# Tools for Managing Seed URIs
## (Detecting Off-Topic Pages)

**Yasmin AlNoamany, Michele C. Weigle, Michael L. Nelson**

Old Dominion University

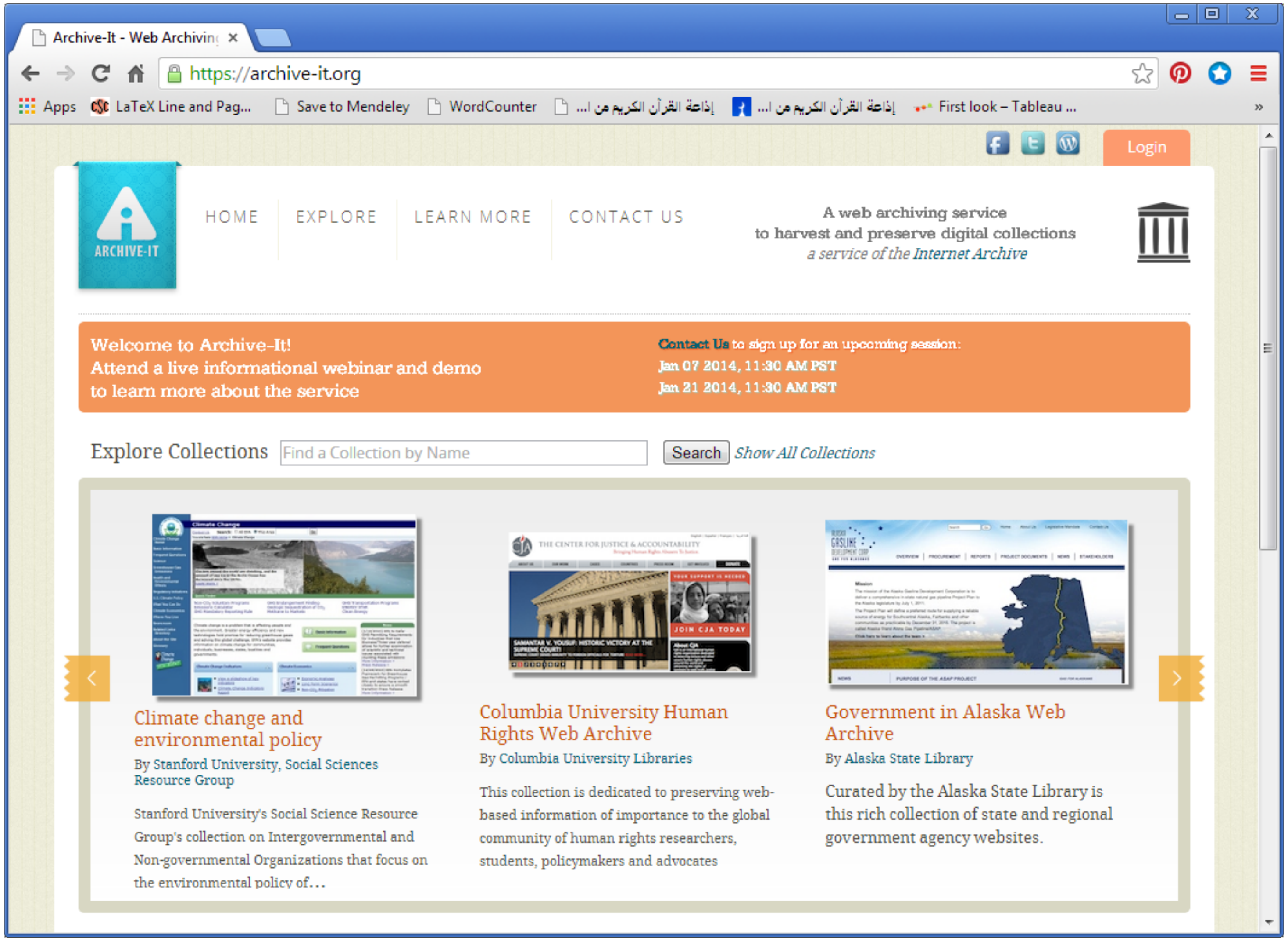
Web Science and Digital Libraries Group

http://ws-dl.cs.odu.edu/, @WebSciDL

Funded by Columbia University Libraries Web Archiving Incentive program

**Web Archiving Collaboration: New Tools and Models**
**June 4-5, 2015**

# Archive-It - Curated Web Collections



A web archiving service
to harvest and preserve digital collections
*a service of the Internet Archive*

Welcome to Archive-It!
Attend a live informational webinar and demo
to learn more about the service

**Contact Us** to sign up for an upcoming session:
Jan 07 2014, 11:30 AM PST
Jan 21 2014, 11:30 AM PST

Explore Collections [Find a Collection by Name] Search *Show All Collections*

**Climate change and environmental policy**
By Stanford University, Social Sciences Resource Group

Stanford University's Social Science Resource Group's collection on Intergovernmental and Non-governmental Organizations that focus on the environmental policy of...

**Columbia University Human Rights Web Archive**
By Columbia University Libraries

This collection is dedicated to preserving web-based information of importance to the global community of human rights researchers, students, policymakers and advocates

**Government in Alaska Web Archive**
By Alaska State Library

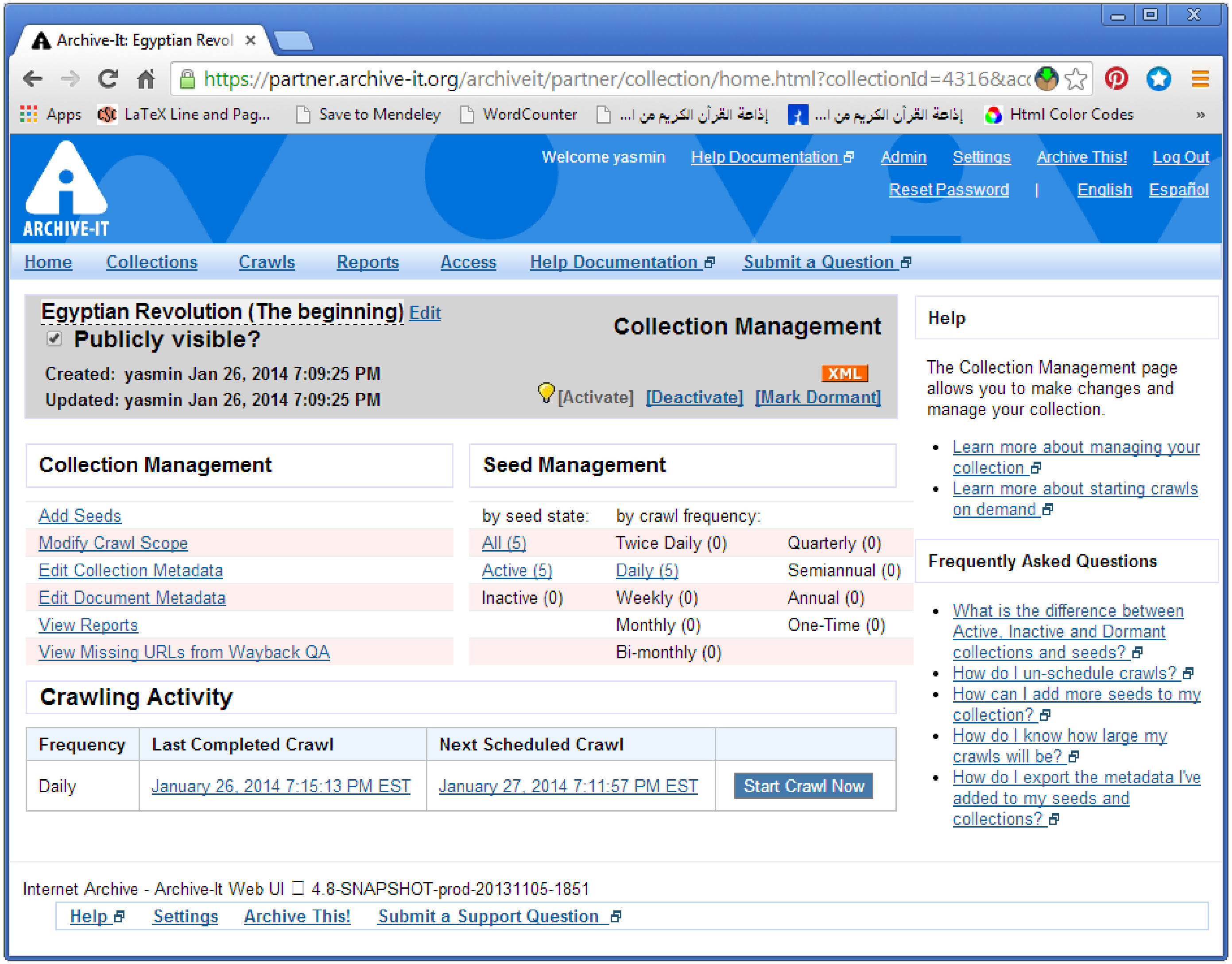
Curated by the Alaska State Library is this rich collection of state and regional government agency websites.

> 3,000 collections

~340 institutions

> 10B archived pages

# Archive-It Interface - Curator's View



Egyptian Revolution (The beginning) Edit

☑ Publicly visible?

Created: yasmin Jan 26, 2014 7:09:25 PM

Updated: yasmin Jan 26, 2014 7:09:25 PM

Collection Management

XML

[Activate] [Deactivate] [Mark Dormant]

**Collection Management**

- Add Seeds
- Modify Crawl Scope
- Edit Collection Metadata
- Edit Document Metadata
- View Reports
- View Missing URLs from Wayback QA

**Seed Management**

| by seed state: | by crawl frequency: | |
|---|---|---|
| All (5) | Twice Daily (0) | Quarterly (0) |
| Active (5) | Daily (5) | Semiannual (0) |
| Inactive (0) | Weekly (0) | Annual (0) |
| | Monthly (0) | One-Time (0) |
| | Bi-monthly (0) | |

**Crawling Activity**

| Frequency | Last Completed Crawl | Next Scheduled Crawl | |
|---|---|---|---|
| Daily | January 26, 2014 7:15:13 PM EST | January 27, 2014 7:11:57 PM EST | Start Crawl Now |

**Help**

The Collection Management page allows you to make changes and manage your collection.

- Learn more about managing your collection
- Learn more about starting crawls on demand

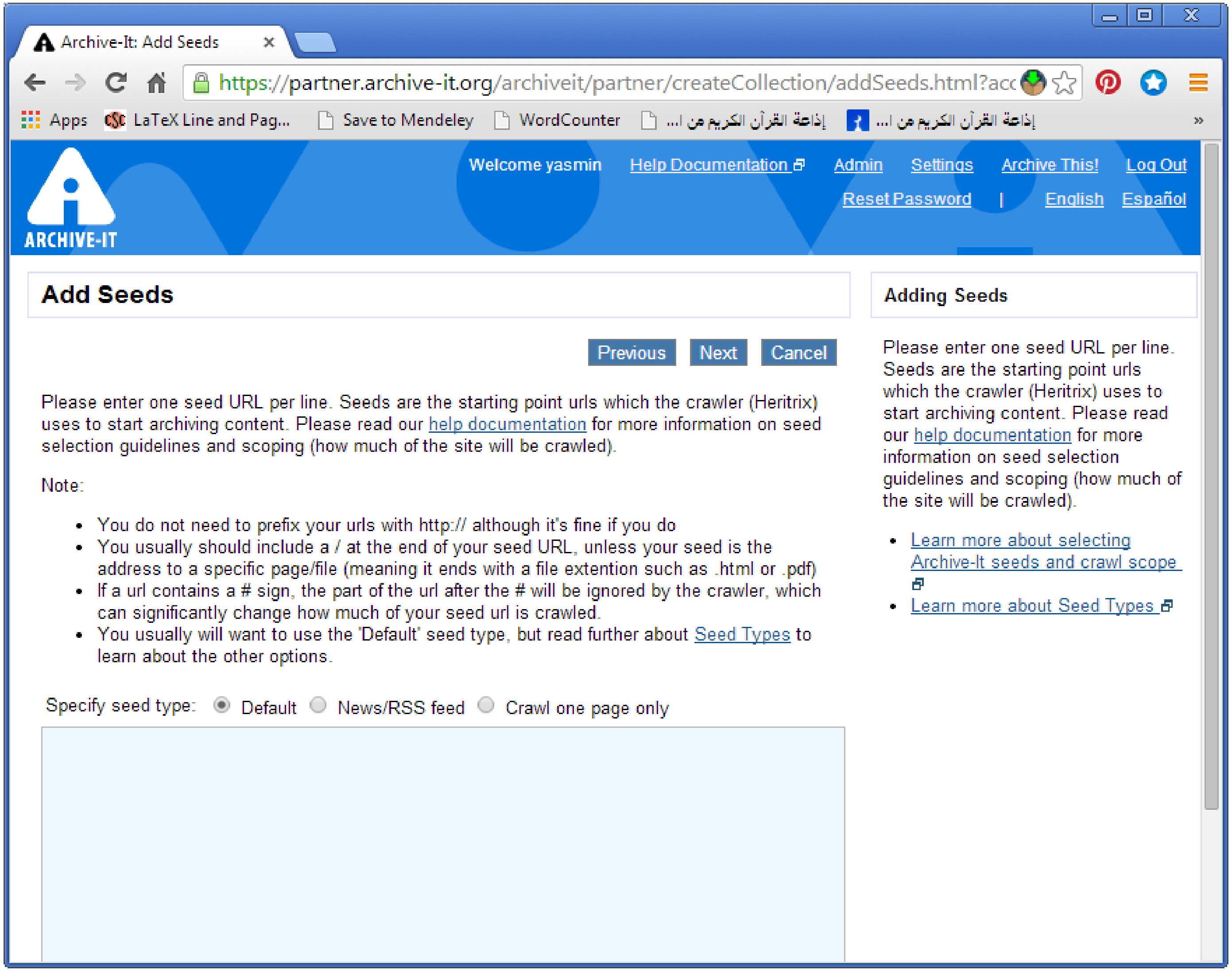
**Frequently Asked Questions**

- What is the difference between Active, Inactive and Dormant collections and seeds?
- How do I un-schedule crawls?
- How can I add more seeds to my collection?
- How do I know how large my crawls will be?
- How do I export the metadata I've added to my seeds and collections?

Internet Archive - Archive-It Web UI ☐ 4.8-SNAPSHOT-prod-20131105-1851

Help Settings Archive This! Submit a Support Question

# The collection curator specifies seed URIs



## Add Seeds

Please enter one seed URL per line. Seeds are the starting point urls which the crawler (Heritrix) uses to start archiving content. Please read our help documentation for more information on seed selection guidelines and scoping (how much of the site will be crawled).

Note:

- You do not need to prefix your urls with http:// although it's fine if you do
- You usually should include a / at the end of your seed URL, unless your seed is the address to a specific page/file (meaning it ends with a file extention such as .html or .pdf)
- If a url contains a # sign, the part of the url after the # will be ignored by the crawler, which can significantly change how much of your seed url is crawled.
- You usually will want to use the 'Default' seed type, but read further about Seed Types to learn about the other options.

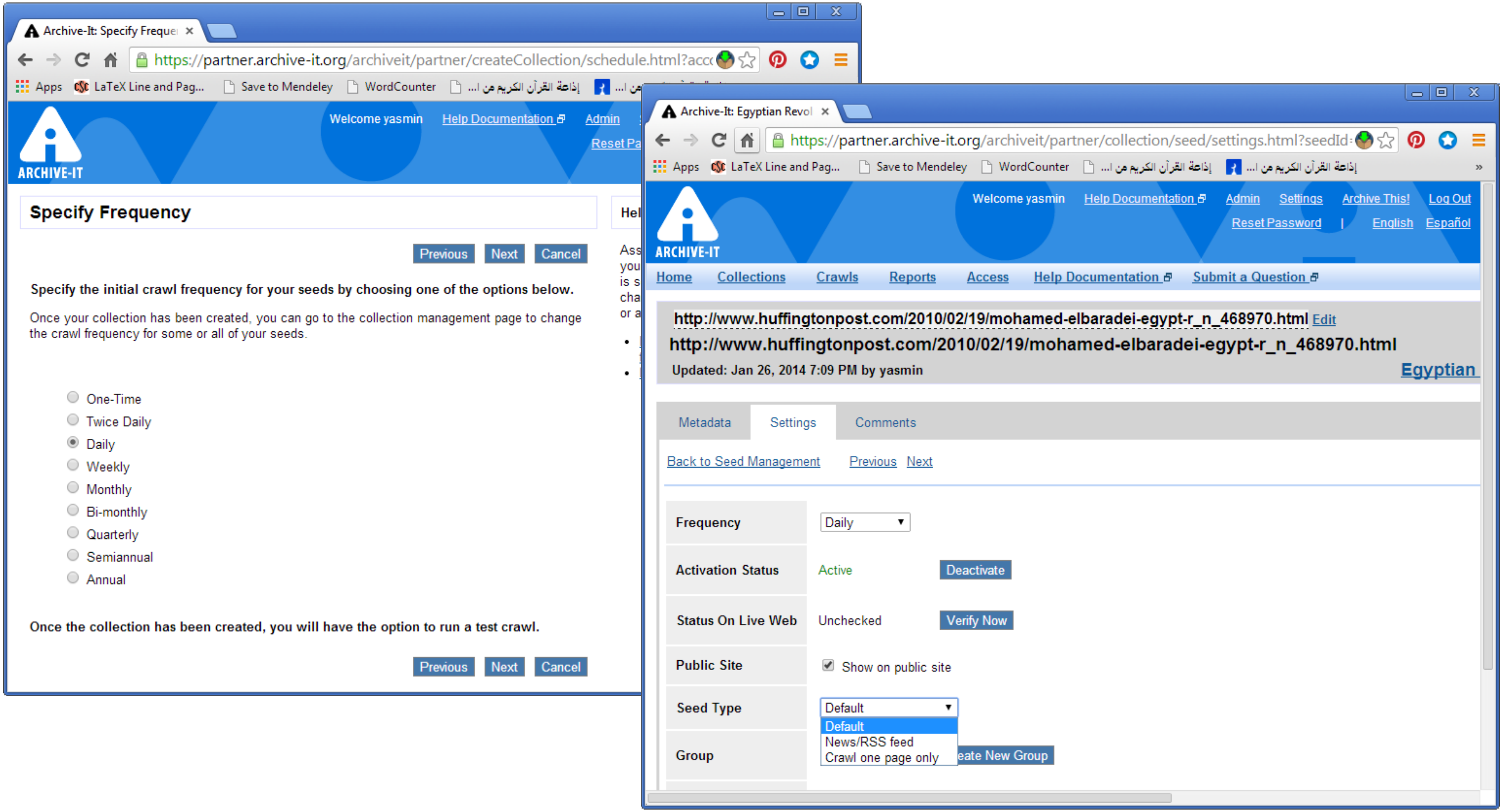
Specify seed type: ● Default ○ News/RSS feed ○ Crawl one page only

## Adding Seeds

Please enter one seed URL per line. Seeds are the starting point urls which the crawler (Heritrix) uses to start archiving content. Please read our help documentation for more information on seed selection guidelines and scoping (how much of the site will be crawled).

- Learn more about selecting Archive-It seeds and crawl scope
- Learn more about Seed Types

# Curators specify the breadth and depth of the crawl



Archive-It: Specify Freque
https://partner.archive-it.org/archiveit/partner/createCollection/schedule.html?acc
Apps
LaTeX Line and Pag...
Save to Mendeley
WordCounter
...إذاعة القرآن الكريم من ا
ARCHIVE-IT
Welcome yasmin
Help Documentation
Admin
Reset Pa

## Specify Frequency

**Specify the initial crawl frequency for your seeds by choosing one of the options below.**

Once your collection has been created, you can go to the collection management page to change the crawl frequency for some or all of your seeds.

- ○ One-Time
- ○ Twice Daily
- ● Daily
- ○ Weekly
- ○ Monthly
- ○ Bi-monthly
- ○ Quarterly
- ○ Semiannual
- ○ Annual

**Once the collection has been created, you will have the option to run a test crawl.**

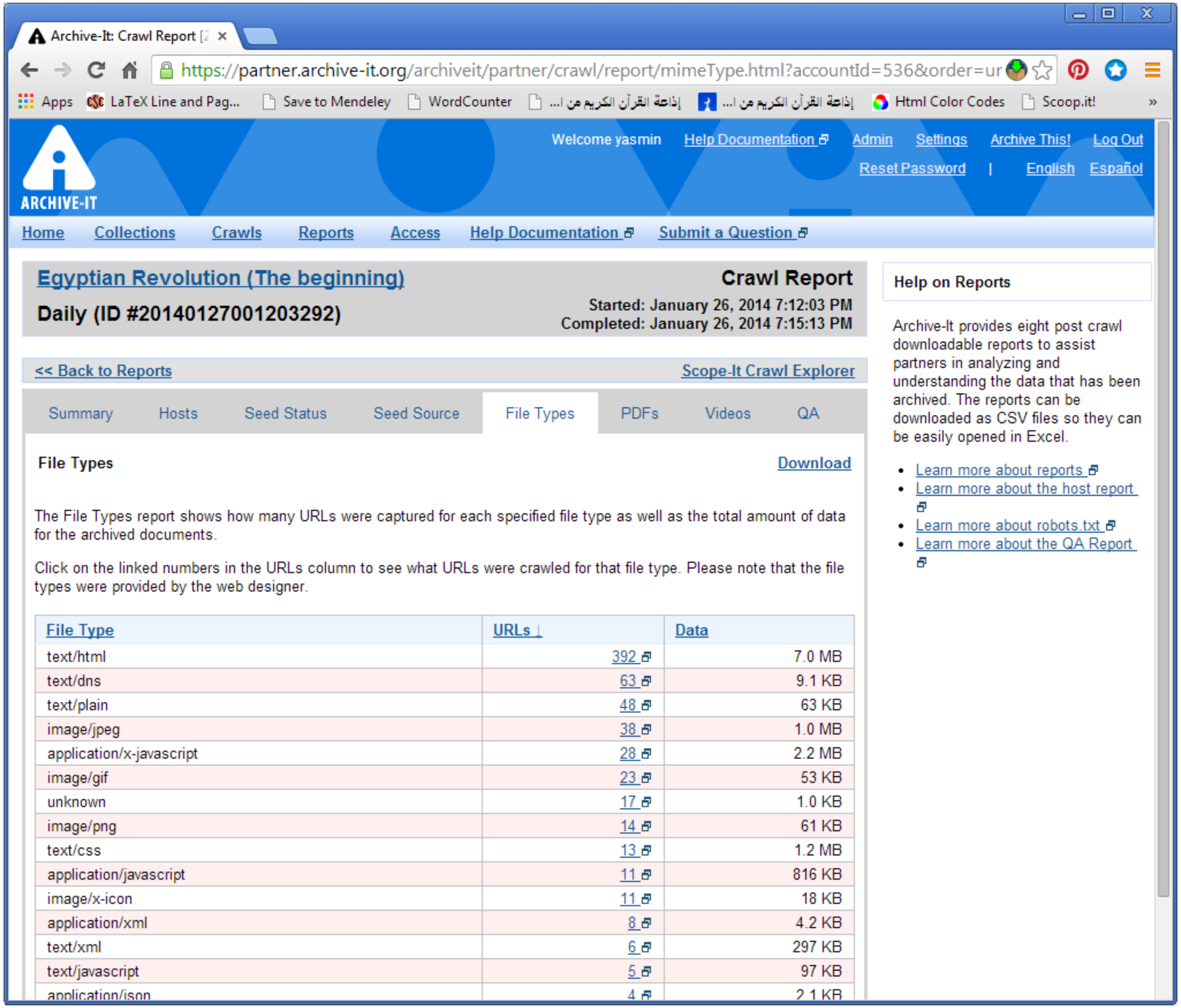
http://www.huffingtonpost.com/2010/02/19/mohamed-elbaradei-egypt-r_n_468970.html Edit

**http://www.huffingtonpost.com/2010/02/19/mohamed-elbaradei-egypt-r_n_468970.html**

**Updated: Jan 26, 2014 7:09 PM by yasmin**

| | | |
|---|---|---|
| **Frequency** | Daily ▾ | |
| **Activation Status** | Active | Deactivate |
| **Status On Live Web** | Unchecked | Verify Now |
| **Public Site** | ☑ Show on public site | |
| **Seed Type** | Default ▾<br>Default<br>News/RSS feed<br>Crawl one page only | |
| **Group** | | eate New Group |

# Current tools measure HTTP events, not “aboutness”



Egyptian Revolution (The beginning)

Daily (ID #20140127001203292)

Crawl Report

Started: January 26, 2014 7:12:03 PM
Completed: January 26, 2014 7:15:13 PM

File Types

The File Types report shows how many URLs were captured for each specified file type as well as the total amount of data for the archived documents.

Click on the linked numbers in the URLs column to see what URLs were crawled for that file type. Please note that the file types were provided by the web designer.

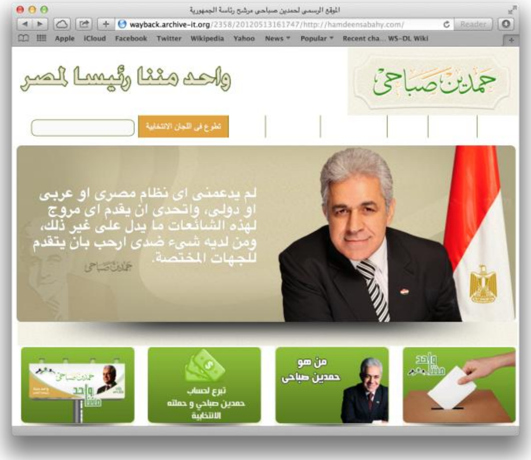
| File Type | URLs ↓ | Data |
|---|---|---|
| text/html | 392 | 7.0 MB |
| text/dns | 63 | 9.1 KB |
| text/plain | 48 | 63 KB |
| image/jpeg | 38 | 1.0 MB |
| application/x-javascript | 28 | 2.2 MB |
| image/gif | 23 | 53 KB |
| unknown | 17 | 1.0 KB |
| image/png | 14 | 61 KB |
| text/css | 13 | 1.2 MB |
| application/javascript | 11 | 816 KB |
| image/x-icon | 11 | 18 KB |
| application/xml | 8 | 4.2 KB |
| text/xml | 6 | 297 KB |
| text/javascript | 5 | 97 KB |
| application/json | 4 | 2.1 KB |

Help on Reports

Archive-It provides eight post crawl downloadable reports to assist partners in analyzing and understanding the data that has been archived. The reports can be downloaded as CSV files so they can be easily opened in Excel.

- Learn more about reports
- Learn more about the host report
- Learn more about robots.txt
- Learn more about the QA Report

# Pages can go off-topic through time

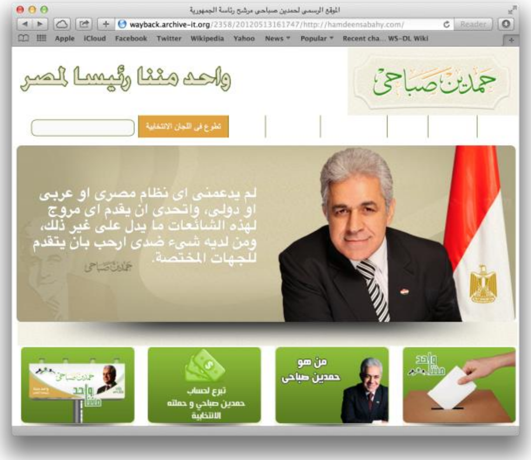


لم يدعمنى اى نظام مصرى او عربى
او دولى، واتحدى ان يقدم اى مروج
لهذه الشائعات ما يدل على غير ذلك،
ومن لديه شىء ضدى ارحب بان يتقدم
للجهات المختصة.

**May 13, 2012: The page started as on-topic.**

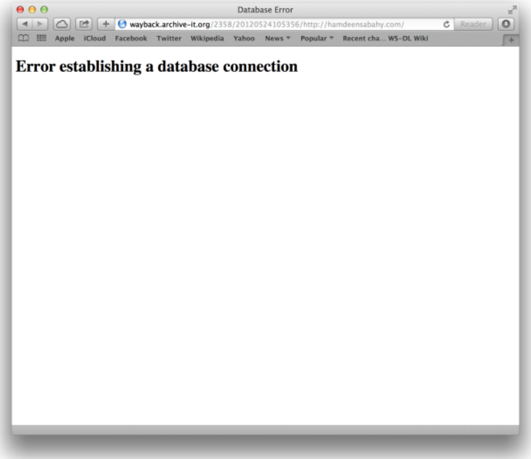
http://wayback.archive-it.org/2358/*/http://hamdeensabahy.com

# Pages can go off-topic through time



لم يدعمنى اى نظام مصرى او عربى
او دولى، واتحدى ان يقدم اى مروج
لهذه الشائعات ما يدل على غير ذلك،
ومن لديه شيء ضدى ارحب بان يتقدم
للجهات المختصة.
حمدين صباحي

**May 13, 2012: The page started as on-topic.**

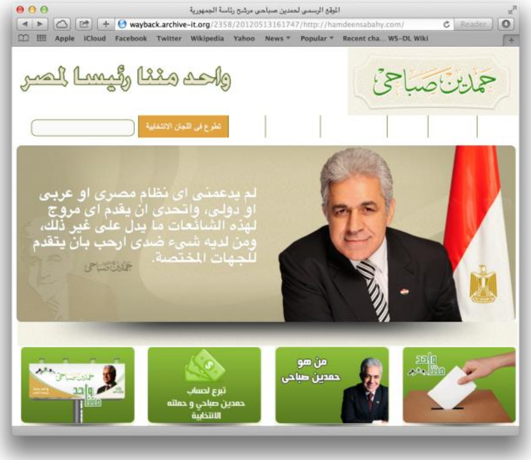


**Error establishing a database connection**

**May 24, 2012: Off-topic due to a database error.**

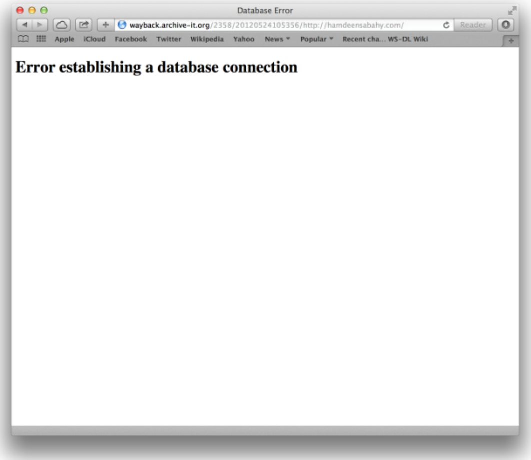
http://wayback.archive-it.org/2358/*/http://hamdeensabahy.com

# Pages can go off-topic through time


الموقع الرسمي لحمدين صباحي مرشح رئاسة الجمهورية
wayback.archive-it.org/2358/20120513161747/http://hamdeensabahy.com/
Apple iCloud Facebook Twitter Wikipedia Yahoo News Popular Recent cha... WS-DL Wiki
Reader
حمدين صباحي
واحد منا رئيسا لمصر
تطوع فى اللجان الانتخابية
من هو
حمدين صباحي
تبرع لحساب
حمدين صباحي و حملته
الانتخابية

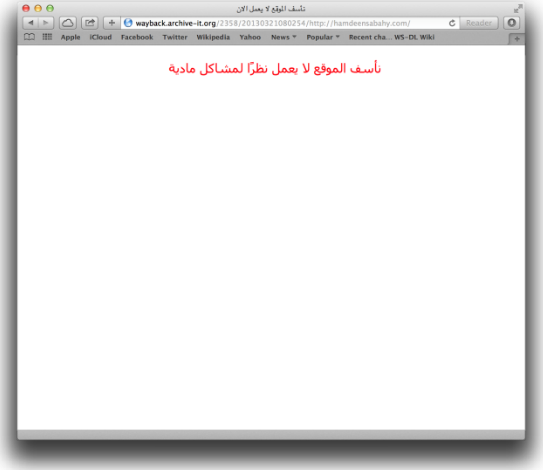
لم يدعمنى اى نظام مصرى او عربى
او دولى، واتحدى ان يقدم اى مروج
لهذه الشائعات ما يدل على غير ذلك،
ومن لديه شيء ضدى ارحب بان يتقدم
للجهات المختصة.

**May 13, 2012: The page started as on-topic.**


Database Error
wayback.archive-it.org/2358/20120524105356/http://hamdeensabahy.com/
Apple iCloud Facebook Twitter Wikipedia Yahoo News Popular Recent cha... WS-DL Wiki
Reader

Error establishing a database connection

**May 24, 2012: Off-topic due to a database error.**

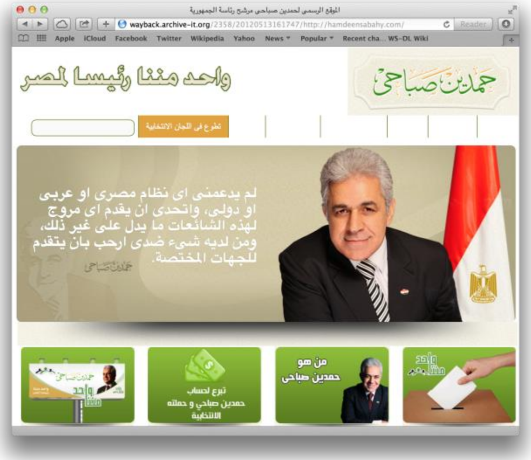

نأسف الموقع لا يعمل الان
wayback.archive-it.org/2358/20130321080254/http://hamdeensabahy.com/
Apple iCloud Facebook Twitter Wikipedia Yahoo News Popular Recent cha... WS-DL Wiki
Reader

نأسف الموقع لا يعمل نظرًا لمشاكل مادية

**Mar. 21, 2013: Not working because of financial problems.**

http://wayback.archive-it.org/2358/*/http://hamdeensabahy.com

# Pages can go off-topic through time

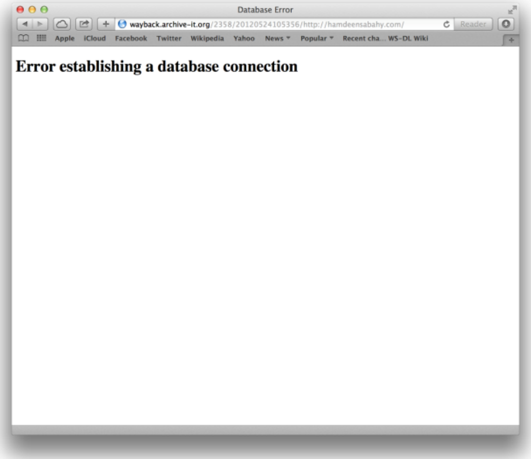


الموقع الرسمى لحمدين صباحى مرشح رئاسة الجمهورية

wayback.archive-it.org/2358/20120513161747/http://hamdeensabahy.com/

Apple iCloud Facebook Twitter Wikipedia Yahoo News Popular Recent cha... WS-DL Wiki
Reader
حمدين صباحى
واحد منا رئيسا لمصر
تطوع فى اللجان الانتخابية
لم يدعمنى اى نظام مصرى او عربى او دولى، واتحدى ان يقدم اى مروج لهذه الشائعات ما يدل على غير ذلك، ومن لديه شيء ضدى ارحب بان يتقدم للجهات المختصة.
حمدين صباحى
تبرع لحساب حمدين صباحي و حملته الانتخابية
من هو حمدين صباحى
واحد منا

May 13, 2012: The page started as on-topic.



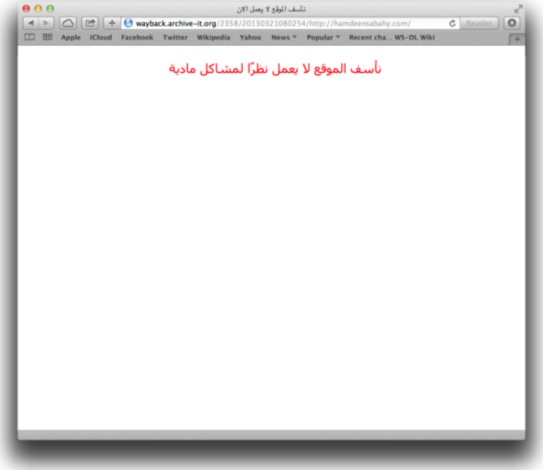
Database Error

wayback.archive-it.org/2358/20120524105356/http://hamdeensabahy.com/

Apple iCloud Facebook Twitter Wikipedia Yahoo News Popular Recent cha... WS-DL Wiki
Reader

Error establishing a database connection

May 24, 2012: Off-topic due to a database error.



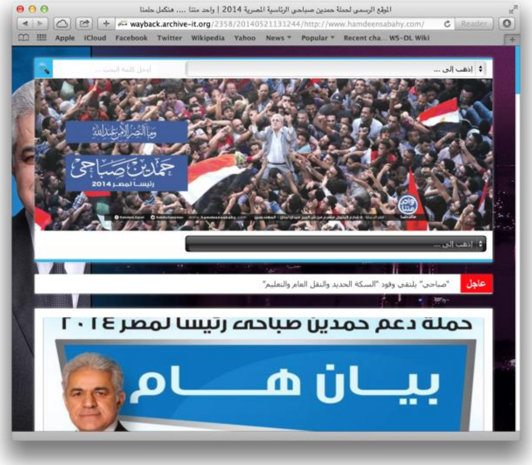
نأسف الموقع لا يعمل الان

wayback.archive-it.org/2358/20130321080254/http://hamdeensabahy.com/

Apple iCloud Facebook Twitter Wikipedia Yahoo News Popular Recent cha... WS-DL Wiki
Reader

نأسف الموقع لا يعمل نظرًا لمشاكل مادية

Mar. 21, 2013: Not working because of financial problems.



الموقع الرسمى لحملة حمدين صباحي الرئاسية المصرية 2014 | واحد منتا .... هنكمل حلمنا

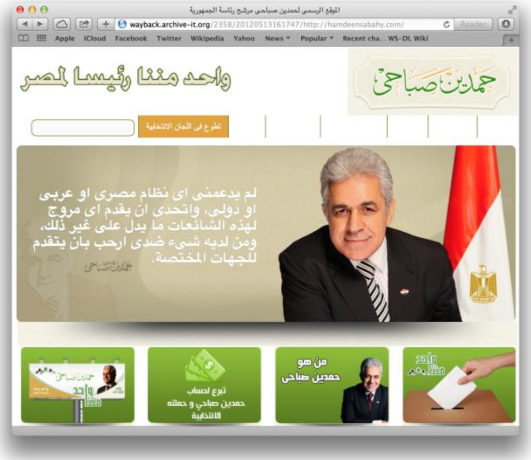
wayback.archive-it.org/2358/20140521131244/http://www.hamdeensabahy.com/

Apple iCloud Facebook Twitter Wikipedia Yahoo News Popular Recent cha... WS-DL Wiki
Reader
ادخل كلمة البحث ...
اذهب الى ...
حمدين صباحى
رئيسا لمصر 2014
اذهب الى ...
عاجل
"صباحي" يلتقى وفود "السكة الحديد والنقل العام والتعليم"
حملة دعم حمدين صباحى رئيسا لمصر ٢٠١٤
بيان هام

May 21, 2013: On-topic again

http://wayback.archive-it.org/2358/*/http://hamdeensabahy.com

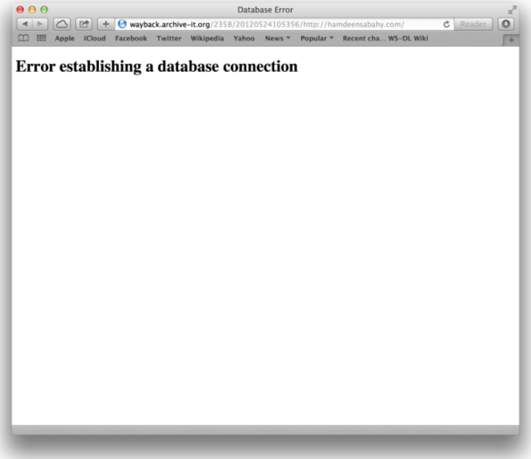
# Pages can go off-topic through time


الموقع الرسمى لحمدين صباحي مرشح رئاسة الجمهورية
wayback.archive-it.org/2358/20120513161747/http://hamdeensabahy.com/
واحد منّنا رئيسا لمصر
حمدين صباحي
تطوع فى اللجان الانتخابية
من هو حمدين صباحي
تبرع لحساب حمدين صباحي و حملته الانتخابية

لم يدعمنى اى نظام مصرى او عربى او دولى، واتحدى ان يقدم اى مروج لهذه الشائعات ما يدل على غير ذلك، ومن لديه شيء ضدى ارحب بان يتقدم للجهات المختصة.

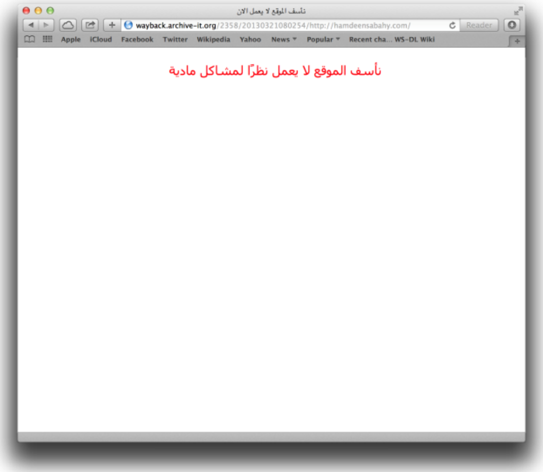
حمدين صباحي

May 13, 2012: The page started as on-topic.


Database Error
wayback.archive-it.org/2358/20120524105356/http://hamdeensabahy.com/

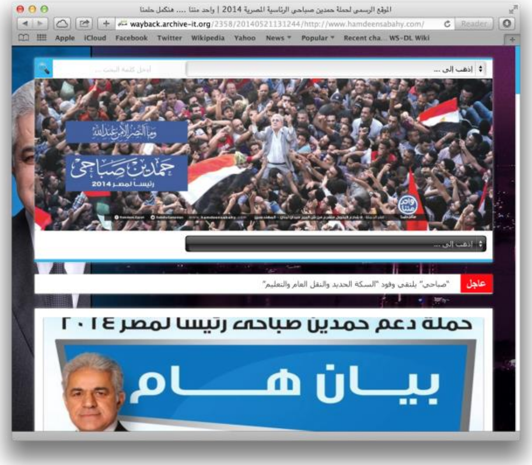
Error establishing a database connection

May 24, 2012: Off-topic due to a database error.

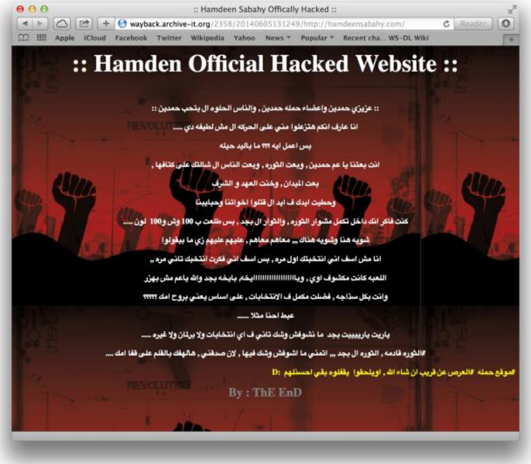

نأسف الموقع لا يعمل الان
wayback.archive-it.org/2358/20130321080254/http://hamdeensabahy.com/

نأسف الموقع لا يعمل نظرًا لمشاكل مادية

Mar. 21, 2013: Not working because of financial problems.


wayback.archive-it.org/2358/20140521131244/http://www.hamdeensabahy.com/
حمدين صباحي
رئيسا لمصر 2014
عاجل
حملة دعم حمدين صباحي رئيسا لمصر ٢٠١٤
بيان هام

May 21, 2013: On-topic again


:: Hamdeen Sabahy Offically Hacked ::
wayback.archive-it.org/2358/20140605131249/http://hamdeensabahy.com/

:: Hamden Official Hacked Website ::

:: عزيزي حمدين واعضاء حمله حمدين, والناس الحلوه ال بتحب حمدين ::

انا عارف انكم هتزعلوا مني على الحركه ال مش لطيفه دي .....

بس اعمل ايه ؟؟؟ ما باليد حيله

انت بعتنا يا عم حمدين, وبعت الثوره, وبعت الناس ال شالتك على كتافها,

بعت الميدان, وخنت العهد و الشرف

وحطيت ايدك ف ايد ال قتلوا اخواتنا وحبايبنا

كنت فاكر انك داخل تكمل مشوار الثوره, والثوار ال بجد, بس طلعت ب 100 وش و100 لون .....

شويه هنا وشويه هناك ,,, معاهم معاهم, عليهم عليهم زي ما بيقولوا

انا مش اسف اني انتخبتك اول مره, بس اسف اني فكرت انتخبك تاني مره ,,

اللعبه كانت مكشوف اوي, ويااااااااااااايخم بايخه بجد والله ياعم مش بهزر

وانت بكل سذاجه, فضلت مكمل ف الانتخابات, على اساس يعني بروح امك ؟؟؟؟؟

عبط احنا مثلا ......

ياريت ياربييييت بجد ما نشوفش وشك تاني ف اي انتخابات ولا برلمان ولا غيره .....

#الثوره قادمه, الثوره ال بجد ,,, اتمنى ما اشوفش وشك فيها, لان صدقني, هالهفك بالقلم على قفا امك ....

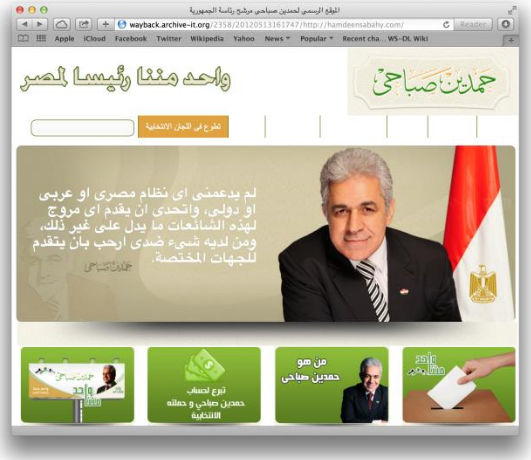
#موقع حمله #العرض عن قريب ان شاء الله, اويلحقوا يقفلوه بقي احسنلهم :D

By : ThE EnD

June 5, 2014: The site has been hacked

http://wayback.archive-it.org/2358/*/http://hamdeensabahy.com

# Over 60% of hamdeensabahy.com archived versions (266) are off-topic

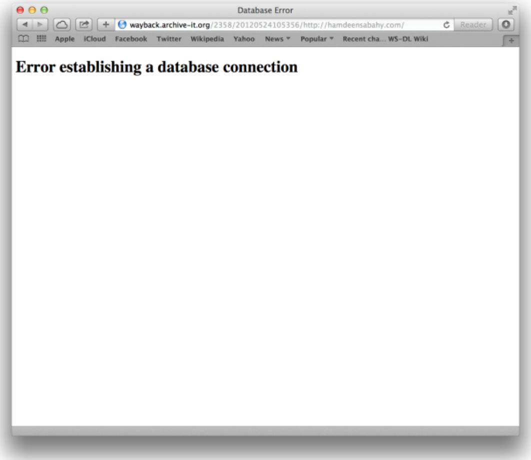


الموقع الرسمى لحمدين صباحى مرشح رئاسة الجمهورية

wayback.archive-it.org/2358/20120513161747/http://hamdeensabahy.com/

Apple iCloud Facebook Twitter Wikipedia Yahoo News Popular Recent cha... WS-DL Wiki
حمدين صباحي
واحد منا رئيسا لمصر
تطوع فى اللجان الانتخابية
لم يدعمنى اى نظام مصرى او عربى او دولى، واتحدى ان يقدم اى مروج لهذه الشائعات ما يدل على غير ذلك، ومن لديه شيء ضدى ارحب بان يتقدم للجهات المختصة.
من هو حمدين صباحي
تبرع لدعم حمدين صباحي و حملته الانتخابية

May 13, 2012: The page started as on-topic.



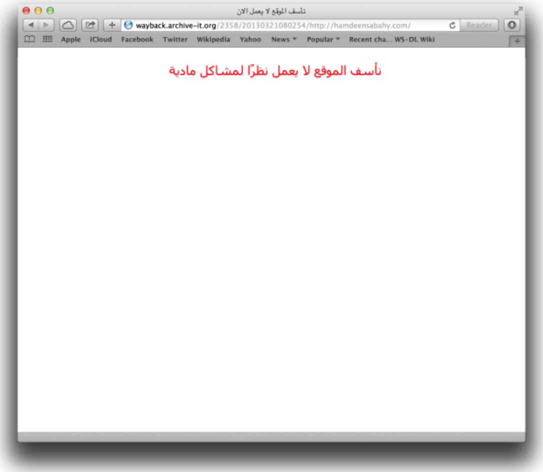
Database Error

wayback.archive-it.org/2358/20120524105356/http://hamdeensabahy.com/

Apple iCloud Facebook Twitter Wikipedia Yahoo News Popular Recent cha... WS-DL Wiki

**Error establishing a database connection**

May 24, 2012: Off-topic due to a database error.



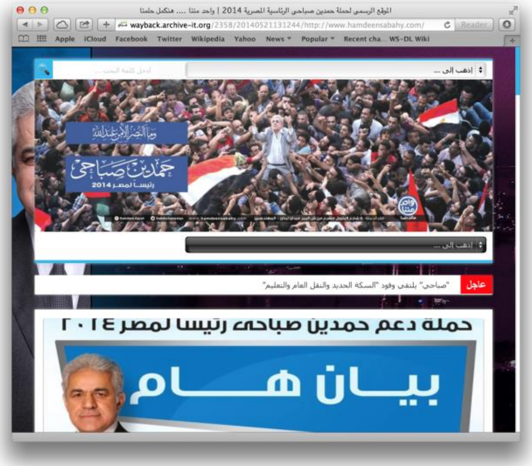
نأسف الموقع لا يعمل الان

wayback.archive-it.org/2358/20130321080254/http://hamdeensabahy.com/

Apple iCloud Facebook Twitter Wikipedia Yahoo News Popular Recent cha... WS-DL Wiki

نأسف الموقع لا يعمل نظرًا لمشاكل مادية

Mar. 21, 2013: Not working because of financial problems.

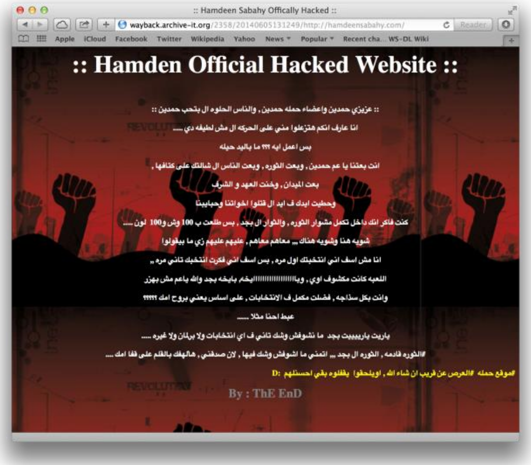


الموقع الرسمى لحملة حمدين صباحي الرئاسية المصرية 2014 | واحد منتا .... هنكمل حلمنا

wayback.archive-it.org/2358/20140521131244/http://www.hamdeensabahy.com/

Apple iCloud Facebook Twitter Wikipedia Yahoo News Popular Recent cha... WS-DL Wiki
إذهب إلى ...
أدخل كلمة البحث ...
وما النصر إلا من عند الله
حمدين صباحي
رئيسا لمصر 2014
عاجل
"صباحي" يلتقي وفود "السكة الحديد والنقل العام والبترول"
حملة دعم حمدين صباحي رئيسا لمصر ٢٠١٤
بيان هام

May 21, 2013: On-topic again



:: Hamdeen Sabahy Offically Hacked ::

wayback.archive-it.org/2358/20140605131249/http://hamdeensabahy.com/

Apple iCloud Facebook Twitter Wikipedia Yahoo News Popular Recent cha... WS-DL Wiki

**:: Hamden Official Hacked Website ::**

:: عزيزي حمدين واعضاء حمله حمدين , والناس الحلوه ال بتحب حمدين ::

انا عارف انكم هتزعلوا مني على الحركه ال مش لطيفه دي .....

بس اعمل ايه ؟؟؟ ما باليد حيله

انت بعتنا يا عم حمدين , وبعت الثوره , وبعت الناس ال شالتك على كتافها ,

بعت الميدان , وخنت العهد و الشرف

وحطيت ايدك ف ايد ال قتلوا اخواتنا وحبايبنا

كنت فاكر انك داخل تكمل مشوار الثوره , والثوار ال بجد , بس طلعت ب 100 وش و100 لون .....

شويه هنا وشويه هناك ,,, معاهم معاهم , عليهم عليهم زي ما بيقولوا

انا مش اسف اني انتخبتك اول مره , بس اسف اني فكرت انتخبك تاني مره ,,

اللعبه كانت مكشوفه اوي , ويا!!!!!!!!!!!!!!يخم, بايخه بجد والله ياعم مش بهزر

وانت بكل سذاجه , فضلت مكمل ف الانتخابات , على اساس يعني بروح امك ؟؟؟؟؟

عبط احنا مثلا ......

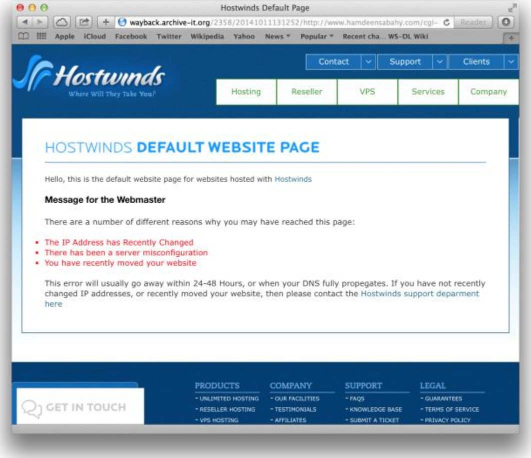
ياريت ياربييييت بجد ما نشوفش وشك تاني ف اي انتخابات ولا برلمان ولا غيره ......

#الثوره قادمه , الثوره ال بجد ,,, اتمني ما اشوفش وشك فيها , لان صدقني , هالهفك بالقلم على قفا امك ....

#موقع حمله #العرض عن قريب ان شاء الله , اويلحقوا يقفلوه بقي احسنلهم :D

By : ThE EnD

June 5, 2014: The site has been hacked



Hostwinds Default Page

wayback.archive-it.org/2358/20141011131252/http://www.hamdeensabahy.com/cgi-

HOSTWINDS **DEFAULT WEBSITE PAGE**

Hello, this is the default website page for websites hosted with Hostwinds

**Message for the Webmaster**

There are a number of different reasons why you may have reached this page:

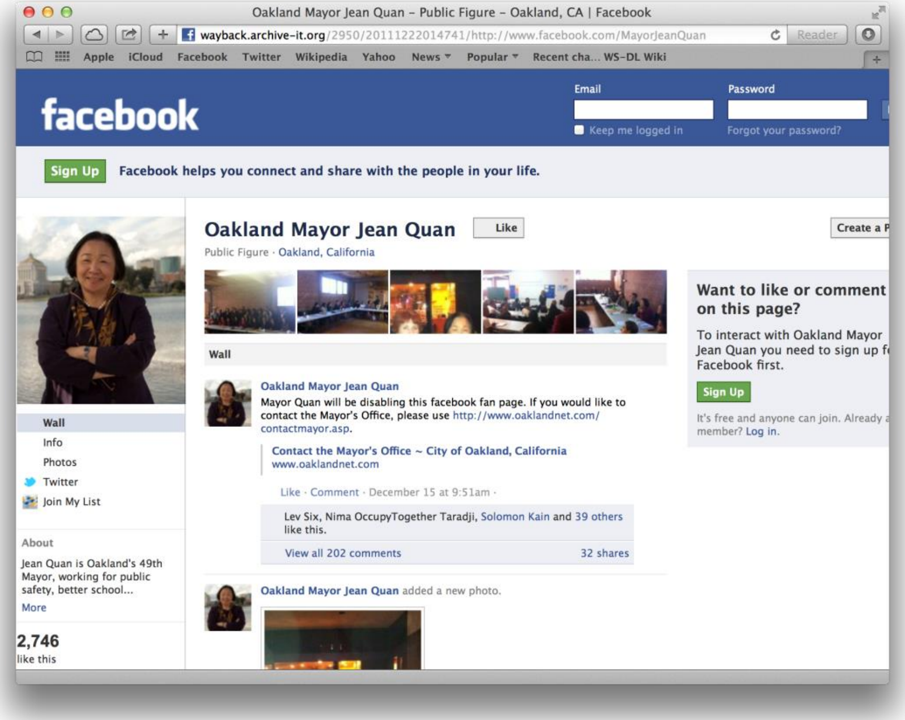
- The IP Address has Recently Changed
- There has been a server misconfiguration
- You have recently moved your website

This error will usually go away within 24-48 Hours, or when your DNS fully propegates. If you have not recently changed IP addresses, or recently moved your website, then please contact the Hostwinds support deparment here

Oct. 10, 2014: The domain has expired.

http://wayback.archive-it.org/2358/*/http://hamdeensabahy.com

# Social media pages can go off-topic



Facebook helps you connect and share with the people in your life.

Oakland Mayor Jean Quan

Public Figure · Oakland, California

Oakland Mayor Jean Quan

Mayor Quan will be disabling this facebook fan page. If you would like to contact the Mayor's Office, please use http://www.oaklandnet.com/contactmayor.asp.

Contact the Mayor's Office ~ City of Oakland, California
www.oaklandnet.com

Like · Comment · December 15 at 9:51am ·

Lev Six, Nima OccupyTogether Taradji, Solomon Kain and 39 others like this.

View all 202 comments 32 shares

Oakland Mayor Jean Quan added a new photo.

About

Jean Quan is Oakland's 49th Mayor, working for public safety, better school...
More

2,746
like this

Want to like or comment on this page?

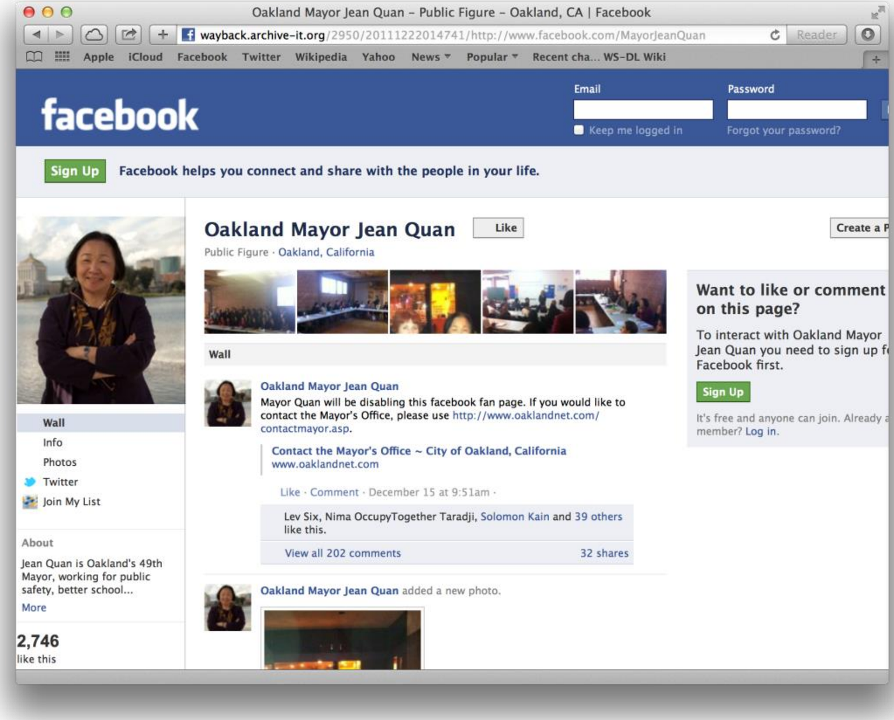
To interact with Oakland Mayor Jean Quan you need to sign up f
Facebook first.

It's free and anyone can join. Already a
member? Log in.

**Dec. 22, 2011: Facebook page was relevant to the Occupy collection**

http://wayback.archive-it.org/2950/*/http://www.facebook.com/MayorJeanQuan

# Social media pages can go off-topic



Oakland Mayor Jean Quan – Public Figure – Oakland, CA | Facebook

wayback.archive-it.org/2950/20111222014741/http://www.facebook.com/MayorJeanQuan

Facebook helps you connect and share with the people in your life.

Oakland Mayor Jean Quan

Public Figure · Oakland, California

Oakland Mayor Jean Quan

Mayor Quan will be disabling this facebook fan page. If you would like to contact the Mayor's Office, please use http://www.oaklandnet.com/contactmayor.asp.

Contact the Mayor's Office ~ City of Oakland, California
www.oaklandnet.com

Like · Comment · December 15 at 9:51am ·

Lev Six, Nima OccupyTogether Taradji, Solomon Kain and 39 others like this.

View all 202 comments 32 shares

Oakland Mayor Jean Quan added a new photo.

About

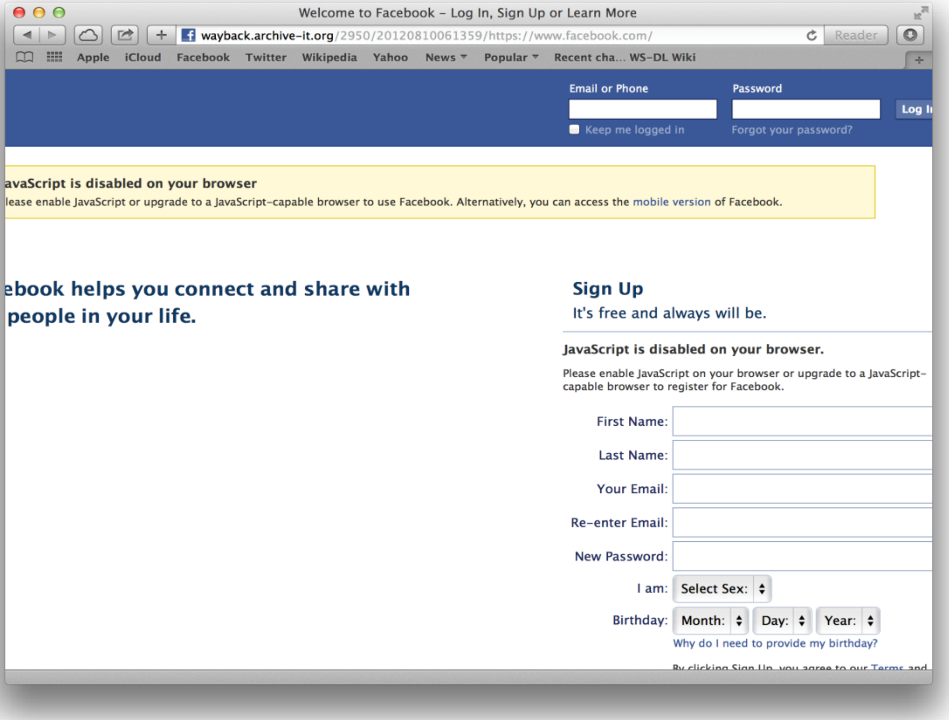
Jean Quan is Oakland's 49th Mayor, working for public safety, better school...
More

2,746
like this

Want to like or comment on this page?

To interact with Oakland Mayor Jean Quan you need to sign up f Facebook first.

It's free and anyone can join. Already a member? Log in.



Welcome to Facebook – Log In, Sign Up or Learn More

wayback.archive-it.org/2950/20120810061359/https://www.facebook.com/

Email or Phone

Password

Keep me logged in

Forgot your password?

avaScript is disabled on your browser

lease enable JavaScript or upgrade to a JavaScript-capable browser to use Facebook. Alternatively, you can access the mobile version of Facebook.

ebook helps you connect and share with people in your life.

Sign Up

It's free and always will be.

JavaScript is disabled on your browser.

Please enable JavaScript on your browser or upgrade to a JavaScript-capable browser to register for Facebook.

First Name:

Last Name:

Your Email:

Re-enter Email:

New Password:

I am: Select Sex:

Birthday: Month: Day: Year:

Why do I need to provide my birthday?

By clicking Sign Up, you agree to our Terms and

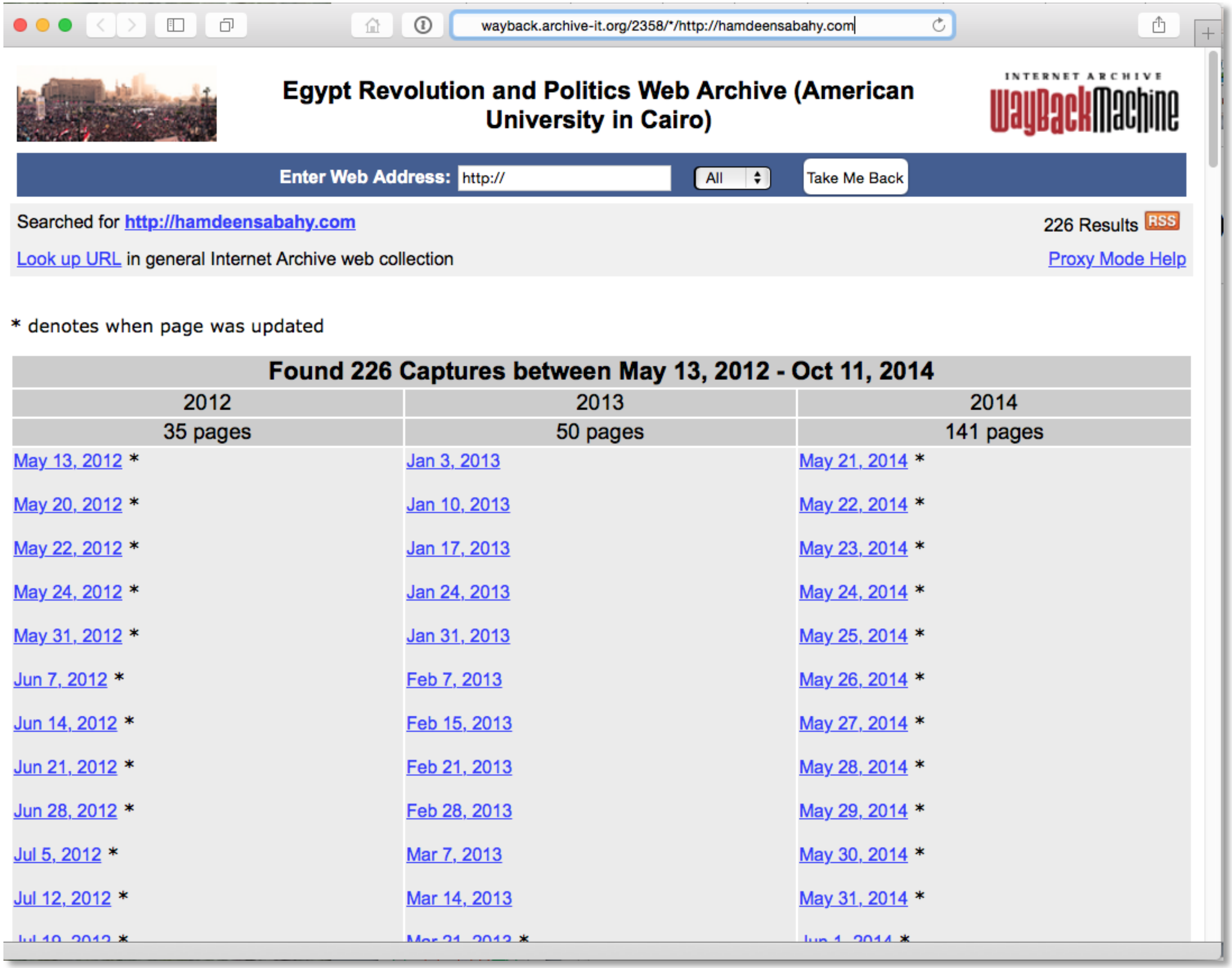
**Dec. 22, 2011: Facebook page was relevant to the Occupy collection**

**Aug. 10, 2012: URI redirects to www.facebook.com**

http://wayback.archive-it.org/2950/*/http://www.facebook.com/MayorJeanQuan

# Classifying web page behavior over time

# A TimeMap is the list of a URI-R's mementos



wayback.archive-it.org/2358/*/http://hamdeensabahy.com
INTERNET ARCHIVE
WayBackMachine
Enter Web Address: http://
All
Take Me Back
RSS

Egypt Revolution and Politics Web Archive (American University in Cairo)

Searched for http://hamdeensabahy.com

226 Results

Look up URL in general Internet Archive web collection

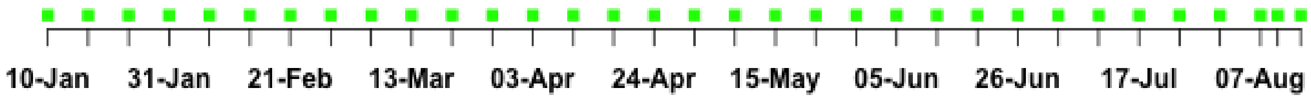
Proxy Mode Help

* denotes when page was updated

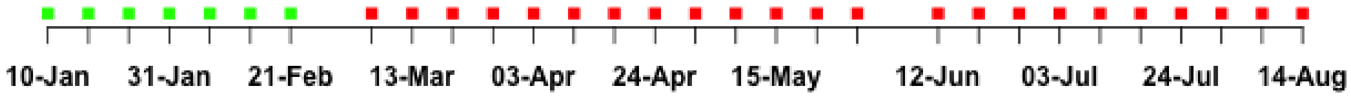
| Found 226 Captures between May 13, 2012 - Oct 11, 2014 | | |
|---|---|---|
| 2012 | 2013 | 2014 |
| 35 pages | 50 pages | 141 pages |
| May 13, 2012 * | Jan 3, 2013 | May 21, 2014 * |
| May 20, 2012 * | Jan 10, 2013 | May 22, 2014 * |
| May 22, 2012 * | Jan 17, 2013 | May 23, 2014 * |
| May 24, 2012 * | Jan 24, 2013 | May 24, 2014 * |
| May 31, 2012 * | Jan 31, 2013 | May 25, 2014 * |
| Jun 7, 2012 * | Feb 7, 2013 | May 26, 2014 * |
| Jun 14, 2012 * | Feb 15, 2013 | May 27, 2014 * |
| Jun 21, 2012 * | Feb 21, 2013 | May 28, 2014 * |
| Jun 28, 2012 * | Feb 28, 2013 | May 29, 2014 * |
| Jul 5, 2012 * | Mar 7, 2013 | May 30, 2014 * |
| Jul 12, 2012 * | Mar 14, 2013 | May 31, 2014 * |
| Jul 19, 2012 * | Mar 21, 2013 * | Jun 1, 2014 * |

# We identified 5 types of TimeMaps

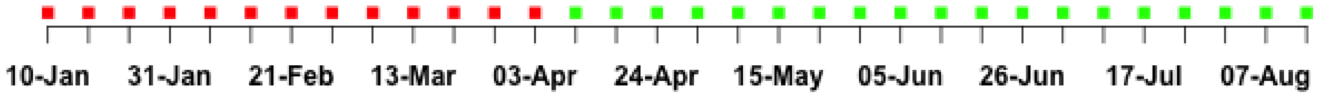
**1. Always On**



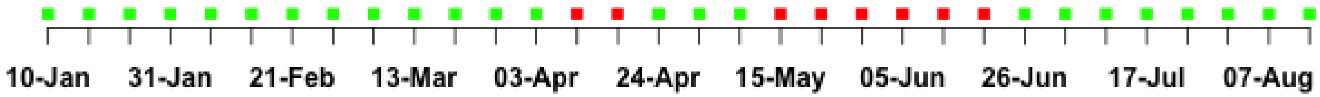
**2. Step Function On**



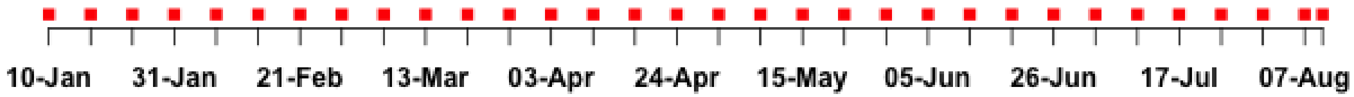
**3. Step Function Off**



**4. Oscillating**

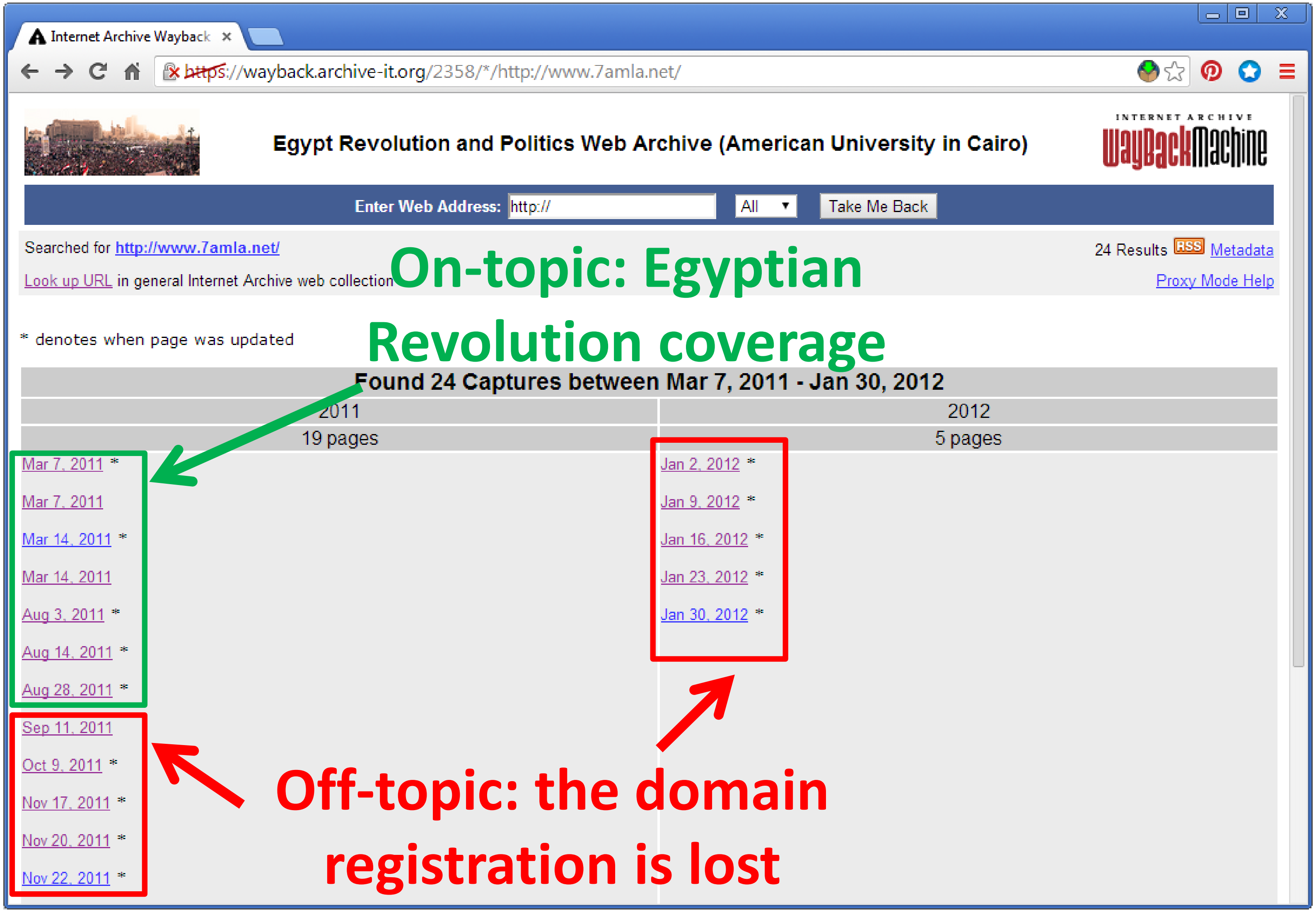


**5. Always Off**



1. wayback.archive-it.org/2950/*/http://occupypsl.org
2. wayback.archive-it.org/2950/*/http://occupygso.tumblr.com
3. wayback.archive-it.org/2950/*/http://occupyashland.com
4. wayback.archive-it.org/2950/*/http://www.indyows.org
5. wayback.archive-it.org/2950/*/http://occupy605.com

# A web page goes off-topic (Step Function On)


Internet Archive Wayback
https://wayback.archive-it.org/2358/*/http://www.7amla.net/
INTERNET ARCHIVE
WaybackMachine
Enter Web Address: http://
All
Take Me Back

Egypt Revolution and Politics Web Archive (American University in Cairo)

Searched for http://www.7amla.net/

Look up URL in general Internet Archive web collection

24 Results RSS Metadata

Proxy Mode Help

* denotes when page was updated

Found 24 Captures between Mar 7, 2011 - Jan 30, 2012

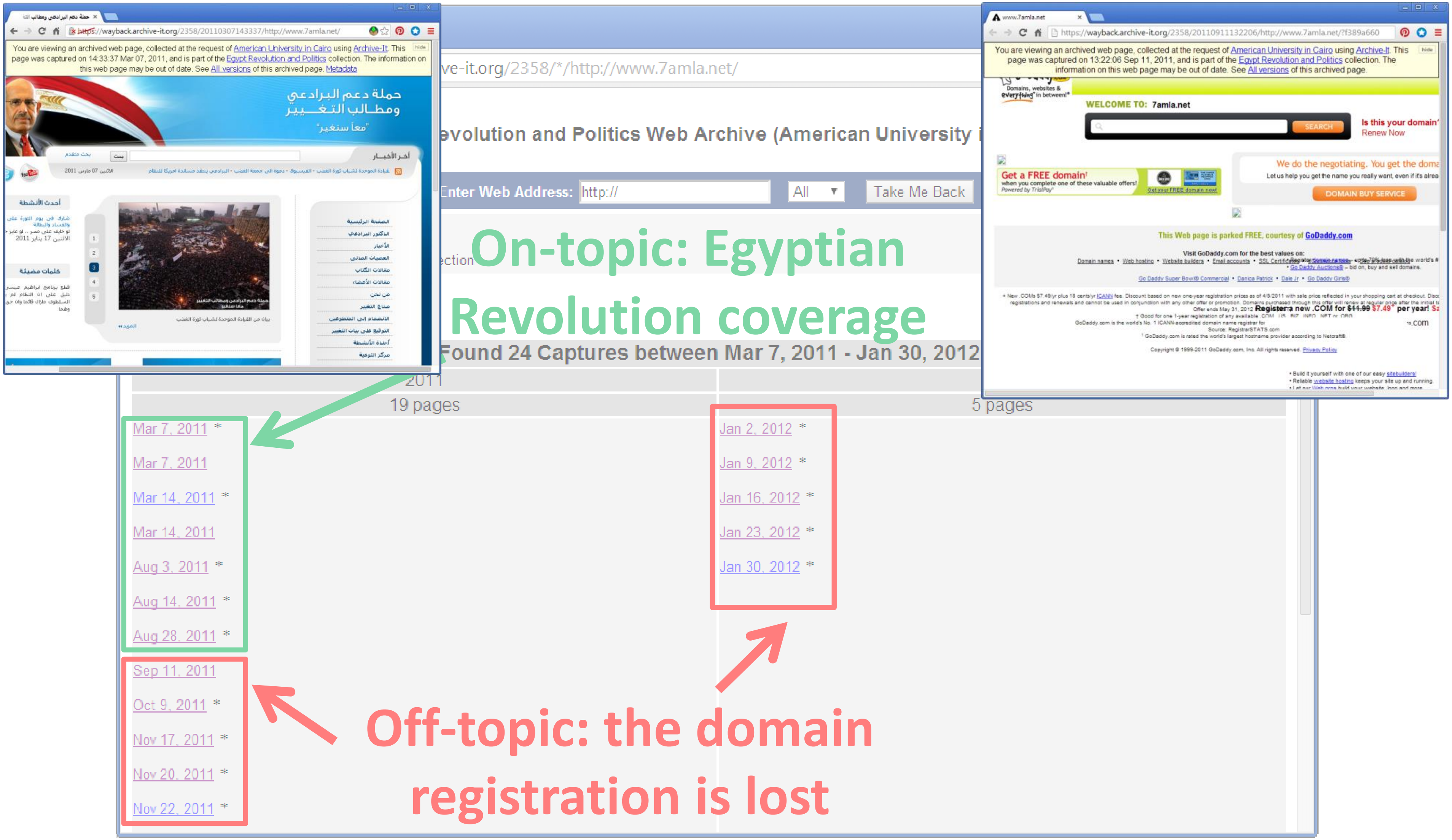
| 2011 | 2012 |
|---|---|
| 19 pages | 5 pages |
| Mar 7, 2011 * | Jan 2, 2012 * |
| Mar 7, 2011 | Jan 9, 2012 * |
| Mar 14, 2011 * | Jan 16, 2012 * |
| Mar 14, 2011 | Jan 23, 2012 * |
| Aug 3, 2011 * | Jan 30, 2012 * |
| Aug 14, 2011 * | |
| Aug 28, 2011 * | |
| Sep 11, 2011 | |
| Oct 9, 2011 * | |
| Nov 17, 2011 * | |
| Nov 20, 2011 * | |
| Nov 22, 2011 * | |

On-topic: Egyptian Revolution coverage

Off-topic: the domain registration is lost

http://wayback.archive-it.org/2358/*/http://www.7amla.net

# A web page goes off-topic (Step Function On)



You are viewing an archived web page, collected at the request of American University in Cairo using Archive-It. This page was captured on 14:33:37 Mar 07, 2011, and is part of the Egypt Revolution and Politics collection. The information on this web page may be out of date. See All versions of this archived page. Metadata

https://wayback.archive-it.org/2358/20110307143337/http://www.7amla.net/
حملة دعم البرادعي ومطالب التغيير
"معاً سنغير"

ve-it.org/2358/*/http://www.7amla.net/

evolution and Politics Web Archive (American University i

Enter Web Address: http:// All Take Me Back

Found 24 Captures between Mar 7, 2011 - Jan 30, 2012

| 2011 | |
|---|---|
| 19 pages | 5 pages |
| Mar 7, 2011 * | Jan 2, 2012 * |
| Mar 7, 2011 | Jan 9, 2012 * |
| Mar 14, 2011 * | Jan 16, 2012 * |
| Mar 14, 2011 | Jan 23, 2012 * |
| Aug 3, 2011 * | Jan 30, 2012 * |
| Aug 14, 2011 * | |
| Aug 28, 2011 * | |
| Sep 11, 2011 | |
| Oct 9, 2011 * | |
| Nov 17, 2011 * | |
| Nov 20, 2011 * | |
| Nov 22, 2011 * | |

You are viewing an archived web page, collected at the request of American University in Cairo using Archive-It. This page was captured on 13:22:06 Sep 11, 2011, and is part of the Egypt Revolution and Politics collection. The information on this web page may be out of date. See All versions of this archived page.

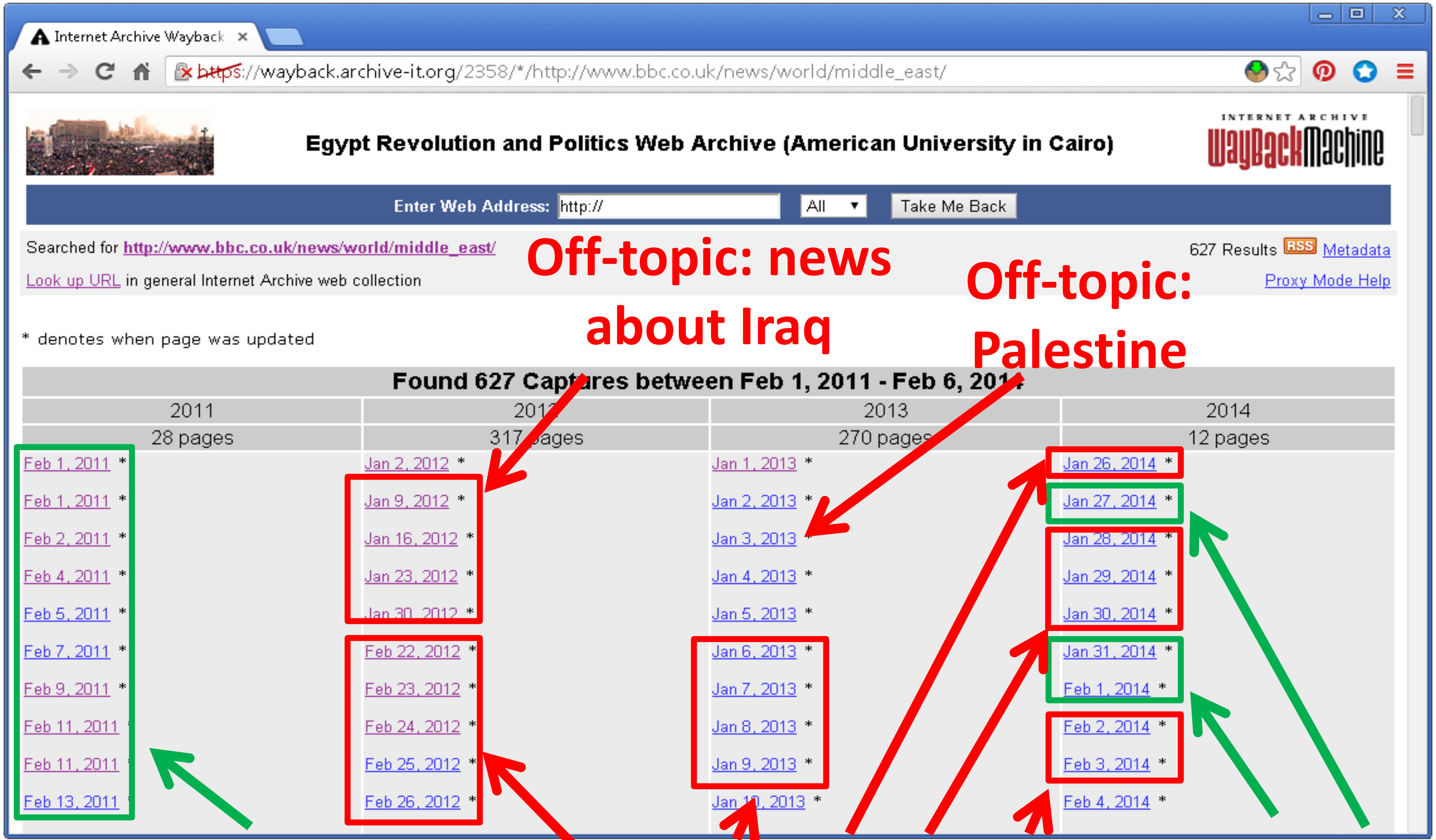
www.7amla.net
https://wayback.archive-it.org/2358/20110911132206/http://www.7amla.net/?f389a660
Domains, websites & everything in between!
WELCOME TO: 7amla.net
SEARCH
Is this your domain
Renew Now
Get a FREE domain!
We do the negotiating. You get the doma
DOMAIN BUY SERVICE
This Web page is parked FREE, courtesy of GoDaddy.com

On-topic: Egyptian Revolution coverage

Off-topic: the domain registration is lost

http://wayback.archive-it.org/2358/*/http://www.7amla.net

# A web page goes off-topic and on-topic many times (Oscillating)



Egypt Revolution and Politics Web Archive (American University in Cairo)

Searched for http://www.bbc.co.uk/news/world/middle_east/

Look up URL in general Internet Archive web collection

* denotes when page was updated

Found 627 Captures between Feb 1, 2011 - Feb 6, 2014

| 2011 | 2012 | 2013 | 2014 |
|---|---|---|---|
| 28 pages | 317 pages | 270 pages | 12 pages |
| Feb 1, 2011 * | Jan 2, 2012 * | Jan 1, 2013 * | Jan 26, 2014 * |
| Feb 1, 2011 * | Jan 9, 2012 * | Jan 2, 2013 * | Jan 27, 2014 * |
| Feb 2, 2011 * | Jan 16, 2012 * | Jan 3, 2013 * | Jan 28, 2014 * |
| Feb 4, 2011 * | Jan 23, 2012 * | Jan 4, 2013 * | Jan 29, 2014 * |
| Feb 5, 2011 * | Jan 30, 2012 * | Jan 5, 2013 * | Jan 30, 2014 * |
| Feb 7, 2011 * | Feb 22, 2012 * | Jan 6, 2013 * | Jan 31, 2014 * |
| Feb 9, 2011 * | Feb 23, 2012 * | Jan 7, 2013 * | Feb 1, 2014 * |
| Feb 11, 2011 * | Feb 24, 2012 * | Jan 8, 2013 * | Feb 2, 2014 * |
| Feb 11, 2011 * | Feb 25, 2012 * | Jan 9, 2013 * | Feb 3, 2014 * |
| Feb 13, 2011 * | Feb 26, 2012 * | Jan 10, 2013 * | Feb 4, 2014 * |

Off-topic: news about Iraq

Off-topic: Palestine

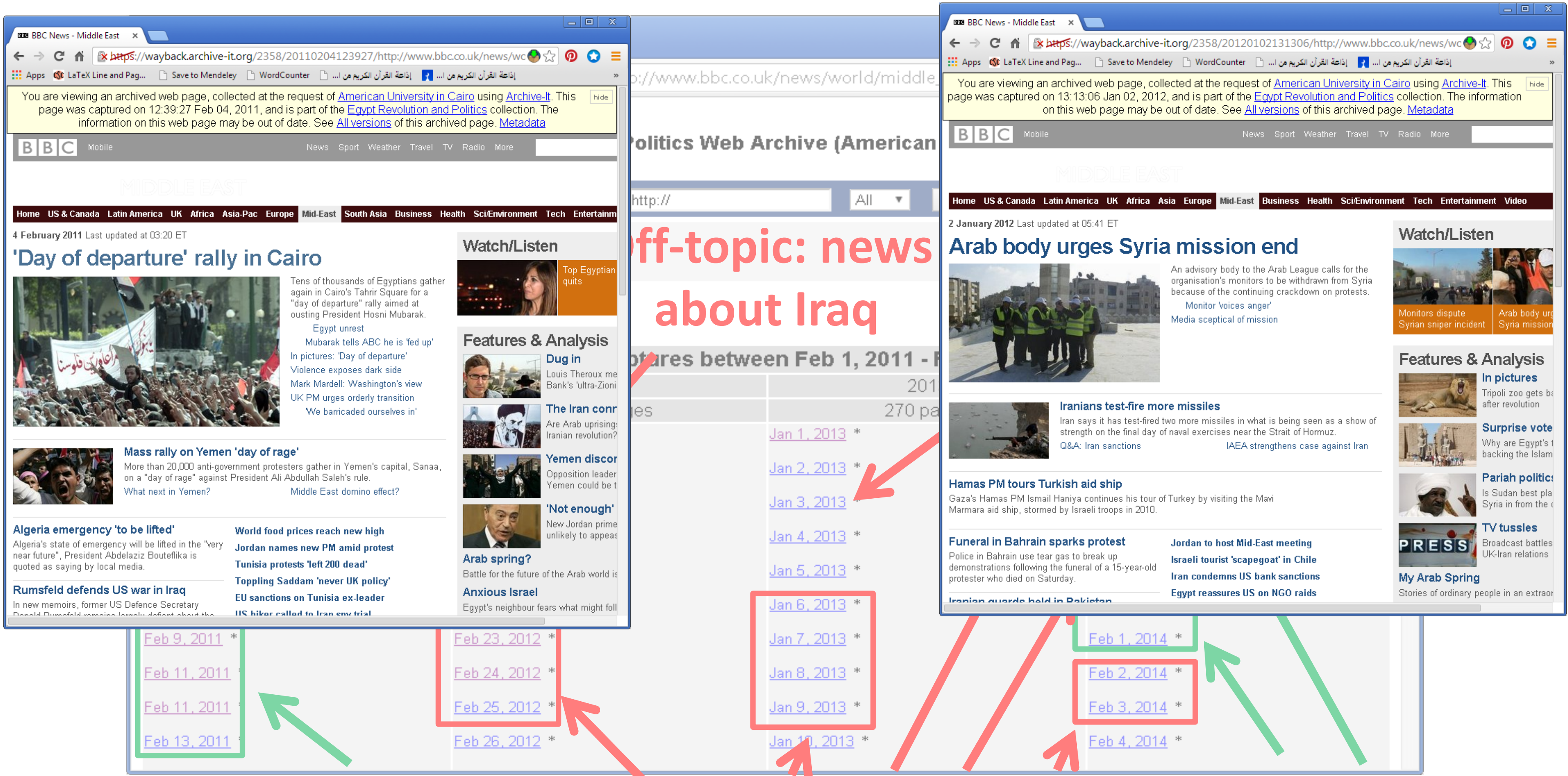
On-topic: Egyptian Revolution coverage

Off-topic: news about Syria

On-topic: Egypt news

http://wayback.archive-it.org/2358/*/http://www.bbc.co.uk/news/world/middle_east/

# A web page goes off-topic and on-topic many times (Oscillating)



BBC News - Middle East

https://wayback.archive-it.org/2358/20110204123927/http://www.bbc.co.uk/news/wc

You are viewing an archived web page, collected at the request of American University in Cairo using Archive-It. This page was captured on 12:39:27 Feb 04, 2011, and is part of the Egypt Revolution and Politics collection. The information on this web page may be out of date. See All versions of this archived page. Metadata

4 February 2011 Last updated at 03:20 ET

'Day of departure' rally in Cairo

Tens of thousands of Egyptians gather again in Cairo's Tahrir Square for a "day of departure" rally aimed at ousting President Hosni Mubarak.

Egypt unrest
Mubarak tells ABC he is 'fed up'
In pictures: 'Day of departure'
Violence exposes dark side
Mark Mardell: Washington's view
UK PM urges orderly transition
'We barricaded ourselves in'

Mass rally on Yemen 'day of rage'

More than 20,000 anti-government protesters gather in Yemen's capital, Sanaa, on a "day of rage" against President Ali Abdullah Saleh's rule.

What next in Yemen? Middle East domino effect?

Algeria emergency 'to be lifted'

Algeria's state of emergency will be lifted in the "very near future", President Abdelaziz Bouteflika is quoted as saying by local media.

Rumsfeld defends US war in Iraq

In new memoirs, former US Defence Secretary

World food prices reach new high
Jordan names new PM amid protest
Tunisia protests 'left 200 dead'
Toppling Saddam 'never UK policy'
EU sanctions on Tunisia ex-leader
US hiker called to Iran spy trial

Watch/Listen
Top Egyptian quits

Features & Analysis
Dug in — Louis Theroux me Bank's 'ultra-Zioni
The Iran conn — Are Arab uprisings Iranian revolution?
Yemen discor — Opposition leader Yemen could be t
'Not enough' — New Jordan prime unlikely to appeas
Arab spring? — Battle for the future of the Arab world is
Anxious Israel — Egypt's neighbour fears what might foll

://www.bbc.co.uk/news/world/middle
Politics Web Archive (American
ptures between Feb 1, 2011 - F
270 pa

Feb 9, 2011 *
Feb 11, 2011 *
Feb 11, 2011 *
Feb 13, 2011 *

Feb 23, 2012 *
Feb 24, 2012 *
Feb 25, 2012 *
Feb 26, 2012 *

Jan 1, 2013 *
Jan 2, 2013 *
Jan 3, 2013 *
Jan 4, 2013 *
Jan 5, 2013 *
Jan 6, 2013 *
Jan 7, 2013 *
Jan 8, 2013 *
Jan 9, 2013 *
Jan 10, 2013 *

Feb 1, 2014 *
Feb 2, 2014 *
Feb 3, 2014 *
Feb 4, 2014 *

Off-topic: news about Iraq

BBC News - Middle East

https://wayback.archive-it.org/2358/20120102131306/http://www.bbc.co.uk/news/wc

You are viewing an archived web page, collected at the request of American University in Cairo using Archive-It. This page was captured on 13:13:06 Jan 02, 2012, and is part of the Egypt Revolution and Politics collection. The information on this web page may be out of date. See All versions of this archived page. Metadata

2 January 2012 Last updated at 05:41 ET

Arab body urges Syria mission end

An advisory body to the Arab League calls for the organisation's monitors to be withdrawn from Syria because of the continuing crackdown on protests.

Monitor voices anger'
Media sceptical of mission

Iranians test-fire more missiles

Iran says it has test-fired two more missiles in what is being seen as a show of strength on the final day of naval exercises near the Strait of Hormuz.

Q&A: Iran sanctions IAEA strengthens case against Iran

Hamas PM tours Turkish aid ship

Gaza's Hamas PM Ismail Haniya continues his tour of Turkey by visiting the Mavi Marmara aid ship, stormed by Israeli troops in 2010.

Funeral in Bahrain sparks protest

Police in Bahrain use tear gas to break up demonstrations following the funeral of a 15-year-old protester who died on Saturday.

Jordan to host Mid-East meeting
Israeli tourist 'scapegoat' in Chile
Iran condemns US bank sanctions
Egypt reassures US on NGO raids

Iranian guards held in Pakistan

Watch/Listen
Monitors dispute Syrian sniper incident
Arab body ur Syria mission

Features & Analysis
In pictures — Tripoli zoo gets b after revolution
Surprise vote — Why are Egypt's backing the Islam
Pariah politic — Is Sudan best pla Syria in from the
TV tussles — Broadcast battles UK-Iran relations
My Arab Spring — Stories of ordinary people in an extraor

On-topic: Egyptian Revolution coverage

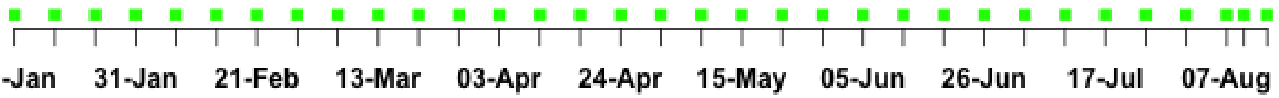
Off-topic: news about Syria

On-topic: Egypt news

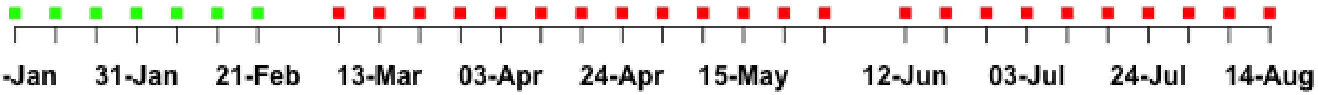
http://wayback.archive-it.org/2358/*/http://www.bbc.co.uk/news/world/middle_east/

# Most TimeMaps are Always On

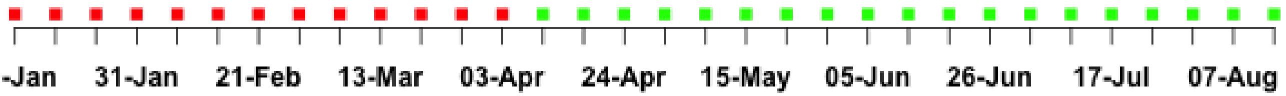
**1. Always On** 74%



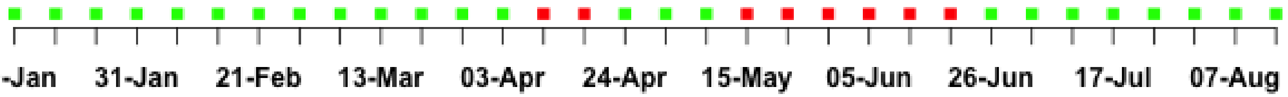
**2. Step Function On** 8-11%



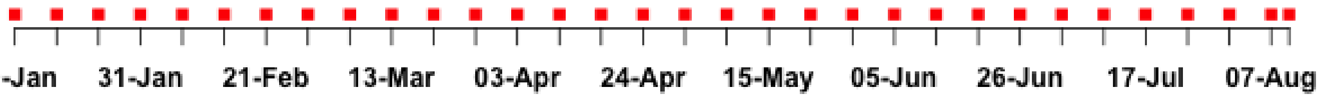
**3. Step Function Off** 0-2%



**4. Oscillating** 6-15%



**5. Always Off** ~0%



1. wayback.archive-it.org/2950/*/http://occupypsl.org
2. wayback.archive-it.org/2950/*/http://occupygso.tumblr.com
3. wayback.archive-it.org/2950/*/http://occupyashland.com
4. wayback.archive-it.org/2950/*/http://www.indyows.org
5. wayback.archive-it.org/2950/*/http://occupy605.com

# Methods for detecting off-topic pages

# Pre-processing

1. Obtain the seed URIs from the front-end interface of Archive-It
2. Obtain the TimeMap of the seed URIs from the CDX file*
3. Extract the HTML of the mementos from the WARC files*
4. Extract the text of the page using the Boilerpipe library
5. Extract terms from the page, using scikit-learn to tokenize, remove stop words, and apply stemming

***locally hosted at ODU**

# Similarity Metrics

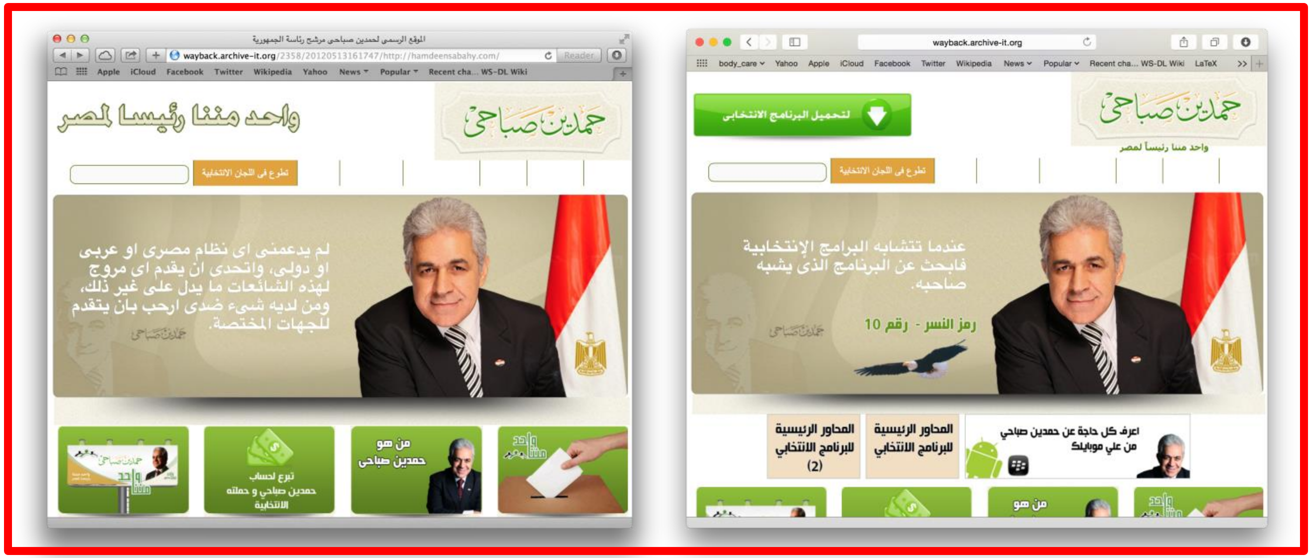
- Textual Content
  - cosine similarity
  - intersection of the most frequent terms
  - Jaccard coefficient

- Semantics
  - Web based kernel function using the search engine (SE)

- Structural
  - the change in number of words
  - the change in content length

# Textual Content

cosine similarity, intersecting the most frequent terms, Jaccard similarity



wayback.archive-it.org/2358/20120513161747/http://hamdeensabahy.com/
واحد منا رئيسا لمصر
حمدين صباحي
wayback.archive-it.org

لم يدعمنى اى نظام مصرى او عربى او دولى، واتحدى ان يقدم اى مروج لهذه الشائعات ما يدل على غير ذلك، ومن لديه شىء ضدى ارحب بان يتقدم للجهات المختصة.

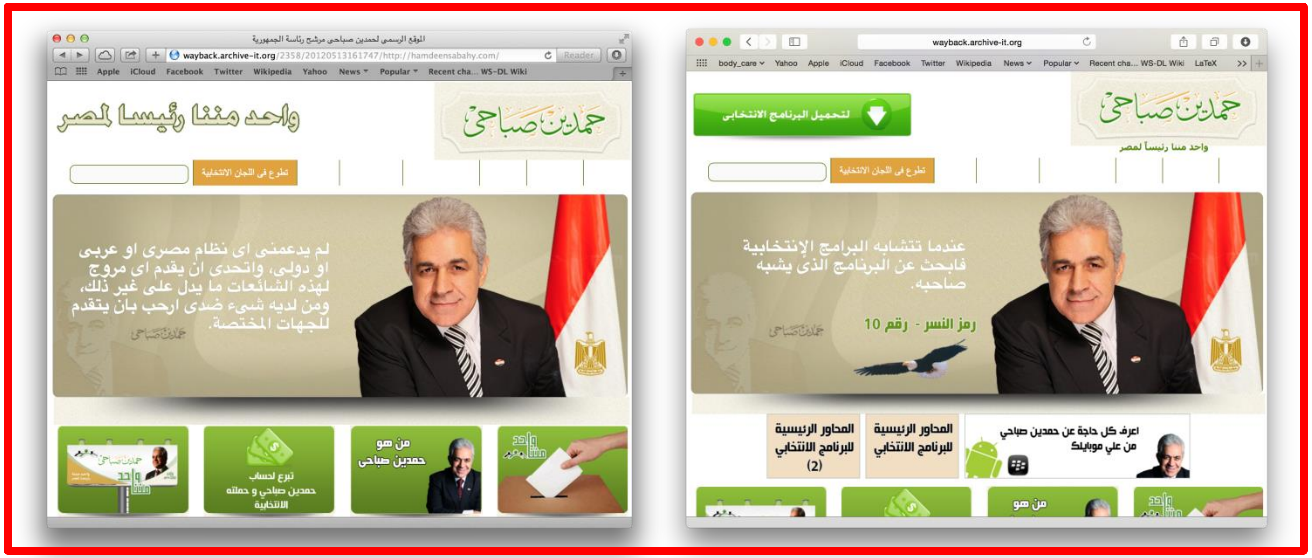
عندما تتشابه البرامج الإنتخابية فابحث عن البرنامج الذى يشبه صاحبه.

زمن النسر - رقم 10

| Method | Similarity |
|---|---|
| cosine | 0.7 |
| TF-Intersection | 0.6 |
| Jaccard | 0.5 |

# Textual Content

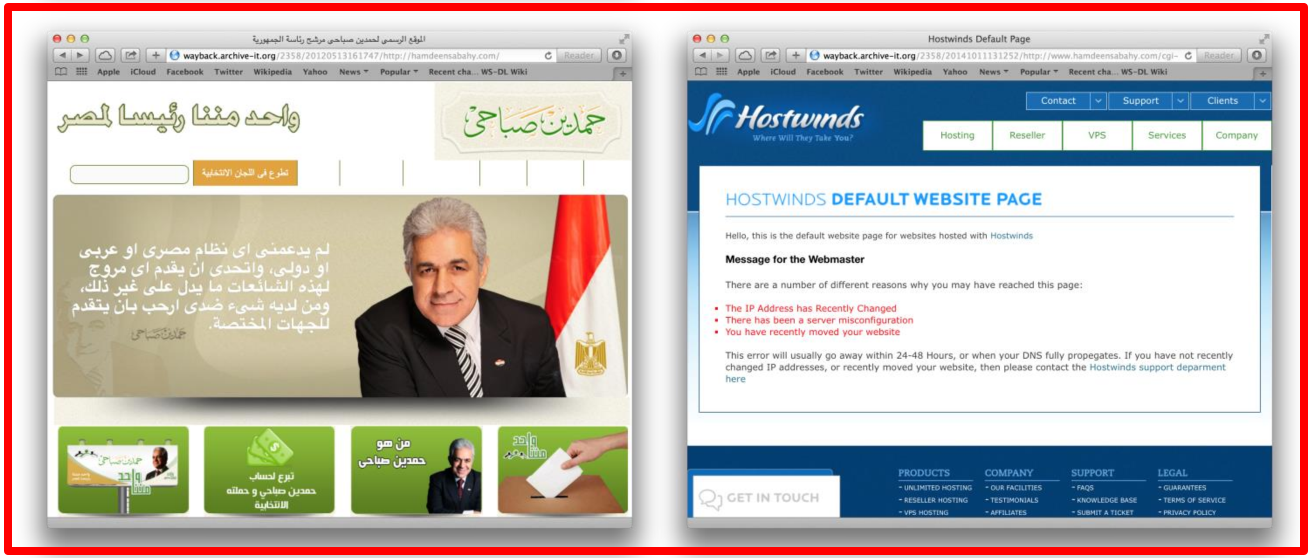
## cosine similarity, intersecting the most frequent terms, Jaccard similarity


wayback.archive-it.org/2358/20120513161747/http://hamdeensabahy.com/
حمدين صباحي
واحد مننا رئيسا لمصر
تطوع فى اللجان الانتخابية
wayback.archive-it.org
تحميل البرنامج الانتخابي
رمز النسر - رقم 10
من هو حمدين صباحي
تبرع لحملة حمدين صباحي و مثلته الانتخابية
المحاور الرئيسية للبرنامج الانتخابي
المحاور الرئيسية للبرنامج الانتخابي (2)
اعرف كل حاجة عن حمدين صباحي من علي موبايلك

لم يدعمني اى نظام مصرى او عربي او دولي، واتحدى ان يقدم اى مروج لهذه الشائعات ما يدل على غير ذلك، ومن لديه شيء ضدى ارحب بان يتقدم للجهات المختصة.

عندما تتشابه البرامج الإنتخابية فابحث عن البرنامج الذى يشبه صاحبه.

| Method | Similarity |
|---|---|
| cosine | 0.7 |
| TF-Intersection | 0.6 |
| Jaccard | 0.5 |


wayback.archive-it.org/2358/20120513161747/http://hamdeensabahy.com/
حمدين صباحي
واحد مننا رئيسا لمصر
Hostwinds Default Page
wayback.archive-it.org/2358/20141011131252/http://www.hamdeensabahy.com/cgi-
Contact
Support
Clients
Hostwinds
Where Will They Take You?
Hosting
Reseller
VPS
Services
Company
GET IN TOUCH
PRODUCTS
COMPANY
SUPPORT
LEGAL

لم يدعمني اى نظام مصرى او عربي او دولي، واتحدى ان يقدم اى مروج لهذه الشائعات ما يدل على غير ذلك، ومن لديه شيء ضدى ارحب بان يتقدم للجهات المختصة.

HOSTWINDS **DEFAULT WEBSITE PAGE**

Hello, this is the default website page for websites hosted with Hostwinds

**Message for the Webmaster**

There are a number of different reasons why you may have reached this page:

- The IP Address has Recently Changed
- There has been a server misconfiguration
- You have recently moved your website

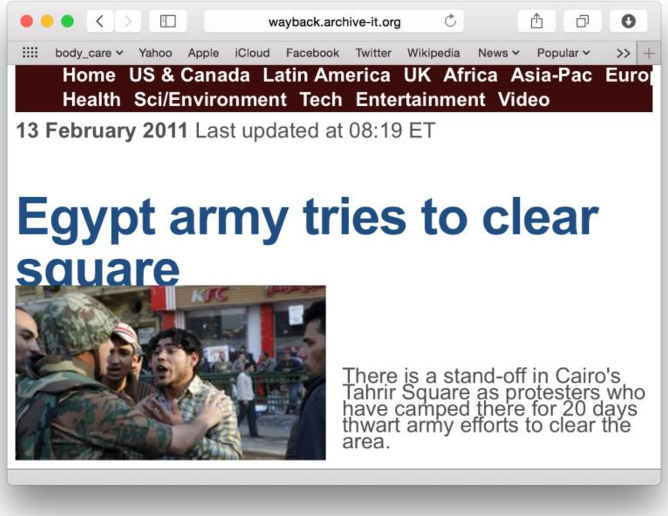
This error will usually go away within 24-48 Hours, or when your DNS fully propegates. If you have not recently changed IP addresses, or recently moved your website, then please contact the Hostwinds support deparment here

| Method | Similarity |
|---|---|
| cosine | 0.0 |
| TF-Intersection | 0.0 |
| Jaccard | 0.0 |

# Semantics of the Text

## Web based kernel function using the search engine (SE)

**Feb. 2011**

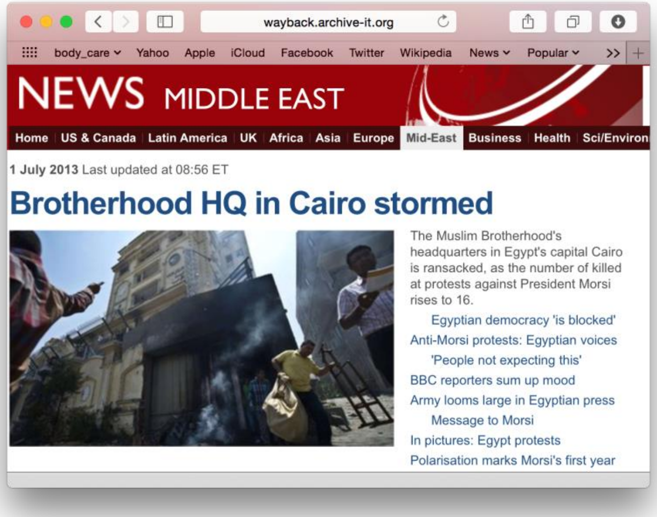

wayback.archive-it.org
body_care
Yahoo
Apple
iCloud
Facebook
Twitter
Wikipedia
News
Popular
>>
Home US & Canada Latin America UK Africa Asia-Pac Europ
Health Sci/Environment Tech Entertainment Video

13 February 2011 Last updated at 08:19 ET

Egypt army tries to clear square

There is a stand-off in Cairo's Tahrir Square as protesters who have camped there for 20 days thwart army efforts to clear the area.

**July 2013**



1 July 2013 Last updated at 08:56 ET

Brotherhood HQ in Cairo stormed

The Muslim Brotherhood's headquarters in Egypt's capital Cairo is ransacked, as the number of killed at protests against President Morsi rises to 16.

Egyptian democracy 'is blocked'

Anti-Morsi protests: Egyptian voices

'People not expecting this'

BBC reporters sum up mood

Army looms large in Egyptian press

Message to Morsi

In pictures: Egypt protests

Polarisation marks Morsi's first year

Tahrir, Egypt, army

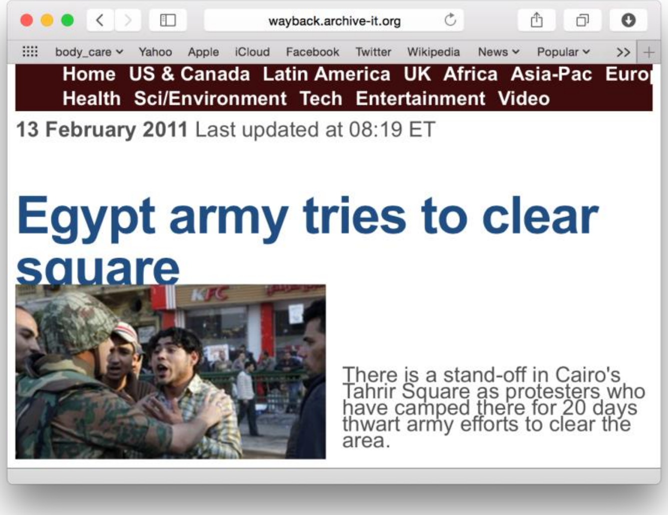
**No term-wise overlap**

Cairo, Morsi, protests

# Semantics of the Text

## Web based kernel function using the search engine (SE)

**Feb. 2011**



wayback.archive-it.org

body_care Yahoo Apple iCloud Facebook Twitter Wikipedia News Popular

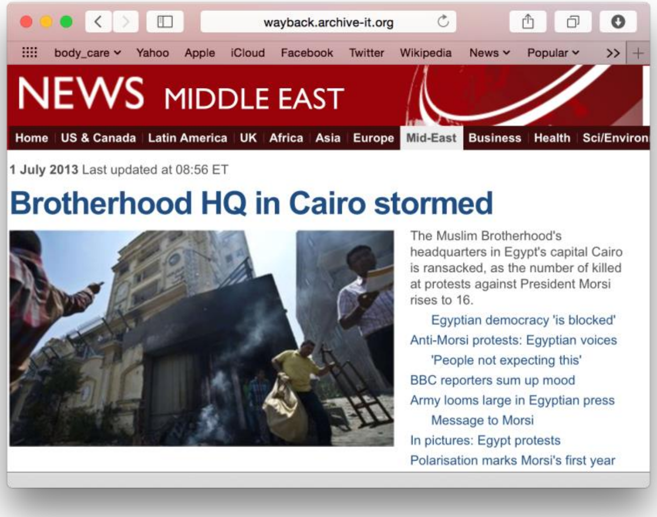
Home US & Canada Latin America UK Africa Asia-Pac Euro
Health Sci/Environment Tech Entertainment Video

**13 February 2011** Last updated at 08:19 ET

**Egypt army tries to clear square**

There is a stand-off in Cairo's Tahrir Square as protesters who have camped there for 20 days thwart army efforts to clear the area.

**July 2013**



wayback.archive-it.org

body_care Yahoo Apple iCloud Facebook Twitter Wikipedia News Popular

NEWS MIDDLE EAST

Home US & Canada Latin America UK Africa Asia Europe Mid-East Business Health Sci/Environ

**1 July 2013** Last updated at 08:56 ET

**Brotherhood HQ in Cairo stormed**

The Muslim Brotherhood's headquarters in Egypt's capital Cairo is ransacked, as the number of killed at protests against President Morsi rises to 16.

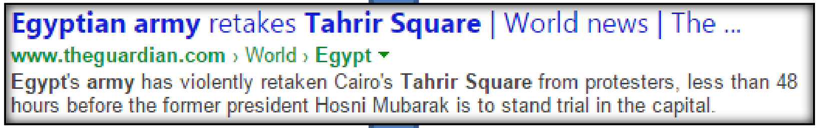
Egyptian democracy 'is blocked'
Anti-Morsi protests: Egyptian voices
'People not expecting this'
BBC reporters sum up mood
Army looms large in Egyptian press
Message to Morsi
In pictures: Egypt protests
Polarisation marks Morsi's first year

Tahrir, Egypt, army

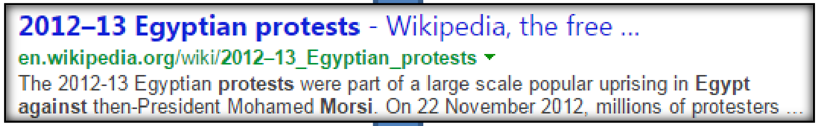
**No term-wise overlap**

Cairo, Morsi, protests

bing



**Egyptian army** retakes **Tahrir Square** | World news | The ...
www.theguardian.com › World › Egypt
Egypt's **army** has violently retaken Cairo's **Tahrir Square** from protesters, less than 48 hours before the former president Hosni Mubarak is to stand trial in the capital.



**2012–13 Egyptian protests** - Wikipedia, the free ...
en.wikipedia.org/wiki/2012–13_Egyptian_protests
The 2012-13 Egyptian **protests** were part of a large scale popular uprising in **Egypt** against then-President Mohamed **Morsi**. On 22 November 2012, millions of protesters ...

Egypt, Tahrir, president, protests, army, Cairo

| Method | Similarity |
|---|---|
| SE-Kernel | 0.7 |

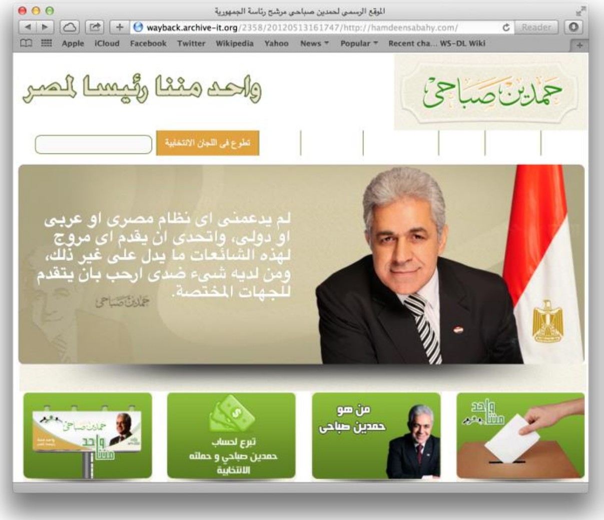
Egypt, protests, Morsi, Cairo, president

Technique inspired by Sahami and Heilman, WWW 2006

# Structural Methods

## no. of words, content-length

100

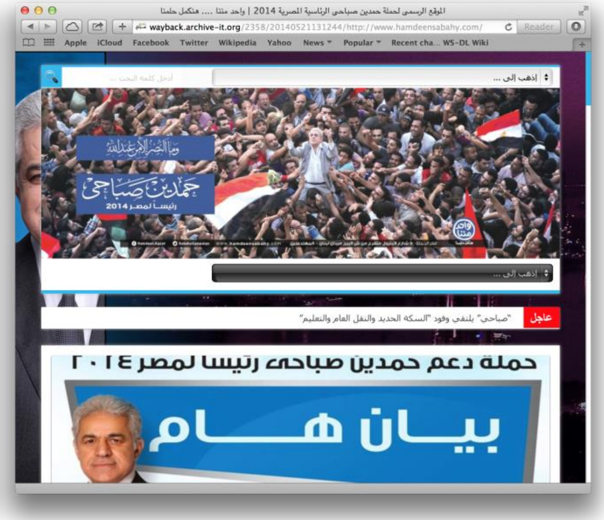


الموقع الرسمي لحمدين صباحي مرشح رئاسة الجمهورية

wayback.archive-it.org/2358/20120513161747/http://hamdeensabahy.com/

لم يدعمنى اى نظام مصرى او عربى او دولى، واتحدى ان يقدم اى مروج لهذه الشائعات ما يدل على غير ذلك، ومن لديه شىء ضدى ارحب بان يتقدم للجهات المختصة.

Apple iCloud Facebook Twitter Wikipedia Yahoo News Popular Recent cha... WS-DL Wiki
Reader
حمدين صباحي
واحد منا رئيسا لمصر
تطوع فى اللجان الانتخابية
حمدين صباحي
من هو حمدين صباحي
تبرع لحساب حمدين صباحي و حملته الانتخابية
واحد منا



الموقع الرسمى لحملة حمدين صباحى الرئاسية المصرية 2014 | واحد منا .... هنكمل حلمنا

wayback.archive-it.org/2358/20140521131244/http://www.hamdeensabahy.com/

عاجل "صباحي" يلتقي وفود "السكة الحديد والنقل العام والتعليم"

Apple iCloud Facebook Twitter Wikipedia Yahoo News Popular Recent cha... WS-DL Wiki
Reader
أدخل كلمة البحث ...
إذهب إلى ...
وما النصر إلا من عند الله
حمدين صباحي
رئيسا لمصر 2014
واحد منا
إذهب إلى ...
حملة دعم حمدين صباحى رئيسا لمصر ٢٠١٤
بيــان هـــــام

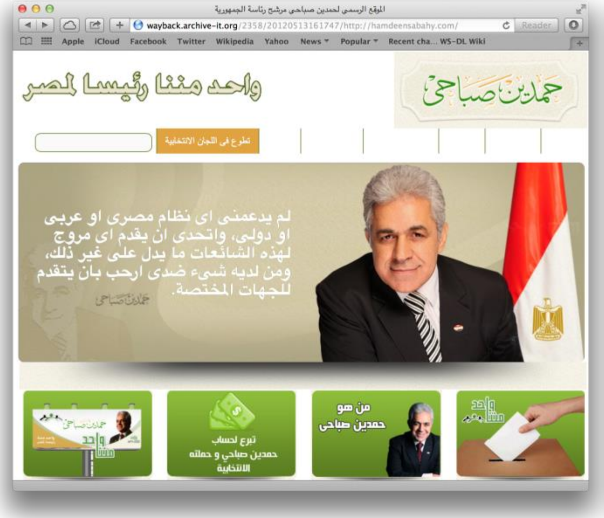
109

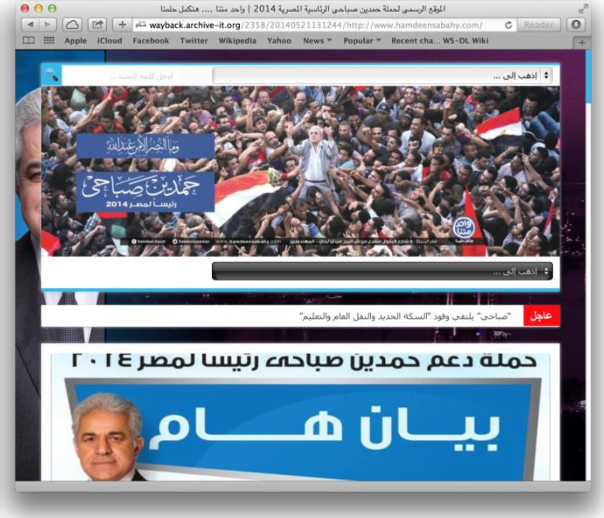
| Method | % change |
|---|---|
| WordCount | 0.09 |

# Structural Methods

## no. of words, content-length

100

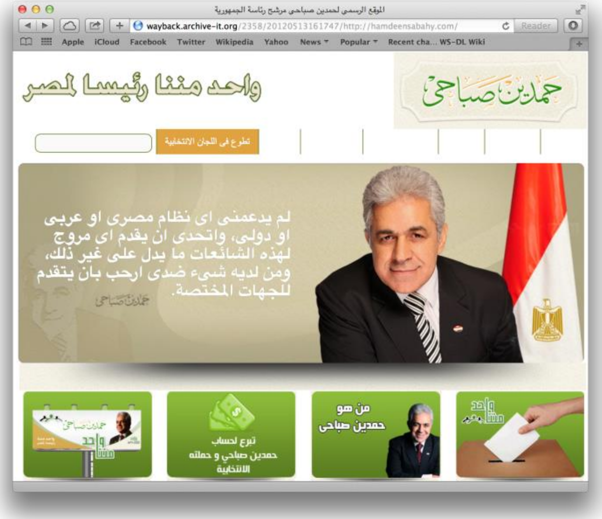

wayback.archive-it.org/2358/20120513161747/http://hamdeensabahy.com/
واحد مننا رئيساً لمصر
حمدين صباحي

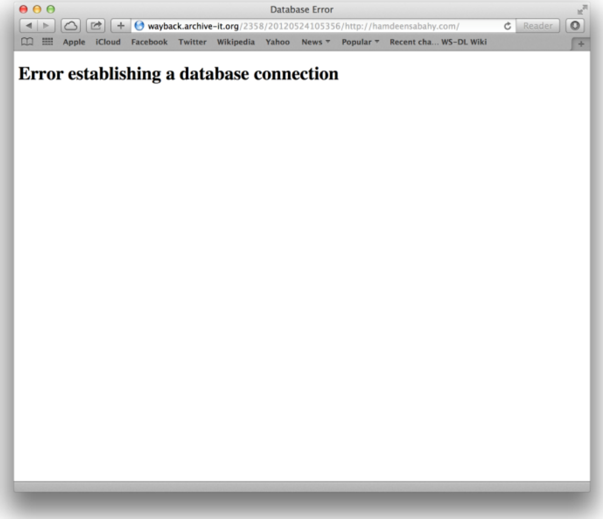

wayback.archive-it.org/2358/20140521131244/http://www.hamdeensabahy.com/
حمدين صباحي
رئيسا لمصر 2014
عاجل
حملة دعم حمدين صباحي رئيسا لمصر ٢٠١٤
بيان هام

109

| Method | % change |
|---|---|
| WordCount | 0.09 |

100


wayback.archive-it.org/2358/20120513161747/http://hamdeensabahy.com/
واحد مننا رئيساً لمصر
حمدين صباحي


Database Error
wayback.archive-it.org/2358/20120524105356/http://hamdeensabahy.com/

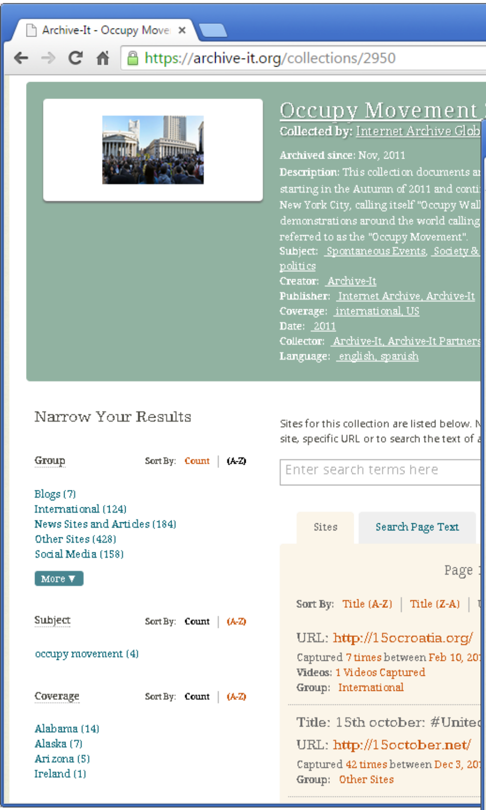
Error establishing a database connection

5

| Method | % change |
|---|---|
| WordCount | -0.95 |

# We built a gold standard data set to evaluate the methods

# We manually labeled 15,760 mementos


Yasmin
Archive-It - Occupy Move ×
https://archive-it.org/collections/2950

Occupy Movement 2011/2012
**Collected by:** Internet Archive Glob

**Archived since:** Nov, 2011
**Description:** This collection documents a
starting in the Autumn of 2011 and conti
New York City, calling itself "Occupy Wall
demonstrations around the world calling
referred to as the "Occupy Movement".
**Subject:** Spontaneous Events, Society &
politics
**Creator:** Archive-It
**Publisher:** Internet Archive, Archive-It
**Coverage:** international, US
**Date:** 2011
**Collector:** Archive-It, Archive-It Partners
**Language:** english, spanish

Narrow Your Results

Group   Sort By: Count | (A-Z)

Blogs (7)
International (124)
News Sites and Articles (184)
Other Sites (428)
Social Media (158)
More ▼

Subject   Sort By: Count | (A-Z)

occupy movement (4)

Coverage   Sort By: Count | (A-Z)

Alabama (14)
Alaska (7)
Arizona (5)
Ireland (1)

Sites for this collection are listed below. N
site, specific URL or to search the text of a

Enter search terms here

Sites | Search Page Text

Page 1

**Sort By:** Title (A-Z) | Title (Z-A) | U

URL: http://15ocroatia.org/
Captured 7 times between Feb 10, 20
**Videos:** 1 Videos Captured
**Group:** International

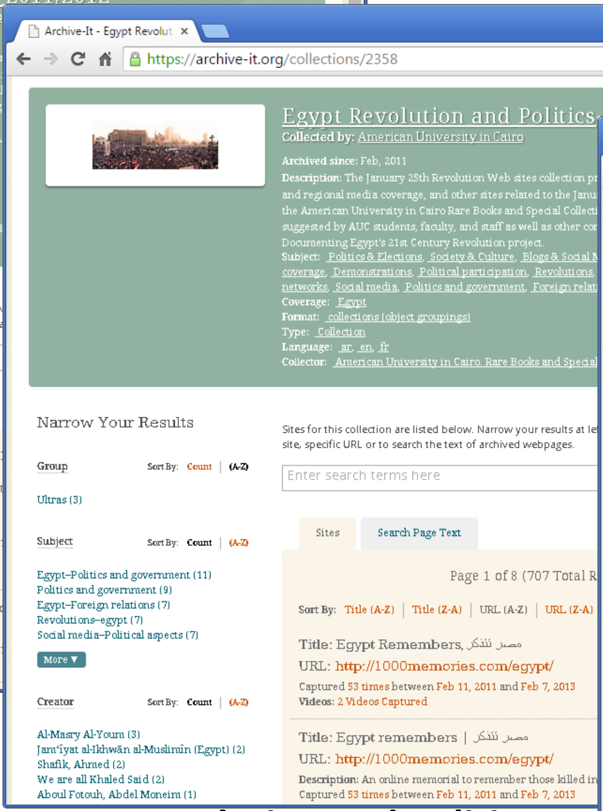
Title: 15th october: #United
URL: http://15october.net/
Captured 42 times between Dec 3, 20
**Group:** Other Sites

**Occupy Movement**

URI-Rs: 255

URI-Ms: 6,570

Off-topic URI-Ms: 458


Yasmin
Archive-It - Egypt Revolut ×
https://archive-it.org/collections/2358

Egypt Revolution and Politics
**Collected by:** American University in Cairo

**Archived since:** Feb, 2011
**Description:** The January 25th Revolution Web sites collection pr
and regional media coverage, and other sites related to the Janu
the American University in Cairo Rare Books and Special Collect
suggested by AUC students, faculty, and staff as well as other co
Documenting Egypt's 21st Century Revolution project.
**Subject:** Politics & Elections, Society & Culture, Blogs & Social M
coverage, Demonstrations, Political participation, Revolutions,
networks, Social media, Politics and government, Foreign relat
**Coverage:** Egypt
**Format:** collections (object groupings)
**Type:** Collection
**Language:** ar, en, fr
**Collector:** American University in Cairo. Rare Books and Special

Narrow Your Results

Group   Sort By: Count | (A-Z)

Ultras (3)

Subject   Sort By: Count | (A-Z)

Egypt–Politics and government (11)
Politics and government (9)
Egypt–Foreign relations (7)
Revolutions–egypt (7)
Social media–Political aspects (7)
More ▼

Creator   Sort By: Count | (A-Z)

Al-Masry Al-Youm (3)
Jamʻīyat al-Ikhwān al-Muslimīn (Egypt) (2)
Shafik, Ahmed (2)
We are all Khaled Said (2)
Aboul Fotouh, Abdel Moneim (1)

Sites for this collection are listed below. Narrow your results at le
site, specific URL or to search the text of archived webpages.

Enter search terms here

Sites | Search Page Text

Page 1 of 8 (707 Total R

**Sort By:** Title (A-Z) | Title (Z-A) | URL (A-Z) | URL (Z-A)

Title: Egypt Remembers, مصر تتذكر
URL: http://1000memories.com/egypt/
Captured 53 times between Feb 11, 2011 and Feb 7, 2013
**Videos:** 2 Videos Captured

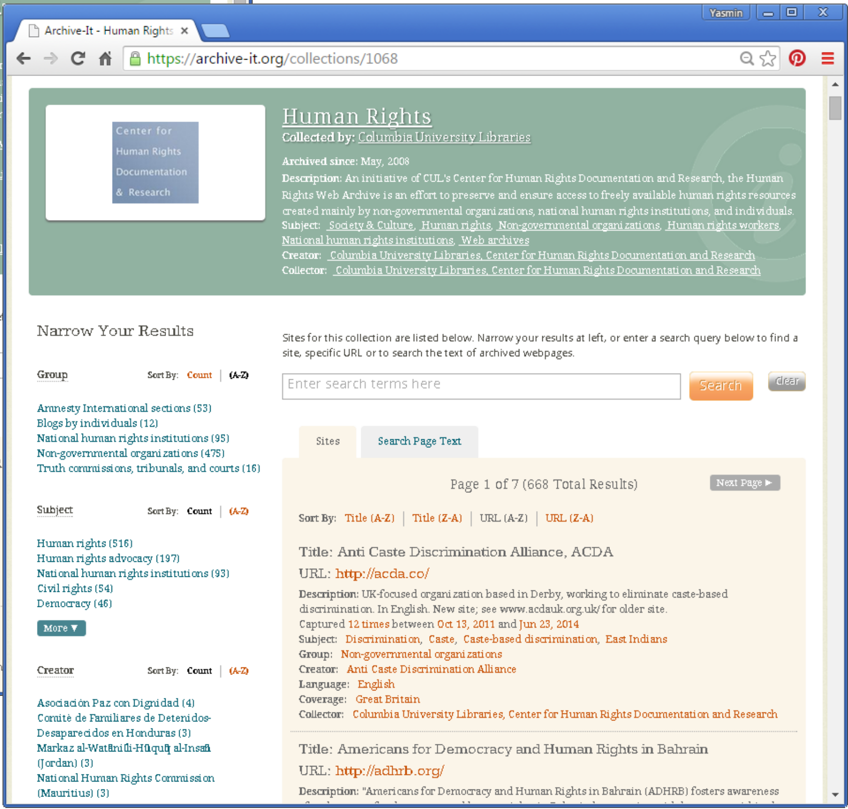
Title: Egypt remembers | مصر تتذكر
URL: http://1000memories.com/egypt/
**Description:** An online memorial to remember those killed in
Captured 53 times between Feb 11, 2011 and Feb 7, 2013

**Egypt Revolution and Politics**

URI-Rs: 136

URI-Ms: 6,886

Off-topic URI-Ms: 384


Yasmin
Archive-It - Human Rights ×
https://archive-it.org/collections/1068
Center for
Human Rights
Documentation
& Research

Human Rights
**Collected by:** Columbia University Libraries

**Archived since:** May, 2008
**Description:** An initiative of CUL's Center for Human Rights Documentation and Research, the Human Rights Web Archive is an effort to preserve and ensure access to freely available human rights resources created mainly by non-governmental organizations, national human rights institutions, and individuals.
**Subject:** Society & Culture, Human rights, Non-governmental organizations, Human rights workers, National human rights institutions, Web archives
**Creator:** Columbia University Libraries, Center for Human Rights Documentation and Research
**Collector:** Columbia University Libraries, Center for Human Rights Documentation and Research

Narrow Your Results

Group   Sort By: Count | (A-Z)

Amnesty International sections (53)
Blogs by individuals (12)
National human rights institutions (95)
Non-governmental organizations (475)
Truth commissions, tribunals, and courts (16)

Subject   Sort By: Count | (A-Z)

Human rights (516)
Human rights advocacy (197)
National human rights institutions (93)
Civil rights (54)
Democracy (46)
More ▼

Creator   Sort By: Count | (A-Z)

Asociación Paz con Dignidad (4)
Comité de Familiares de Detenidos-Desaparecidos en Honduras (3)
Markaz al-Waṭanī lī-Ḥuqūq al-Insān (Jordan) (3)
National Human Rights Commission (Mauritius) (3)

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

Enter search terms here   Search   Clear

Sites | Search Page Text

Page 1 of 7 (668 Total Results)   Next Page ▶

**Sort By:** Title (A-Z) | Title (Z-A) | URL (A-Z) | URL (Z-A)

Title: Anti Caste Discrimination Alliance, ACDA
URL: http://acda.co/
**Description:** UK-focused organization based in Derby, working to eliminate caste-based discrimination. In English. New site; see www.acdauk.org.uk/ for older site.
Captured 12 times between Oct 13, 2011 and Jun 23, 2014
**Subject:** Discrimination, Caste, Caste-based discrimination, East Indians
**Group:** Non-governmental organizations
**Creator:** Anti Caste Discrimination Alliance
**Language:** English
**Coverage:** Great Britain
**Collector:** Columbia University Libraries, Center for Human Rights Documentation and Research

Title: Americans for Democracy and Human Rights in Bahrain
URL: http://adhrb.org/
**Description:** "Americans for Democracy and Human Rights in Bahrain (ADHRB) fosters awareness

**Columbia Univ. Human Rights collection**

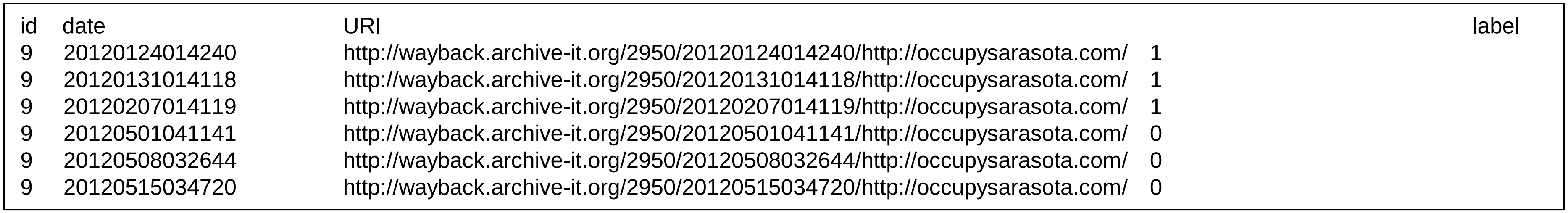
URI-Rs: 198

URI-Ms: 2,304

Off-topic URI-Ms: 94

# Manually labeled set is available

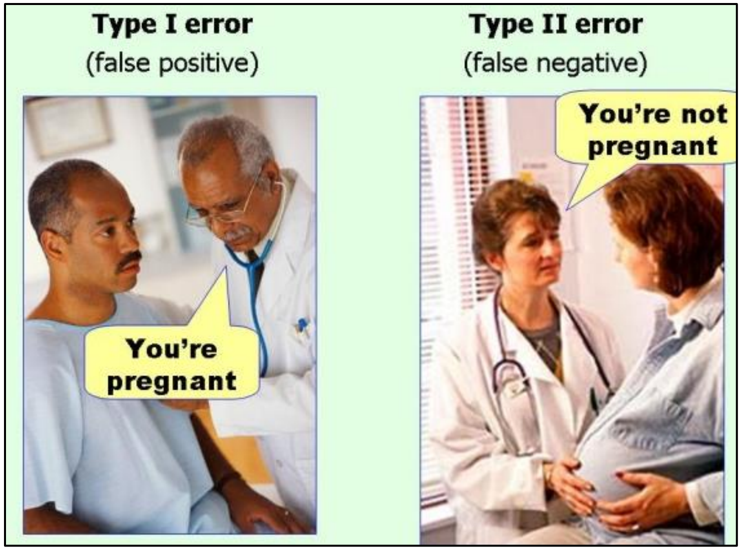
https://github.com/yasmina85/OffTopic-Detection/tree/master/collections_dataset



| id | date | URI | label |
|---|---|---|---|
| 9 | 20120124014240 | http://wayback.archive-it.org/2950/20120124014240/http://occupysarasota.com/ | 1 |
| 9 | 20120131014118 | http://wayback.archive-it.org/2950/20120131014118/http://occupysarasota.com/ | 1 |
| 9 | 20120207014119 | http://wayback.archive-it.org/2950/20120207014119/http://occupysarasota.com/ | 1 |
| 9 | 20120501041141 | http://wayback.archive-it.org/2950/20120501041141/http://occupysarasota.com/ | 0 |
| 9 | 20120508032644 | http://wayback.archive-it.org/2950/20120508032644/http://occupysarasota.com/ | 0 |
| 9 | 20120515034720 | http://wayback.archive-it.org/2950/20120515034720/http://occupysarasota.com/ | 0 |

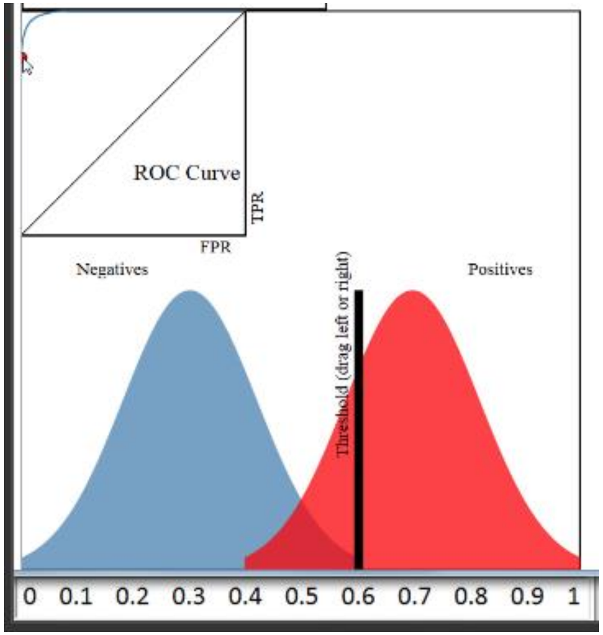
Future work: convert to annotated/extended TimeMap format

# Evaluation Metrics



- False positives (FP)
  - on-topic labeled as off-topic
- False negatives (FN)
  - off-topic labeled as on-topic
- Accuracy (ACC)
  - proportion of correct classifications
  - (TP + TN)/(TP + FP + FN + TN)

- $F_1$ score
  - weighted average of precision and recall
  - 2TP/(2TP + FP + FN)
- AUC
  - area under the ROC curve
  - ROC - plots false positive rate vs. true positive rate



# Evaluation Setup

- Assumed first memento was on-topic
- Tested each method on each labeled memento at 21 different thresholds
- Combined two methods ('OR') to find best combination method
  - 15 combinations
  - 6,615 tests (15 combinations x 21 thresholds x 21 thresholds)
- Averaged the results at each threshold over the three collections

# Cosine Similarity performed well

| Similarity Measure | Threshold | FP | FN | FP+FN | ACC | F1 ▼ | AUC |
|---|---|---|---|---|---|---|---|
| Cosine\|WordCount | 0.10\|-0.85 | 24 | 10 | 34 | 0.987 | 0.906 | 0.968 |
| Cosine\|SEKernel | 0.10\|0.00 | 6 | 35 | 40 | 0.990 | 0.901 | 0.934 |
| Cosine | 0.15 | 31 | 22 | 53 | 0.983 | 0.881 | 0.961 |
| WordCount\|SEKernel | -0.80\|0.00 | 14 | 27 | 42 | 0.985 | 0.818 | 0.885 |
| WordCount | -0.85 | 6 | 44 | 50 | 0.982 | 0.806 | 0.870 |
| SEKernel | 0.05 | 64 | 83 | 147 | 0.965 | 0.683 | 0.865 |
| Bytes | -0.65 | 28 | 133 | 161 | 0.962 | 0.584 | 0.746 |
| Jaccard | 0.05 | 74 | 86 | 159 | 0.962 | 0.538 | 0.809 |
| TF-Intersection | 0.00 | 49 | 104 | 153 | 0.967 | 0.537 | 0.740 |

# Finding off-topic pages in other Archive-It collections

# Applied best method to 11 Archive-It collections

- Cosine|Word Count with 0.10|-0.85 thresholds
- Collection Characteristics
  - governmental, event-based, theme-based
  - time spans of 1 week - 7 years
  - 35 - 1459 URI-Rs
  - 118 - 10,283 URI-Ms

# Average precision of 0.92 on 11 Archive-It collections

| ID | Collection | URI-Rs | URI-Ms | Off-topic URI-Ms | Affected URI-Rs | TP | FP | P |
|---|---|---|---|---|---|---|---|---|
| 2893 | Global Food Crisis | 65 | 3063 | 22 | 7 | 22 | 0 | 1.000 |
| 1084 | Government in Alaska | 68 | 506 | 16 | 4 | 16 | 0 | 1.000 |
| 2966 | Virginia Tech Shootings | 239 | 1670 | 24 | 2 | 24 | 0 | 1.000 |
| 2017 | Wikileaks 2010 Document | 35 | 2360 | 107 | 8 | 107 | 0 | 1.000 |
| 2323 | Jasmine Revolution 2011 | 231 | 4076 | 114 | 31 | 107 | 7 | 0.939 |
| 1827 | IT Historical Resource | 1459 | 10,283 | 59 | 34 | 45 | 14 | 0.763 |
| 1475 | Human Rights Document | 147 | 1530 | 54 | 20 | 39 | 15 | 0.722 |
| 1826 | Maryland State Document | 69 | 184 | 0 | 0 | - | - | - |
| 694 | April 16 Archive | 35 | 118 | 0 | 0 | - | - | - |
| 2535 | Brazilian School Shooting | 476 | 1092 | 0 | 0 | - | - | - |
| 2823 | Russia Plane Crash | 65 | 447 | 0 | 0 | - | - | - |

# Summary

- We presented six methods for measuring similarity between pages:
  - cosine similarity
  - Jaccard similarity
  - intersection of the most 20 frequent terms
  - Web-based kernel function
  - change in number of words
  - change in content length
- We tested the approaches on a gold standard data set from three Archive-It collections
- We evaluated best approach on 11 diverse Archive-It collections

# Findings

- Combining cosine similarity at threshold 0.10 and change in size using word count at threshold −0.85 gives the best performance

- Cosine similarity at threshold = 0.15 is the best single method

- Using the combined method, we achieved 0.92 average precision on 11 Archive-It collections

# Tool for detecting off-topic pages

- A python command-line tool for suggesting off-topic pages in web archives
  - Cosine Similarity
  - default threshold is 0.15
  - operates on live TimeMaps

Available at

https://github.com/yasmina85/OffTopic-Detection

# Detecting off-topic pages in an Archive-It Collection (Maryland State Docs)

```
% python detect_off_topic.py -i 1826 -th 0.15

extracting seed list
…
http://agroecol.umd.edu/Research/index.cfm
http://casademaryland.org
…
50 URIs are extracted from collection https://archive-it.org/collections/1826
Downloading timemap using uri http://wayback.archive-
it.org/1826/timemap/link/http://agroecol.umd.edu/Research/index.cfm
Downloading timemap using uri http://wayback.archive-it.org/1826/timemap/link/http://casademaryland.org
…
Downloading 4 mementos out of 306
Downloading 14 mementos out of 306
…
Detecting off-topic mementos

Similarity     memento_uri
0.0            http://wayback.archive-it.org/1826/20131220205908/http://www.mncppc.org/commission_home.html/
0.0            http://wayback.archive-it.org/1826/20141118195815/http://www.mncppc.org/commission_home.html/
```

This was run live after we did the evaluation, so now there are off-topic mementos

# Detecting off-topic pages in a single TimeMap

```
% python detect_off_topic.py -t https://wayback.archive-it.org/2358/timemap/link/http://hamdeensabahy.com/

Downloading 0 mementos out of 270
http://wayback.archive-it.org/2358/20140524131241/http://www.hamdeensabahy.com/
http://wayback.archive-it.org/2358/20130321080254/http://hamdeensabahy.com/
http://wayback.archive-it.org/2358/20130621131337/http://www.hamdeensabahy.com/
…
Downloading 270 mementos out of 270
…
Extracting text from the html
…
Detecting off-topic mementos

Similarity      memento_uri
0.0509170839413         http://wayback.archive-it.org/2358/20140524131241/http://www.hamdeensabahy.com/
0.0                     http://wayback.archive-it.org/2358/20130321080254/http://hamdeensabahy.com/
0.0368021561791         http://wayback.archive-it.org/2358/20130621131337/http://www.hamdeensabahy.com/
0.12899637517           http://wayback.archive-it.org/2358/20140602131307/http://hamdeensabahy.com/
…
```

# Future Work

- Enhancements to the detection tool
  - add WordCount method
  - allow input of local CDX and WARC files

- Determine collection aboutness
  - detect off-topic URI-Rs in a collection

# Tools for Managing Seed URIs
## (Detecting Off-Topic Pages)

**Yasmin AlNoamany, Michele C. Weigle, Michael L. Nelson**

Old Dominion University

Web Science and Digital Libraries Group

http://ws-dl.cs.odu.edu/, @WebSciDL

Python Tool: https://github.com/yasmina85/OffTopic-Detection

**Web Archiving Collaboration: New Tools and Models**
**June 4-5, 2015**