

6-4-2015

Tools Managing Seed URLs (Detecting Off-Topic Pages)

Yasmin AlNoamany
Old Dominion University

Michele C. Weigle
Old Dominion University, mweigle@odu.edu

Michael L. Nelson
Old Dominion University, mnelson@odu.edu

Follow this and additional works at: https://digitalcommons.odu.edu/computerscience_presentations

 Part of the [Archival Science Commons](#), and the [Computer Sciences Commons](#)

Recommended Citation

AlNoamany, Yasmin; Weigle, Michele C.; and Nelson, Michael L., "Tools Managing Seed URLs (Detecting Off-Topic Pages)" (2015). *Computer Science Presentations*. 32.
https://digitalcommons.odu.edu/computerscience_presentations/32

This Book is brought to you for free and open access by the Computer Science at ODU Digital Commons. It has been accepted for inclusion in Computer Science Presentations by an authorized administrator of ODU Digital Commons. For more information, please contact digitalcommons@odu.edu.



OLD DOMINION
UNIVERSITY

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 hosts curated web collections

The screenshot shows the Archive-It website interface. At the top, there's a navigation bar with 'HOME', 'EXPLORE', 'LEARN MORE', and 'CONTACT US'. Below this is a banner for a webinar: 'Welcome to Archive-It! Attend a live informational webinar and demo to learn more about the service'. To the right of the banner, it says 'Contact Us to sign up for an upcoming session:' with dates 'Jan 07 2014, 11:30 AM PST' and 'Jan 21 2014, 11:30 AM PST'. Below the banner is a search bar for 'Explore Collections' with a 'Search' button and a link to 'Show All Collections'. The main content area displays three collection thumbnails. The first is 'Climate change and environmental policy' by Stanford University, Social Sciences Resource Group. The second is 'Columbia University Human Rights Web Archive' by Columbia University Libraries. The third is 'Government in Alaska Web Archive' by Alaska State Library. Each thumbnail includes a small image and a brief description of the collection's focus.

> 3,000
collections

~340
institutions

> 10B archived
pages

Curator's view of Archive-It

The screenshot shows the Archive-It curator interface for the 'Egyptian Revolution' collection. The page is titled 'Egyptian Revolution (The beginning)' and is publicly visible. It features a 'Collection Management' section with options to activate, deactivate, or mark dormant the collection. There are also sections for 'Seed Management' and 'Crawling Activity'.

Collection Management

Publicly visible?

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

[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

[Previous](#) [Next](#) [Cancel](#)

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 extension 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

The image displays two screenshots from the Archive-IT web interface, illustrating how curators specify crawl parameters.

Left Screenshot: Specify Frequency

The URL is <https://partner.archive-it.org/archiveit/partner/createCollection/schedule.html?acc...>

The page title is "Specify Frequency". It includes navigation buttons: "Previous", "Next", and "Cancel".

Text: "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."

Options (radio buttons):

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

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

Buttons: "Previous", "Next", "Cancel"

Right Screenshot: Egyptian Revolt

The URL is <https://partner.archive-it.org/archiveit/partner/collection/seed/settings.html?seedId...>

The page title is "Egyptian Revolt". It includes navigation buttons: "Home", "Collections", "Crawls", "Reports", "Access", "Help Documentation", "Submit a Question".

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

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

Buttons: "Deactivate", "Verify Now", "Create New Group"

Settings:

- Frequency: Daily
- Activation Status: Active
- Status On Live Web: Unchecked
- Public Site: Show on public site
- Seed Type: Default
- Group: Crawl one page only

Current tools measure HTTP events, not "aboutness"

The screenshot shows a web browser window displaying a Crawl Report from Archive-It. The report is for a 'Daily' crawl (ID #20140127001203292) titled 'Egyptian Revolution (The beginning)'. It was started on January 26, 2014, at 7:12:03 PM and completed at 7:15:13 PM. The report is viewed in the 'File Types' tab, which shows a table of file types and their corresponding data sizes. A 'Help on Reports' sidebar is visible on the right, explaining that 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 and easily opened in Excel. Links are provided for learning more about reports, host reports, robots.txt, and QA reports.

Egyptian Revolution (The beginning) **Crawl Report**
Daily (ID #20140127001203292)
Started: January 26, 2014 7:12:03 PM
Completed: January 26, 2014 7:15:13 PM

[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](#)

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/ison	4	2.1 KB

Pages can go off-topic through time

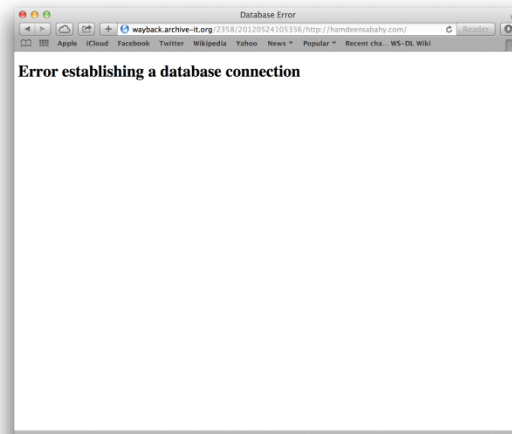


May 13, 2012: The page started as on-topic.

Pages can go off-topic through time



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

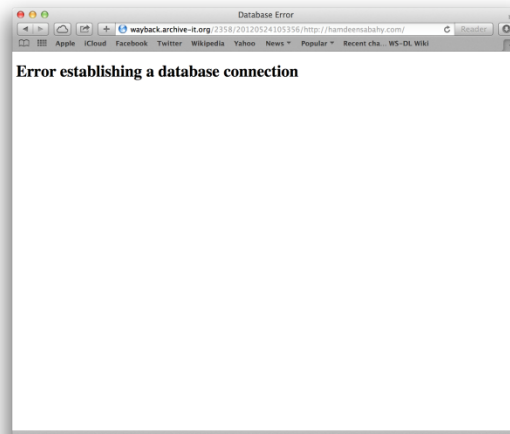


May 24, 2012: Off-topic due to a database error.

Pages can go off-topic through time



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



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

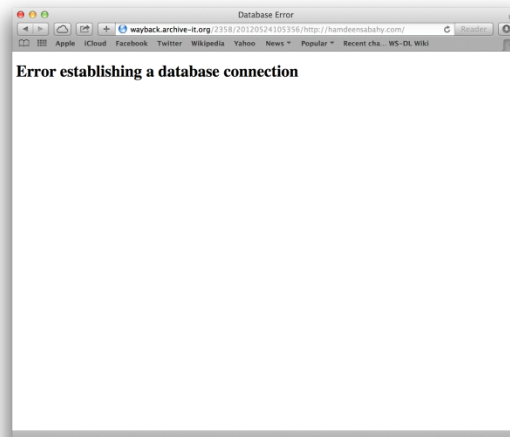


Mar. 21, 2013: Not working because of financial problems.

Pages can go off-topic through time



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



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



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



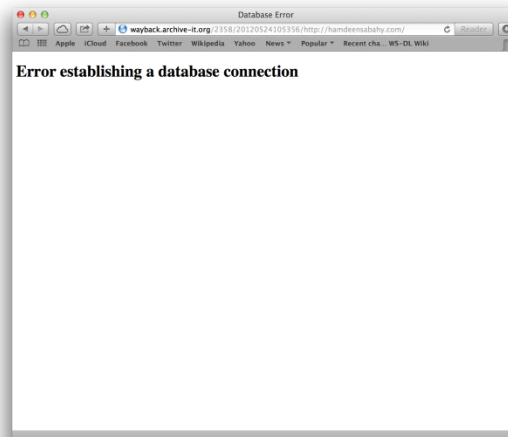
May 21, 2013: On-topic again

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

Pages can go off-topic through time



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



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



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



May 21, 2013: On-topic again



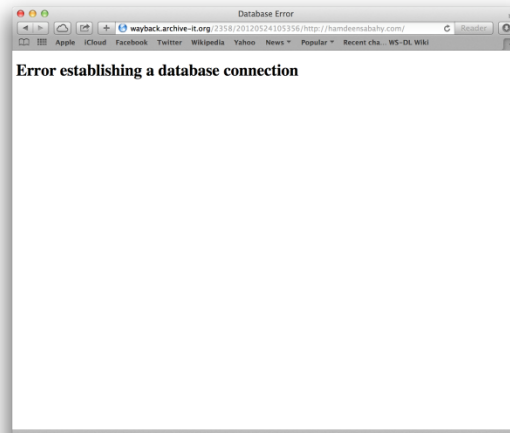
June 5, 2014: The site has been hacked

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

Over 60% of archived versions of hamdeensabahy.com are off-topic



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



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



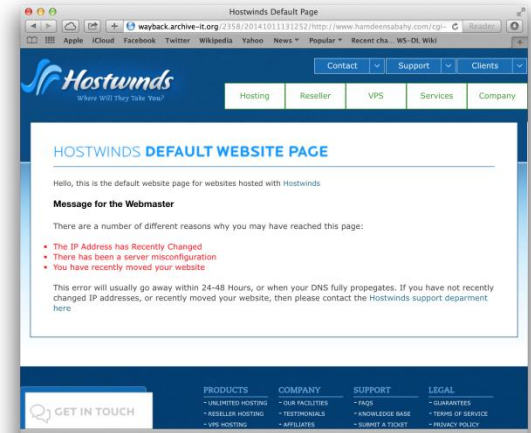
Mar. 21, 2013: Not working because of financial problems.



May 21, 2013: On-topic again



June 5, 2014: The site has been hacked



Oct. 10, 2014: The domain has expired.

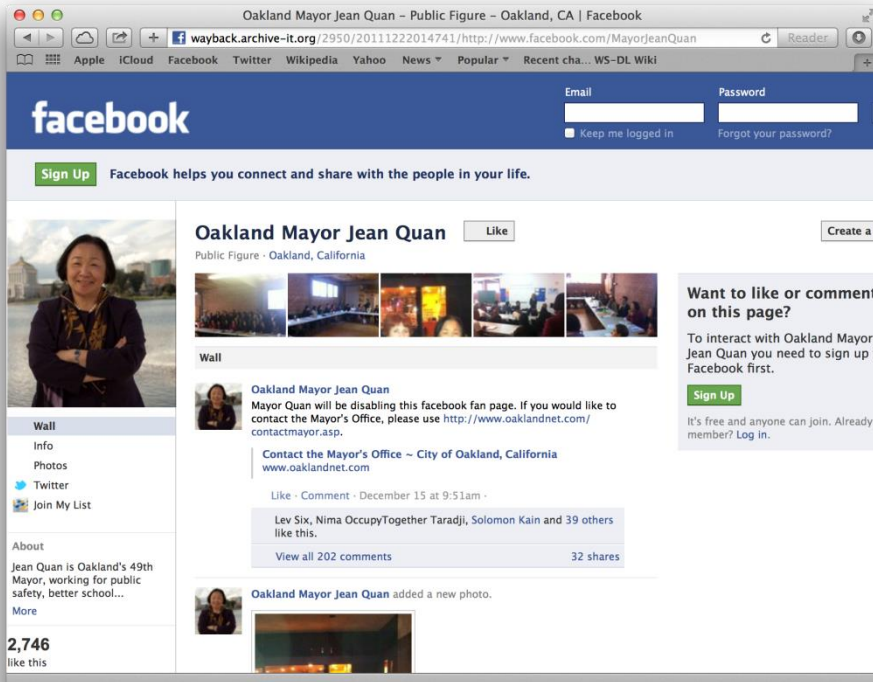
http://wayback.archive-it.org/2358/*/<http://hamdeensabahy.com>

Social media pages can go off-topic

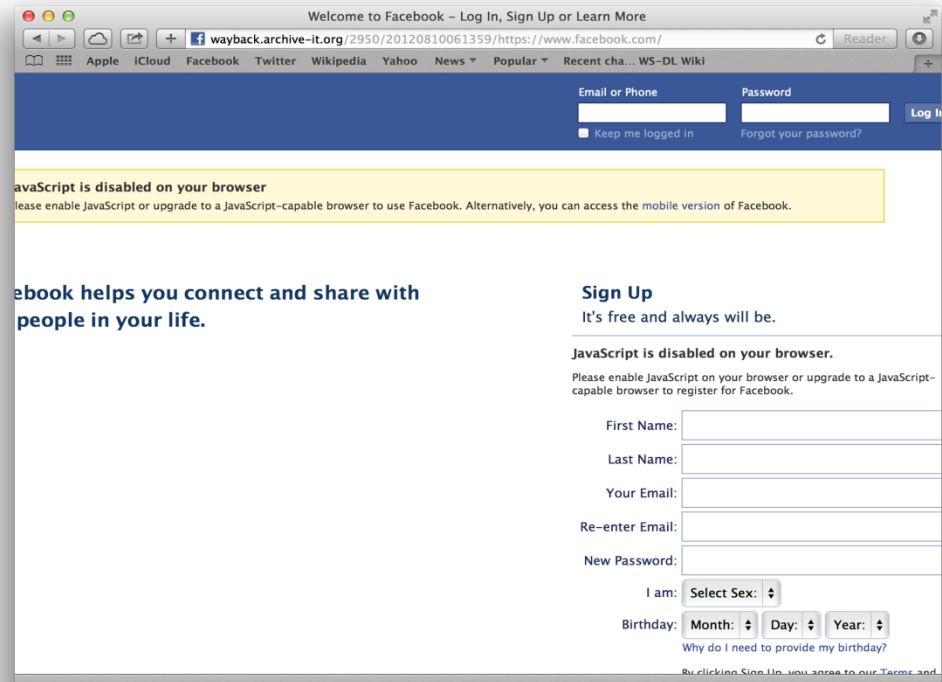


Dec. 22, 2011: Facebook page was relevant to the Occupy collection

Social media pages can go off-topic



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



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

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

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

INTERNET ARCHIVE
WayBackMachine

Enter Web Address: All

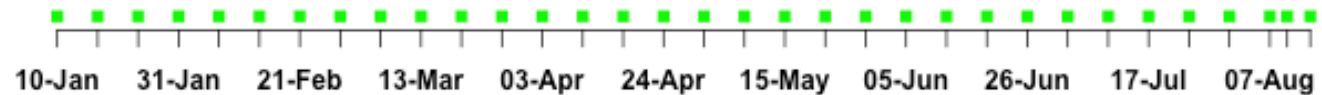
Searched for <http://hamdeensabahy.com> 226 Results [RSS](#)
[Look up URL](#) in general Internet Archive web collection [Proxy Mode Help](#)

* denotes when page was updated

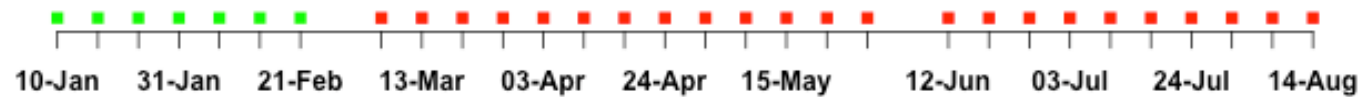
Found 226 Captures between May 13, 2012 - Oct 11, 2014		
2012	2013	2014
35 pages	50 pages	141 pages
May 13, 2012 *	Jan 3, 2013	May 21, 2014 *
May 20, 2012 *	Jan 10, 2013	May 22, 2014 *
May 22, 2012 *	Jan 17, 2013	May 23, 2014 *
May 24, 2012 *	Jan 24, 2013	May 24, 2014 *
May 31, 2012 *	Jan 31, 2013	May 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 classes of TimeMaps

1. Always On



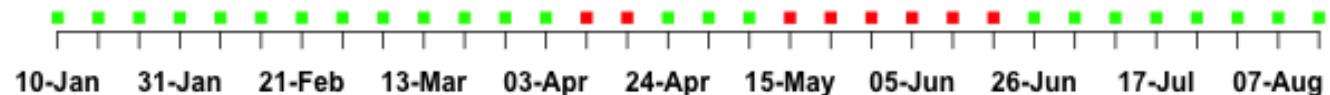
2. Step Function On



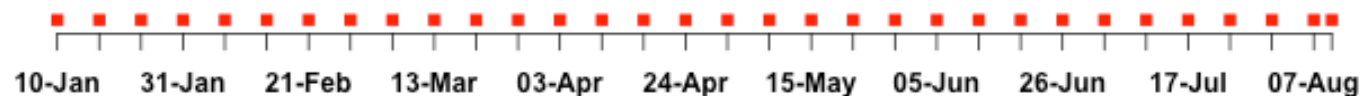
3. Step Function Off



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 x

← → ↻ ⌂ https://wayback.archive-it.org/2358/*/http://www.7amla.net/ ☆ 📌 🌐 ☰

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

Enter Web Address: All ▾ Take Me Back

Searched for <http://www.7amla.net/> 24 Results [RSS](#) [Metadata](#)
[Look up URL](#) in general Internet Archive web collection [Proxy Mode Help](#)

* denotes when page was updated

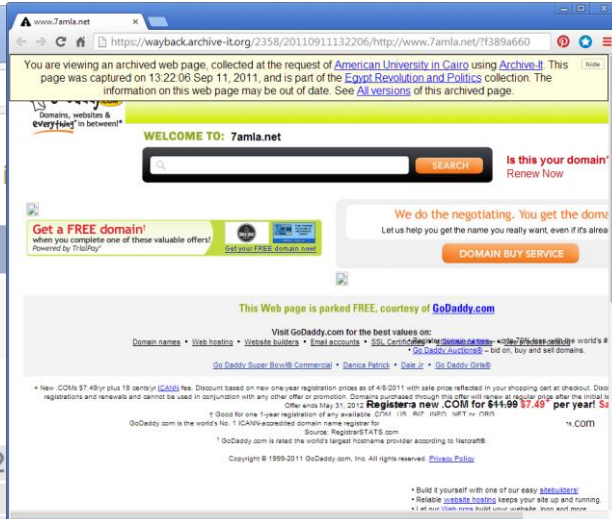
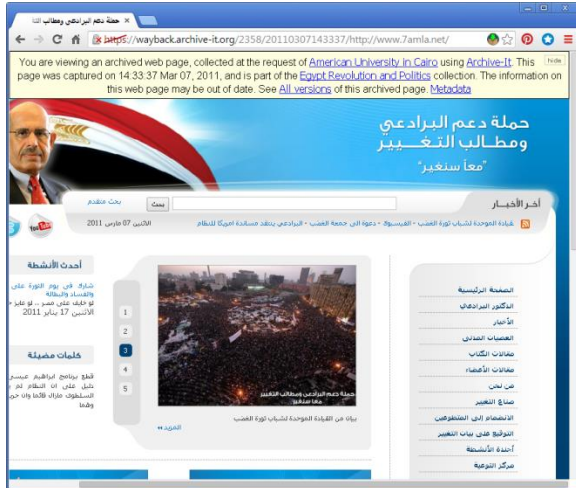
On-topic: Egyptian Revolution coverage

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

2011	2012
19 pages	5 pages

Off-topic: the domain registration is lost

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



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

Internet Archive Wayback Machine

Enter Web Address: All Take Me Back

Searched for http://www.bbc.co.uk/news/world/middle_east/ 627 Results [RSS](#) [Metadata](#)
[Look up URL](#) in general Internet Archive web collection [Proxy Mode Help](#)

* 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 * Feb 1, 2011 * Feb 2, 2011 * Feb 4, 2011 * Feb 5, 2011 * Feb 7, 2011 * Feb 9, 2011 * Feb 11, 2011 * Feb 11, 2011 * Feb 13, 2011 *	Jan 2, 2012 * Jan 9, 2012 * Jan 16, 2012 * Jan 23, 2012 * Jan 30, 2012 * Feb 22, 2012 * 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 *	Jan 26, 2014 * Jan 27, 2014 * Jan 28, 2014 * Jan 29, 2014 * Jan 30, 2014 * Jan 31, 2014 * Feb 1, 2014 * Feb 2, 2014 * Feb 3, 2014 * 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

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

Politics Web Archive (American University in Cairo)

Off-topic: news about Iraq

On-topic: Egyptian Revolution coverage

Off-topic: news about Syria

On-topic: Egypt news

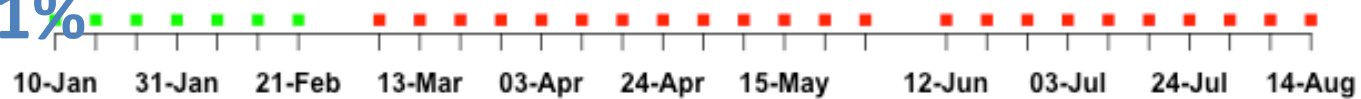
Date	Topic
Jan 1, 2013	Off-topic
Jan 2, 2013	Off-topic
Jan 3, 2013	Off-topic
Jan 4, 2013	Off-topic
Jan 5, 2013	Off-topic
Jan 6, 2013	Off-topic
Jan 7, 2013	Off-topic
Jan 8, 2013	Off-topic
Jan 9, 2013	Off-topic
Jan 10, 2013	Off-topic
Feb 1, 2014	On-topic
Feb 2, 2014	On-topic
Feb 3, 2014	On-topic
Feb 4, 2014	On-topic

Most TimeMaps are Always On

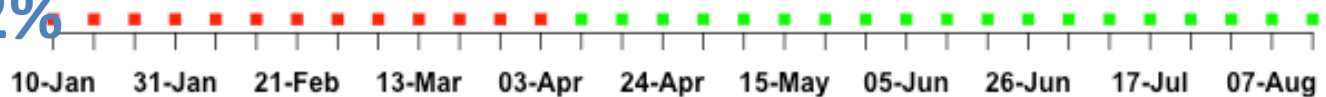
1. Always On **74%**



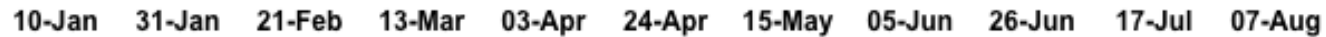
2. Step Function On **8-11%**



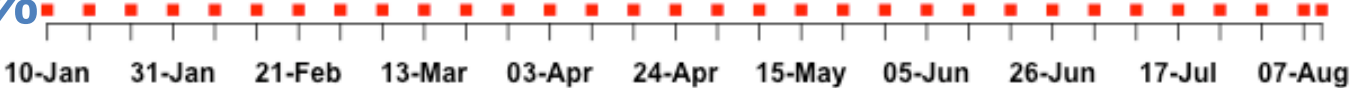
3. Step Function Off **0-2%**



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

From Archive-It collection to terms

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

We investigated 6 similarity metrics

- Textual Content
 - cosine similarity of TF-IDF
 - intersection of the 20 most frequent terms
 - Jaccard similarity coefficient
- Semantics
 - Web-based kernel function using a 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



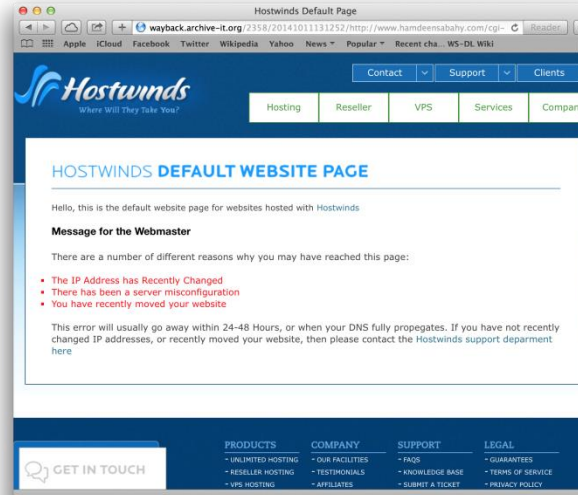
Method	Similarity
cosine	0.7
TF-Intersection	0.6
Jaccard	0.5

Textual Content

cosine similarity, intersecting the most frequent terms,
Jaccard similarity



Method	Similarity
cosine	0.7
TF-Intersection	0.6
Jaccard	0.5



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



Tahrir, Egypt, army

No term-wise overlap

July 2013

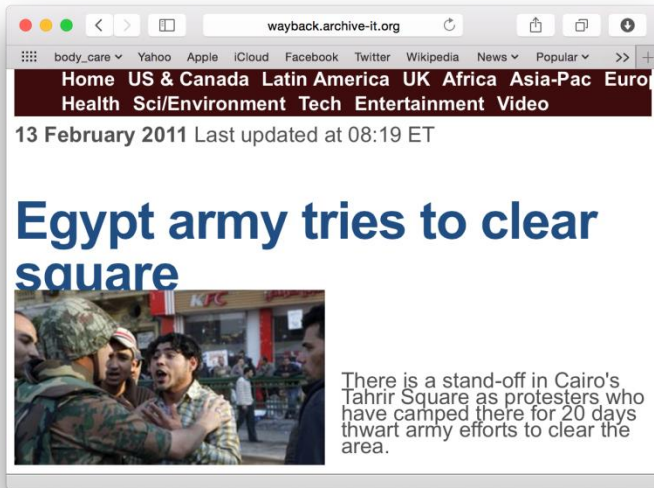


Cairo, Morsi, protests

Semantics of the Text

Web based kernel function using the search engine (SE)

Feb. 2011



July 2013



Tahrir, Egypt, army

No term-wise overlap

Cairo, Morsi, protests

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

Structural Methods

no. of words, content-length

100



109



Method	% change
WordCount	0.09

Structural Methods

no. of words, content-length

100



109



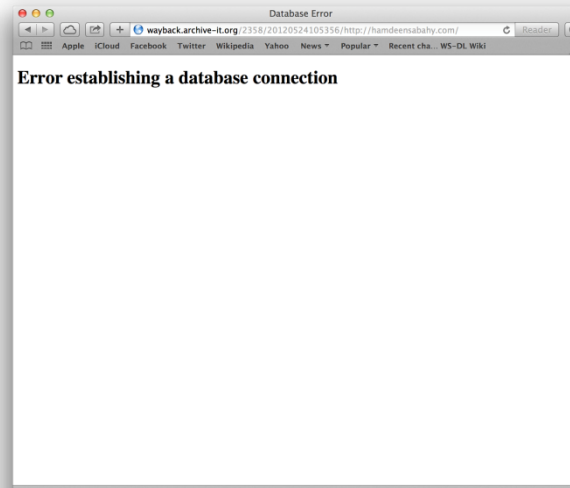
Method % change

WordCount 0.09

100



5



Method % change

WordCount -0.95

We built a gold standard data set to
evaluate the methods

We manually labeled 15,760 mementos

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

Occupy Movement 2011/2012

Collected by: [Internet Archive Global](#)

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

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

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)

Sort By: [Title \(A-Z\)](#) | [Title \(Z-A\)](#) | [URL \(A-Z\)](#) | [URL \(Z-A\)](#)

Page 1 of 8 (707 Total Results)

URL: <http://15october.net/>

Captured 7 times between Feb 10, 2011 and Feb 7, 2013

Videos: 1 Videos Captured

Group: [International](#)

Title: [15th october: #United](#)

URL: <http://15october.net/>

Captured 42 times between Dec 3, 2010 and Feb 7, 2013

Group: [Other Sites](#)

Archive-It - Egypt Revolut x
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 of news 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 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

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

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-Youni (3)

Jam'iyyat al-Ikhwan al-Muslimin (Egypt) (2)

Shafik, Ahmed (2)

We are all Khaled Said (2)

Aboul Fotouh, Abdel Moneim (1)

Sort By: [Title \(A-Z\)](#) | [Title \(Z-A\)](#) | [URL \(A-Z\)](#) | [URL \(Z-A\)](#)

Page 1 of 8 (707 Total Results)

Title: [Egypt Remembers](#), مهنن نلتكر

URL: <http://1000memories.com/egypt/>

Captured 53 times between Feb 11, 2011 and Feb 7, 2013

Videos: 2 Videos Captured

Title: [Egypt remembers](#) | مهنن نلتكر

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

Archive-It - Human Rights x
https://archive-it.org/collections/1068

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

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

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 Desaparecidos y Desaparecidos en Honduras (3)

Markaz al-Watniyya-Hiqaqat al-insan (Jordan) (3)

National Human Rights Commission (Mauritius) (3)

Sort By: [Title \(A-Z\)](#) | [Title \(Z-A\)](#) | [URL \(A-Z\)](#) | [URL \(Z-A\)](#)

Page 1 of 7 (668 Total Results)

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

Occupy Movement
URI-Rs: 255
URI-Ms: 6,570
Off-topic URI-Ms: 458

Egypt Revolution and Politics
URI-Rs: 136
URI-Ms: 6,886
Off-topic URI-Ms: 384

Columbia Univ. Human Rights collection
URI-Rs: 198
URI-Ms: 2,304
Off-topic URI-Ms: 94

Example of manually labeled set

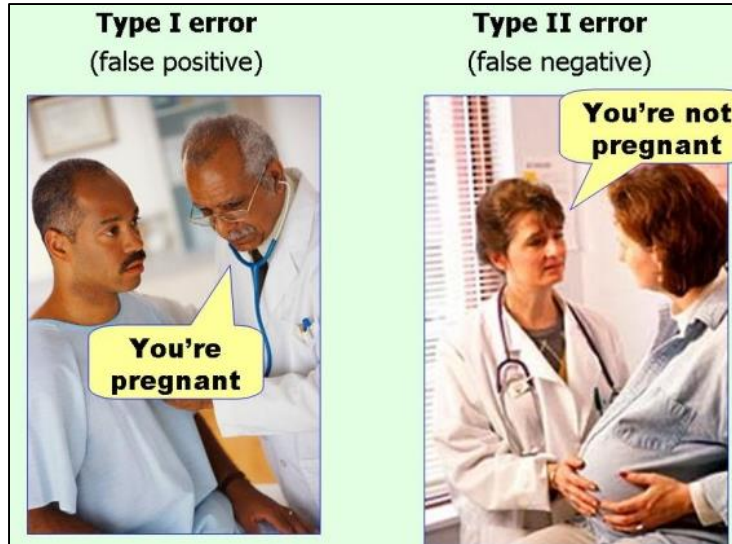
id	date	URI	label
9	20120124014240	http://wayback.archive-it.org/2950/20120124014240/http://occupysarasota.com/	1
9	20120131014118	http://wayback.archive-it.org/2950/20120131014118/http://occupysarasota.com/	1
9	20120207014119	http://wayback.archive-it.org/2950/20120207014119/http://occupysarasota.com/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

Evaluated 6 methods at 21 thresholds

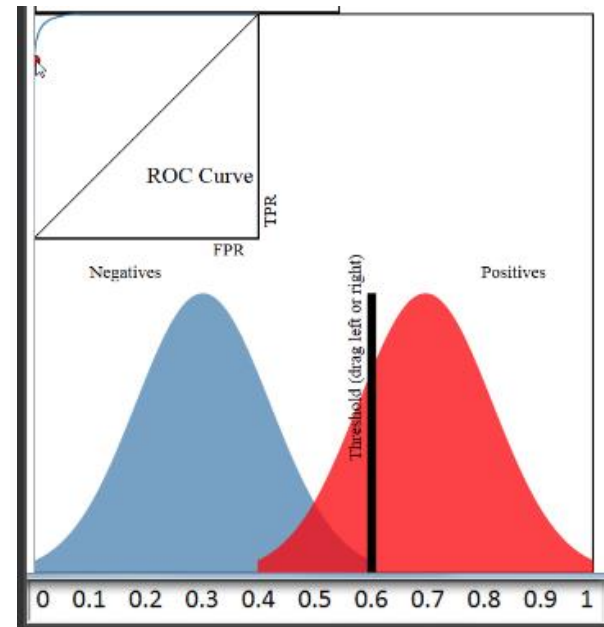
- Assumed first memento was on-topic
- Combined two methods ('OR') to find best combination method
 - 15 combinations
 - 6,615 tests (15 combinations x 21 thresholds x 21 thresholds)
- Averaged the results at each threshold over the three collections

Evaluated based on 5 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



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 investigated six methods for measuring similarity between mementos in a TimeMap:
 - cosine similarity of TF-IDF
 - Jaccard similarity
 - intersection of the 20 most 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/
```

```
...
```

We're continuing work on this

- Enhancements to the detection tool
 - add the other similarity methods (WordCount first)
 - allow input of local CDX and WARC files
- Investigate characteristics of collections and TimeMaps that affect choosing thresholds
- Detect off-topic seeds (URI-Rs) in a collection
 - determine collection aboutness



OLD DOMINION
UNIVERSITY

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