# CS 751: Assignment 2

Bharath Kongara

Spring 2015

# Contents

## 1 Question 1

Choose 100 URIs from Assignment1 and generate WARC files of those URIs using:

- wget
- WARCreate
- Heritrix (stand-alone or via WAIL)
- and webrecorder.io

Describe the resulting WARC files: quantitatively compare and contrast the results of the WARC files of the same URI as generated by different tools

- choose interesting examples

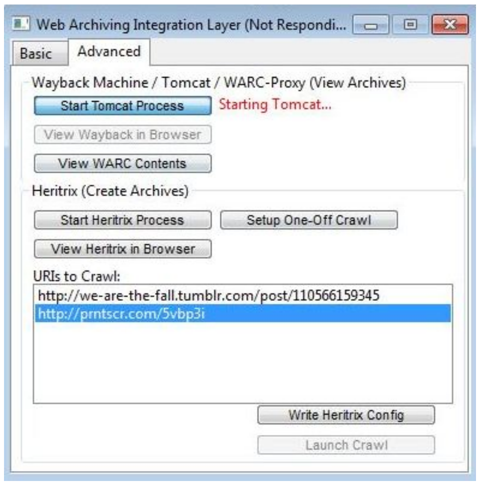
Demonstrate playback of 2-3 WARCs in the (Wayback Machine (via WAIL or stand-alone) or pywb) and (webrecorder.io)

- "`https://github.com/iipc/openwayback`"
- "`https://github.com/ikreymer/pywb`"

### 1.1 Solution

The following steps were taken to setup tools and generate WARC files:

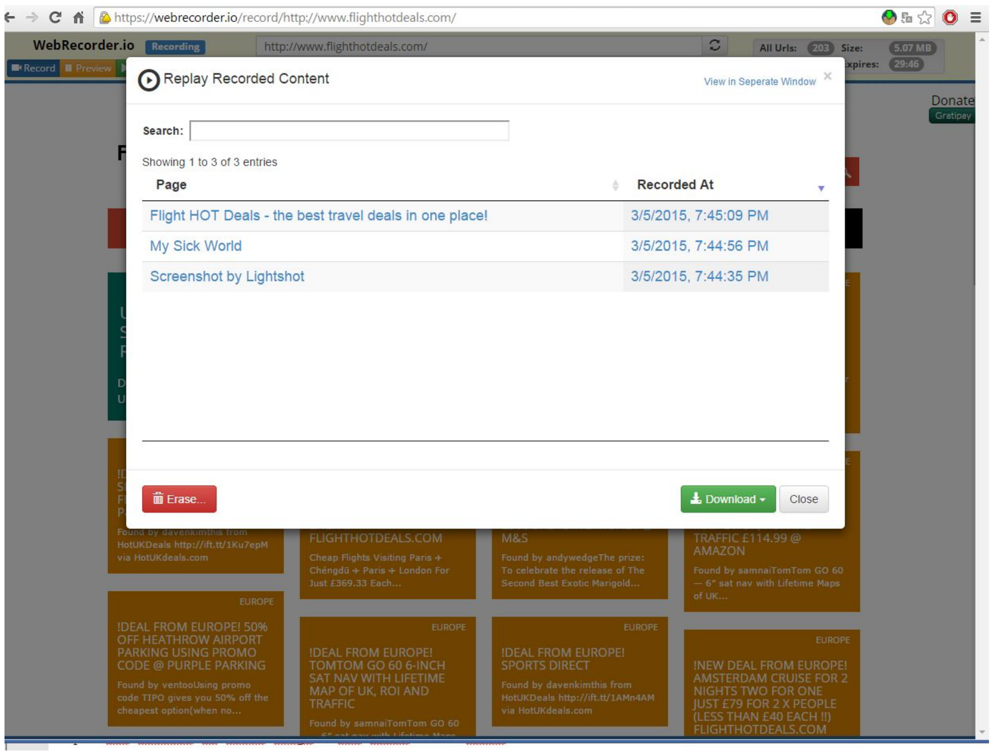
- Installed wget using 'brew install wget'.
- WARC file for each URI is generated using the command 'wget –warc-file=outputfilename URI'.
- Downloaded WARCreate chrome extension from chrome web store and added it to the chrome browser. Generated WARC for each URI manually by providing input to WARCreate.
- Installed WAIL.
- The figure below shows how to generate WARC files for multiple URIs using single heritrix instance.



```
http://we-are-the-fall.tumblr.com/post/110566159345
http://prntscr.com/5vbp3i
```

Figure 1: Generating WARC files for multiple URIs using WAIL

- Of the 100 URI selected, when any of the URIs had a parameter then WAIL wasn't able to generate WARC file for it.
- Used "`https://webrecorder.io/`" . Generating WARC files for multiple URIs is done as shown in the figure below.



Showing 1 to 3 of 3 entries

| Page | Recorded At |
|---|---|
| Flight HOT Deals - the best travel deals in one place! | 3/5/2015, 7:45:09 PM |
| My Sick World | 3/5/2015, 7:44:56 PM |
| Screenshot by Lightshot | 3/5/2015, 7:44:35 PM |

Figure 2: Generating WARC files for multiple URIs using webrecorder

- Quantitative comparison of WARC files generated by WAIL,webrecorder,WARCreater and wget is done on the basis of WARC file sizes.
- The comparison sizes are WAIL = 40 MB , webrecorder.io = 20.48 MB, WARCreate = 15.36 MB, wget = 4 MB.
- The WAIL size is more compared to others because WAIL software crawls data of the links in the website.
- This comparison is shown in the graph below.
- Installed pywb to playback WARC files.
- pywb requires cdx for each WARC to playback, so generated .cdx file for each WARC file.
- I have put all my WARC files into a new folder and changed the archive paths of config.xml to point to my WARC files.
- Play back for two WARC files using pywb is shown in the below figure.
- The WARC file which was in the play back does not have some images and some URLs.
- The WARC file which is in the play back does not contain certain CSS archives and some links
- WebRecorder.io takes WARC file to replay the archived file. Play back for two WARC files using WebRecorder.io is shown in the below figure
- The WARC file in this play back worked perfectly without any faults.
- The WARC file in this play back did not have certain images.

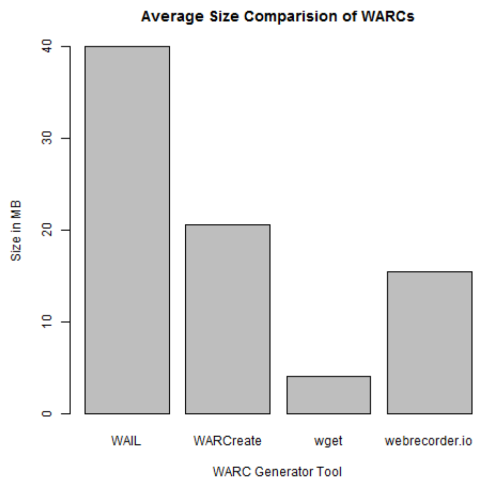
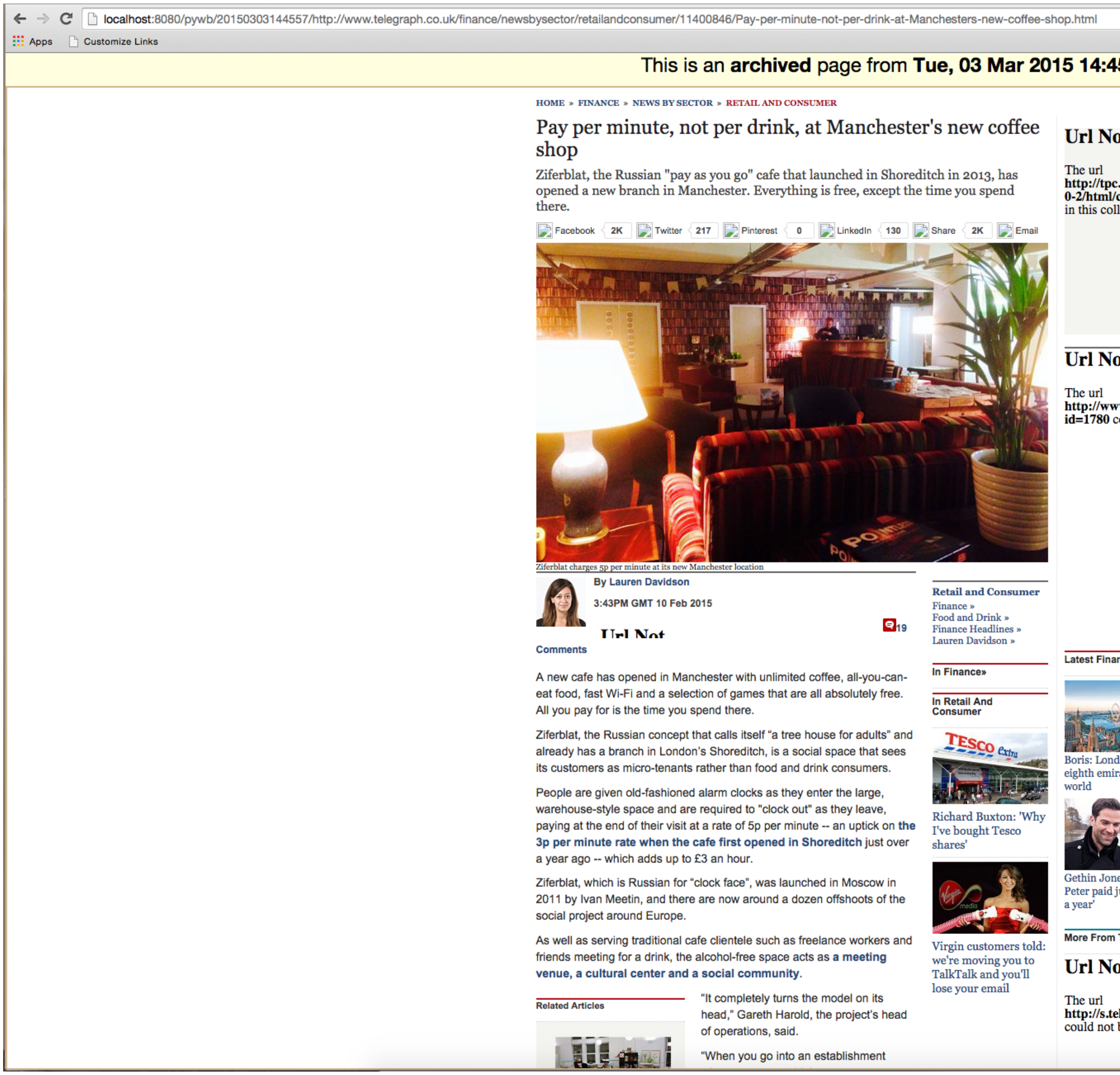


Figure 3: Average Size Comparision of WARC


localhost:8080/pywb/20150303144557/http://www.telegraph.co.uk/finance/newsbysector/retailandconsumer/11400846/Pay-per-minute-not-per-drink-at-Manchesters-new-coffee-shop.html

Apps Customize Links

This is an **archived** page from **Tue, 03 Mar 2015 14:4**

HOME » FINANCE » NEWS BY SECTOR » RETAIL AND CONSUMER

Pay per minute, not per drink, at Manchester's new coffee shop

Ziferblat, the Russian "pay as you go" cafe that launched in Shoreditch in 2013, has opened a new branch in Manchester. Everything is free, except the time you spend there.

Facebook 2K Twitter 217 Pinterest 0 LinkedIn 130 Share 2K Email

Ziferblat charges 5p per minute at its new Manchester location

By Lauren Davidson

3:43PM GMT 10 Feb 2015

19 Comments

A new cafe has opened in Manchester with unlimited coffee, all-you-can-eat food, fast Wi-Fi and a selection of games that are all absolutely free. All you pay for is the time you spend there.

Ziferblat, the Russian concept that calls itself "a tree house for adults" and already has a branch in London's Shoreditch, is a social space that sees its customers as micro-tenants rather than food and drink consumers.

People are given old-fashioned alarm clocks as they enter the large, warehouse-style space and are required to "clock out" as they leave, paying at the end of their visit at a rate of 5p per minute -- an uptick on **the 3p per minute rate when the cafe first opened in Shoreditch** just over a year ago -- which adds up to £3 an hour.

Ziferblat, which is Russian for "clock face", was launched in Moscow in 2011 by Ivan Meetin, and there are now around a dozen offshoots of the social project around Europe.

As well as serving traditional cafe clientele such as freelance workers and friends meeting for a drink, the alcohol-free space acts as **a meeting venue, a cultural center and a social community**.

Related Articles

"It completely turns the model on its head," Gareth Harold, the project's head of operations, said.

"When you go into an establishment

Retail and Consumer
Finance »
Food and Drink »
Finance Headlines »
Lauren Davidson »

In Finance»

In Retail And Consumer

Richard Buxton: 'Why I've bought Tesco shares'

Virgin customers told: we're moving you to TalkTalk and you'll lose your email

Url No
The url
http://tpc.
0-2/html/c
in this coll

Url No
The url
http://www
id=1780 c

Latest Finar

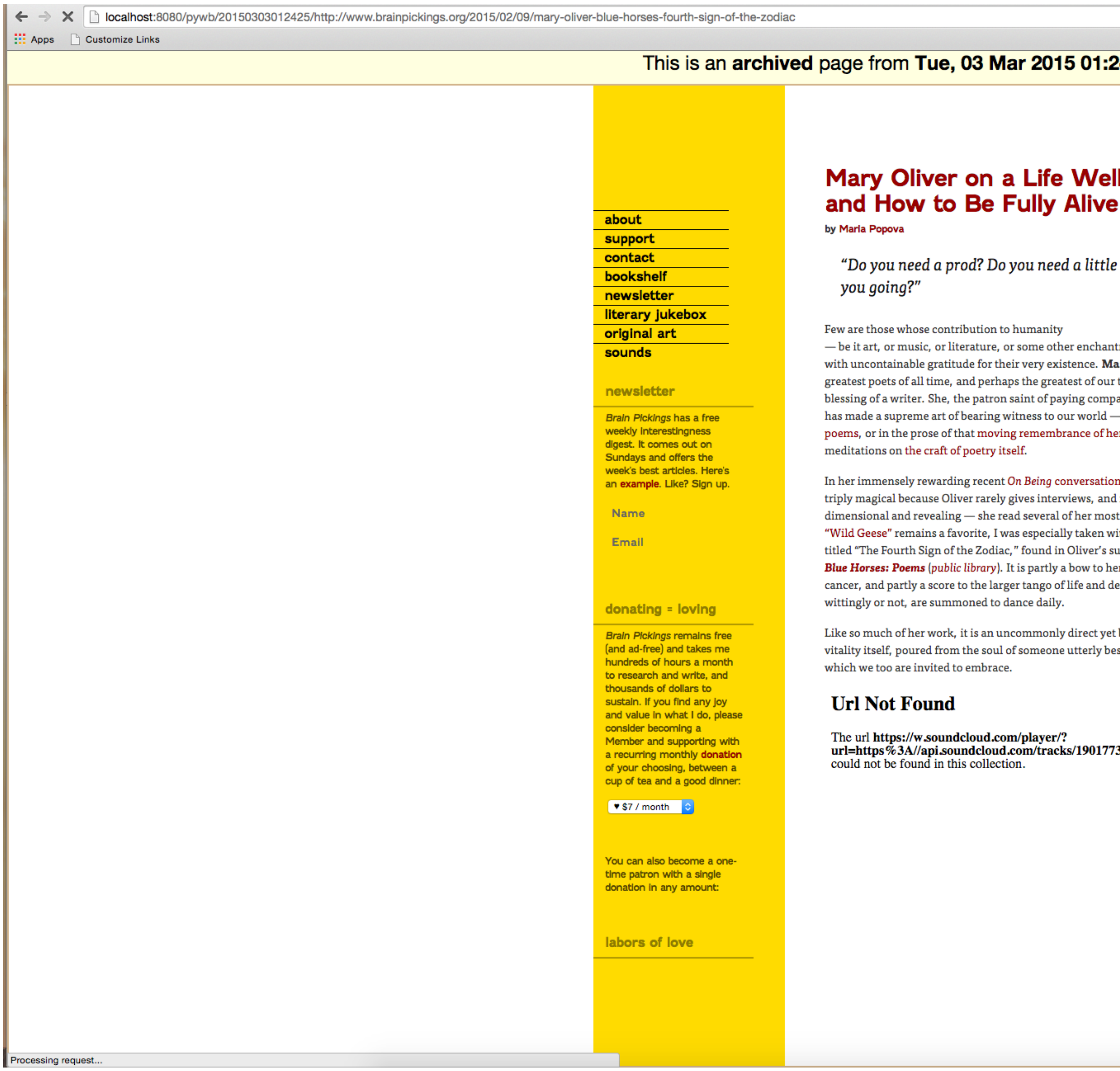
Boris: Lond
eighth emir
world

Gethin Jone
Peter paid j
a year'

More From

Url No
The url
http://s.tel
could not

Figure 4: Play back WARC file using pywb



This is an **archived** page from **Tue, 03 Mar 2015 01:2**

**Mary Oliver on a Life Well**
**and How to Be Fully Alive**

by Maria Popova

*"Do you need a prod? Do you need a little*
*you going?"*

Few are those whose contribution to humanity
— be it art, or music, or literature, or some other enchant
with uncontainable gratitude for their very existence. **Ma**
greatest poets of all time, and perhaps the greatest of our t
blessing of a writer. She, the patron saint of paying compa
has made a supreme art of bearing witness to our world —
poems, or in the prose of that moving remembrance of he
meditations on the craft of poetry itself.

In her immensely rewarding recent *On Being* conversation
triply magical because Oliver rarely gives interviews, and
dimensional and revealing — she read several of her most
"Wild Geese" remains a favorite, I was especially taken wi
titled "The Fourth Sign of the Zodiac," found in Oliver's su
***Blue Horses: Poems*** (*public library*). It is partly a bow to he
cancer, and partly a score to the larger tango of life and de
wittingly or not, are summoned to dance daily.

Like so much of her work, it is an uncommonly direct yet l
vitality itself, poured from the soul of someone utterly bes
which we too are invited to embrace.

*Brain Pickings* has a free weekly interestingness digest. It comes out on Sundays and offers the week's best articles. Here's an example. Like? Sign up.

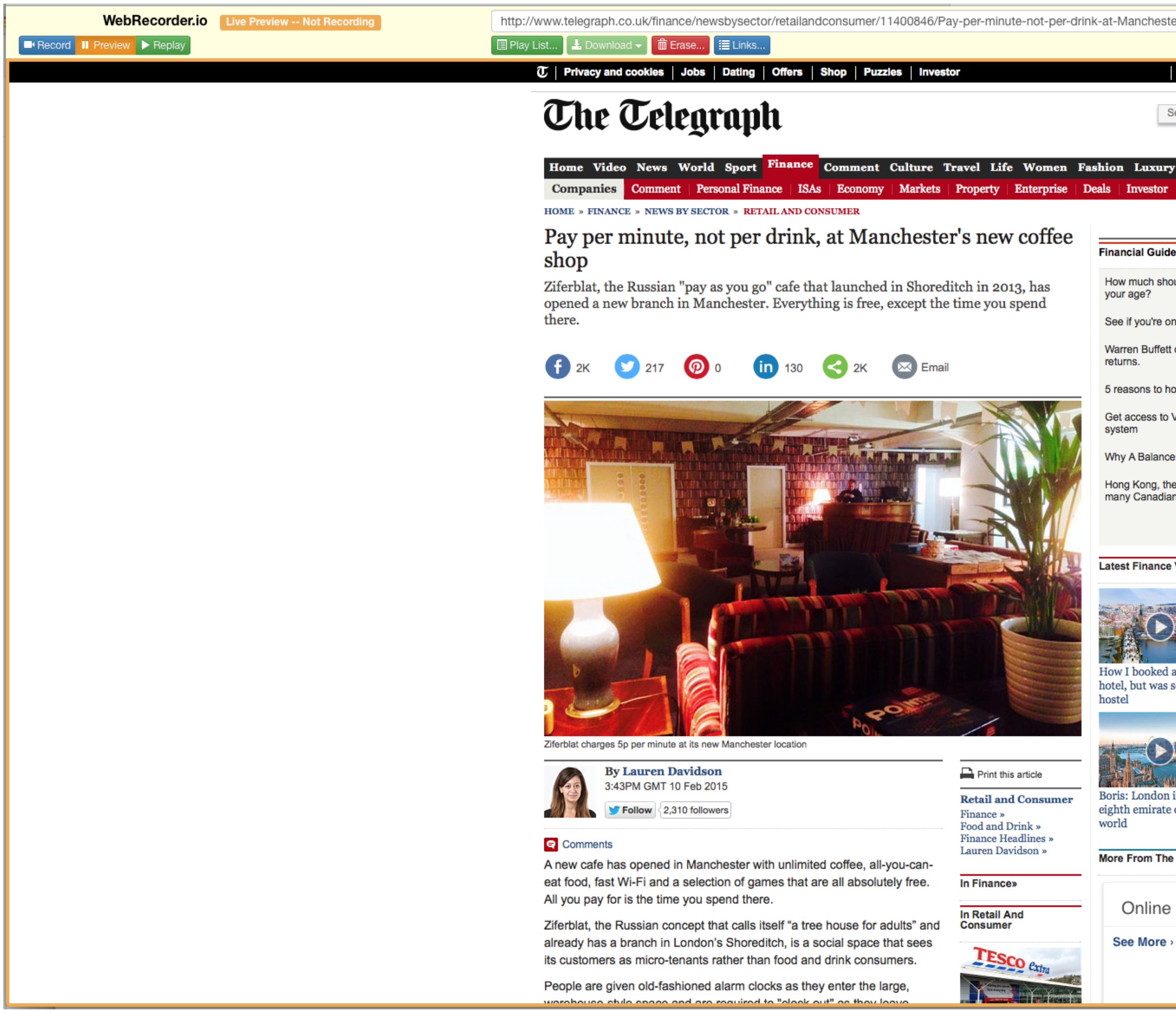
*Brain Pickings* remains free (and ad-free) and takes me hundreds of hours a month to research and write, and thousands of dollars to sustain. If you find any joy and value in what I do, please consider becoming a Member and supporting with a recurring monthly donation of your choosing, between a cup of tea and a good dinner:

You can also become a one-time patron with a single donation in any amount:

**Url Not Found**

The url **https://w.soundcloud.com/player/?**
**url=https%3A//api.soundcloud.com/tracks/1901773**
could not be found in this collection.

Figure 5: Play back WARC file using pywb


WebRecorder.io
Live Preview -- Not Recording
Record
Preview
Replay
Play List...
Download
Erase...
Links...
Privacy and cookies | Jobs | Dating | Offers | Shop | Puzzles | Investor
Home Video News World Sport Finance Comment Culture Travel Life Women Fashion Luxury
Companies Comment Personal Finance ISAs Economy Markets Property Enterprise Deals Investor
2K 217 0 130 2K Email
Follow 2,310 followers
Print this article
Comments

http://www.telegraph.co.uk/finance/newsbysector/retailandconsumer/11400846/Pay-per-minute-not-per-drink-at-Mancheste

The Telegraph

HOME » FINANCE » NEWS BY SECTOR » RETAIL AND CONSUMER

Pay per minute, not per drink, at Manchester's new coffee shop

Ziferblat, the Russian "pay as you go" cafe that launched in Shoreditch in 2013, has opened a new branch in Manchester. Everything is free, except the time you spend there.

Ziferblat charges 5p per minute at its new Manchester location

By Lauren Davidson
3:43PM GMT 10 Feb 2015

A new cafe has opened in Manchester with unlimited coffee, all-you-can-eat food, fast Wi-Fi and a selection of games that are all absolutely free. All you pay for is the time you spend there.

Ziferblat, the Russian concept that calls itself "a tree house for adults" and already has a branch in London's Shoreditch, is a social space that sees its customers as micro-tenants rather than food and drink consumers.

People are given old-fashioned alarm clocks as they enter the large,

Retail and Consumer
Finance »
Food and Drink »
Finance Headlines »
Lauren Davidson »

In Finance»

In Retail And Consumer

Financial Guide

How much shou your age?

See if you're on

Warren Buffett returns.

5 reasons to ho

Get access to V system

Why A Balance

Hong Kong, the many Canadian

Latest Finance

How I booked a hotel, but was s hostel

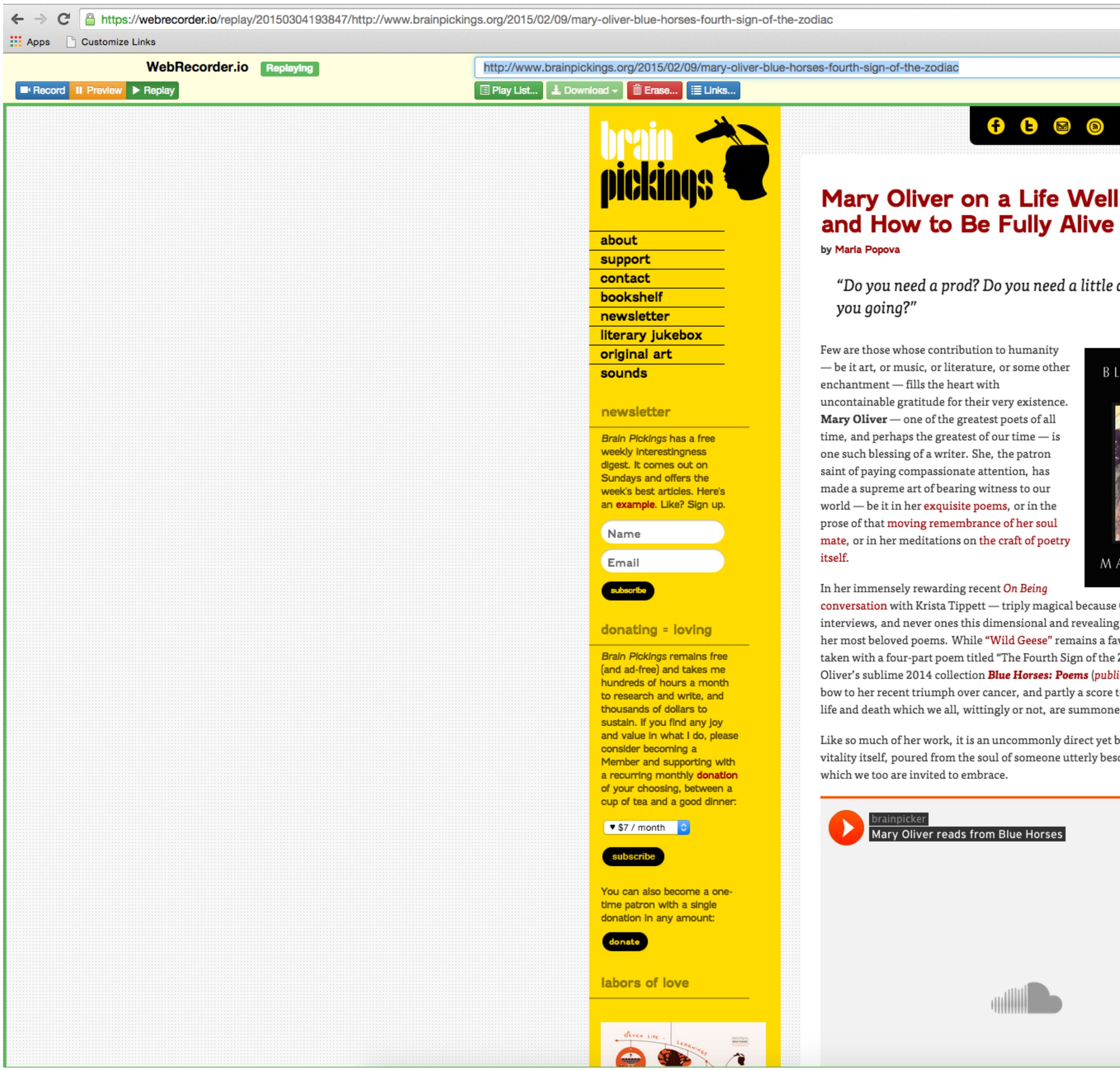
Boris: London i eighth emirate world

More From The

Online

See More ›

Figure 6: Play back WARC file using webrecorder


https://webrecorder.io/replay/20150304193847/http://www.brainpickings.org/2015/02/09/mary-oliver-blue-horses-fourth-sign-of-the-zodiac

Apps  Customize Links

WebRecorder.io  Replaying

http://www.brainpickings.org/2015/02/09/mary-oliver-blue-horses-fourth-sign-of-the-zodiac

Record  Preview  Replay  Play List...  Download  Erase...  Links...

brain pickings

about
support
contact
bookshelf
newsletter
literary jukebox
original art
sounds

newsletter

*Brain Pickings* has a free weekly interestingness digest. It comes out on Sundays and offers the week's best articles. Here's an example. Like? Sign up.

Name
Email
subscribe

donating = loving

*Brain Pickings* remains free (and ad-free) and takes me hundreds of hours a month to research and write, and thousands of dollars to sustain. If you find any joy and value in what I do, please consider becoming a Member and supporting with a recurring monthly donation of your choosing, between a cup of tea and a good dinner:

$7 / month
subscribe

You can also become a one-time patron with a single donation in any amount:

donate

labors of love

Mary Oliver on a Life Well
and How to Be Fully Alive

by Maria Popova

*"Do you need a prod? Do you need a little c*
*you going?"*

Few are those whose contribution to humanity — be it art, or music, or literature, or some other enchantment — fills the heart with uncontainable gratitude for their very existence. **Mary Oliver** — one of the greatest poets of all time, and perhaps the greatest of our time — is one such blessing of a writer. She, the patron saint of paying compassionate attention, has made a supreme art of bearing witness to our world — be it in her exquisite poems, or in the prose of that moving remembrance of her soul mate, or in her meditations on the craft of poetry itself.

In her immensely rewarding recent *On Being* conversation with Krista Tippett — triply magical because
interviews, and never ones this dimensional and revealing
her most beloved poems. While "Wild Geese" remains a fav
taken with a four-part poem titled "The Fourth Sign of the Z
Oliver's sublime 2014 collection ***Blue Horses: Poems*** *(publi*
bow to her recent triumph over cancer, and partly a score t
life and death which we all, wittingly or not, are summone

Like so much of her work, it is an uncommonly direct yet b
vitality itself, poured from the soul of someone utterly besc
which we too are invited to embrace.

brainpicker
Mary Oliver reads from Blue Horses

Figure 7: Play back WARC file using webrecorder

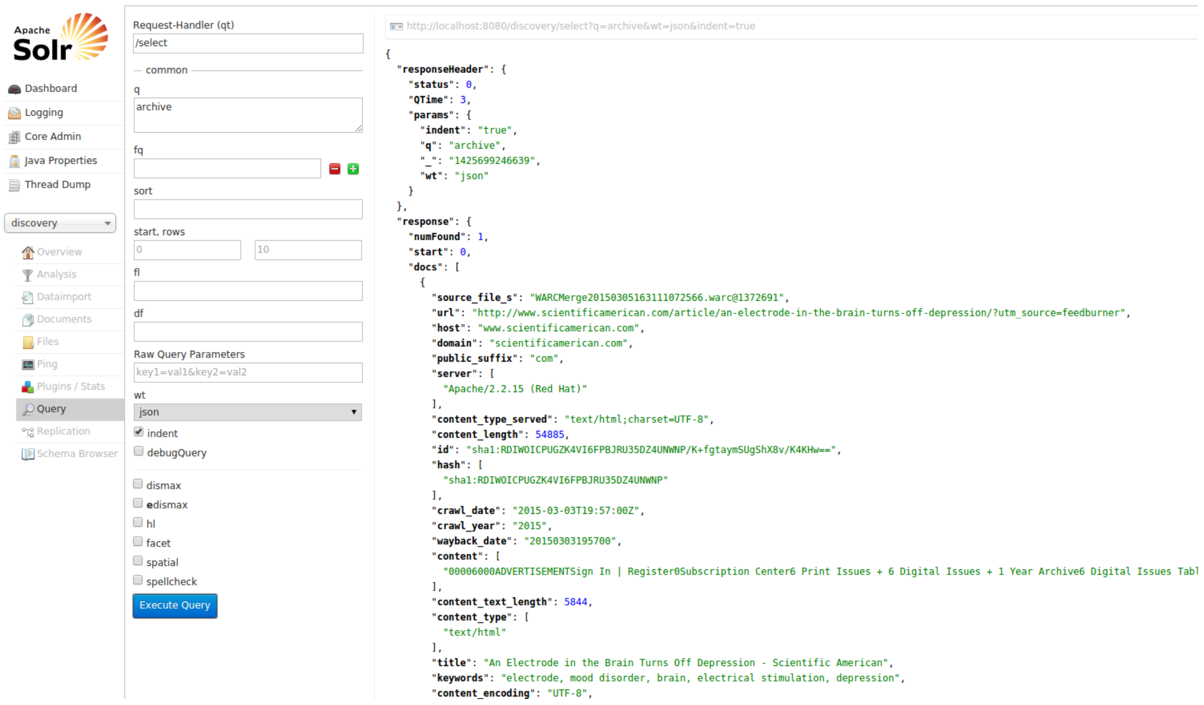
## 2 Question 2

- 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

### 2.1 Solution

The following steps were taken to configure SOLR and process documents :

- The pre-requisites for SOLR are Maven 3, Java 7 so I installed them.
- I faced an issue while installing SOLR which is jetty dependency not found.
- I got it working by adding the dependency to the pom.xml
- The command "mvn jetty:run-exploded" starts the SOLR instance.
- I indexed the WARC file by using the command "java -jar path of jar -s "`http://localhost:8080/discovery`" -t Path of WARC".
- I tried to set-up Shine for SOLR as the front end instead of the default UI for SOLR but after trying for two days and multiple email exchanges with the author of the tool I wasn't able to. So I used the default SOLR front end.
- In SOLR we can perform the following queries.
  - Here we are demonstrating how to retrieve the names and ids of all documents with "`http://localhost:8080/solr/select?q=inStock:false&wt=json&fl=id,name`"
  - Here we are using the functional query idf(field,term). This function returns the inverse document frequency for the given term, using the similarity for the field. "`http://localhost:8080/solr/select/?fl=score,id&defType=func&q=mul(tf(text,memory),idf(text,memory))`"
  - The functional query being used here is tf(field,term) and it returns the inverse document frequency factor for the given term using the similarity for the field. "`http://localhost:8080/solr/select/?fl=score,id&defType=func&q=mul(tf($f,$t),idf($f,$t))&f=text&t=memory`"
  - termfreq(field,term) is the functional query that is being used and it returns the number of times the term term appers for that field in the document. "`http://localhost:8080/solr/select/?fl=score,id&q=DDR&sort=termfreq(text,memory)desc`"
  - Here norm(field) is the functional query that is being used. It returns the norm stored in the index, the product of the index time boost and then length normalization factor. "`http://localhost:8983/solr/select/?fl=score,id&q=DDR&sort=norm(text)asc`".
- Below are the snapshots of the queries that I ran on SOLR after indexing.

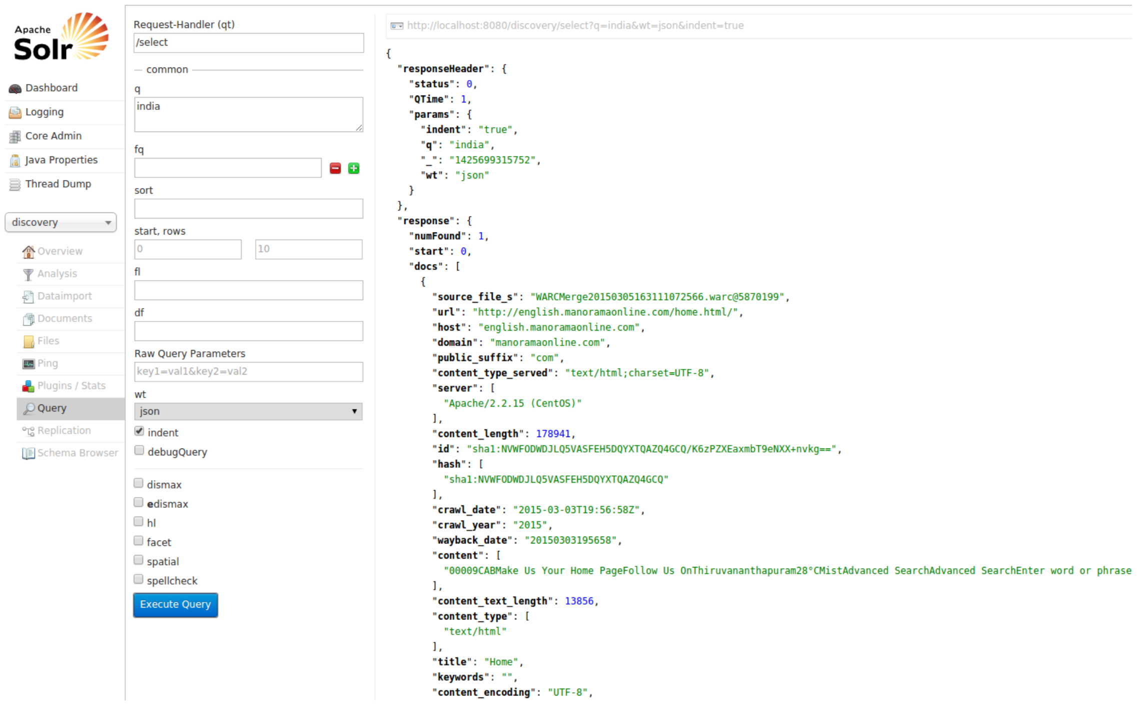


Request-Handler (qt): /select
q: archive
start, rows: 0, 10
Raw Query Parameters: key1=val1&key2=val2
wt: json
[✓] indent
[ ] debugQuery

http://localhost:8080/discovery/select?q=archive&wt=json&indent=true

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 3,
    "params": {
      "indent": "true",
      "q": "archive",
      "_": "1425699246639",
      "wt": "json"
    }
  },
  "response": {
    "numFound": 1,
    "start": 0,
    "docs": [
      {
        "source_file_s": "WARCMerge20150305163111072566.warc@1372691",
        "url": "http://www.scientificamerican.com/article/an-electrode-in-the-brain-turns-off-depression/?utm_source=feedburner",
        "host": "www.scientificamerican.com",
        "domain": "scientificamerican.com",
        "public_suffix": "com",
        "server": [
          "Apache/2.2.15 (Red Hat)"
        ],
        "content_type_served": "text/html;charset=UTF-8",
        "content_length": 54885,
        "id": "sha1:RDIWOICPUGZK4VI6FPBJRU35DZ4UNWNP/K+fgtaymSUgShX8v/K4KHw==",
        "hash": [
          "sha1:RDIWOICPUGZK4VI6FPBJRU35DZ4UNWNP"
        ],
        "crawl_date": "2015-03-03T19:57:00Z",
        "crawl_year": "2015",
        "wayback_date": "20150303195700",
        "content": [
          "00006000ADVERTISEMENTSign In | Register0Subscription Center6 Print Issues + 6 Digital Issues + 1 Year Archive6 Digital Issues Tabl
        ],
        "content_text_length": 5844,
        "content_type": [
          "text/html"
        ],
        "title": "An Electrode in the Brain Turns Off Depression - Scientific American",
        "keywords": "electrode, mood disorder, brain, electrical stimulation, depression",
        "content_encoding": "UTF-8",
```

Figure 8: Query 1 on SOLR

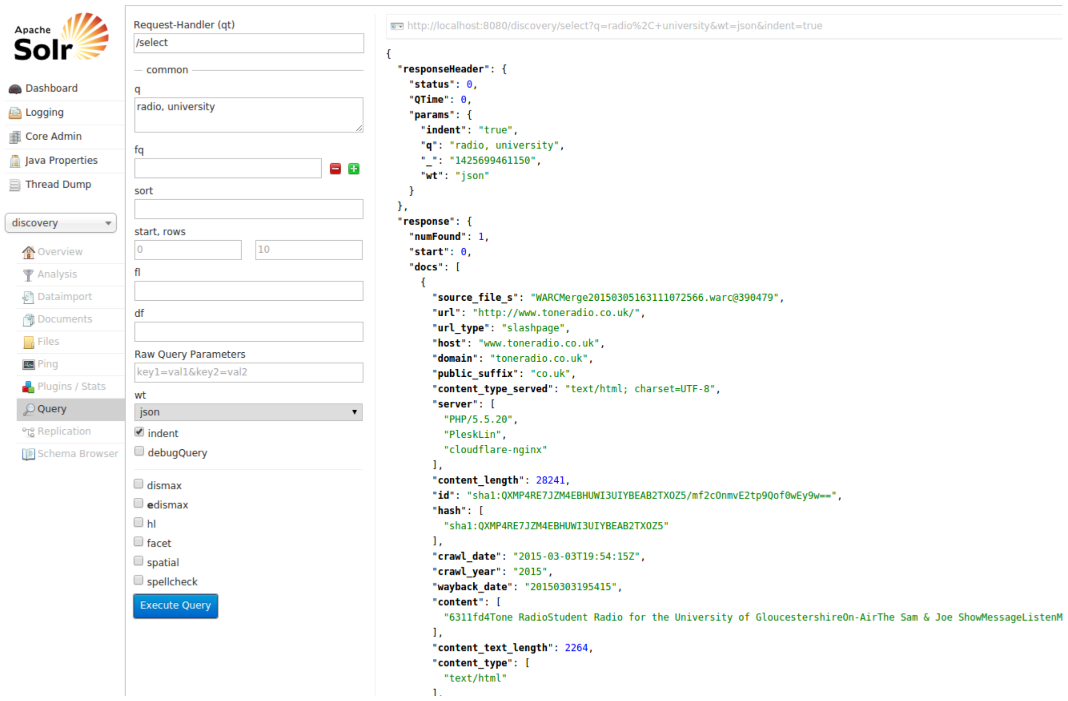


Request-Handler (qt): /select
q: india
start, rows: 0, 10
Raw Query Parameters: key1=val1&key2=val2
wt: json
[✓] indent
[ ] debugQuery

http://localhost:8080/discovery/select?q=india&wt=json&indent=true

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 1,
    "params": {
      "indent": "true",
      "q": "india",
      "_": "1425699315752",
      "wt": "json"
    }
  },
  "response": {
    "numFound": 1,
    "start": 0,
    "docs": [
      {
        "source_file_s": "WARCMerge20150305163111072566.warc@5870199",
        "url": "http://english.manoramaonline.com/home.html/",
        "host": "english.manoramaonline.com",
        "domain": "manoramaonline.com",
        "public_suffix": "com",
        "content_type_served": "text/html;charset=UTF-8",
        "server": [
          "Apache/2.2.15 (CentOS)"
        ],
        "content_length": 178941,
        "id": "sha1:NVWFODWDJLQ5VASFEH5DQYXTQAZQ4GCQ/K6zPZXEaxmbT9eNXX+nvkg==",
        "hash": [
          "sha1:NVWFODWDJLQ5VASFEH5DQYXTQAZQ4GCQ"
        ],
        "crawl_date": "2015-03-03T19:56:58Z",
        "crawl_year": "2015",
        "wayback_date": "20150303195658",
        "content": [
          "00009CABMake Us Your Home PageFollow Us OnThiruvananthapuram28°CMistAdvanced SearchAdvanced SearchEnter word or phrase
        ],
        "content_text_length": 13856,
        "content_type": [
          "text/html"
        ],
        "title": "Home",
        "keywords": "",
        "content_encoding": "UTF-8",
```

Figure 9: Query 2 on SOLR



Request-Handler (qt): /select

q: radio, university

start, rows: 0 10

wt: json

[✓ indent]

debugQuery

```
http://localhost:8080/discovery/select?q=radio%2C+university&wt=json&indent=true
```

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 0,
    "params": {
      "indent": "true",
      "q": "radio, university",
      "_": "1425699461150",
      "wt": "json"
    }
  },
  "response": {
    "numFound": 1,
    "start": 0,
    "docs": [
      {
        "source_file_s": "WARCMerge20150305163111072566.warc@390479",
        "url": "http://www.toneradio.co.uk/",
        "url_type": "slashpage",
        "host": "www.toneradio.co.uk",
        "domain": "toneradio.co.uk",
        "public_suffix": "co.uk",
        "content_type_served": "text/html; charset=UTF-8",
        "server": [
          "PHP/5.5.20",
          "PleskLin",
          "cloudflare-nginx"
        ],
        "content_length": 28241,
        "id": "sha1:QXMP4RE7JZM4EBHUWI3UIYBEAB2TXOZ5/mf2cOnmvE2tp9Qof0wEy9w==",
        "hash": [
          "sha1:QXMP4RE7JZM4EBHUWI3UIYBEAB2TXOZ5"
        ],
        "crawl_date": "2015-03-03T19:54:15Z",
        "crawl_year": "2015",
        "wayback_date": "20150303195415",
        "content": [
          "6311fd4Tone RadioStudent Radio for the University of GloucestershireOn-AirThe Sam & Joe ShowMessageListenM
        ],
        "content_text_length": 2264,
        "content_type": [
          "text/html"
        ],
```

Figure 10: Query 3 on SOLR

# Bibliography

[1] Download warcreate. https://chrome.google.com/webstore/detail/warcreate/kenncghfghgolcbmckhiljgaabnpcaaa?hl=en-US.

[2] Generating simple bar plots in r. http://www.statmethods.net/graphs/bar.html.

[3] Playback using webrecorder. https://webrecorder.io/.

[4] Setting up pywb. https://github.com/ikreymer/pywb.

[5] Using wail. http://matkelly.com/wail/.

[6] Warc merge. https://github.com/maturban/WARCMerge.