National Science Foundation - Internet Archive research

Rutgers University, School of Communication and Information
Prof. Matthew Weber
2013-2018
Coding support: Hai Nguyen
Technical support: Eric Marshall, Kristina Plazonic, Ben Bakelaar

---

Batch 1

Processed between 2015-2017

DATA SETS

/media
/newsecosystem
/ows
/sandy

INPUT DATA

ARC file format
http://archive.org/web/researcher/ArcFileFormat.php

WARC file format
https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml

Processed with Apache Pig script
Hai Nguyen
https://en.wikipedia.org/wiki/Pig_(programming_tool)

Code:

```
REGISTER /home/hdn11/Scripts/archive-commons-jar-with-dependencies.jar;
REGISTER /home/hdn11/Scripts/webcrawler.jar;
REGISTER /home/hdn11/Scripts/commons-validator-1.4.0.jar;

*** For WARC data
titles = LOAD '/nsfia/input/warc/files' USING
org.archive.hadoop.ArchiveJSONViewLoader(
       'Envelope.Payload-Metadata.HTTP-Response-Metadata.HTML-Metadata',
       'Envelope.WARC-Header-Metadata.WARC-Target-URI',
       'Envelope.WARC-Header-Metadata.WARC-Date',
       'Envelope.WARC-Header-Metadata.Content-Type',
       'Envelope.WARC-Header-Metadata.Content-Length') AS
       (links:chararray, target:chararray, date:chararray, contenttype:chararray,
contentlength:chararray);

*** For ARC data
titles = LOAD '/nsfia/input/...' USING org.archive.hadoop.ArchiveJSONViewLoader(
```

```
       'Envelope.Payload-Metadata.HTTP-Response-Metadata.HTML-Metadata',
       'Envelope.ARC-Header-Metadata.ARC-Target-URI',
       'Envelope.ARC-Header-Metadata.ARC-Date',
       'Envelope.ARC-Header-Metadata.Content-Type',
       'Envelope.ARC-Header-Metadata.Content-Length') AS
       (links:chararray, target:chararray, date:chararray, contenttype:chararray,
contentlength:chararray);

nonnulls = filter titles by links is not null;
paths = foreach nonnulls generate org.sci.historycrawl.parser($0,$1,$2),$2,$3,$4;
i6 = foreach paths generate bagwati.url,$1,$2,$3;
i7 = foreach i6 generate flatten($0) as words,
       org.sci.historycrawl.formatdate(SUBSTRING($1,0,10)), $2,$3;
i8 = foreach i7 generate org.sci.historycrawl.getsourceURL($0),
       org.sci.historycrawl.getdstURL($0),
       org.sci.historycrawl.getText($0),$1,$2,(long)$3;
i9 = group i8 by ($0,$1,$3);
i10 = foreach i9 generate FLATTEN(group), FLATTEN(TOP(1,0,i8.$2)), COUNT(i8),
       FLATTEN(TOP(1,0,i8.$4)), SUM(i8.$5);
i11 = filter i10 by $0 is not null;
i12 = filter i11 by $1 is not null;
store i12 INTO '/nsfia/output/...' using PigStorage();
```

OUTPUT DATA

Fields

1 Source URL
2 Target URL (clipped)
3 Date (YYYYMMDD)
4 Text within hyperlink
5 Content-type
6 Content-length
7 # of links

MEDIA dataset

01 MEDIA_2008_2012_part01.tar.gz
02 MEDIA_2008_2012_part02.tar.gz
03 MEDIA_2008_2012_part03.tar.gz
04 MEDIA_2008_2012_part04.tar.gz

05 MEDIA_2008_2012_part05.tar.gz
06 MEDIA_2008_2012_part06.tar.gz
07 MEDIA_2008_2012_part07.tar.gz
08 MEDIA_2008_2012_part08.tar.gz
09 MEDIA_2008_2012_part09.tar.gz
10 MEDIA_2008_2012_part10.tar.gz
11 MEDIA_2008_2012_part11.tar.gz
12 MEDIA_2008_2012_part12.tar.gz
13 MEDIA_2008_2012_part13.tar.gz
14 MEDIA_2008_2012_part14.tar.gz
15 MEDIA_2008_2012_part15.tar.gz
16 MEDIA_2008_2012_part16.tar.gz
17 MEDIA_2008_2012_part17.tar.gz
18 MEDIA_2008_2012_part18.tar.gz

NEWSECOSYSTEM dataset

01 NEWSECOSYSTEM_2015_2015_part01.tar.gz
02 NEWSECOSYSTEM_2015_2015_part02.tar.gz
03 NEWSECOSYSTEM_2015_2015_part03.tar.gz
04 NEWSECOSYSTEM_2015_2015_part04.tar.gz

OWS (Occupy Wall Street) dataset

01 NSFIA_OWS_2010_2012_part01.tar.gz
02 NSFIA_OWS_2010_2012_part02.tar.gz
03 NSFIA_OWS_2010_2012_part03.tar.gz

SANDY (Hurricane Sandy) dataset

01 NSFIA_SANDY_2003_2012-all.tar.gz

Batch 2

Processed between 2018

DATA SETS

/earlyweb
/house
/senate

INPUT DATA

ARC file format
http://archive.org/web/researcher/ArcFileFormat.php

WARC file format
https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml

Processed with Sparq/Scala script
Kristina Plazonic
Data Scientist
Office of Advanced Research Computing
Office of Information Technology

Note that Senate and house dataset links were provided raw due to most links being internal (same SLD)

OUTPUT DATA

Fields

1 Source URL (only SLD, no pages/sub-directories)
2 Target URL (same as above)
3 Date (YYYYMMDD)
4 Sum of content length
5 # of links

EARLYWEB dataset

EARLYWEB_1996_2000_part01.tar
EARLYWEB_1996_2000_part02.tar
EARLYWEB_1996_2000_part03.tar
EARLYWEB_1996_2000_part04.tar
EARLYWEB_1996_2000_part05.tar
EARLYWEB_1996_2000_part06.tar
EARLYWEB_1996_2000_part07.tar
EARLYWEB_1996_2000_part08.tar
EARLYWEB_1996_2000_part09.tar
EARLYWEB_1996_2000_part10.tar
EARLYWEB_1996_2000_part11.tar
EARLYWEB_1996_2000_part12.tar
EARLYWEB_1996_2000_part13.tar
EARLYWEB_1996_2000_part14.tar
EARLYWEB_1996_2000_part15.tar
EARLYWEB_1996_2000_part16.tar

HOUSE dataset

HOUSE_2009_part01.tar.gz
HOUSE_2009_part02.tar.gz
HOUSE_2009_part03.tar.gz
HOUSE_2009_part04.tar.gz
HOUSE_2009_part05.tar.gz
HOUSE_2009_part06.tar.gz
HOUSE_2009_part07.tar.gz
HOUSE_2009_part08.tar.gz

SENATE dataset

SENATE_2009_part01.tar.gz
SENATE_2009_part02.tar.gz
SENATE_2009_part03.tar.gz

ARC

SAMPLE ARC FILE FORMAT HEADER (JSON)


```
{
   "Envelope":{
      "Format":"ARC",
      "ARC-Header-Metadata":{
         "Date":"20130724000000",
         "Content-Length":"76",
         "Content-Type":"text/plain",
"Target-URI":"filedesc://NSF-HOUSEGOV-2006-2012-EXTRACTION-PART-00222-000109.arc.gz",
         "IP-Address":"0.0.0.0"
      },
      "ARC-Header-Length":"107",
      "Payload-Metadata":{
         "Trailing-Slop-Length":"1",
         "Actual-Content-Type":"alexa/filedesc",
         "Block-Digest":"sha1:KRKKRWMWT3AKLJMKNCQ7ROCGWVDA4YRV",
         "Actual-Content-Length":"76",
         "Filedesc-Metadata":{
            "Format":"URL IP-address Archive-date Content-type Archive-length",
            "Organization":"InternetArchive",
            "Major-Version":"1",
            "Minor-Version":"0"
         }
      }
   },
   "Container":{
      "Compressed":true,
      "Gzip-Metadata":{
         "Footer-Length":"8",
         "Deflate-Length":"170",
         "Header-Length":"10",
         "Inflated-CRC":"1680566000",
         "Inflated-Length":"184"
      },
      "Offset":"0",
      "Filename":"NSF-HOUSEGOV-2006-2012-EXTRACTION-PART-00222-000109.arc.gz"
   }
}
```

WARC

SAMPLE WARC FILE FORMAT HEADER (JSON)


```
{
   "Envelope":{
      "Format":"WARC",
      "WARC-Header-Length":"286",
      "Block-Digest":"sha1:FJDC7W6E3472YPPWOGPCB42HZPHW54TN",
      "Actual-Content-Length":"349",
      "WARC-Header-Metadata":{
         "WARC-Type":"warcinfo",
"WARC-Filename":"NSF-HOUSEGOV-2006-2012-WARCS-EXTRACTION-PART-00238-000012.warc.gz",
         "WARC-Date":"2013-07-24T00:00:00.000Z",
         "Content-Length":"349",
         "WARC-Record-ID":"<urn:uuid:urn:uuid:d60290cf-3915-4202-953d-5b0243401a71>",
         "Content-Type":"application/warc-fields"
      },
      "Payload-Metadata":{
         "Trailing-Slop-Length":"4",
         "Actual-Content-Type":"application/warc-fields",
         "Actual-Content-Length":"345",
         "WARC-Info-Metadata":{
            "robots":"ignore",
            "software":"archive-commons-0.0.1-SNAPSHOT-2011-06-28 03:04:05
Extractor",
            "audience":"NSF Project",
            "http-header-from":"archive-crawler-agent@lists.sourceforge.net",
            "http-header-user-agent":"Mozilla/5.0 (compatible; archive.org_bot/1.10.0
+http://www.archives.gov/crawl.html)",
            "format":"WARC File Format 1.0",
            "publisher":"Internet Archive"
         }
      }
   },
   "Container":{
      "Compressed":true,
      "Gzip-Metadata":{
         "Footer-Length":"8",
         "Deflate-Length":"451",
         "Header-Length":"10",
         "Inflated-CRC":"1827023597",
         "Inflated-Length":"635"
      },
      "Offset":"0",
```

```
      "Filename":"NSF-HOUSEGOV-2006-2012-WARCS-EXTRACTION-PART-00238-000012.warc.gz"
   }
}
```