VT Web Archiving

Anthony Rinaldi and Dev Mehta
CS 4624
Clients: Mohamed Magdy and Tarek Kanan
Blacksburg, VA
5/6/2014

Project Goals

- Setup a web-crawler with Heritrix
- Archive files from vt.edu
- Integrate with Wayback
- Set-up Search with Solr (Stretch)

Problems Encountered

- Older version of software.
- Finding documentation to configure Heritrix.
  - Only crawl vt.edu pages.
  - Crawl all vt.edu pages.
- Issues with CentOS firewalling.

Work Accomplished

- Working set-up of Heritrix that successfully crawls vt.edu web-pages.
  - Customized configuration to increase crawl depth.
  - Reject non-domain based URLs.
- Working set-up of Wayback machine:
  - Processes warc files from Heritrix.
  - Front-end for Heritrix-based crawls.

Lessons Learned

- Sometimes, documentation leaves much to be desired.
- Crawls can be extremely large if not configured properly.

Demo

Heritrix:

- https://administrator:mQW8GzEsZAr8SxAketPY@webarchive.cc.vt.edu:12222/

Wayback:

- http://webarchive.cc.vt.edu/

Engine

Memory:

162426 KiB used; 232960 KiB current heap; 232960 KiB max heap run garbage collector

Jobs Directory:

/apps/nas/heritrix-3.2.0/jobs

Job Directories

(7) detected rescan

vteduDevRun1 «Finished: ABORTED» 2 launches
/apps/nas/heritrix-3.2.0/jobs/vteduDevRun1/crawler-beans.cxml
(last at 2014-05-05T23:35:39.836Z)
vteduDevRun «Finished: ABORTED» 1 launches
/apps/nas/heritrix-3.2.0/jobs/vteduDevRun/crawler-beans.cxml
(last at 2014-05-05T22:44:36.438Z)
vtedurun «Finished: FINISHED» 1 launches
/apps/nas/heritrix-3.2.0/jobs/vtedurun/crawler-beans.cxml
(last at 2014-04-28T18:44:38.186Z)
vtedu-2 0 launches
/apps/nas/heritrix-3.2.0/jobs/vtedu-2/profile-crawler-beans.cxml
1 launches

Job *vteduDevRun1*

(2 launches, last 3d3h ago)

build | launch | pause | unpause | checkpoint | terminate | teardown

Job Log more

```
2014-05-07T01:11:32.815Z INFO FINISHED 20140505233541
2014-05-07T01:11:24.301Z WARNING unable to tally host stats for 0.0.0.0 (in thread 'org.archive.crawler.frontier
2014-05-07T01:10:31.907Z INFO STOPPING 20140505233541
2014-05-07T01:10:27.275Z INFO PAUSING 20140505233541
2014-05-06T15:59:52.912Z WARNING unable to tally host stats for http://0.0.0.0/moat-content-tag.js (in thread 'T
```

Job is Finished: ABORTED

Totals
2,520,583 downloaded + 7,325,359 queued = 9,845,942 total
547 GiB crawled (547 GiB novel, 0 B dupByHash, 764 B notModified)
Alerts
18 tail alert log...
Rates
0 URIs/sec (27.35 avg); 0 KB/sec (6,228 avg)

Job is Finished: ABORTED

Totals
2,520,583 downloaded + 7,325,359 queued = 9,845,942 total
547 GiB crawled (547 GiB novel, 0 B dupByHash, 764 B notModified)
Alerts
18 tail alert log...
Rates
0 URIs/sec (27.35 avg); 0 KB/sec (6,228 avg)
Load
0 active of 0 threads; 99.36 congestion ratio; 167,453 deepest queue; 88 average depth
Elapsed
1d1h35m49s15ms
Threads
n/a
Frontier
FINISH - 90,029 URI queues: 903 active (0 in-process; 69 ready; 834 snoozed); 81,963 inactive; 0 ineligible; 0 retired; 7,163 exhausted
Memory
167277 KiB used; 232960 KiB current heap; 232960 KiB max heap

Crawl Log more

```
2014-05-07T01:10:38.678Z   200   13305044 http://lmld.org/wp-content/uploads/2013/01/006edited.jpg LLEXRRXXXXEEE
2014-05-07T01:10:31.357Z   200      57403 http://www.colorado.edu/news/releases/2013/09/18/profiles/cu_homepage/
2014-05-07T01:10:30.353Z   200     194698 http://weibo.com/p/10151501_2832143/home?from=page_101515&mod=rank&cof
2014-05-07T01:10:30.305Z   200     111871 http://www.abowlfulloflemons.net/2013/08/sponsor-shout-out-august-2013
2014-05-07T01:10:30.297Z   301         26 http://www.shape.com/lifestyle/sex-and-love/blogs/healthy-eating/fit-f
```

```
<?xml version="1.0" encoding="UTF-8"?>
<!--
  HERITRIX 3 CRAWL JOB CONFIGURATION FILE

  This is a relatively minimal configuration suitable for many crawls.

  Commented-out beans and properties are provided as an example; values
  shown in comments reflect the actual defaults which are in effect
  if not otherwise specified specification. (To change from the default
  behavior, uncomment AND alter the shown values.)
 -->
<beans xmlns="http://www.springframework.org/schema/beans"
       xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
       xmlns:context="http://www.springframework.org/schema/context"
       xmlns:aop="http://www.springframework.org/schema/aop"
       xmlns:tx="http://www.springframework.org/schema/tx"
       xsi:schemaLocation="http://www.springframework.org/schema/beans http://www.springframework.org/schema/beans/spring-beans-3.0.xsd
          http://www.springframework.org/schema/aop http://www.springframework.org/schema/aop/spring-aop-3.0.xsd
          http://www.springframework.org/schema/tx http://www.springframework.org/schema/tx/spring-tx-3.0.xsd
          http://www.springframework.org/schema/context http://www.springframework.org/schema/context/spring-context-3.0.xsd">

 <context:annotation-config/>

<!--
  OVERRIDES
  Values elsewhere in the configuration may be replaced ('overridden')
  by a Properties map declared in a PropertiesOverrideConfigurer,
  using a dotted-bean-path to address individual bean properties.
  This allows us to collect a few of the most-often changed values
  in an easy-to-edit format here at the beginning of the model
  configuration.
 -->
 <!-- overrides from a text property list -->
 <bean id="simpleOverrides" class="org.springframework.beans.factory.config.PropertyOverrideConfigurer">
  <property name="properties">
   <value>
# This Properties map is specified in the Java 'property list' text format
# http://java.sun.com/javase/6/docs/api/java/util/Properties.html#load%28java.io.Reader%29

metadata.operatorContactUrl=http://www.vt.edu
metadata.jobName=basic
metadata.description=Basic crawl starting with useful defaults

##..more?..##
   </value>
  </property>
 </bean>

 <!-- overrides from declared <prop> elements, more easily allowing
      multiline values or even declared beans -->
 <bean id="longerOverrides" class="org.springframework.beans.factory.config.PropertyOverrideConfigurer">
  <property name="properties">
   <props>
    <prop key="seeds.textSource.value">
```

save changes /apps/nas/heritrix-3.2.0/jobs/vteduDevRun1/crawler-beans.cxml view

Paged view: /apps/nas/heritrix-3.2.0/jobs/vteduDevRun1/20140505233541/logs/crawl.log

« end ‹ later bytes 832554776-832515808/832554776 100% earlier › start »

forward | **reversed** + 128 - lines

«EOF»

```
2014-05-07T01:10:38.678Z   200    13305044 http://lmld.org/wp-content/uploads/2013/01/006edited.jpg LLEXRRXXXXEEEEEEEEEEEEEEEEE http://lmld.org/page/18/ image/jpeg #006
2014-05-07T01:10:31.357Z   200       57403 http://www.colorado.edu/news/releases/2013/09/18/profiles/cu_homepage/themes/cu_960_responsive/js/profiles/cu_homepage/modules
2014-05-07T01:10:30.353Z   200      194698 http://weibo.com/p/10151501_2832143/home?from=page_101515&mod=rank&coflag=page_right_bangdan LEEERXXRXXXXX http://weibo.com/p/
2014-05-07T01:10:30.305Z   200      111871 http://www.abowlfulloflemons.net/2013/08/sponsor-shout-out-august-2013.html LLEXRRXRXRXXXXXEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
2014-05-07T01:10:30.297Z   301          26 http://www.shape.com/lifestyle/sex-and-love/blogs/healthy-eating/fit-foodies LLEXXEXRRXXRXX http://www.shape.com/lifestyle/sex
2014-05-07T01:10:30.297Z   200     1262106 http://nyheter24.se/modette/lindahallberg/files/2012/09/kraft.jpg 9+XXEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE http://
2014-05-07T01:10:30.267Z   200       35874 http://photoblog.mrussellphotography.com/august-2008-part-ii/ 9+EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEX http://phot
2014-05-07T01:10:29.475Z   200      125865 http://www.cddc.vt.edu/host/imev/record.php?recID=0.936 LLLLLLLL http://www.cddc.vt.edu/host/imev/Results.php?startsWith=G-g t
2014-05-07T01:10:29.462Z   200      558841 http://beingfrugalbychoice.blogspot.com/2012/03/homemade-vicks-vapor-shower-disks.html?showComment0751333427271674 LLEXRRXXXX
2014-05-07T01:10:29.445Z   200       70406 http://www.sfenvironment.org/downloads/library/sites/all/modules/contrib/panels/js/misc/sites/all/themes/sfe/css/fields.css LI
2014-05-07T01:10:29.435Z   200      157283 http://www.2littlesuperheroes.com/2013/03/how-to-prevent-your-outdoor-chalkboard.html/ LLEXRRXXXXRX http://www.2littlesuperhe
2014-05-07T01:10:29.410Z   200       41604 http://urbancomfort.typepad.com/.a/6a01156f70f21e970c0147e2731ff7970b-800wi LLEXRRXXXXEEEEEEEEEEEEEEEEEEEEEEEEEEEE http://urb
2014-05-07T01:10:29.408Z   301         114 http://www.ivillage.com/bowl-cut/6-b-231702 LLEXRREXXXXXXXXRXX http://www.ivillage.com/suris-bob-paxs-mohawk-and-more-coolest-
2014-05-07T01:10:29.408Z   302           0 http://blogs.hbr.org/2012/04/stop-documenting-start-experienc/trackback/ LLXXRXXRRXRXREEEEEX http://blogs.hbr.org/2012/04/stop
2014-05-07T01:10:29.400Z   200        1214 http://pinyourhome.com/2013/02/10/kitchen-booth-seating-with-storage-dont-normally-go-for-built-in-but-trade-the-check-pillows
2014-05-07T01:10:29.400Z   200        5057 http://monikastamp.wordpress.com/2012/10/25/butterfly-creation/feed/ LLEXRRXRXRXXXXEEEEEEEEEEE http://monikastamp.wordpress.c
2014-05-07T01:10:29.399Z   200       99620 http://www.freshdesignblog.com/wp-content/uploads/2013/05/trailfinders-australian-chelsea-garden-design-2013.jpeg LLEXRRXXXXRE
2014-05-07T01:10:29.174Z   200    11245554 http://indico.phys.vt.edu/getFile.py/access?contribId=64&sessionId=53&resId=0&materialId=slides&confId=1 LLLLLLLLLRLX http://i
2014-05-07T01:10:29.174Z   200        6897 http://www.vogue.fr/uploads/images/thumbs/201344/22_333642009_north_80x80.jpg LLEXRRXXXXRXXRXEE http://www.vogue.fr/photo/diap
2014-05-07T01:10:28.502Z   200        9929 http://www.shelterness.com/pictures/how-to-make-mossy-terra-cota-flower-pots-1-160x120.jpg LLEXRRXRXRXXXXXEEE http://www.shelt
2014-05-07T01:10:28.496Z   200       10813 https://vine.co/v/MrzLiK0EUQr LEEERXXXEX https://vine.co/v/MrzLiK0EUQr/embed/simple text/html #021 20140507011028405+85 sha1:6
2014-05-07T01:10:28.475Z   200       53437 https://mozorg.cdn.mozilla.net/media/img/firefox/os/have-it-all/fox-tail-detail.png LXXRREXREE https://mozorg.cdn.mozilla.net/
2014-05-07T01:10:28.456Z   200       87403 http://thepapermama.com/page/297/ 256+EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE http://thepapermama.com/page/296/ tex
2014-05-07T01:10:28.418Z -9998           - http://www.crateandbarrel.com/Account/Login.aspx?r=%2fkitchen-and-food%2fcookware-bakeware%2f LLEXRRXXXXRXXX http://www.crate
2014-05-07T01:10:28.417Z   200       11535 http://www.visitthecapitol.gov/video_filter/dashboard/sites/all/modules/date/date_api/sites/all/modules/views_slideshow/js/mis
2014-05-07T01:10:28.414Z   404       32080 http://dailysavings.allyou.com/2013/10/31/halloween-restaurant-freebies/community/blog/dailysavings LLEXRRXRXRXXXXEEEEEEEEEEE
2014-05-07T01:10:28.407Z   200       25840 http://greenlitebites.com/images/2010/08/20100809_beanTomatoSpinach21-300x199.jpg LLEXRRXXXXEEEEEEEEEEEEEEEEEEEEEEEEEEE ht
2014-05-07T01:10:28.405Z   200      172430 http://wamu.org/news/14/05/01/misc/sites/all/modules/date/date_repeat_field/sites/all/modules/date/date_api/date.css LLLLXXXX
2014-05-07T01:10:28.404Z   404         316 http://www.x3dom.org/wp-content/themes/images/arrow-down.gif LLLLLRLXXLEEXEE http://www.x3dom.org/wp-content/themes/x3dom/inst
2014-05-07T01:10:28.137Z   200       65514 http://www.iab.net/events_training/webinars LLEXXXRREXRXX http://www.iab.net/ text/html #004 20140507011027659+425 sha1:U63ERA
2014-05-07T01:10:28.128Z   200       92954 http://www.eat-drink-smile.com/wp-content/uploads/2011/10/taste-of-tn-party1-580x372.jpg 45+EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
2014-05-07T01:10:28.126Z -9998           - http://www.restorationhardware.com/modal/mini-cart.jsp LLEXRRXXXXEX http://www.restorationhardware.com/assets/js/base/global.j
2014-05-07T01:10:28.125Z   200      389432 http://hellosociety.com/blog/wp-content/uploads/2013/05/popcorn_bakerbydesign.jpg LLEXRRXRXRXXXXEEEEEEEEEEEEE http://hellosc
2014-05-07T01:10:28.100Z   404         162 http://cdn.20minutos.es/mmedia/especiales/plan-avanza/css/img/mov_vermas.png XXEXXXEE http://cdn.20minutos.es/mmedia/especiale
2014-05-07T01:10:28.099Z   200      178906 http://jilliankirbybaby.com/blog/wp-content/uploads/2011/08/JKB_0396ebwweb.jpg LLEXRRXXRXXEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
2014-05-07T01:10:28.092Z   302           0 http://iowagirleats.com/2012/07/06/daily-escape-cannon-beach-oregon/feed/ 9+RXXXXXEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
2014-05-07T01:10:28.084Z   200       16861 http://a.fastcompany.net/multisite_files/fastcompany/imagecache/slideshow_small/slideshow/2013/11/3022049-slide-s-5-reunited-s
2014-05-07T01:10:27.731Z   200       42297 http://cms.web.cern.ch/news/modules/user/modules/field/theme/sites/all/modules/admin_menu/admin_devel/admin_devel.js LEEERXXX
2014-05-07T01:10:27.720Z   404       14706 http://www.psd2html.com/408/159-733 LLLEXEXREX http://www.psd2html.com/js/1399424447/main-all.js text/html #020 20140507011027
2014-05-07T01:10:27.690Z   200      120149 http://www.hollywoodreporter.com/movies?latest_news=1436 1398+EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE http://www.hc
2014-05-07T01:10:27.663Z   200       32711 http://intothegloss.com/wp-content/uploads3/2013/12/photo-2-50x50.jpg LLEXRRXXXXEEEEEEEEEEEEEEEEE http://intothegloss.com/2013
2014-05-07T01:10:27.659Z   200       20536 http://decorating-time.com/2013/12/11/very-light-grey-backsplash/ LLEXRRXXXXREEEEEEEEEEEEEEEEEEEEEEEEEEEEEE http://decorati
```

Enter Web Address: http:// All Take Me Back Adv. Search

This is the new Wayback Machine prototype. Any URL in ARC files accessible to this sevice can be searched above.

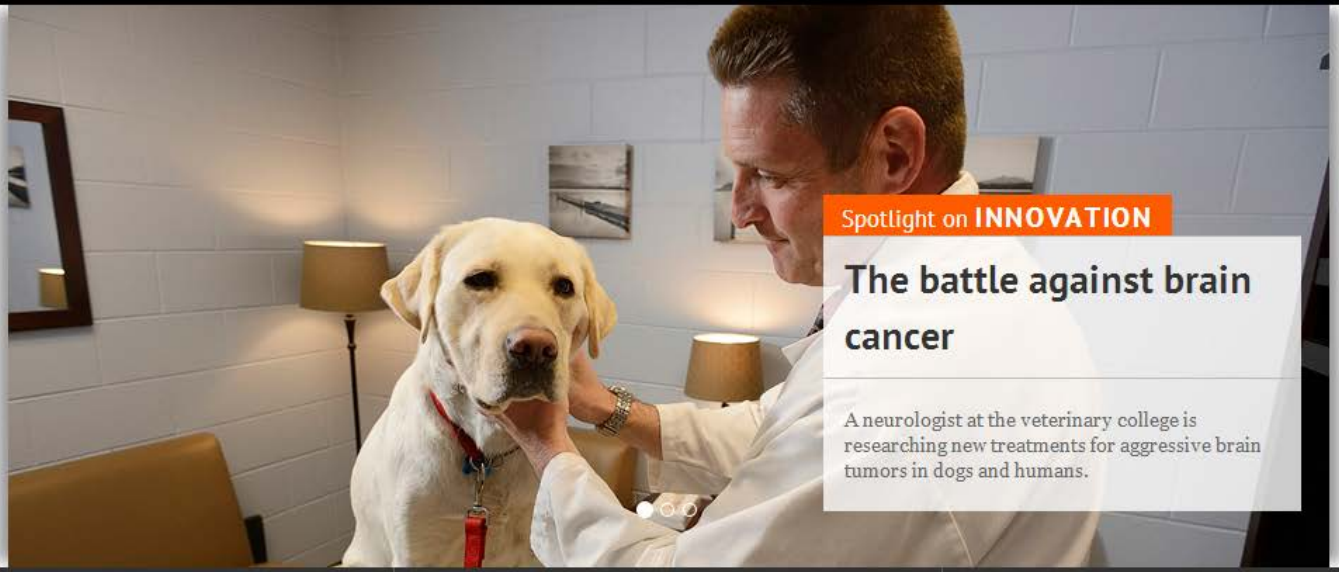
Home | Help

Search Results for Jan 1, 1996 - Dec 31, 2014

Jan 1996 - Dec 1997	Jan 1998 - Dec 1999	Jan 2000 - Dec 2001	Jan 2002 - Dec 2003	Jan 2004 - Dec 2005	Jan 2006 - Dec 2007	Jan 2008 - Dec 2009	Jan 2010 - Dec 2011	Jan 2012 - Dec 2013	Jan 2014 - Dec 2015
0 pages	0 pages	0 pages	0 pages	0 pages	0 pages	0 pages	0 pages	0 pages	17 pages
									Apr 23, 2014 *
									Apr 23, 2014 *
									Apr 23, 2014 *
									Apr 28, 2014 *
									Apr 28, 2014 *
									Apr 28, 2014 *
									Apr 28, 2014 *
									Apr 28, 2014 *
									Apr 28, 2014 *
									May 5, 2014 *
									May 5, 2014 *
									May 5, 2014 *
									May 5, 2014 *
									May 5, 2014 *
									May 5, 2014 *
									May 6, 2014
									May 6, 2014 *

VirginiaTech
Invent the Future®

Search pages and people



Questions?