

Impact of HTTP Cookie Violations in Web Archives

Sawood Alam, Plinio Vargas, Michele C. Weigle, Michael L. Nelson
(arXiv:1906.07141, 2019)

CS895 - Web Archiving Forensics, Fall 2020

Old Dominion University

Department of Computer Science

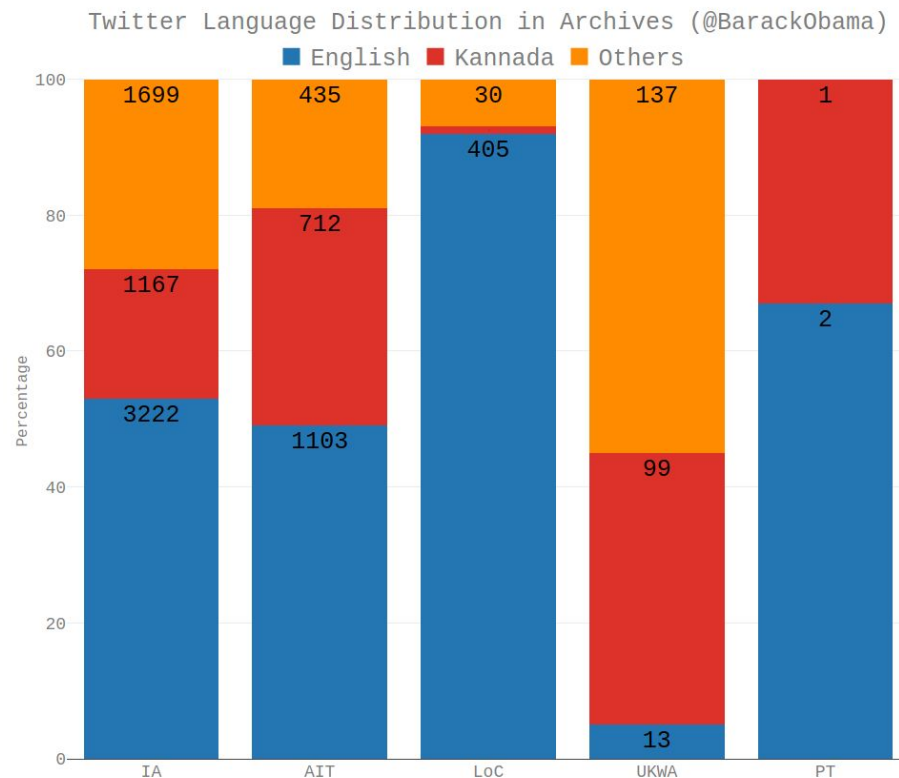
Kritika Garg, kgarg001@odu.edu

Ex-President Obama's archived Twitter page in Urdu

The screenshot shows a web browser displaying the archived Twitter profile of Barack Obama (@BarackObama) in Urdu. The browser's address bar shows the URL <https://web.archive.org/web/20170909164231/https://twitter.com/BarackObama>. The page features a large banner image with the text "Why is this capture not in English?" overlaid. Below the banner, there are statistics in Urdu: فالو کریں (Followers) 3, پسندیدہ (Likes) 10, فالورز (Retweets) 9.46 کروڑ (Crores), فالونگ (Following) 6.28 لاکھ (Lakhs), and ٹویٹس (Tweets) 15.5 ہزار (Thousands). The main tweet is from @BarackObama, dated 19 گھنٹے (19 hours) ago, with the text "Proud of these McKinley Tech students—inspiring young minds that make me hopeful about our future." Below the tweet is a video thumbnail. The right sidebar shows the profile picture, name "Barack Obama", bio "Dad, husband, President, citizen.", location "Washington, DC", website "obama.org", and a date range "مارچ کو شامل ہوئے 2007" (Included in March 2007).

<https://ws-dl.blogspot.com/2018/03/2018-03-21-cookies-are-why-your.html>

Kannada is a prominent language in archives



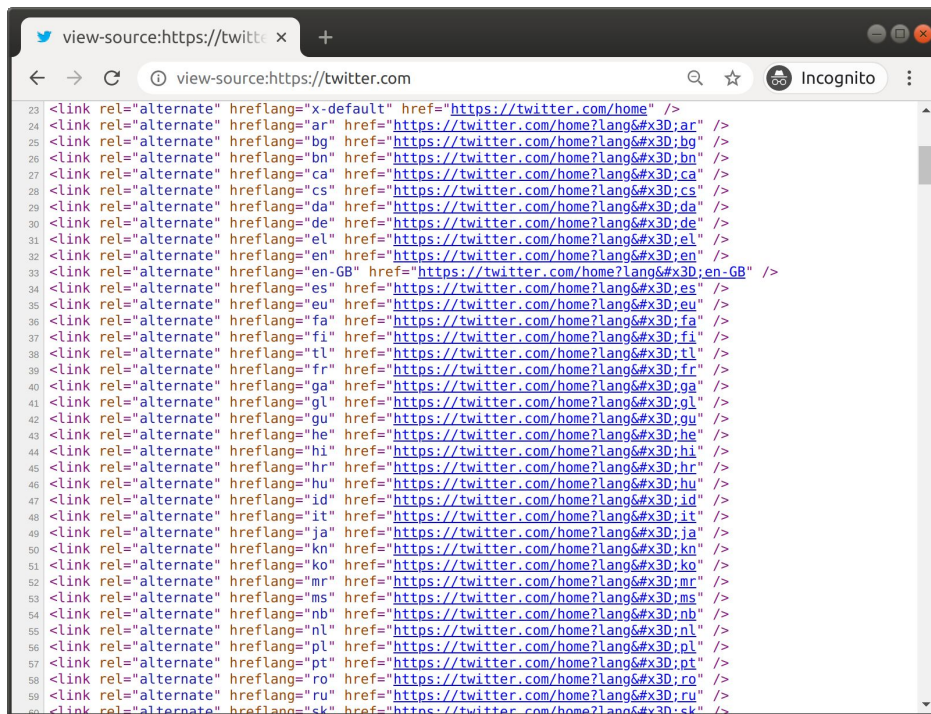
Overall Twitter language distribution of over 9,000 @BarackObama mementos:

- 53% English
- 22% Kannada
- 25% Other languages

<https://ws-dl.blogspothttps://blog.dshr.org/2018/04/all-your-tweets-are-belong-to-kannada.html.com/>

<https://ws-dl.blogspot.com/2018/03/2018-03-21-cookies-are-why-you.html>

Twitter's list of alternate links for 47 language



```
23 <link rel="alternate" hreflang="x-default" href="https://twitter.com/home" />
24 <link rel="alternate" hreflang="ar" href="https://twitter.com/home?lang#x3D:ar" />
25 <link rel="alternate" hreflang="bg" href="https://twitter.com/home?lang#x3D:bg" />
26 <link rel="alternate" hreflang="bn" href="https://twitter.com/home?lang#x3D:bn" />
27 <link rel="alternate" hreflang="ca" href="https://twitter.com/home?lang#x3D:ca" />
28 <link rel="alternate" hreflang="cs" href="https://twitter.com/home?lang#x3D:cs" />
29 <link rel="alternate" hreflang="da" href="https://twitter.com/home?lang#x3D:da" />
30 <link rel="alternate" hreflang="de" href="https://twitter.com/home?lang#x3D:de" />
31 <link rel="alternate" hreflang="el" href="https://twitter.com/home?lang#x3D:el" />
32 <link rel="alternate" hreflang="en" href="https://twitter.com/home?lang#x3D:en" />
33 <link rel="alternate" hreflang="en-GB" href="https://twitter.com/home?lang#x3D:en-GB" />
34 <link rel="alternate" hreflang="es" href="https://twitter.com/home?lang#x3D:es" />
35 <link rel="alternate" hreflang="eu" href="https://twitter.com/home?lang#x3D:eu" />
36 <link rel="alternate" hreflang="fa" href="https://twitter.com/home?lang#x3D:fa" />
37 <link rel="alternate" hreflang="fi" href="https://twitter.com/home?lang#x3D:fi" />
38 <link rel="alternate" hreflang="fr" href="https://twitter.com/home?lang#x3D:fr" />
39 <link rel="alternate" hreflang="ga" href="https://twitter.com/home?lang#x3D:ga" />
40 <link rel="alternate" hreflang="gl" href="https://twitter.com/home?lang#x3D:gl" />
41 <link rel="alternate" hreflang="gu" href="https://twitter.com/home?lang#x3D:gu" />
42 <link rel="alternate" hreflang="he" href="https://twitter.com/home?lang#x3D:he" />
43 <link rel="alternate" hreflang="hi" href="https://twitter.com/home?lang#x3D:hi" />
44 <link rel="alternate" hreflang="hr" href="https://twitter.com/home?lang#x3D:hr" />
45 <link rel="alternate" hreflang="hu" href="https://twitter.com/home?lang#x3D:hu" />
46 <link rel="alternate" hreflang="id" href="https://twitter.com/home?lang#x3D:id" />
47 <link rel="alternate" hreflang="it" href="https://twitter.com/home?lang#x3D:it" />
48 <link rel="alternate" hreflang="ja" href="https://twitter.com/home?lang#x3D:ja" />
49 <link rel="alternate" hreflang="kn" href="https://twitter.com/home?lang#x3D:kn" />
50 <link rel="alternate" hreflang="ko" href="https://twitter.com/home?lang#x3D:ko" />
51 <link rel="alternate" hreflang="mr" href="https://twitter.com/home?lang#x3D:mr" />
52 <link rel="alternate" hreflang="ms" href="https://twitter.com/home?lang#x3D:ms" />
53 <link rel="alternate" hreflang="nb" href="https://twitter.com/home?lang#x3D:nb" />
54 <link rel="alternate" hreflang="nl" href="https://twitter.com/home?lang#x3D:nl" />
55 <link rel="alternate" hreflang="pl" href="https://twitter.com/home?lang#x3D:pl" />
56 <link rel="alternate" hreflang="pt" href="https://twitter.com/home?lang#x3D:pt" />
57 <link rel="alternate" hreflang="ro" href="https://twitter.com/home?lang#x3D:ro" />
58 <link rel="alternate" hreflang="ru" href="https://twitter.com/home?lang#x3D:ru" />
59 <link rel="alternate" hreflang="sk" href="https://twitter.com/home?lang#x3D:sk" />
```

The page source of twitter provides a list of alternate links for each language.

List: <https://gist.github.com/ibnesayeed/c7e5773318d6ea041984fb2433bf1d1e>
<https://developer.twitter.com/en/docs/twitter-for-websites/supported-languages>

Twitter's language negotiation

#"lang" query parameter

```
$ curl -s https://twitter.com/?lang=hi | grep "<html"
```

```
<html dir="ltr" lang="hi">
```

#Accept-Language request header

```
$ curl --silent -H "Accept-Language: hi" https://twitter.com/ | grep "<html"
```

```
<html dir="ltr" lang="hi">
```

```
$ curl --silent -A "googlebot" -c /tmp/twitter.cookie.old https://twitter.com/?lang=hi | grep "<html"
```

```
<html lang="hi" data-scribe-reduced-action-queue="true">
```

```
$ curl --silent -A "googlebot" -b /tmp/twitter.cookie.old https://twitter.com/ | grep "<html"
```

```
<html lang="hi" data-scribe-reduced-action-queue="true">
```

Session Cookies

```
$ cat /tmp/twitter.cookie.old
# Netscape HTTP Cookie File
# https://curl.haxx.se/docs/http-cookies.html
# This file was generated by libcurl! Edit at your own risk.

#HttpOnly_.twitter.com TRUE / TRUE 1604589198 fm 0
#HttpOnly_.twitter.com TRUE / TRUE 0 _twitter_sess
BAh7CSIKZmxhc2hJQzonQWN0aW9uQ29udHJvbGxlcjo6Rmxhc2g6OkZsYXNo%250ASGFzaHsABjoKQHVzZWR7A
DoPY3JIYXRIZF9hdGwrCHMa%252BJh1AToMY3NyZl9p%250AZCIIOWFiY2VkZGJiMjM4M2U0ZmYzZjUwY2I0NzMy
ODg2Yzk6B2IkliUyOTU1%250AZDdhNTMyMGUyZjFjYjdIMTRiMjFmOWFIMmNjMA%253D%253D--1ccc619bf24402d
1bbc3ef90dc2ee5d9bf30606e
.twitter.com TRUE / TRUE 1667661198 personalization_id "v1_N5t2gb9AIH2GDksMP3a2Tg=="
twitter.com FALSE / FALSE 0 lang hi
.twitter.com TRUE / TRUE 1667661198 guest_id v1%3A160458919793691525
.twitter.com TRUE / TRUE 1604610798 ct0 e948c8891fb71e0be1c0b7bbe24f7659
```

Sticky cookies are the real culprit

- The crawler's use of session cookies while archiving the Twitter pages impacts the language displayed in the resulting memento.
- Traditional crawlers such as Heritrix uses cookies for language negotiation with Twitter.
 - Twitter's list of 47 alternate language links get added to the frontier queue of the crawler while crawling a Twitter page.
 - Once any of the link is loaded, the language gets set based on cookie for the consecutive links in the frontier queue as well, until the language is overwritten or session expires.

```
<link rel="alternate" hreflang="fr" href="https://twitter.com/?lang=fr">  
<link rel="alternate" hreflang="en" href="https://twitter.com/?lang=en">
```

Cookies & Crawlers

Replicated the crawling process to verify the impact of the Twitter cookies on archiving Twitter page in Internet Archive.

Used Heratrix to crawl the following seeds in this particular sequence:

1. <https://twitter.com/?lang=ar>
2. https://twitter.com/phonedude_mln/

WARC request and response headers for 1st seed.

```
WARC/1.0
WARC-Type: request
WARC-Target-URI: https://twitter.com/?lang=ar
WARC-Date: 2018-03-16T21:58:44Z
WARC-Concurrent-To: <urn:uuid:7dbc3a67-5cf8-4375-8343-c0f6b03039f4>
WARC-Record-ID: <urn:uuid:473273f6-48fa-4dd3-a5f0-81caf9786e07>
Content-Type: application/http; msgtype=request
Content-Length: 301
```

```
GET /?lang=ar HTTP/1.0
User-Agent: Mozilla/5.0 (compatible; heritrix/3.2.0
+http://cs.odu.edu/)
Connection: close
Accept:
text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Host: twitter.com
Cookie: guest_id=v1%3A152123752160566016;
personalization_id=v1_uAUfoUV9+DkWI8mETqFuFg==
```

```
WARC/1.0
WARC-Type: response
WARC-Target-URI: https://twitter.com/?lang=ar
WARC-Date: 2018-03-16T21:58:44Z
WARC-Payload-Digest: sha1:FCOPDBN2U5LXU7FEUUGQ4WXYGR7OP5J1
WARC-IP-Address: 104.244.42.129
WARC-Record-ID: <urn:uuid:7dbc3a67-5cf8-4375-8343-c0f6b03039f4>
Content-Type: application/http; msgtype=response
Content-Length: 151985
```

```
HTTP/1.0 200 OK
cache-control: no-cache, no-store, must-revalidate, pre-check=0, post-check=0
content-length: 150665
content-type: text/html; charset=utf-8
date: Fri, 16 Mar 2018 21:58:44 GMT
expires: Tue, 31 Mar 1981 05:00:00 GMT
last-modified: Fri, 16 Mar 2018 21:58:44 GMT
pragma: no-cache
server: tsa b
set-cookie: fm=0; Expires=Fri, 16 Mar 2018 21:58:34 UTC; Path=/; Domain=.twitter.com;
Secure; HTTPOnly
set-cookie:
  twitter sess=BAh7CSIKZmxhc2hJQzonQWN0aW9uQ29udHJvbGxlcjo6Rmxhc2g6OkZsYXNo%250ASGFzaHsAB
  jQKQHVzZWR7ADoPY3JlYXRlZF9hdGwrCGKB0jBiAToMY3NyZl9p%250AZC1lZmQ1MTY4ZjQ3NmExZWQ1NjUyNDRm
  MzhhZGNIbmFhZjQ6B2lkIiU0OTQ0%250AZDMxMDY4NjJhYjM4NjBkMzI4MDE0NjYyOGM5ZA%253D--f5716
  56f1526d7ff1b363d527822ebd4495a1fa3; Path=/; Domain=.twitter.com; Secure; HTTPOnly
set-cookie: lang=ar; Path=/
set-cookie: ct0=10558ec97ee83fe0f2bc6de552ed4b0e; Expires=Sat, 17 Mar 2018 03:58:44 UTC;
Path=/; Domain=.twitter.com; Secure
status: 200 OK
strict-transport-security: max-age=631138519
x-connection-hash: 2a2fc89f51b930202ab24be79b305312
x-content-type-options: nosniff
x-frame-options: SAMEORIGIN
x-response-time: 100
x-transaction: 001495f800dc517f
x-twitter-response-tags: BouncerCompliant
x-ua-compatible: IE=edge,chrome=1
x-xss-protection: 1; mode=block; report=https://twitter.com/i/xss_report

<!DOCTYPE html>
<html lang="ar" data-scribe-reduced-action-queue="true">
```

```

WARC/1.0
WARC-Type: request
WARC-Target-URI: https://twitter.com/phonedude_mln/
WARC-Date: 2018-03-16T21:58:48Z
WARC-Concurrent-To: <urn:uuid:634dea88-6994-4bd4-af05-5663d24c3727>
WARC-Record-ID: <urn:uuid:eeef134ed-f3dc-459b-95e7-624b4d747bc1>
Content-Type: application/http; msgtype=request
Content-Length: 655

GET /phonedude mln/ HTTP/1.0
User-Agent: Mozilla/5.0 (compatible; heritrix/3.2.0
+http://cs.odu.edu/)
Connection: close
Accept:
text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Host: twitter.com
Cookie: lang=ar;
_twitter_sess=Bah7CSiKZmxhc2hJQzonQWN0aW9uQ29udHJvbGxlcjo6Rmxhc2g
6OkZsYXNo%250ASGFzaHsABjokQHvzZWR7ADoPY3JlYXRlZF9ndGwcrCGKB0jBiATo
MY3NyZl9p%250AZCilZmQ1MTY4ZjZjQ3NmExZWQ1NjUyYndRMzhhZGNIbMmFhZjZjQ6B2l
kIiU0OTQ0%250AZDMxMDY4NjJhYjM4NjBkMzI4MDE0NjYyOGM5ZA%253D%253D--f
571656f1526d7ff1b363d52782ebd4495a1fa3;
ct0=10558ec97ee83fe0f2bc6de552ed4b0e;
guest_id=v1%3A152123752160566016;
personalization_id=v1 uAUfoUV9+DkWI8mETgfuFg==

```

```
HTTP/1.0 200 OK
cache-control: no-cache, no-store, must-revalidate, pre-check=0, post-check=0
content-length: 516921
content-type: text/html; charset=utf-8
date: Fri, 16 Mar 2018 21:58:48 GMT
expires: Tue, 31 Mar 1981 05:00:00 GMT
last-modified: Fri, 16 Mar 2018 21:58:48 GMT
pragma: no-cache
server: tsa_b
set-cookie: fm=0; Expires=Fri, 16 Mar 2018 21:58:38 UTC; Path=/;
Domain=.twitter.com; Secure; HTTPOnly
set-cookie:
_twitter_sess=Bah7CSIK2mxc2hJQzonQWN0aW9uQ29udHJvbGxlcjo6Rmxhc2g6OkZsYXNo%250ASGFz
aHsABjokQHvZWR7AdoPY3JlYXRlZj9hdGwrcGKB0jBiAtOMY3NyZl9p%250AZCilZmQ1MTY4ZjQ3NmExZW
Q1NjUyNDRmMzhhZGNiMmFhZjQ6B2lkIiU0OTQ0%250AZDMxMDY4NjJhYjM4NjBkMzI4MDEONjYyOGM5ZA%2
53D%253D--f571656f1526d7ff1b363d527822ebd4495alfa3; Path=/; Domain=.twitter.com;
Secure; HTTPOnly
status: 200 OK
strict-transport-security: max-age=631138519
x-connection-hash: ef102c969c74f3abf92966e5ffddb6ba
x-content-type-options: nosniff
x-frame-options: SAMEORIGIN
x-response-time: 335
x-transaction: 0014986c00687fa3
x-twitter-response-tags: BouncerCompliant
x-ua-compatible: IE=edge,chrome=1
x-xss-protection: 1; mode=block; report=https://twitter.com/i/xss_report

<!DOCTYPE html>
<html lang="ar" data-scribe-reduced-action-queue="true">..
```

Replay the captured WARC using PyWb

<https://twitter.com/?lang=ar>



https://twitter.com/phonedude_mln/



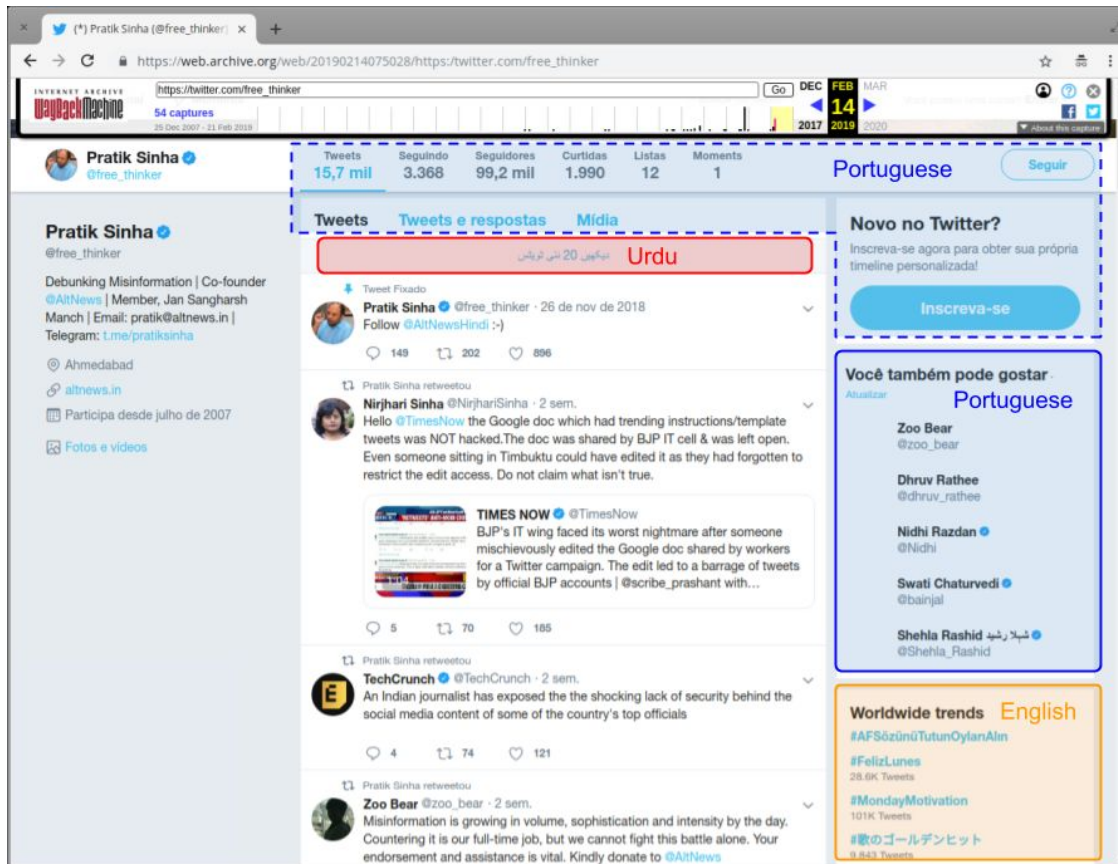
<https://github.com/webrecorder/pywb>

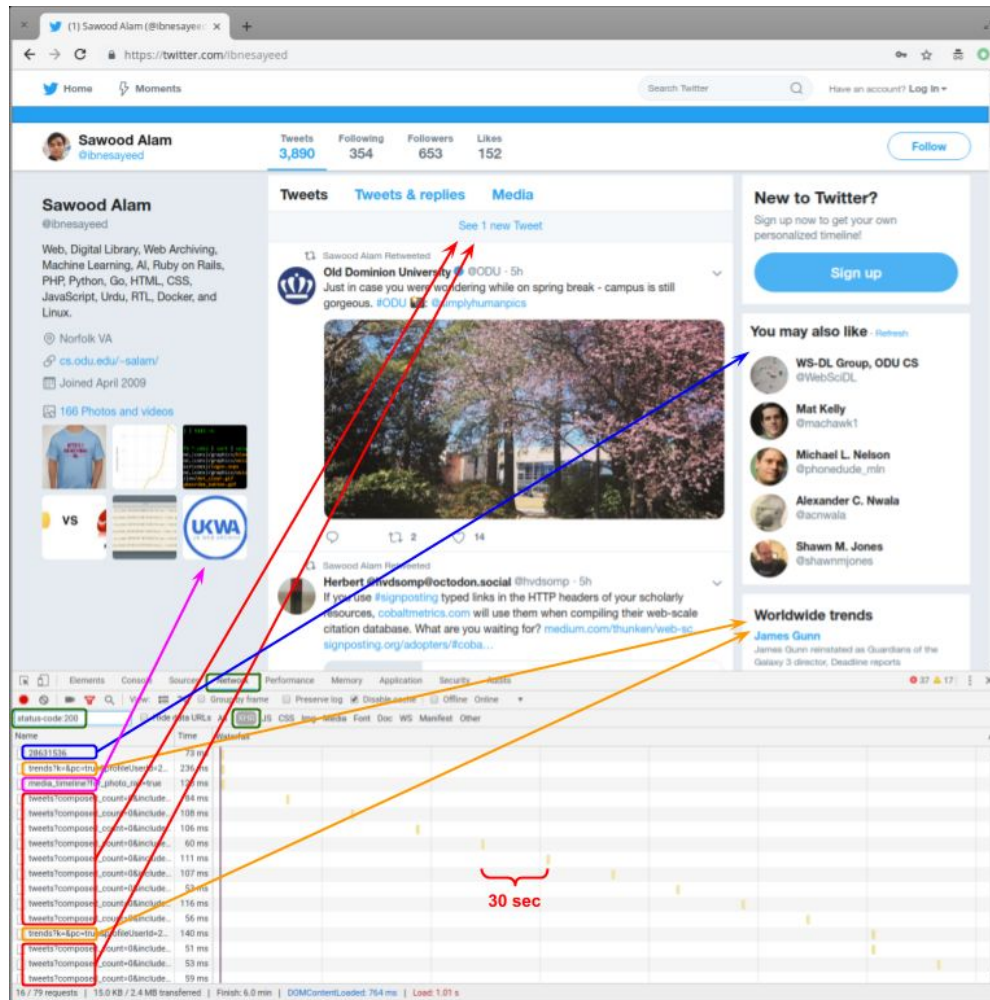
why was Kannada so dominant?

- Kanada was the last language in the list.
- The languages gets overwritten when the next language specific link in the frontier queue is loaded.
- Similarly, Bengali was last language in the list for year 2015 making it a dominant language for archival activities between July 2014 to July 2015.

```
<link rel="alternate" hreflang="x-default" href="https://twitter.com/">  
<link rel="alternate" hreflang="fr" href="https://twitter.com/?lang=fr">  
<link rel="alternate" hreflang="en" href="https://twitter.com/?lang=en">  
<link rel="alternate" hreflang="ar" href="https://twitter.com/?lang=ar">  
----- (43 links truncated)  
<link rel="alternate" hreflang="kn" href="https://twitter.com/?lang=kn">
```

Defaced composite mementos





Anatomy of a Twitter timeline of old interface.

- Bio and initial 20 tweets are embedded in the HTML.
- “You might like” section and media timeline are lazily loaded.
- New tweets after every 30 sec for active page.
- Global trends are loaded after every 5 min.

Defaced composite mementos

The screenshot shows a web browser displaying a Twitter profile page for Pratik Sinha (@free_thinker). The page is a composite of various elements, some of which are highlighted with callouts:

- memento is archived in Portuguese**: A callout pointing to the top right of the page, indicating the language of the archived version.
- Embedded in HTML**: A callout pointing to the "Portuguese" language selector in the top right.
- closest archived copy in Urdu (2019:02:27T22:04:50)**: A callout pointing to a tweet by Pratik Sinha in Urdu, which is highlighted with a red border.
- Media not archived**: A callout pointing to the left sidebar of the profile page, indicating that the media content is not archived.
- Loaded right after main page**: A callout pointing to the right sidebar of the page, indicating that the content is loaded after the main page.
- closest archived copy in English (2019:02:11T13:21:45)**: A callout pointing to the "Worldwide trends" section in the bottom right, which is highlighted with an orange border.

Prior cookie may impact subsequent XHR responses

```
$ curl --silent -b /tmp/twitter.cookie "https://twitter.com/itrends?k=&pc=true&profileUserId=28631536&show_context=true&src=module" | jq
{
  "module html": "<div class=\"flex-module trends-container context-trends-container\">\n  <div class=\"flex-module-header\">\n    \n  </div>\n  <div class=\"flex-module-inner\">\n    <ul class=\"trend-items js-trends\">\n      <li class=\"trend-item js-trend-item context-trend-item\" data-trend-name=\"#PiDay\" data-trends-id=\"1025618545345384837\" data-trend-token=:location request:hashtag trend:taxi_country_source:moments_metadescription:moments_badge:\n      <a class=\"pretty-link js-nav js-tooltip u-linkComplex \">\n        .\n        .\n        .\n        .\n        .\n      </div>\n    </div>\n  </div>\n  \"personalized\": false,\n  \"woeid\": 1\n}
```


Proposed solutions

- Cookies with short expiration time.
- Isolating sessions by sandboxing the crawl jobs from same domain.
- Advertise content negotiation in “Vary” header.
- Utilize cookies during replay (renaming cookies).

Extras

No session cookie in new Twitter UI

#Accept-Language request header

```
$ curl --silent -H "Accept-Language: hi" https://twitter.com/ | grep "<html"
<html dir="ltr" lang="hi">
```

#"lang" query parameter

```
$ curl --silent -c /tmp/twitter.cookie https://twitter.com/?lang=hi | grep "<html"
<html dir="rtl" lang="hi">
$ curl --silent -b /tmp/twitter.cookie https://twitter.com/ | grep "<html"
```

#Saved Cookie (No lang, no session information)

```
$ cat /tmp/twitter.cookie
# Netscape HTTP Cookie File
# https://curl.haxx.se/docs/http-cookies.html
# This file was generated by libcurl! Edit at your own risk.
```

```
.twitter.com TRUE / TRUE 1667659100 personalization_id "v1_k4XhIqGYDP2KKq+3uw0hEg=="
.twitter.com TRUE / TRUE 1667659100 guest_id v1%3A160458710070234309
```

Twitter language list

```
<link rel="alternate" hreflang="x-default" href="https://twitter.com/home" />
<link rel="alternate" hreflang="ar" href="https://twitter.com/home?lang&#x3D;ar" />
<link rel="alternate" hreflang="bg" href="https://twitter.com/home?lang&#x3D;bg" />
<link rel="alternate" hreflang="bn" href="https://twitter.com/home?lang&#x3D;bn" />
<link rel="alternate" hreflang="ca" href="https://twitter.com/home?lang&#x3D;ca" />
<link rel="alternate" hreflang="cs" href="https://twitter.com/home?lang&#x3D;cs" />
<link rel="alternate" hreflang="da" href="https://twitter.com/home?lang&#x3D;da" />
<link rel="alternate" hreflang="de" href="https://twitter.com/home?lang&#x3D;de" />
<link rel="alternate" hreflang="el" href="https://twitter.com/home?lang&#x3D;el" />
<link rel="alternate" hreflang="en" href="https://twitter.com/home?lang&#x3D;en" />
<link rel="alternate" hreflang="en-GB" href="https://twitter.com/home?lang&#x3D;en-GB" />
-----
<link rel="alternate" hreflang="ro" href="https://twitter.com/home?lang&#x3D;ro" />
<link rel="alternate" hreflang="ru" href="https://twitter.com/home?lang&#x3D;ru" />
<link rel="alternate" hreflang="sk" href="https://twitter.com/home?lang&#x3D;sk" />
<link rel="alternate" hreflang="sr" href="https://twitter.com/home?lang&#x3D;sr" />
<link rel="alternate" hreflang="sv" href="https://twitter.com/home?lang&#x3D;sv" />
<link rel="alternate" hreflang="ta" href="https://twitter.com/home?lang&#x3D;ta" />
<link rel="alternate" hreflang="th" href="https://twitter.com/home?lang&#x3D;th" />
<link rel="alternate" hreflang="tr" href="https://twitter.com/home?lang&#x3D;tr" />
<link rel="alternate" hreflang="uk" href="https://twitter.com/home?lang&#x3D;uk" />
<link rel="alternate" hreflang="ur" href="https://twitter.com/home?lang&#x3D;ur" />
<link rel="alternate" hreflang="vi" href="https://twitter.com/home?lang&#x3D;vi" />
<link rel="alternate" hreflang="zh" href="https://twitter.com/home?lang&#x3D;zh" />
<link rel="alternate" hreflang="zh-Hant" href="https://twitter.com/home?lang&#x3D;zh-Hant" />
```

Now sorted in alphabetical order.

- “zh-Hant” traditional chinese is the last language.