Assignment 2 : CS 751

HARISH

March 6, 2015

0.1 Part 1

Choose 100 URIs from A1 Generate WARC files of those URIs using:wget,WARCreate,Heritrix (stand-alone or via WAIL),webrecorder.io

WARC using webrecorder.io I have chosen 100 URIs from Assignment1. In order to generate WARC files in webrecorder.io , we use "https://webrecorder.io/" and record the data using the options provided in the website . The WARC for all the 100 URIs are generated . It can be downloaded in zip format . The file 100interestingurlss.txt contains the 100 urls.

WARC using WARCreate The WARC files using WARCreate can be achieved via chrome plugin 'WARCreate' . The WARC files are downloaded in zipped format (.gz).

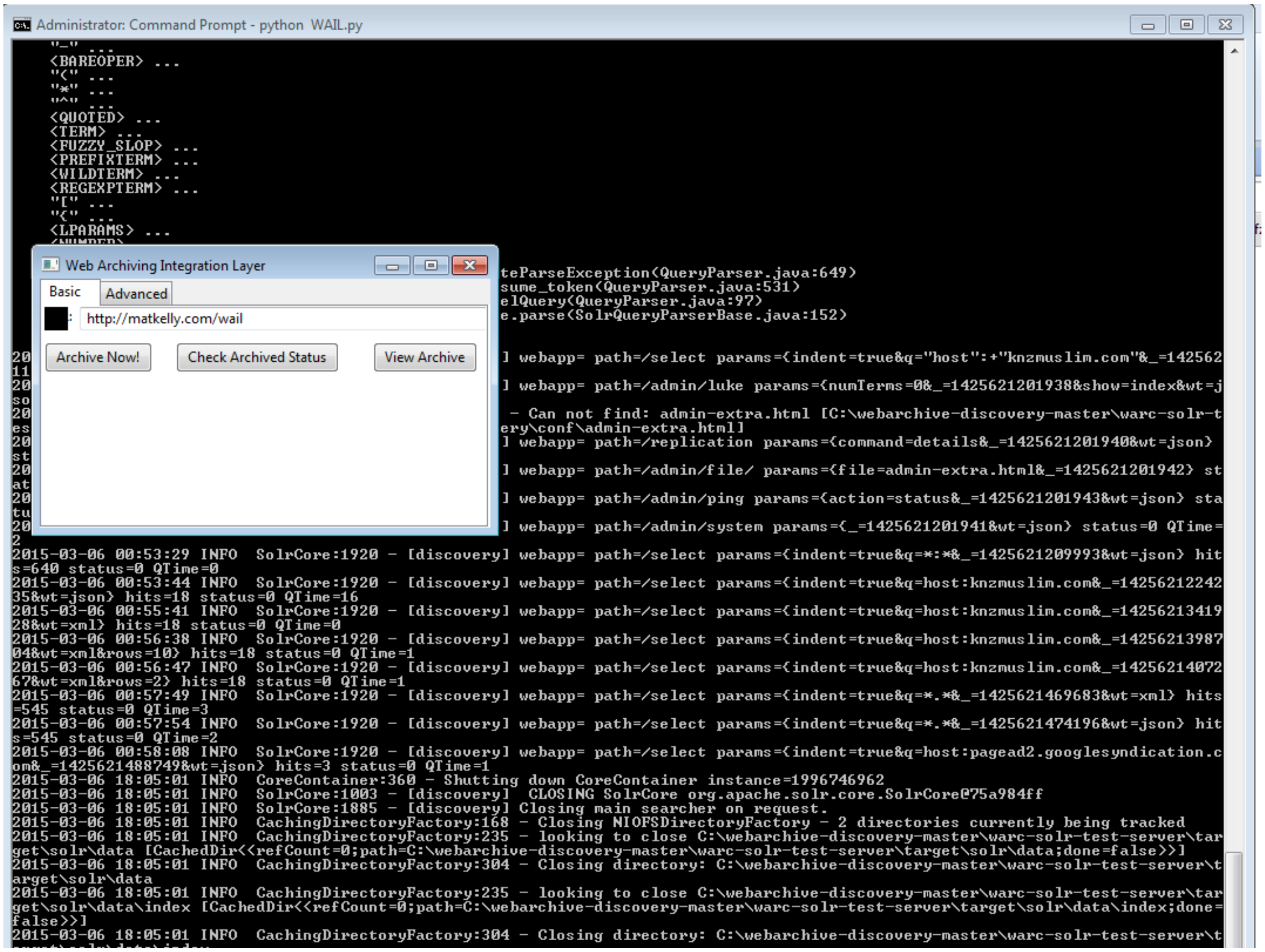
WARC using wget I have written a shell script using wget command to download warc files for all the URIs .

WARC using WAIL I have used following steps to set up WAIL in the system.
1. Install application from webpage, download: Windows 7+, version 0.2013.2.19
2. Unzip
3. Install wx python module: http://www.wxpython.org/download.php (I installed the last one 64-bit, python 2.7)
4. Moved "WAIL..." to C: and rename "WAIL" (remove extra characters)
5. Moved every file folder to the path C:WAIL (That is make direct path to folder, no intermediate folders)
6. Modified BDBCollection.xml as mentioned in google groups.
7. In WAIL.py changed catalina-start.bat .

I followed the below step for adding multiple URIs to WAIL , to crawl all links at once

WAIL : Advanced : Setup One-Off Crawl : Click first line to add URI0 : Click first line to add URI1 ... : Clicked Launch Crawl

WAIL Instance


Administrator: Command Prompt - python WAIL.py
Web Archiving Integration Layer
Basic
Advanced
Archive Now!
Check Archived Status
View Archive

[redacted]: http://matkelly.com/wail

```
    "_" ...
    <BAREOPER> ...
    "(" ...
    "*" ...
    "^" ...
    <QUOTED> ...
    <TERM> ...
    <FUZZY_SLOP> ...
    <PREFIXTERM> ...
    <WILDTERM> ...
    <REGEXPTERM> ...
    "[" ...
    "{" ...
    <LPARAMS> ...
    <NUMBER>
                                                   teParseException(QueryParser.java:649)
                                                   sume_token(QueryParser.java:531)
                                                   elQuery(QueryParser.java:97)
                                                   e.parse(SolrQueryParserBase.java:152)

20                                                 ] webapp= path=/select params={indent=true&q="host":+"knzmuslim.com"&_=142562
11
20                                                 ] webapp= path=/admin/luke params={numTerms=0&_=1425621201938&show=index&wt=j
so
20                                                  - Can not find: admin-extra.html [C:\webarchive-discovery-master\warc-solr-t
es                                                 ery\conf\admin-extra.html]
20                                                 ] webapp= path=/replication params={command=details&_=1425621201940&wt=json}
st
20                                                 ] webapp= path=/admin/file/ params={file=admin-extra.html&_=1425621201942} st
at
20                                                 ] webapp= path=/admin/ping params={action=status&_=1425621201943&wt=json} sta
tu
20                                                 ] webapp= path=/admin/system params={_=1425621201941&wt=json} status=0 QTime=
2
2015-03-06 00:53:29 INFO  SolrCore:1920 - [discovery] webapp= path=/select params={indent=true&q=*:*&_=1425621209993&wt=json} hit
s=640 status=0 QTime=0
2015-03-06 00:53:44 INFO  SolrCore:1920 - [discovery] webapp= path=/select params={indent=true&q=host:knzmuslim.com&_=14256212242
35&wt=json} hits=18 status=0 QTime=16
2015-03-06 00:55:41 INFO  SolrCore:1920 - [discovery] webapp= path=/select params={indent=true&q=host:knzmuslim.com&_=14256213419
28&wt=xml} hits=18 status=0 QTime=0
2015-03-06 00:56:38 INFO  SolrCore:1920 - [discovery] webapp= path=/select params={indent=true&q=host:knzmuslim.com&_=14256213987
04&wt=xml&rows=10} hits=18 status=0 QTime=1
2015-03-06 00:56:47 INFO  SolrCore:1920 - [discovery] webapp= path=/select params={indent=true&q=host:knzmuslim.com&_=14256214072
67&wt=xml&rows=2} hits=18 status=0 QTime=1
2015-03-06 00:57:49 INFO  SolrCore:1920 - [discovery] webapp= path=/select params={indent=true&q=*.*&_=1425621469683&wt=xml} hits
=545 status=0 QTime=3
2015-03-06 00:57:54 INFO  SolrCore:1920 - [discovery] webapp= path=/select params={indent=true&q=*.*&_=1425621474196&wt=json} hit
s=545 status=0 QTime=2
2015-03-06 00:58:08 INFO  SolrCore:1920 - [discovery] webapp= path=/select params={indent=true&q=host:pagead2.googlesyndication.c
om&_=1425621488749&wt=json} hits=3 status=0 QTime=1
2015-03-06 18:05:01 INFO  CoreContainer:360 - Shutting down CoreContainer instance=1996746962
2015-03-06 18:05:01 INFO  SolrCore:1003 - [discovery]  CLOSING SolrCore org.apache.solr.core.SolrCore@75a984ff
2015-03-06 18:05:01 INFO  SolrCore:1885 - [discovery] Closing main searcher on request.
2015-03-06 18:05:01 INFO  CachingDirectoryFactory:168 - Closing NIOFSDirectoryFactory - 2 directories currently being tracked
2015-03-06 18:05:01 INFO  CachingDirectoryFactory:235 - looking to close C:\webarchive-discovery-master\warc-solr-test-server\tar
get\solr\data [CachedDir<<refCount=0;path=C:\webarchive-discovery-master\warc-solr-test-server\target\solr\data;done=false>>]
2015-03-06 18:05:01 INFO  CachingDirectoryFactory:304 - Closing directory: C:\webarchive-discovery-master\warc-solr-test-server\t
arget\solr\data
2015-03-06 18:05:01 INFO  CachingDirectoryFactory:235 - looking to close C:\webarchive-discovery-master\warc-solr-test-server\tar
get\solr\data\index [CachedDir<<refCount=0;path=C:\webarchive-discovery-master\warc-solr-test-server\target\solr\data\index;done=
false>>]
2015-03-06 18:05:01 INFO  CachingDirectoryFactory:304 - Closing directory: C:\webarchive-discovery-master\warc-solr-test-server\t
```

Figure 1: Web Archiving Integration Layer

Describe the resulting WARC files: quantitatively compare and contrast the results of the WARC files of the same URI as generated by different tools In order to quantitatively compare and contrast the results of the WARC files , I have taken 2 URIs and compared their warc file size generated by 4 different tools.
I have created a bar plot which contains the size differences of the warc files for same URIs. I have also pasted the Rscript(sizeComparision.R located in the same folder) which i have used to plot this graph .

File Size Comparison

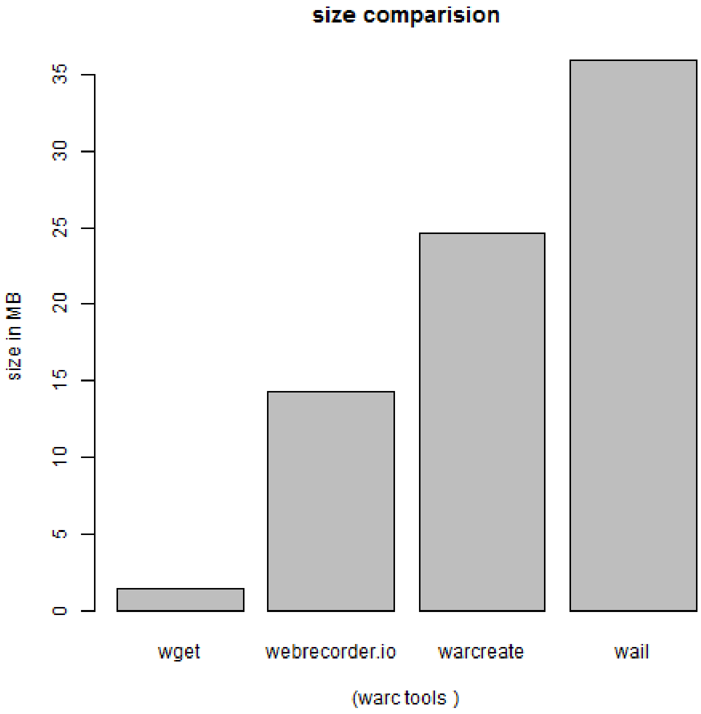


Figure 2: warc size comparison 1

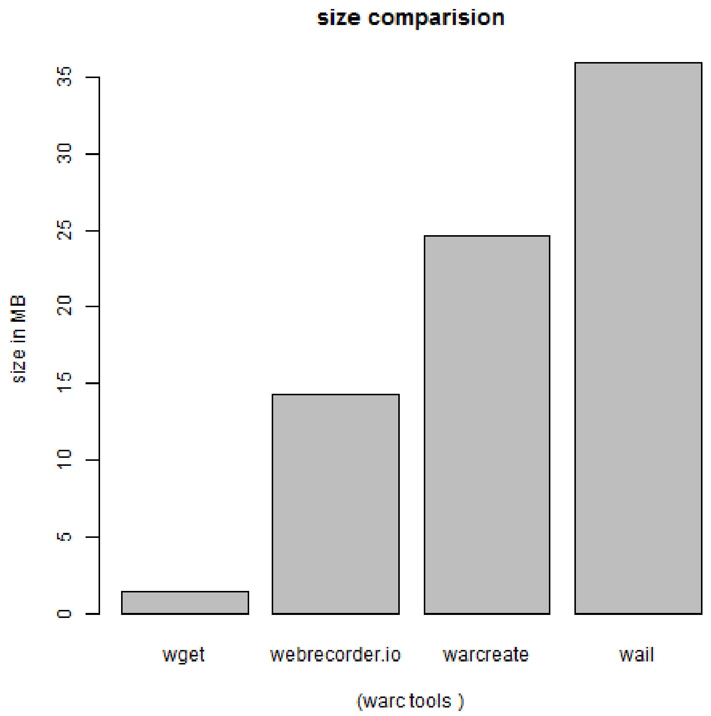


Figure 3: Histogram 2

The size of same URI warc file depends on the tool we use to generate it . WAIL warc size is comparatively long when compared to other warc file tools because wail crawls for the links present in the URI and gets the data of those links as well. wget warc file size is comparatively smaller than other warc files size.

Demonstrate playback of 2-3 WARCs in the (Wayback Machine (via WAIL or stand-alone) or pywb) and (webrecorder.io) I have played the same WARC file using pywb and webrecorder.io.

I have used this link to setup pywb "https://github.com/ikreymer/pywb ". Pywb takes the warc and uses the following command to generate cdx file.

"cdx-indexer –sort myarchive/cdx myarchive/warcs"

After setting up pywb . Run the wayback server using wayback command .We can search for the URIs using "http://localhost:8080/pywb"

I have included playback screen shots generated using pwyb(Figure 4,5) .Also ,I have included the playback screen shots generated using webrecorder.io (Figure 6 , 7 ,8)

0.2 Part 2

Ingest the 100 URIs from their resulting WARC files into a SOLR instance see the code + tutorial at: https://github.com/ukwa/webarchive-discovery Demonstrate several functioning queries on the files (a full front-end is not required) describe the configuration choices you made in setting up SOLR and processing the documents

I have set up SOLR instance using the link "https://github.com/ukwa/webarchive-discovery "

The pre-requisites for this are Maven 3 and Oracle java 7 .

Following are the commands used to run SOLR instance.
$cdwarc - solr - test - server$
mvn jetty:run-exploded

This will fire up a suitable Solr instance, with a UI on port 8080 .
Indexing a WARC file

Downloaded warc-indexer from "https://oss.sonatype.org/content/repositories/snapshots/uk/b
indexer/2.0.1-SNAPSHOT/"

Copied the downloaded folder to webdiscovery.

Merged all the 100 WARC files using the WARC merger posted in google groups.

Indexed the warc file using the command

–java -jar warc-indexer-2.0.1-SNAPSHOT-jar-with-dependencies.jar -s http://localhost:8080/dis
-t warc-indexer/src/test/resources/wikipedia-mona-lisa/flashfrozen-jwat-recompressed.warc.gz

Querying SOLR The queries i have used are

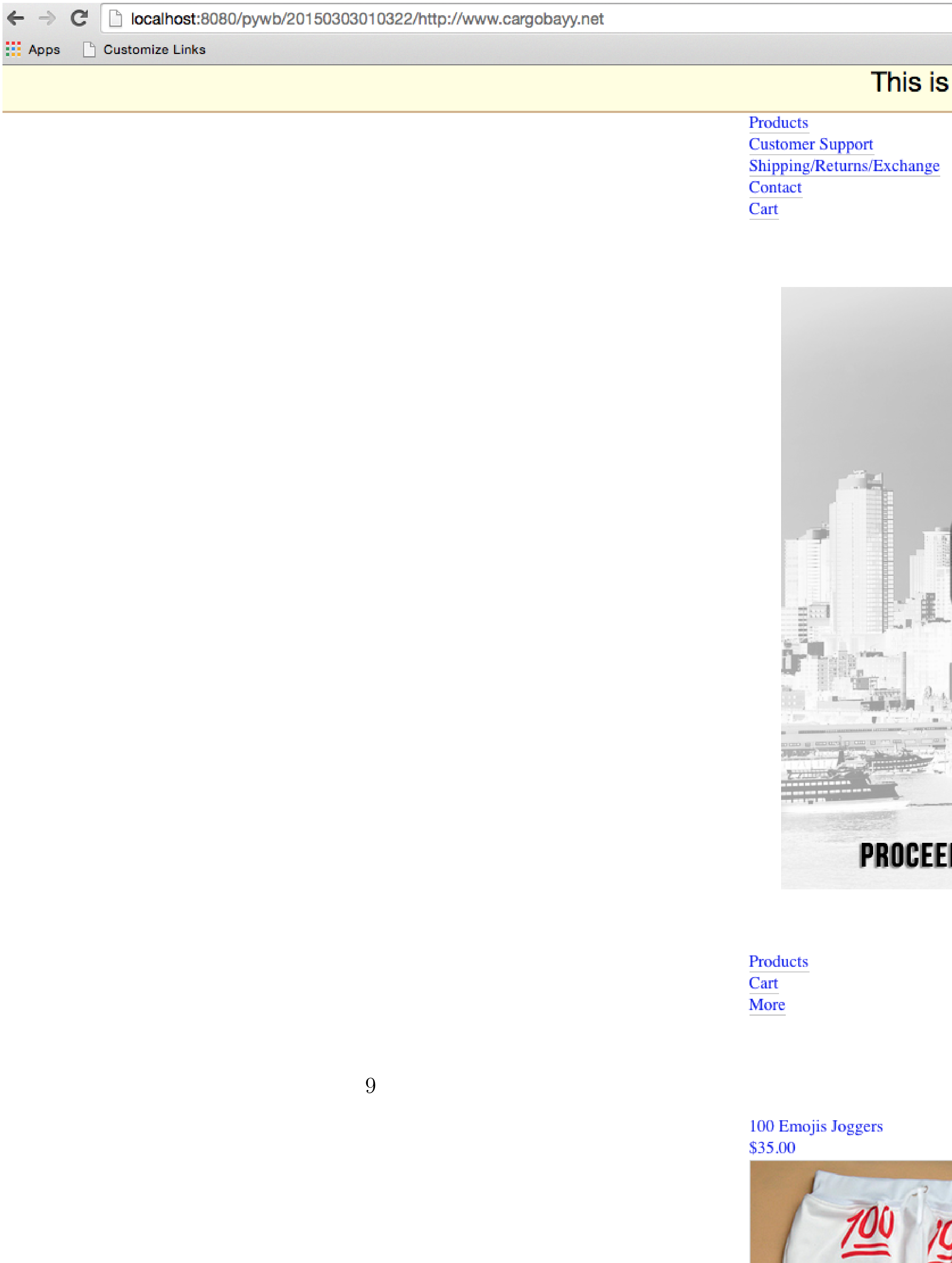
1 To search the host name using xml content , change host name data changes automatically (shown in the figure 9 ,17

2 To get json data , set up wt to json (shown in figure 10)

3 To get xml data , set up wt to xml (shown in figure 16)

4 To retrieve only two rows from the query , set up row count to 2 (shown in figure 16)

5 To search based on host name in json script(shown in figure 10)



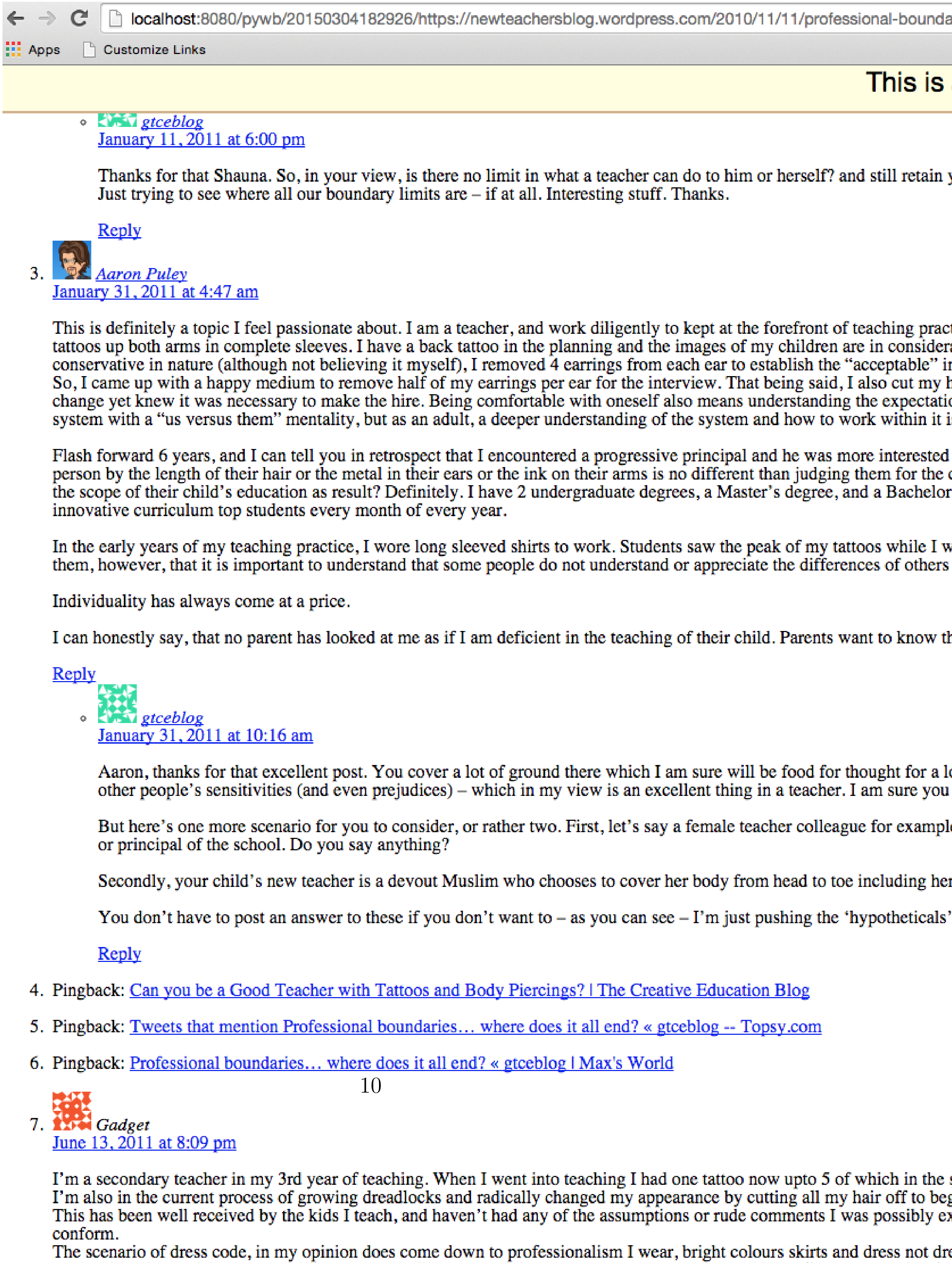
localhost:8080/pywb/20150303010322/http://www.cargobayy.net

This is

Products
Customer Support
Shipping/Returns/Exchange
Contact
Cart

Products
Cart
More

100 Emojis Joggers
$35.00



localhost:8080/pywb/20150304182926/https://newteachersblog.wordpress.com/2010/11/11/professional-bounda
Apps
Customize Links
This is

- *gtceblog*
  January 11, 2011 at 6:00 pm

  Thanks for that Shauna. So, in your view, is there no limit in what a teacher can do to him or herself? and still retain y
  Just trying to see where all our boundary limits are – if at all. Interesting stuff. Thanks.

  Reply

3. *Aaron Puley*
   January 31, 2011 at 4:47 am

   This is definitely a topic I feel passionate about. I am a teacher, and work diligently to kept at the forefront of teaching pract
   tattoos up both arms in complete sleeves. I have a back tattoo in the planning and the images of my children are in considera
   conservative in nature (although not believing it myself), I removed 4 earrings from each ear to establish the "acceptable" in
   So, I came up with a happy medium to remove half of my earrings per ear for the interview. That being said, I also cut my h
   change yet knew it was necessary to make the hire. Being comfortable with oneself also means understanding the expectatio
   system with a "us versus them" mentality, but as an adult, a deeper understanding of the system and how to work within it i

   Flash forward 6 years, and I can tell you in retrospect that I encountered a progressive principal and he was more interested
   person by the length of their hair or the metal in their ears or the ink on their arms is no different than judging them for the c
   the scope of their child's education as result? Definitely. I have 2 undergraduate degrees, a Master's degree, and a Bachelor
   innovative curriculum top students every month of every year.

   In the early years of my teaching practice, I wore long sleeved shirts to work. Students saw the peak of my tattoos while I w
   them, however, that it is important to understand that some people do not understand or appreciate the differences of others

   Individuality has always come at a price.

   I can honestly say, that no parent has looked at me as if I am deficient in the teaching of their child. Parents want to know th

   Reply

   - *gtceblog*
     January 31, 2011 at 10:16 am

     Aaron, thanks for that excellent post. You cover a lot of ground there which I am sure will be food for thought for a lo
     other people's sensitivities (and even prejudices) – which in my view is an excellent thing in a teacher. I am sure you

     But here's one more scenario for you to consider, or rather two. First, let's say a female teacher colleague for exampl
     or principal of the school. Do you say anything?

     Secondly, your child's new teacher is a devout Muslim who chooses to cover her body from head to toe including her

     You don't have to post an answer to these if you don't want to – as you can see – I'm just pushing the 'hypotheticals'

     Reply

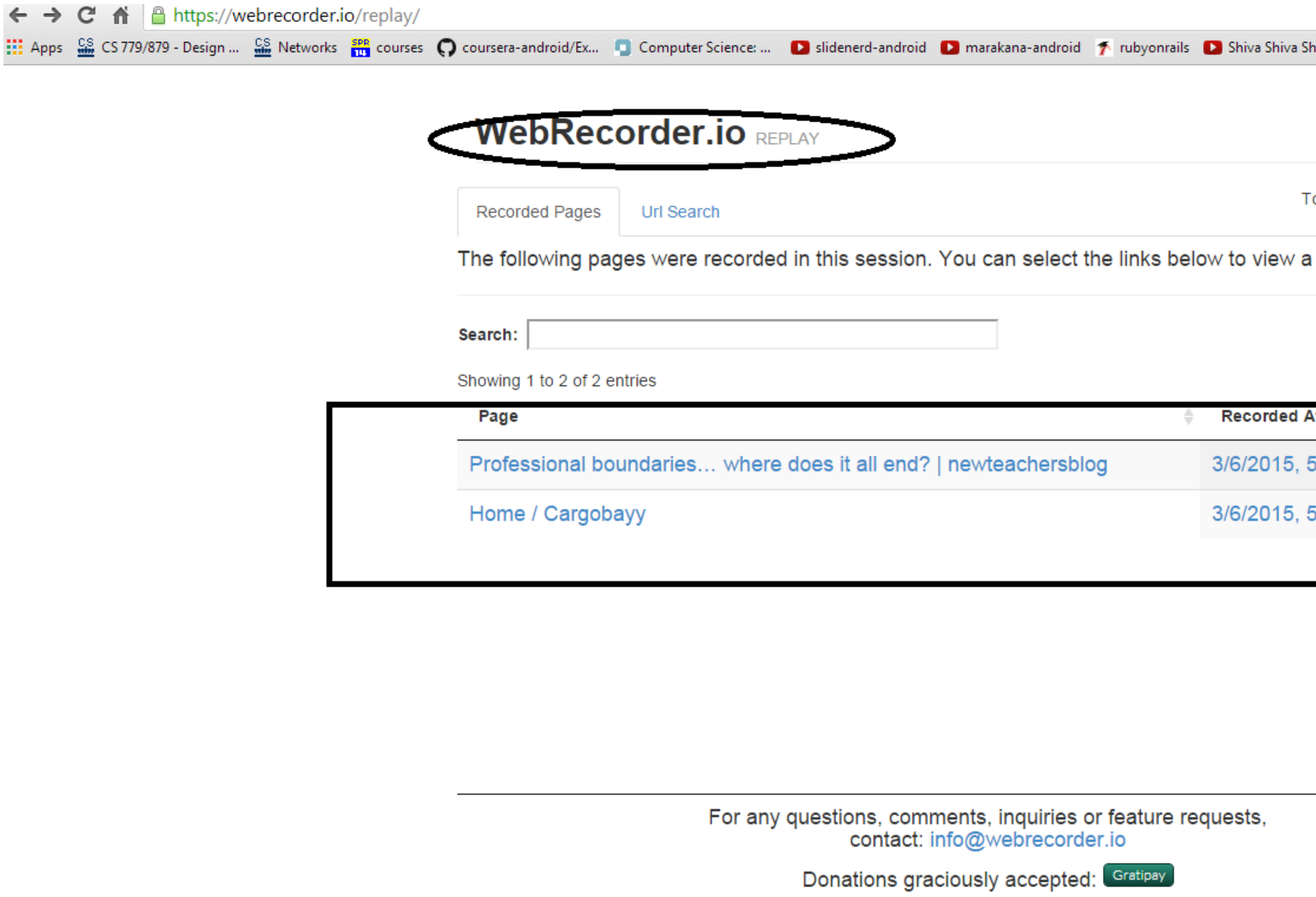
4. Pingback: Can you be a Good Teacher with Tattoos and Body Piercings? | The Creative Education Blog

5. Pingback: Tweets that mention Professional boundaries… where does it all end? « gtceblog -- Topsy.com

6. Pingback: Professional boundaries… where does it all end? « gtceblog | Max's World

7. *Gadget*
   June 13, 2011 at 8:09 pm

   I'm a secondary teacher in my 3rd year of teaching. When I went into teaching I had one tattoo now upto 5 of which in the s
   I'm also in the current process of growing dreadlocks and radically changed my appearance by cutting all my hair off to beg
   This has been well received by the kids I teach, and haven't had any of the assumptions or rude comments I was possibly ex
   conform.
   The scenario of dress code, in my opinion does come down to professionalism I wear, bright colours skirts and dress not dre



The following pages were recorded in this session. You can select the links below to view a

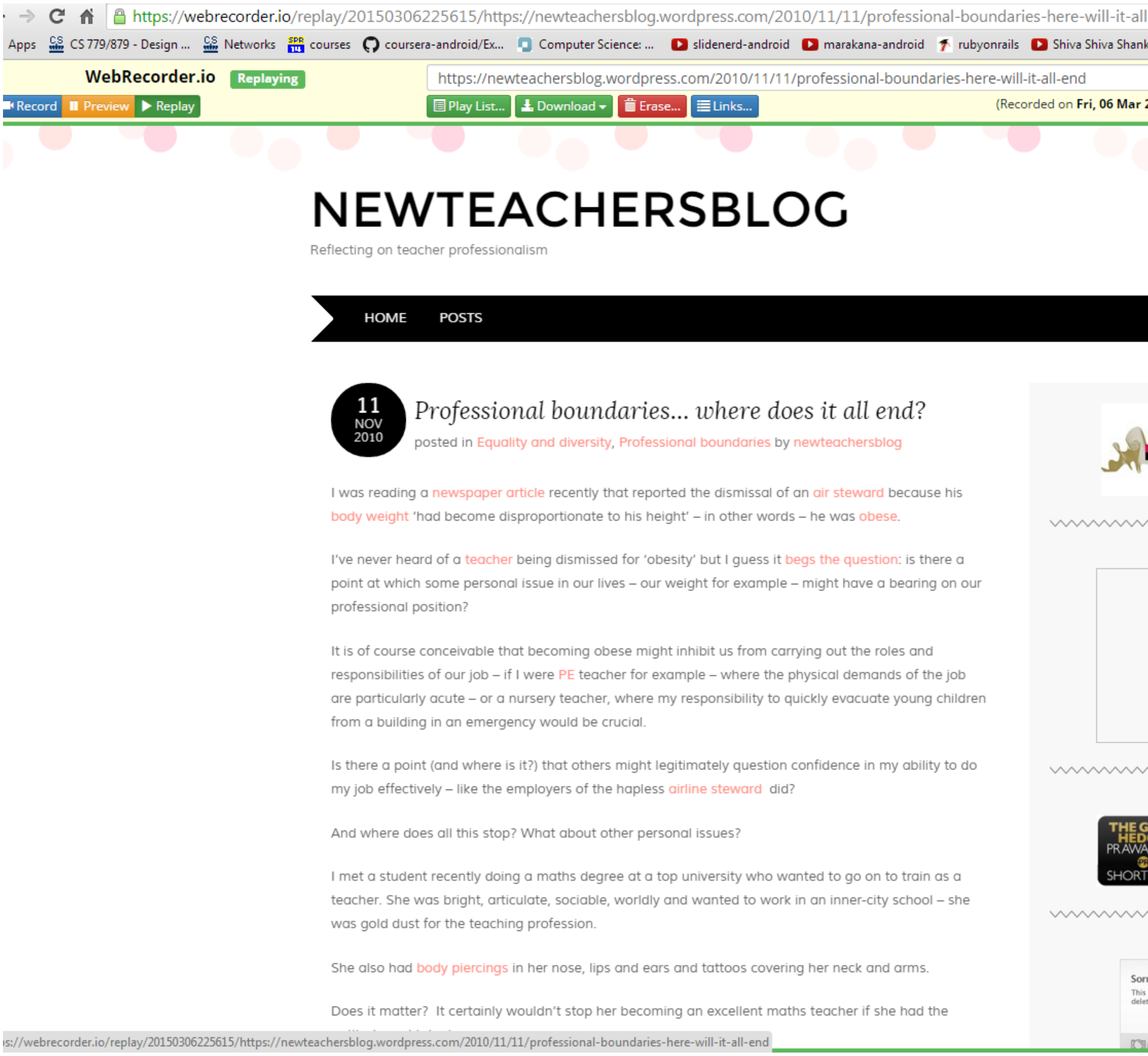
Showing 1 to 2 of 2 entries

Page	Recorded A
Professional boundaries… where does it all end? | newteachersblog	3/6/2015, 5
Home / Cargobayy	3/6/2015, 5

For any questions, comments, inquiries or feature requests,
contact: info@webrecorder.io

Donations graciously accepted:

Figure 6: webrecorder



https://webrecorder.io/replay/20150306225615/https://newteachersblog.wordpress.com/2010/11/11/professional-boundaries-here-will-it-all
Apps
CS 779/879 - Design ...
Networks
courses
coursera-android/Ex...
Computer Science: ...
slidenerd-android
marakana-android
rubyonrails
Shiva Shiva Shank
WebRecorder.io
Replaying
https://newteachersblog.wordpress.com/2010/11/11/professional-boundaries-here-will-it-all-end
Record
Preview
Replay
Play List...
Download
Erase...
Links...
(Recorded on Fri, 06 Mar 2
NEWTEACHERSBLOG
Reflecting on teacher professionalism
HOME
POSTS
11
NOV
2010

Professional boundaries... where does it all end?

posted in Equality and diversity, Professional boundaries by newteachersblog

I was reading a newspaper article recently that reported the dismissal of an air steward because his body weight 'had become disproportionate to his height' – in other words – he was obese.

I've never heard of a teacher being dismissed for 'obesity' but I guess it begs the question: is there a point at which some personal issue in our lives – our weight for example – might have a bearing on our professional position?

It is of course conceivable that becoming obese might inhibit us from carrying out the roles and responsibilities of our job – if I were PE teacher for example – where the physical demands of the job are particularly acute – or a nursery teacher, where my responsibility to quickly evacuate young children from a building in an emergency would be crucial.

Is there a point (and where is it?) that others might legitimately question confidence in my ability to do my job effectively – like the employers of the hapless airline steward did?

And where does all this stop? What about other personal issues?

I met a student recently doing a maths degree at a top university who wanted to go on to train as a teacher. She was bright, articulate, sociable, worldly and wanted to work in an inner-city school – she was gold dust for the teaching profession.

She also had body piercings in her nose, lips and ears and tattoos covering her neck and arms.

Does it matter? It certainly wouldn't stop her becoming an excellent maths teacher if she had the

THE G
HED
PRAWA
SHORT
Sor
This
dele
s://webrecorder.io/replay/20150306225615/https://newteachersblog.wordpress.com/2010/11/11/professional-boundaries-here-will-it-all-end

Figure 7: webrecorder plaback file

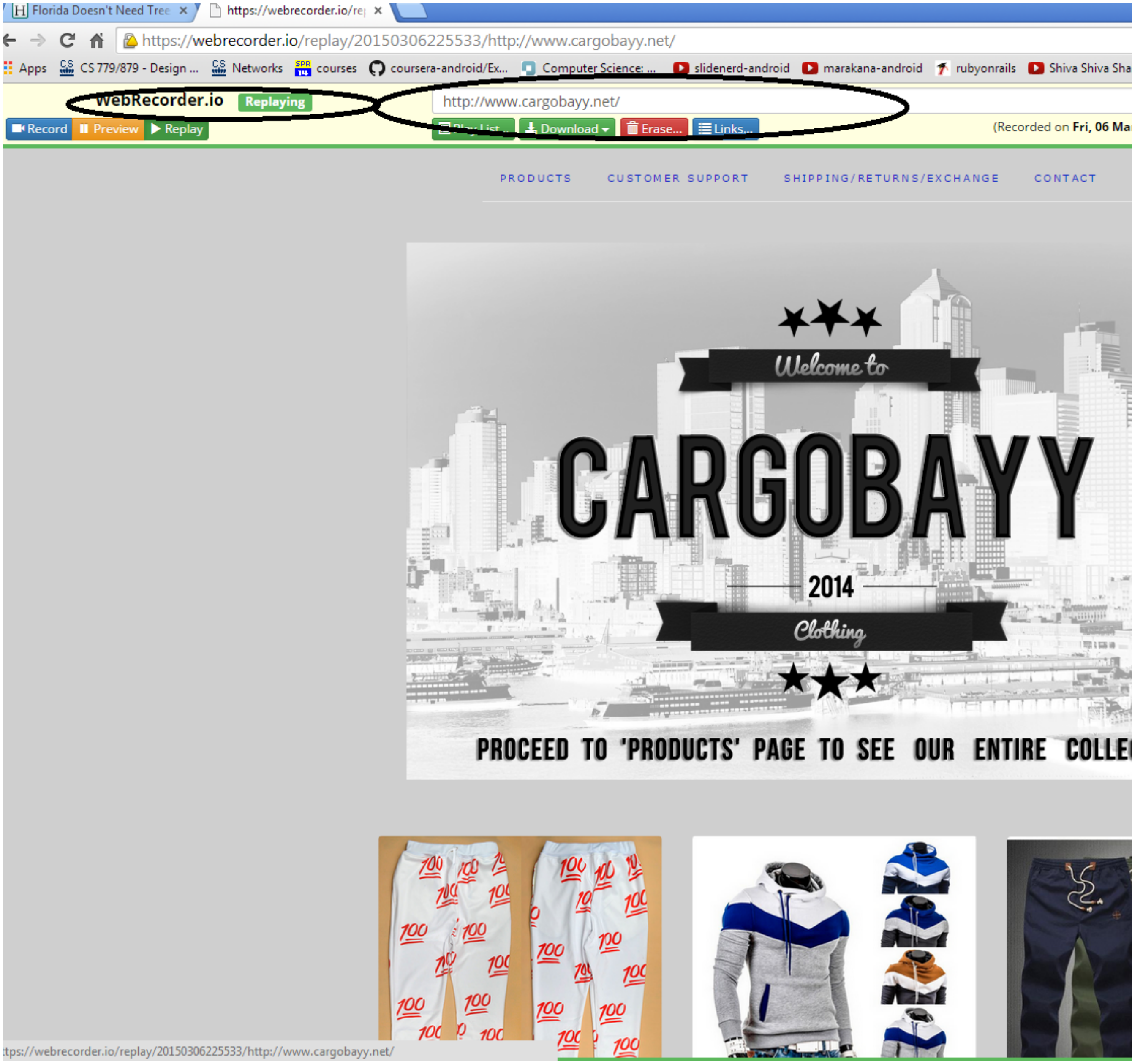
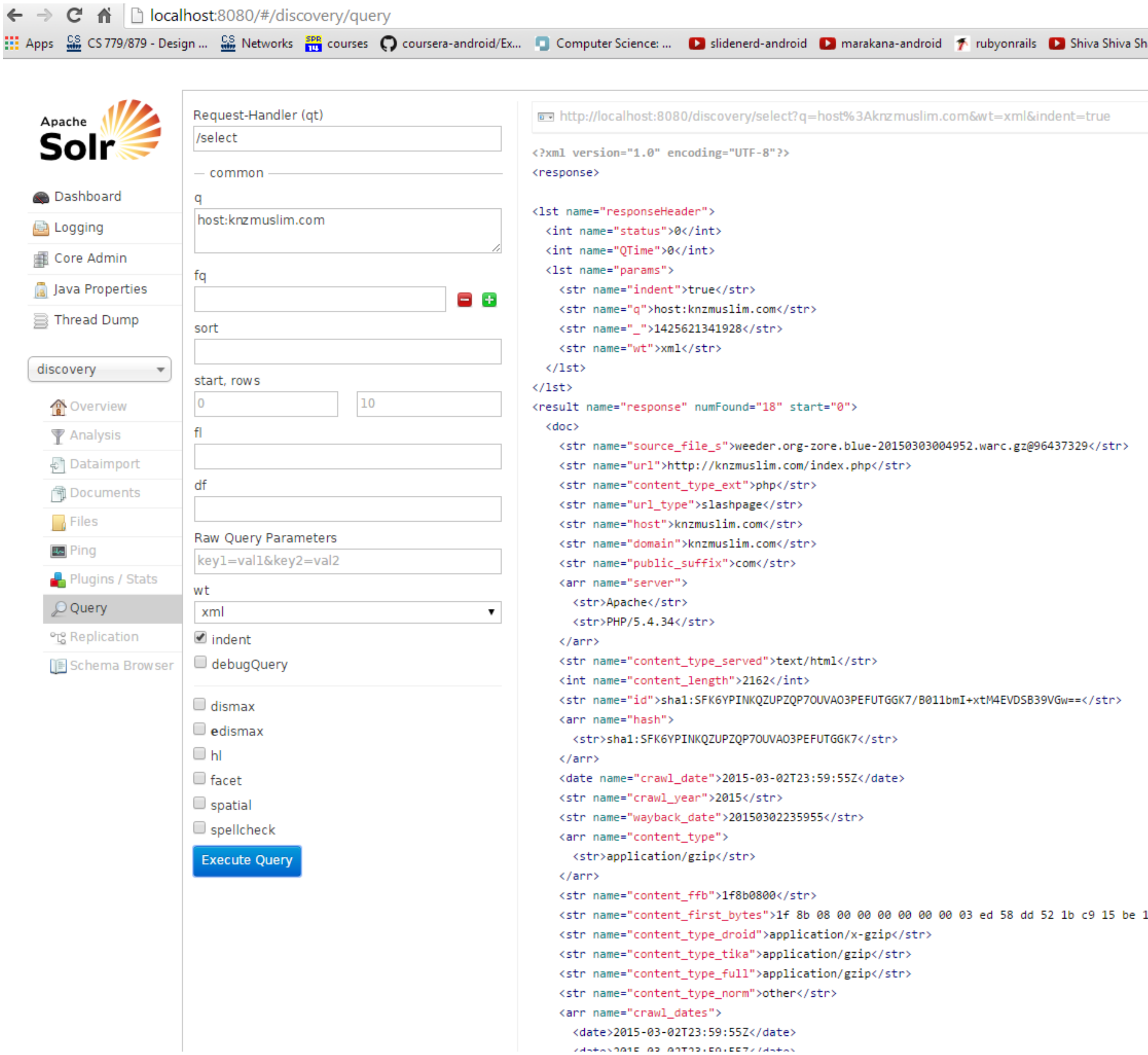


Figure 8: webrecorder playback file 2



localhost:8080/#/discovery/query

Request-Handler (qt): /select

common

q: host:knzmuslim.com

fq:

sort:

start, rows: 0 10

fl:

df:

Raw Query Parameters: key1=val1&key2=val2

wt: xml

☑ indent
□ debugQuery

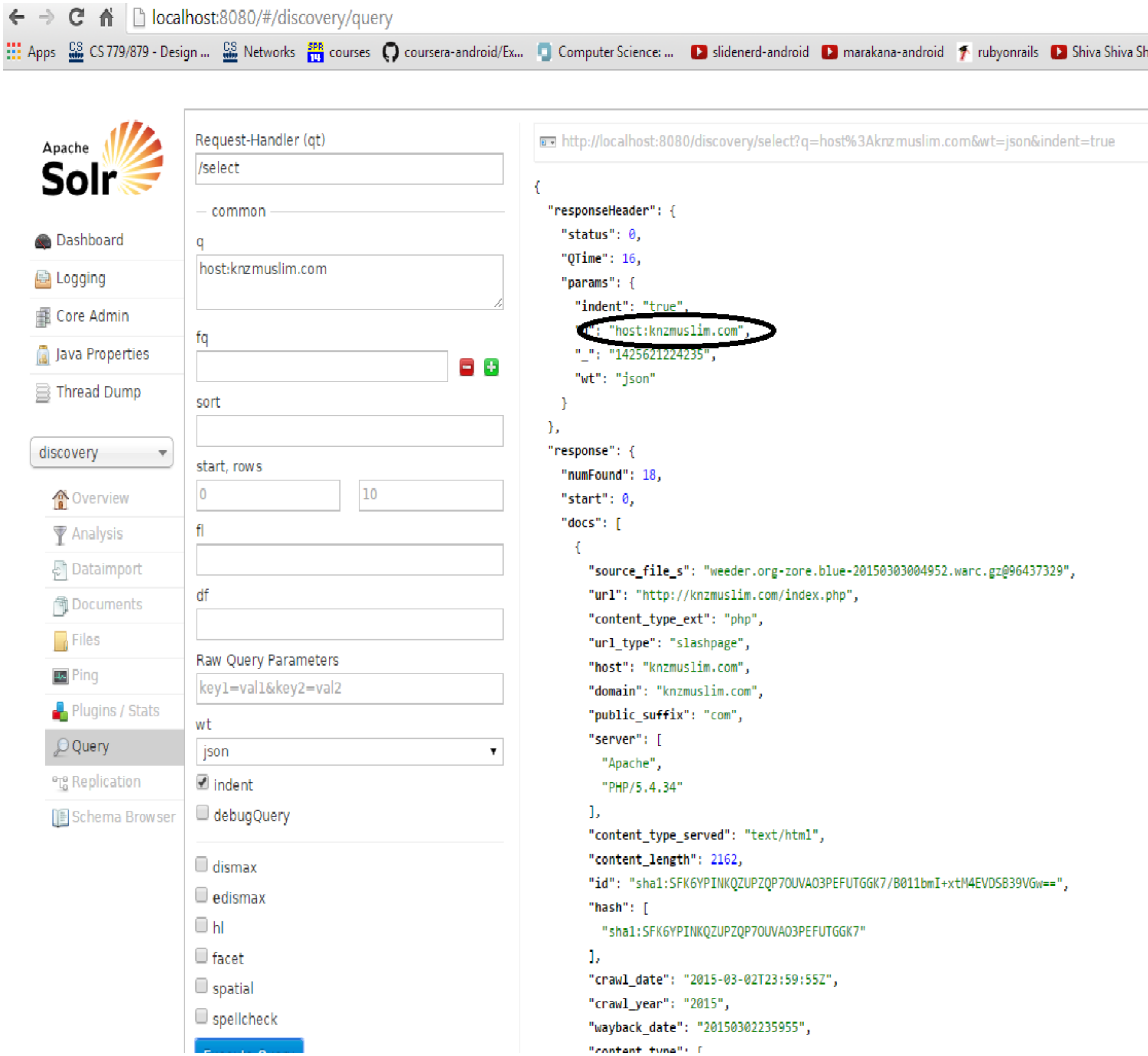
□ dismax
□ edismax
□ hl
□ facet
□ spatial
□ spellcheck

http://localhost:8080/discovery/select?q=host%3Aknzmuslim.com&wt=xml&indent=true

```
<?xml version="1.0" encoding="UTF-8"?>
<response>

<lst name="responseHeader">
  <int name="status">0</int>
  <int name="QTime">0</int>
  <lst name="params">
    <str name="indent">true</str>
    <str name="q">host:knzmuslim.com</str>
    <str name="_">1425621341928</str>
    <str name="wt">xml</str>
  </lst>
</lst>
<result name="response" numFound="18" start="0">
  <doc>
    <str name="source_file_s">weeder.org-zore.blue-20150303004952.warc.gz@96437329</str>
    <str name="url">http://knzmuslim.com/index.php</str>
    <str name="content_type_ext">php</str>
    <str name="url_type">slashpage</str>
    <str name="host">knzmuslim.com</str>
    <str name="domain">knzmuslim.com</str>
    <str name="public_suffix">com</str>
    <arr name="server">
      <str>Apache</str>
      <str>PHP/5.4.34</str>
    </arr>
    <str name="content_type_served">text/html</str>
    <int name="content_length">2162</int>
    <str name="id">sha1:SFK6YPINKQZUPZQP7OUVAO3PEFUTGGK7/B011bmI+xtM4EVDSB39VGw==</str>
    <arr name="hash">
      <str>sha1:SFK6YPINKQZUPZQP7OUVAO3PEFUTGGK7</str>
    </arr>
    <date name="crawl_date">2015-03-02T23:59:55Z</date>
    <str name="crawl_year">2015</str>
    <str name="wayback_date">20150302235955</str>
    <arr name="content_type">
      <str>application/gzip</str>
    </arr>
    <str name="content_ffb">1f8b0800</str>
    <str name="content_first_bytes">1f 8b 08 00 00 00 00 00 00 03 ed 58 dd 52 1b c9 15 be 1
    <str name="content_type_droid">application/x-gzip</str>
    <str name="content_type_tika">application/gzip</str>
    <str name="content_type_full">application/gzip</str>
    <str name="content_type_norm">other</str>
    <arr name="crawl_dates">
      <date>2015-03-02T23:59:55Z</date>
      <date>2015-03-02T23:59:55Z</date>
```

Figure 9: Querying solrInstance



Request-Handler (qt): /select

common

q: host:knzmuslim.com

fq:

sort:

start, rows: 0 10

fl:

df:

Raw Query Parameters: key1=val1&key2=val2

wt: json

☑ indent

□ debugQuery

□ dismax

□ edismax

□ hl

□ facet

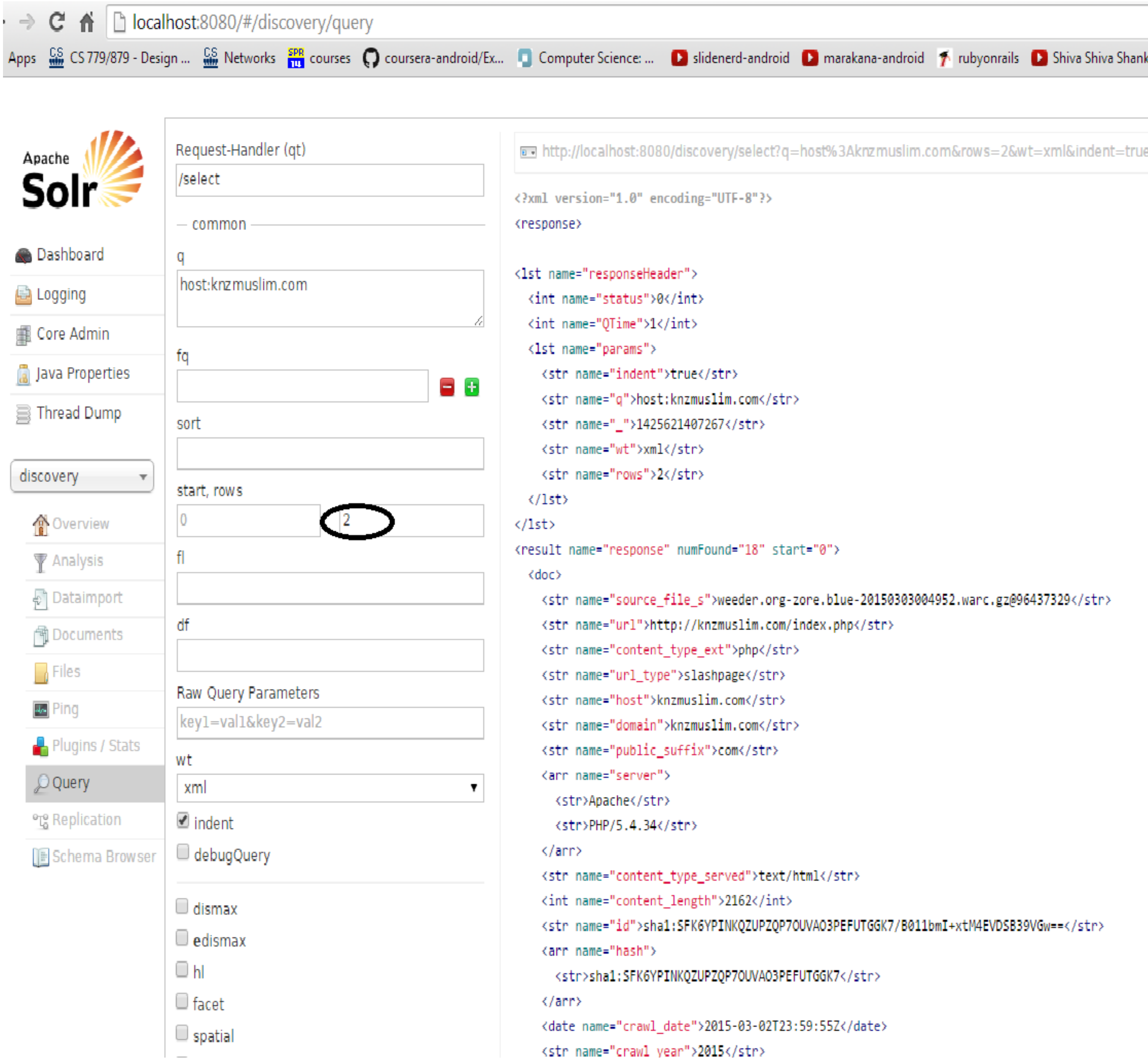
□ spatial

□ spellcheck

http://localhost:8080/discovery/select?q=host%3Aknzmuslim.com&wt=json&indent=true

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 16,
    "params": {
      "indent": "true",
      "q": "host:knzmuslim.com",
      "_": "1425621224235",
      "wt": "json"
    }
  },
  "response": {
    "numFound": 18,
    "start": 0,
    "docs": [
      {
        "source_file_s": "weeder.org-zore.blue-20150303004952.warc.gz@96437329",
        "url": "http://knzmuslim.com/index.php",
        "content_type_ext": "php",
        "url_type": "slashpage",
        "host": "knzmuslim.com",
        "domain": "knzmuslim.com",
        "public_suffix": "com",
        "server": [
          "Apache",
          "PHP/5.4.34"
        ],
        "content_type_served": "text/html",
        "content_length": 2162,
        "id": "sha1:SFK6YPINKQZUPZQP7OUVAO3PEFUTGGK7/B011bmI+xtM4EVDSB39VGw==",
        "hash": [
          "sha1:SFK6YPINKQZUPZQP7OUVAO3PEFUTGGK7"
        ],
        "crawl_date": "2015-03-02T23:59:55Z",
        "crawl_year": "2015",
        "wayback_date": "20150302235955",
        "content_type": [
```

Figure 10: Querying solrInstance



Request-Handler (qt): /select

common

q: host:knzmuslim.com

fq:

sort:

start, rows: 0 | 2

fl:

df:

Raw Query Parameters: key1=val1&key2=val2

wt: xml

☑ indent

□ debugQuery

□ dismax

□ edismax

□ hl

□ facet

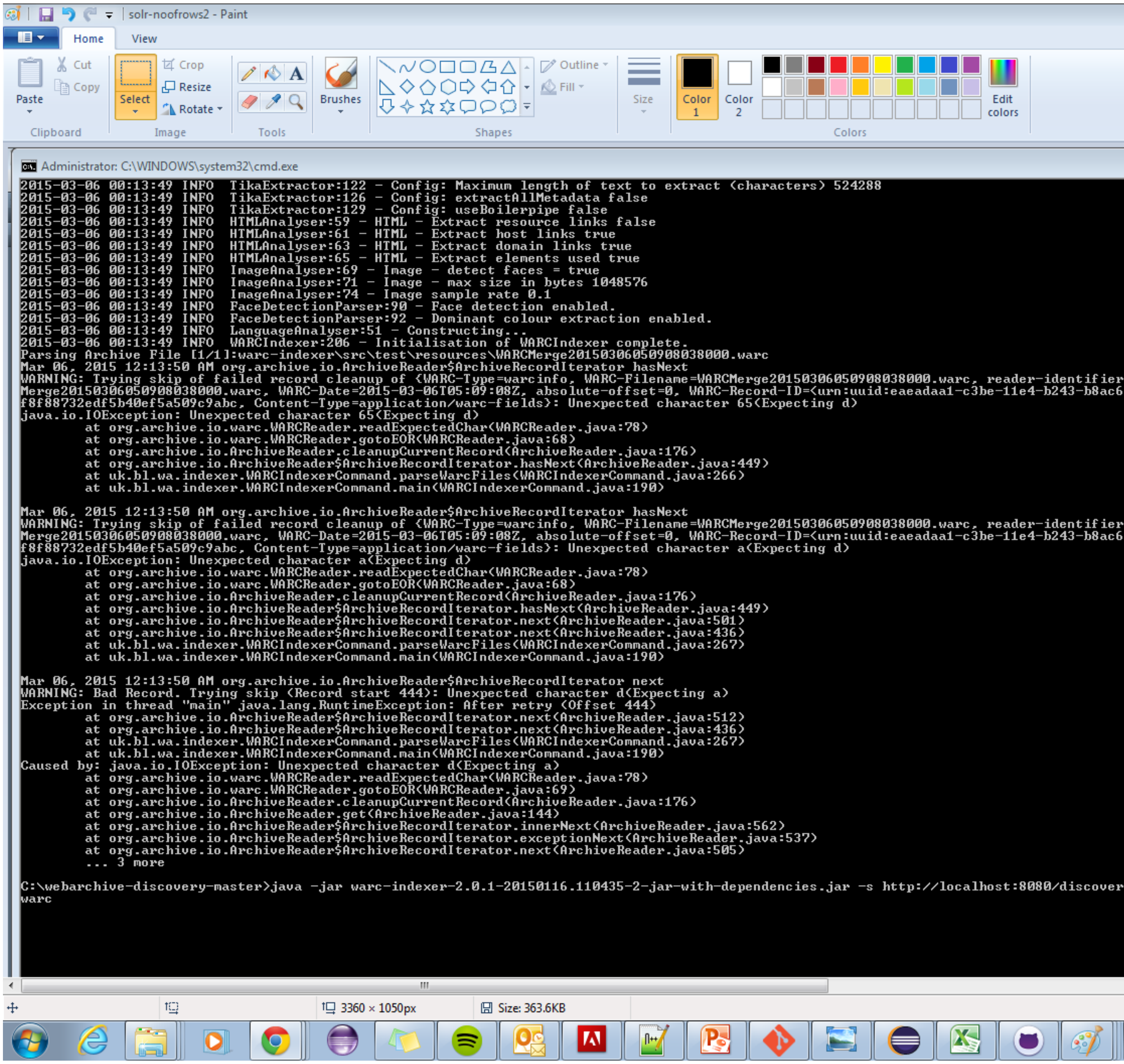
□ spatial

http://localhost:8080/discovery/select?q=host%3Aknzmuslim.com&rows=2&wt=xml&indent=true

```
<?xml version="1.0" encoding="UTF-8"?>
<response>

<lst name="responseHeader">
  <int name="status">0</int>
  <int name="QTime">1</int>
  <lst name="params">
    <str name="indent">true</str>
    <str name="q">host:knzmuslim.com</str>
    <str name="_">1425621407267</str>
    <str name="wt">xml</str>
    <str name="rows">2</str>
  </lst>
</lst>
<result name="response" numFound="18" start="0">
  <doc>
    <str name="source_file_s">weeder.org-zore.blue-20150303004952.warc.gz@96437329</str>
    <str name="url">http://knzmuslim.com/index.php</str>
    <str name="content_type_ext">php</str>
    <str name="url_type">slashpage</str>
    <str name="host">knzmuslim.com</str>
    <str name="domain">knzmuslim.com</str>
    <str name="public_suffix">com</str>
    <arr name="server">
      <str>Apache</str>
      <str>PHP/5.4.34</str>
    </arr>
    <str name="content_type_served">text/html</str>
    <int name="content_length">2162</int>
    <str name="id">sha1:SFK6YPINKQZUPZQP7OUVAO3PEFUTGGK7/B011bmI+xtM4EVDSB39VGw==</str>
    <arr name="hash">
      <str>sha1:SFK6YPINKQZUPZQP7OUVAO3PEFUTGGK7</str>
    </arr>
    <date name="crawl_date">2015-03-02T23:59:55Z</date>
    <str name="crawl_year">2015</str>
```

Figure 11: Querying solrInstance



```
2015-03-06 00:13:49 INFO  TikaExtractor:122 - Config: Maximum length of text to extract (characters) 524288
2015-03-06 00:13:49 INFO  TikaExtractor:126 - Config: extractAllMetadata false
2015-03-06 00:13:49 INFO  TikaExtractor:129 - Config: useBoilerpipe false
2015-03-06 00:13:49 INFO  HTMLAnalyser:59 - HTML - Extract resource links false
2015-03-06 00:13:49 INFO  HTMLAnalyser:61 - HTML - Extract host links true
2015-03-06 00:13:49 INFO  HTMLAnalyser:63 - HTML - Extract domain links true
2015-03-06 00:13:49 INFO  HTMLAnalyser:65 - HTML - Extract elements used true
2015-03-06 00:13:49 INFO  ImageAnalyser:69 - Image - detect faces = true
2015-03-06 00:13:49 INFO  ImageAnalyser:71 - Image - max size in bytes 1048576
2015-03-06 00:13:49 INFO  ImageAnalyser:74 - Image sample rate 0.1
2015-03-06 00:13:49 INFO  FaceDetectionParser:90 - Face detection enabled.
2015-03-06 00:13:49 INFO  FaceDetectionParser:92 - Dominant colour extraction enabled.
2015-03-06 00:13:49 INFO  LanguageAnalyser:51 - Constructing...
2015-03-06 00:13:49 INFO  WARCIndexer:206 - Initialisation of WARCIndexer complete.
Parsing Archive File [1/1]:warc-indexer\src\test\resources\WARCMerge20150306050908038000.warc
Mar 06, 2015 12:13:50 AM org.archive.io.ArchiveReader$ArchiveRecordIterator hasNext
WARNING: Trying skip of failed record cleanup of {WARC-Type=warcinfo, WARC-Filename=WARCMerge20150306050908038000.warc, reader-identifier
Merge20150306050908038000.warc, WARC-Date=2015-03-06T05:09:08Z, absolute-offset=0, WARC-Record-ID=<urn:uuid:eaeadaa1-c3be-11e4-b243-b8ac6
f8f88732edf5b40ef5a509c9abc, Content-Type=application/warc-fields}: Unexpected character 65(Expecting d)
java.io.IOException: Unexpected character 65(Expecting d)
        at org.archive.io.warc.WARCReader.readExpectedChar(WARCReader.java:78)
        at org.archive.io.warc.WARCReader.gotoEOR(WARCReader.java:68)
        at org.archive.io.ArchiveReader.cleanupCurrentRecord(ArchiveReader.java:176)
        at org.archive.io.ArchiveReader$ArchiveRecordIterator.hasNext(ArchiveReader.java:449)
        at uk.bl.wa.indexer.WARCIndexerCommand.parseWarcFiles(WARCIndexerCommand.java:266)
        at uk.bl.wa.indexer.WARCIndexerCommand.main(WARCIndexerCommand.java:190)

Mar 06, 2015 12:13:50 AM org.archive.io.ArchiveReader$ArchiveRecordIterator hasNext
WARNING: Trying skip of failed record cleanup of {WARC-Type=warcinfo, WARC-Filename=WARCMerge20150306050908038000.warc, reader-identifier
Merge20150306050908038000.warc, WARC-Date=2015-03-06T05:09:08Z, absolute-offset=0, WARC-Record-ID=<urn:uuid:eaeadaa1-c3be-11e4-b243-b8ac6
f8f88732edf5b40ef5a509c9abc, Content-Type=application/warc-fields}: Unexpected character a(Expecting d)
java.io.IOException: Unexpected character a(Expecting d)
        at org.archive.io.warc.WARCReader.readExpectedChar(WARCReader.java:78)
        at org.archive.io.warc.WARCReader.gotoEOR(WARCReader.java:68)
        at org.archive.io.ArchiveReader.cleanupCurrentRecord(ArchiveReader.java:176)
        at org.archive.io.ArchiveReader$ArchiveRecordIterator.hasNext(ArchiveReader.java:449)
        at org.archive.io.ArchiveReader$ArchiveRecordIterator.next(ArchiveReader.java:501)
        at org.archive.io.ArchiveReader$ArchiveRecordIterator.next(ArchiveReader.java:436)
        at uk.bl.wa.indexer.WARCIndexerCommand.parseWarcFiles(WARCIndexerCommand.java:267)
        at uk.bl.wa.indexer.WARCIndexerCommand.main(WARCIndexerCommand.java:190)

Mar 06, 2015 12:13:50 AM org.archive.io.ArchiveReader$ArchiveRecordIterator next
WARNING: Bad Record. Trying skip (Record start 444): Unexpected character d(Expecting a)
Exception in thread "main" java.lang.RuntimeException: After retry (Offset 444)
        at org.archive.io.ArchiveReader$ArchiveRecordIterator.next(ArchiveReader.java:512)
        at org.archive.io.ArchiveReader$ArchiveRecordIterator.next(ArchiveReader.java:436)
        at uk.bl.wa.indexer.WARCIndexerCommand.parseWarcFiles(WARCIndexerCommand.java:267)
        at uk.bl.wa.indexer.WARCIndexerCommand.main(WARCIndexerCommand.java:190)
Caused by: java.io.IOException: Unexpected character d(Expecting a)
        at org.archive.io.warc.WARCReader.readExpectedChar(WARCReader.java:78)
        at org.archive.io.warc.WARCReader.gotoEOR(WARCReader.java:69)
        at org.archive.io.ArchiveReader.cleanupCurrentRecord(ArchiveReader.java:176)
        at org.archive.io.ArchiveReader.get(ArchiveReader.java:144)
        at org.archive.io.ArchiveReader$ArchiveRecordIterator.innerNext(ArchiveReader.java:562)
        at org.archive.io.ArchiveReader$ArchiveRecordIterator.exceptionNext(ArchiveReader.java:537)
        at org.archive.io.ArchiveReader$ArchiveRecordIterator.next(ArchiveReader.java:505)
        ... 3 more

C:\webarchive-discovery-master>java -jar warc-indexer-2.0.1-20150116.110435-2-jar-with-dependencies.jar -s http://localhost:8080/discover
warc
```

Figure 12: Querying solrInstance