>
<td>2016</td>
</tr>
</tbody>
</table>
Focused Crawler – Collecting / Archiving

START

Manually Curate Seeds

URLs Queue

Download Page

Calculate Relevancy

Relevant?

Yes

All URLs?

Yes

Process Page & Convert into WARC Format

Append Result warc.gz Event File

No

No

Discard

No
WARC Libraries

• Wget (Version 1.14 or later)

```
wget
  --mirror
  --warc-file=YOUR_FILENAME
  --warc-cdx
  --page-requisites
  --html-extension
  --convert-links
  --execute robots=off
  --directory-prefix=. 
  --span-hosts
  --domains=example.com, www.example.com, cdn.example.com
  --user-agent=Mozilla (mailto:archiver@petekeen.net)
  --wait=10
  --random-wait
```
WARC Libraries

- Wpull

Note

When resuming downloads with `--warc-file` and `--database`, Wpull will overwrite the WARC file by default. This occurs because Wpull simply maintains a list of URLs that are fetched and not fetched. You should either rename the existing file manually or use the additional option `--warc-append` or move the files `--warc-move`.

- `--warc-append`
- `--warc-move`: Move WARC files out of the way for resuming a crashed crawl.
WARC Libraries

- **WARCIO**: WARC (and ARC) Streaming Library
  - Python 2.7+ and 3.3+
  - Post-Processing: Read / Write WARC format

```python
from warcio.warcwriter import WARCWriter
from warcio.statusandheaders import StatusAndHeaders

import requests

with open('example.warc.gz', 'wb') as output:
    writer = WARCWriter(output, gzip=True)

    resp = requests.get('http://example.com/',
                        headers={'Accept-Encoding': 'identity'},
                        stream=True)

    # get raw headers from urllib3
    headers_list = resp.raw.headers.items()

    http_headers = StatusAndHeaders('200 OK', headers_list, protocol='HTTP/1.0')

    record = writer.create_warc_record('http://example.com/', 'response',
                                        payload=resp.raw,
                                        http_headers=http_headers)

    writer.write_record(record)
```
### Ten Events Collections

- **Naming Convention**
  - `[location]_[year].warc.gz`

<table>
<thead>
<tr>
<th>ID</th>
<th>Date</th>
<th>Time</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>islam 788109490</td>
<td>Apr 18</td>
<td>13:45</td>
<td>townville_2016.warc.gz</td>
</tr>
<tr>
<td>islam 758426598</td>
<td>Apr 18</td>
<td>13:43</td>
<td>umpqua_2015.warc.gz</td>
</tr>
<tr>
<td>islam 696778729</td>
<td>Apr 18</td>
<td>13:38</td>
<td>omaha_NE_2011.warc.gz</td>
</tr>
<tr>
<td>islam 661944603</td>
<td>Apr 18</td>
<td>13:40</td>
<td>Sandy_2012.warc.gz</td>
</tr>
<tr>
<td>islam 454081122</td>
<td>Apr 18</td>
<td>13:29</td>
<td>VT_2007.warc.gz</td>
</tr>
<tr>
<td>islam 279509497</td>
<td>Apr 18</td>
<td>13:41</td>
<td>Sparks_2013.warc.gz</td>
</tr>
<tr>
<td>islam 15635547</td>
<td>Apr 18</td>
<td>13:36</td>
<td>dunbar_chicago_2009.warc.gz</td>
</tr>
<tr>
<td>islam 14708587</td>
<td>Apr 18</td>
<td>13:41</td>
<td>reynolds_2014.warc.gz</td>
</tr>
<tr>
<td>islam 5705504</td>
<td>Apr 18</td>
<td>13:36</td>
<td>huntsville_2010.warc.gz</td>
</tr>
<tr>
<td>islam 531513</td>
<td>Apr 18</td>
<td>13:35</td>
<td>NIU_2008.warc.gz</td>
</tr>
</tbody>
</table>
Tools for Data Processing

• ArchiveSpark
  • Apache Spark framework for Web Archives
  • Easy data extraction
  • Input: WARC and CDX files

• CDX Writer
  • Python script to create CDX files of WARC files
  • Format: CDX N b a m s k r M S V g
    • e.g., edu,vt,cnre)/ 20170422005601 http://cnre.vt.edu text/html
      200 BT3ILJXROIILHBKQPNYDUCUVZRDKG3OA - - 9478 20104749
      data/Virginia-Tech-Shooting_20070416.warc.gz
Data Preprocessing

• Webpage Cleaning
  • Extract Raw Text
    • payload.string.html.body.text
  • Remove jQuery & JavaScript
    • \{ WPGroHo.syncProfileData( hash, id ); \}, ...
  • Remove tags
    • <br>, <p>, ...
  • Remove markers
    • *, |+, ...
  • Remove stopwords
    • a, about, the, ...

Global Events Team Final Presentation
Data Processing

• Entity Extraction
  • Basic Parsing
    • event name and date
  • Stanford NER (Integrated model)
    • entities, shooter name
• Regular Expression
  • event date
  • shooter name and age
  • number of victims
  • weapon list

• Score Function
  • $tf \times df$
### HBase

- Build-in ImportTsv Utility
  - Import Data into HBase

<table>
<thead>
<tr>
<th>Table Name</th>
<th>globalevents</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Row Key</strong></td>
<td>Event_Date + Event Hash Value</td>
</tr>
<tr>
<td><strong>Column Family</strong></td>
<td>event</td>
</tr>
<tr>
<td>event: name</td>
<td>Virginia Tech Shooting</td>
</tr>
<tr>
<td>event: date</td>
<td>20070416</td>
</tr>
<tr>
<td>event: shooter_age</td>
<td>23-year-old</td>
</tr>
<tr>
<td>event: shooting_victims</td>
<td>32 victims</td>
</tr>
<tr>
<td>event: entities</td>
<td>Virginia;Tech;VA;University;…</td>
</tr>
<tr>
<td>event: entities_count</td>
<td>146900;62415;13940;7732;…</td>
</tr>
<tr>
<td>event: entities_url</td>
<td>url1,url2,url3,url4,url6;url2,url3,url4,url5;url1,url3,url4,url5,url6;…</td>
</tr>
</tbody>
</table>
• Key Stages
  • Initialization
    • Create Spark Session
    • Create NLP Core
    • Create Storage
  • Processing
    • Extract Event Name/Date/URL
    • Extract Name Entities
    • Extract Other Event Features
  • Export and Import
    • Generate TSV file
    • Import TSV file into HBase
Global Events Viewer

• Efficient visualization of long-term global events
  • Show representative terms -> link to corresponding URLs
  • Visualize events’ trends over time (time series)
• Java 7 Spring Boot Web application
  • Build system - Gradle
  • Embedded Tomcat Web server
  • Backend - HBase, in-memory
  • Frontend - D3.js, Bootstrap

https://github.com/dedocibula/global-events-viewer
• Key Components
  • WordCloud, Range Selection, URL List, Trends
Problem Faced

**Data Collection**
- Encoding problems (UTF-8, ASCII and others)
- Get more relevant seeds for old events

**Data Processing**
- Lack of documentation (ArchiveSpark)
- Version conflict (CDX Writer, Kernel in Jupyter)
- JVM issue (Spark)

**Data Visualization**
- Spring boot IntelliJ setup
- JQuery UI
Lessons Learned

Data Collection
- WARCIO
- Focused Crawler

Data Processing
- ArchiveSpark
- Spark & Scala (Map/Reduce Process)

Data Visualization
- D3 WordCloud
- D3 Dynamic Line Charts
# Future Work

## Data Collection
- Wayback Machine
- Automatic Routine for Focused Crawler
- Event Extension (Sources, Time, Space)

## Data Processing
- Standalone Mode -> Cluster Mode
- Name Entity Recognizer
- Automatic Processing (CDX Writer and HBase)

## Data Visualization
- Localization – Datamaps
- Weapons
## Acknowledgement

### Projects

<table>
<thead>
<tr>
<th>Project ID</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSF IIS - 1319578</td>
<td>III: Small: Integrated Digital Event Archiving and Library (IDEAL)</td>
</tr>
<tr>
<td>NSF IIS - 1619028</td>
<td>III: Small: Collaborative Research: Global Event and Trend Archive Research (GETAR)</td>
</tr>
</tbody>
</table>

### Organizations

- Internet Archive
- L3S Research Center

### Persons

- **Instructor**: Dr. Edward A. Fox
- **Alumnus**: Dr. Mohamed Magdy Farag
- **Labmates**: Prashant Chandrasekar, Xuan Zhang
Thank you!

Questions?