Week 11 – Web Usage Data

Yasmin AlNoamany

L&S 88-2

University of California, Berkeley

Announcements

- Project Milestone 3: Status Update
  - Due Nov. 13th
- Assignment 2
- Thanksgiving week?

Previous lecture

- Lecture by Alberto Cairo
- Lecture by Ian and Zan from Google
- Guests' lectures evaluation

Today's lecture

- Intro to Web usage mining
- Applications of Web usage mining
- Web analytics versus Web usage mining
- Web usage data
- HTTP response codes
- Analyzing usage data

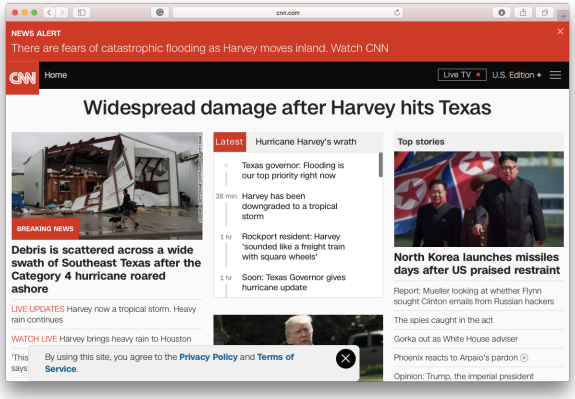
Types of Web Mining

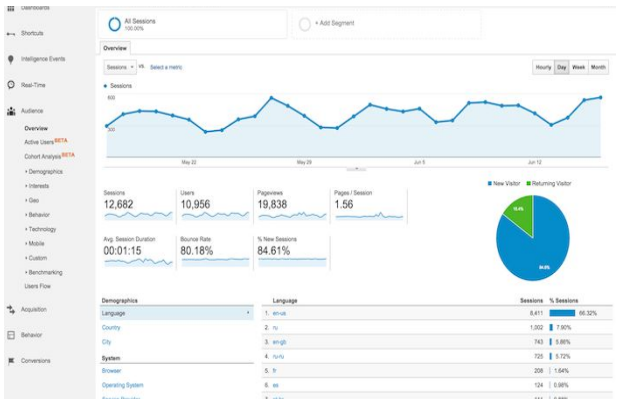
Web Mining

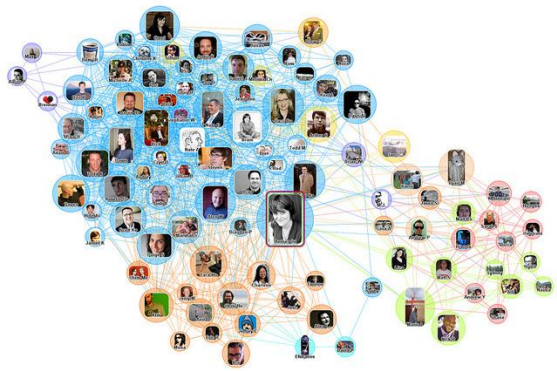
Web Content Mining

Web Usage Mining

Web Structure Mining



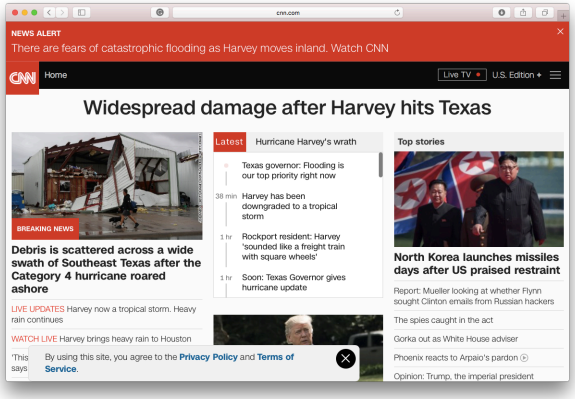




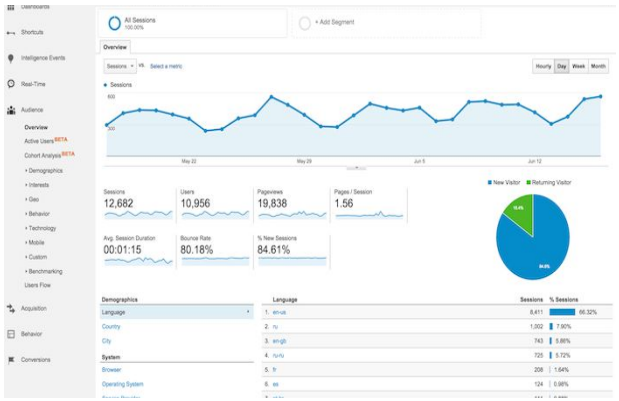
Types of Web Mining

Web Mining

Web Content Mining



Web Usage Mining



Web Structure Mining

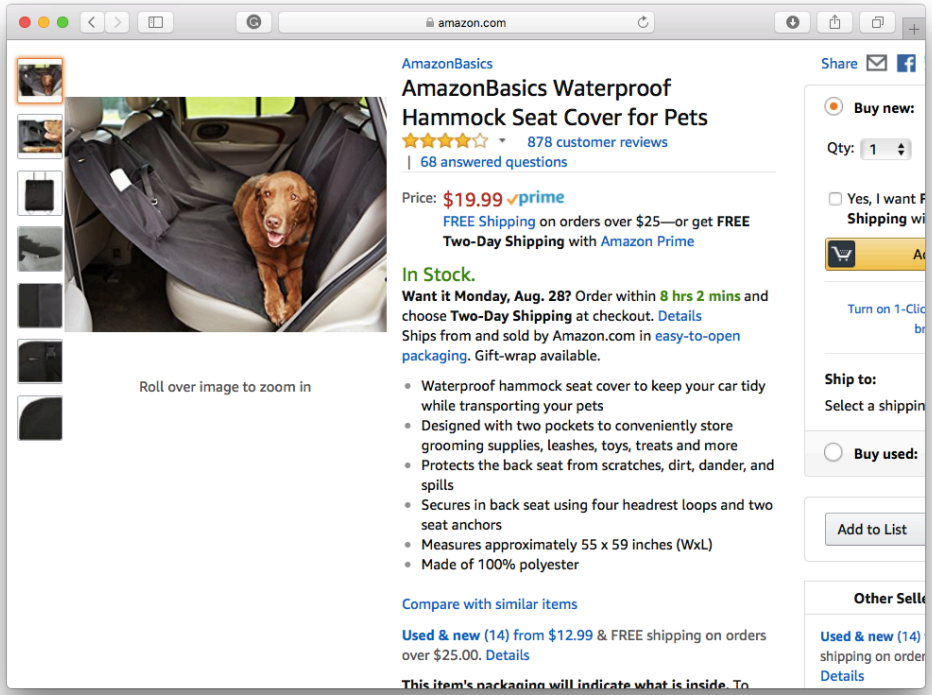
Web usage mining

- Extracting interesting patterns from user interactions with resources on one or more Web sites
- Data
  - clickstream data from Web/application server access logs
  - e-commerce and product-oriented user events (e.g., shopping cart changes, product click-throughs, purchases, etc.)
  - user profiles data, user ratings, user contributed data (tags, comments, reviews)

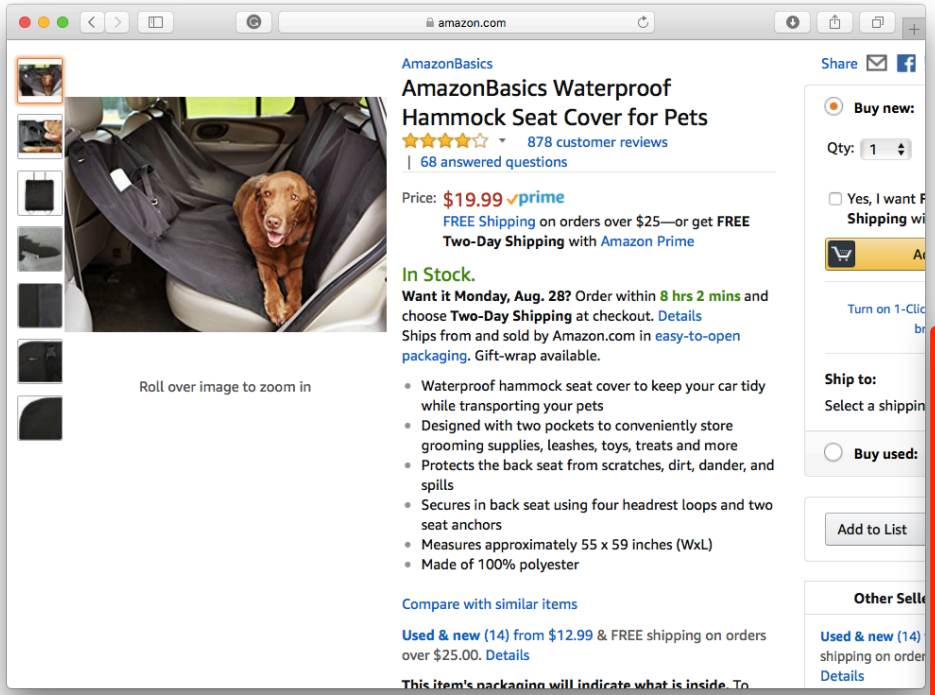
Applications

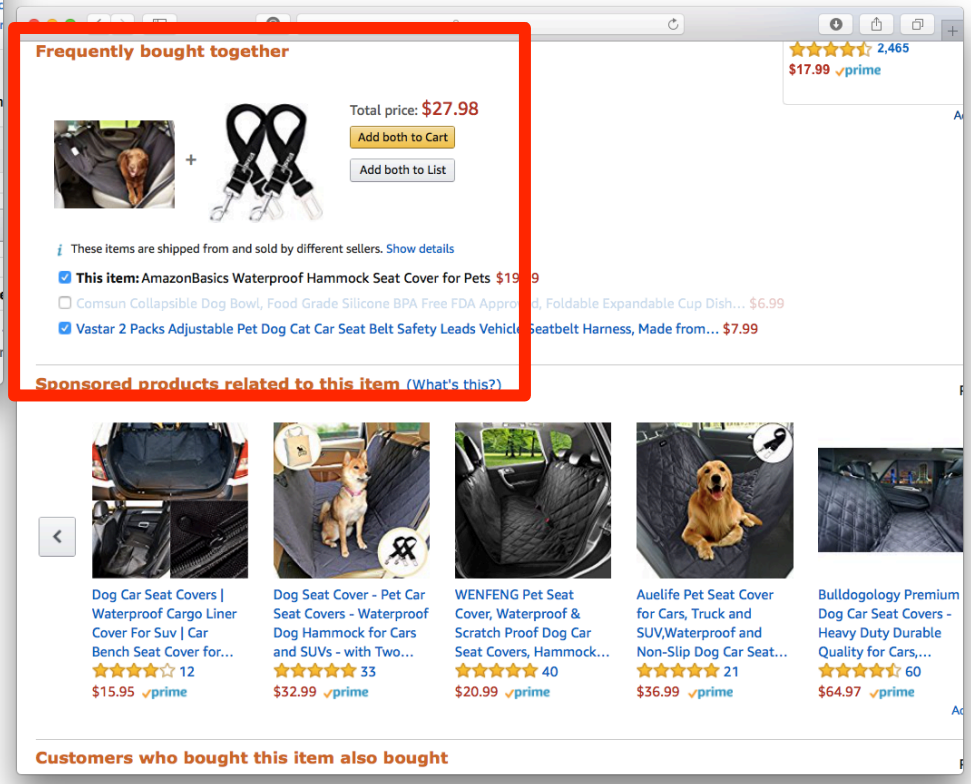
- User and customer behavior modeling
- Web site optimization
- E-customer relationship management
- Web marketing
- Targeted advertising
- Recommender systems

Applications

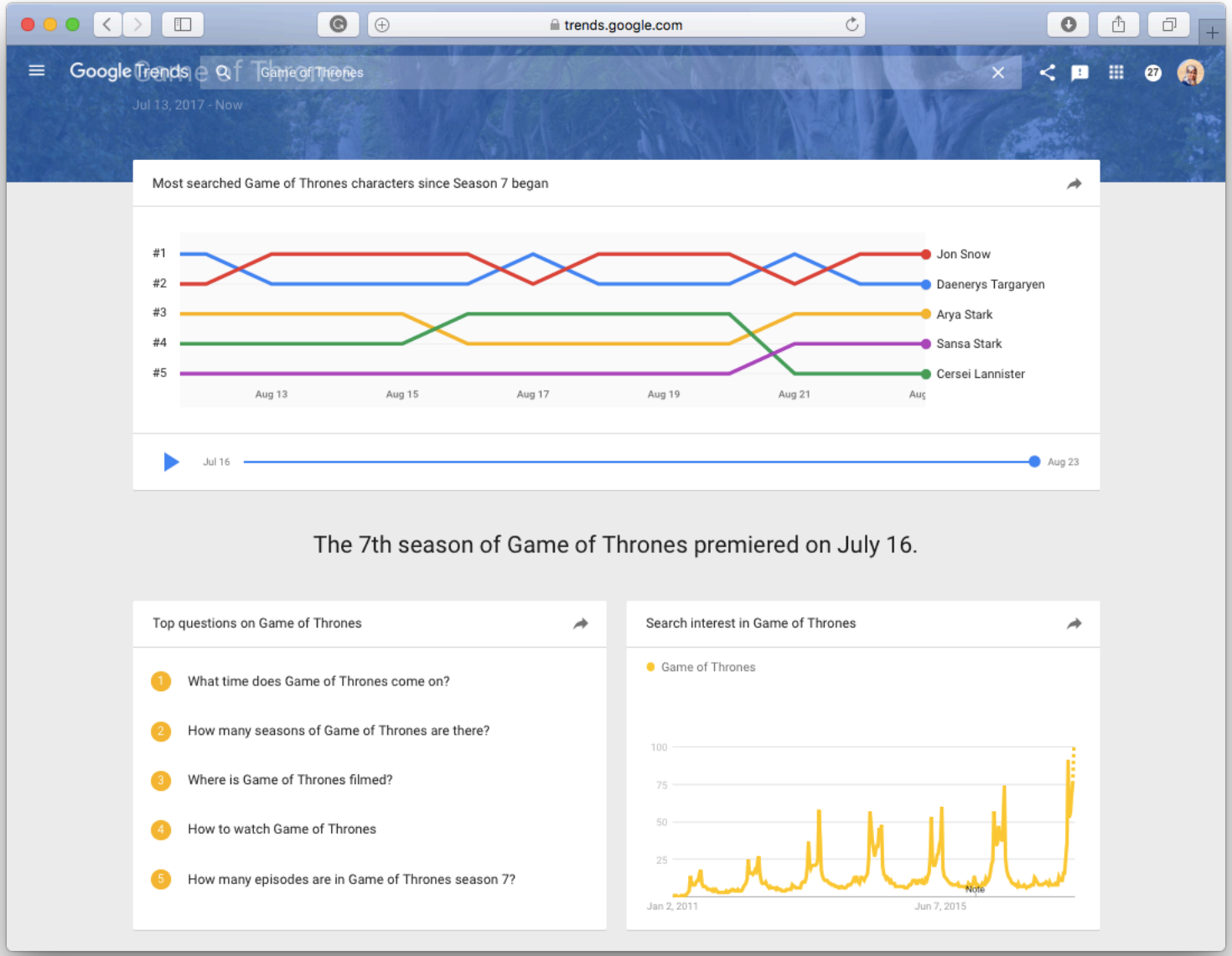


Applications





Using the Web to understand the Web

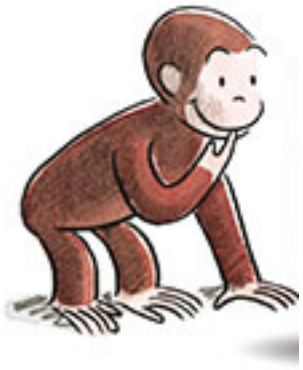


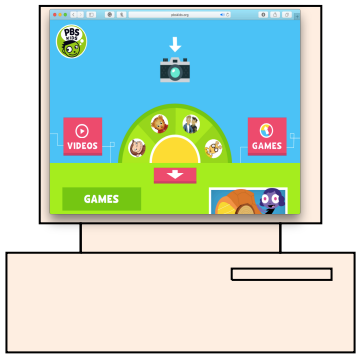
Web analytics versus Web usage mining

- Web Analytics
  - refers to the measurement, analysis, and reporting of user behavior on the Web
  - usually involves descriptive statistics from clickstream and other user behavior data at different levels of aggregations across predetermined dimensions such as time, content/product categories, referring sites, etc.
  - e.g., Google Analytics

- Web Usage Mining
  - goes beyond basic analytics to
    - discover patterns in usage data
    - identify and characterize important customer segments
    - find affinities across pages or products
    - build models to predict future behavior

How the Web works





request

Get http://pbskids.org

Web Server

Response

HTTP/1.1 200 OK

HTTP responses

- 1xx: Informational - Request received, continuing process
  - 100 Continue
- 2xx: Success - The action was successfully received, understood, and accepted
  - 200 OK
- 3xx: Redirection - Further action must be taken in order to complete the request
  - 303 Moved Temporarily
- 4xx: Client Error - The request contains bad syntax or cannot be fulfilled
  - 404 Page Not found, 403 Unauthorized
- 5xx: Server Error - The server failed to fulfill an apparently valid request
  - 501 Internal Server Error

HTTP responses
Why this is important?

- Multiple errors from assignment 2:

TweepError: Twitter error response: status code = 401

HTTP responses

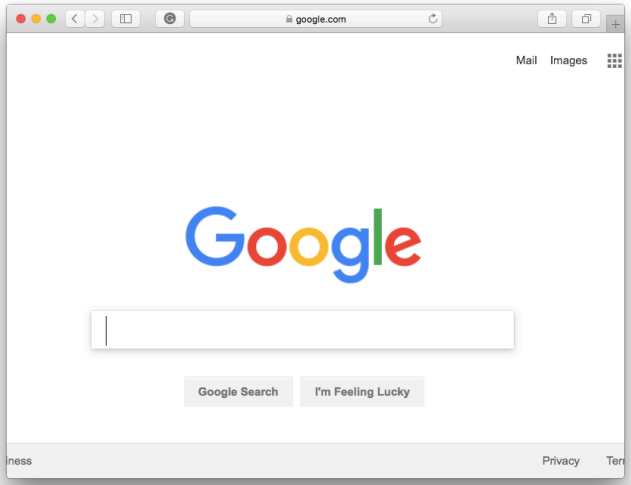
- 1xx: Informational - Request received, continuing process
  - 100 Continue
- 2xx: Success - The action was successfully received, understood, and accepted
  - 200 OK
- 3xx: Redirection - Further action must be taken in order to complete the request
  - 303 Moved Temporarily
- 4xx: Client Error - The request contains bad syntax or cannot be fulfilled
  - 404 Page Not found, 403 Unauthorized
- 5xx: Server Error - The server failed to fulfill an apparently valid request
  - 500 Internal Server Error

HTTP responses

- 1xx: Informational - Request received, continuing process
  - 180 Ringing
- 2xx: Success - The action was successfully received, understood, and accepted
  - 200 OK
- 3xx: Redirection - Further action must be taken in order to complete the request
  - 303 Moved Temporarily
- 4xx: Client Error - The request contains bad syntax or cannot be fulfilled
  - 404 Page Not found, 403 Unauthorized
- 5xx: Server Error - The server failed to fulfill an apparently valid request
  - 501 Internal Server Error

HTTP responses

- 1xx: Informational - Request received, continuing process
  - 180 Ringing
- 2xx: Success - The action was successfully received, understood, and accepted
  - 200 OK
- 3xx: Redirection - Further action must be taken in order to complete the request
  - 303 Moved Temporarily
- 4xx: Client Error - The request contains bad syntax or cannot be fulfilled
  - 404 Page Not found, 403 Unauthorized
- 5xx: Server Error - The server failed to fulfill an apparently valid request
  - 501 Internal Server Error



HTTP responses

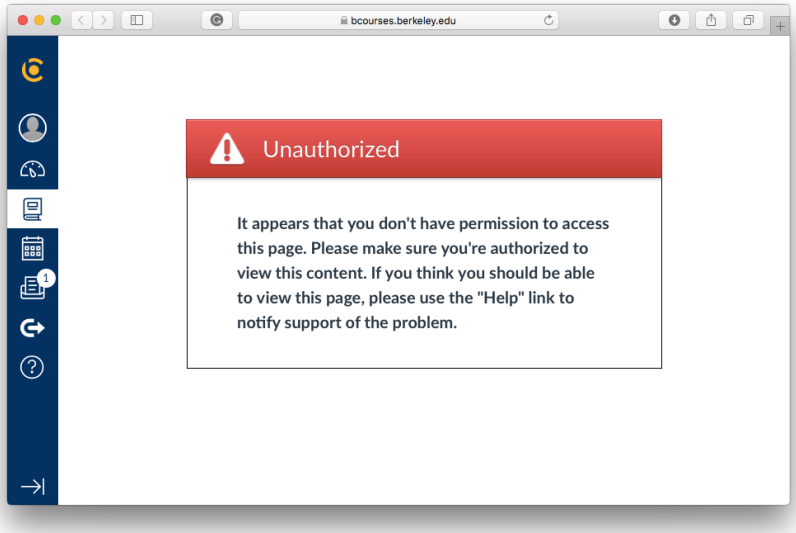
- 1xx: Informational - Request received, continuing process
  - 180 Ringing
- 2xx: Success - The action was successfully received, understood, and accepted
  - 200 OK
- 3xx: Redirection - Further action must be taken in order to complete the request
  - 302 Moved Temporarily
- 4xx: Client Error - The request contains bad syntax or cannot be fulfilled
  - 404 Page Not found, 403 Unauthorized
- 5xx: Server Error - The server failed to fulfill an apparently valid request
  - 501 Internal Server Error

HTTP responses

- 1xx: Informational - Request received, continuing process
  - 180 Ringing
- 2xx: Success - The action was successfully received, understood, and accepted
  - 200 OK
- 3xx: Redirection - Further action must be taken in order to complete the request
  - 302 Moved Temporarily
- 4xx: Client Error - The request contains bad syntax or cannot be fulfilled
  - 404 Not found, 401 Unauthorized
- 5xx: Server Error - The server failed to fulfill an apparently valid request
  - 501 Internal Server Error

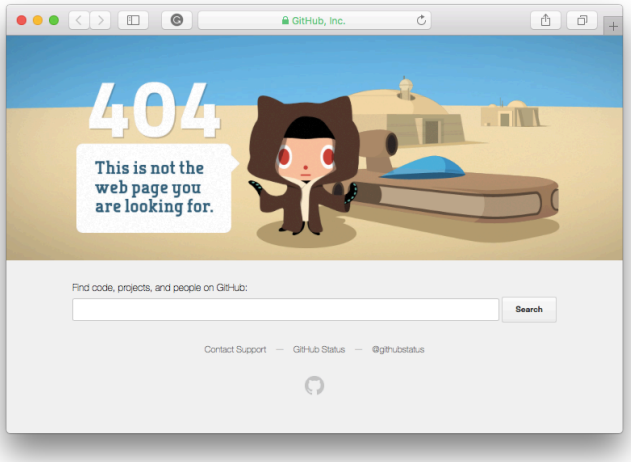
HTTP responses

- 1xx: Informational - Request received, continuing process
  - 180 Ringing
- 2xx: Success - The action was successfully received, understood, and accepted
  - 200 OK
- 3xx: Redirection - Further action must be taken in order to complete the request
  - 303 Moved Temporarily
- 4xx: Client Error - The request contains bad syntax or cannot be fulfilled
  - 404 Not found, 401 Unauthorized
- 5xx: Server Error - The server failed to fulfill an apparently valid request
  - 501 Internal Server Error



HTTP responses

- 1xx: Informational - Request received, continuing process
  - 180 Ringing
- 2xx: Success - The action was successfully received, understood, and accepted
  - 200 OK
- 3xx: Redirection - Further action must be taken in order to complete the request
  - 303 Moved Temporarily
- 4xx: Client Error - The request contains bad syntax or cannot be fulfilled
  - 404 Not found, 401 Unauthorized
- 5xx: Server Error - The server failed to fulfill an apparently valid request
  - 501 Internal Server Error



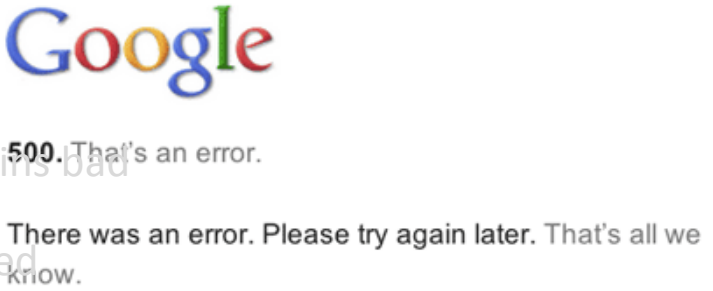
HTTP responses

- 1xx: Informational - Request received, continuing process
  - 180 Ringing
- 2xx: Success - The action was successfully received, understood, and accepted
  - 200 OK
- 3xx: Redirection - Further action must be taken in order to complete the request
  - 303 Moved Temporarily
- 4xx: Client Error - The request contains bad syntax or cannot be fulfilled
  - 404 Not found, 401 Unauthorized
- 5xx: Server Error - The server failed to fulfill an apparently valid request
  - 501 Internal Server Error

https://i0.wp.com/s3.amazonaws.com/production-wordpress-assets/blog/wp-content/uploads/2016/11/29074529/500-internal-server-error.png?fit=604%2C237&ssl=1

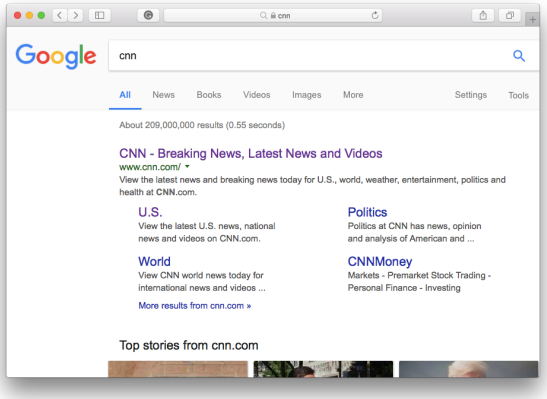
HTTP responses

- 1xx: Informational - Request received, continuing process
  - 180 Ringing
- 2xx: Success - The action was successfully received, understood, and accepted
  - 200 OK
- 3xx: Redirection - Further action must be taken in order to complete the request
  - 303 Moved Temporarily
- 4xx: Client Error - The request contains bad syntax or cannot be fulfilled
  - 404 Not found, 401 Unauthorized
- 5xx: Server Error - The server failed to fulfill an apparently valid request
  - 501 Internal Server Error



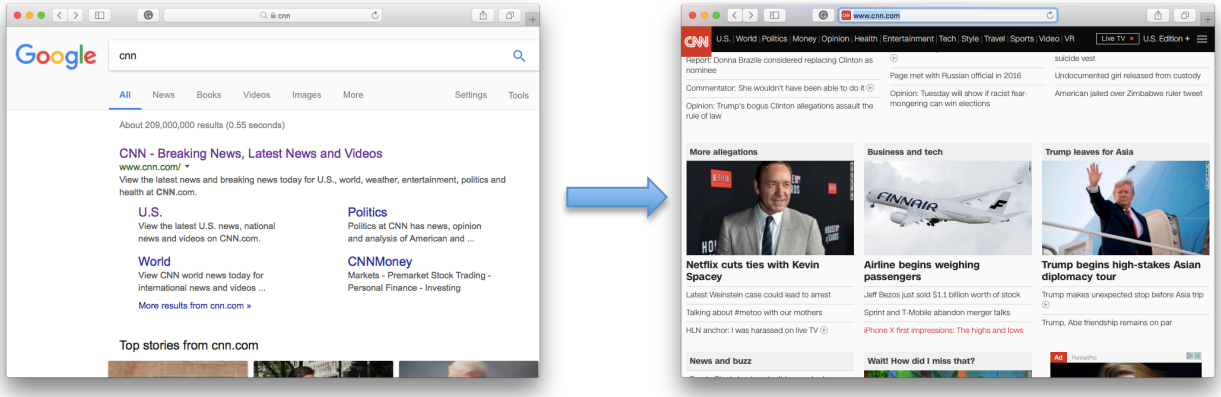
https://i0.wp.com/s3.amazonaws.com/production-wordpress-assets/blog/wp-content/uploads/2016/11/29074529/500-internal-server-error.png?fit=604%2C237&ssl=1

What happen when I surf the internet?



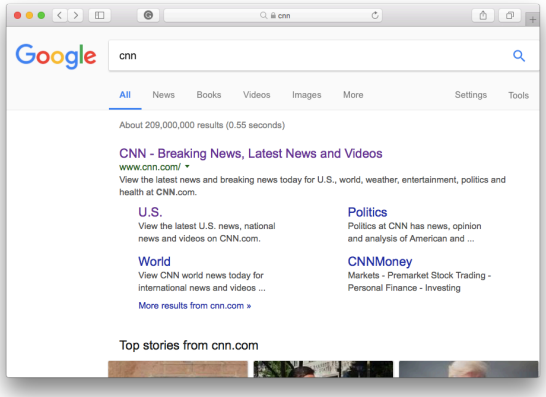

```
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433
”https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7
(KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7”
```

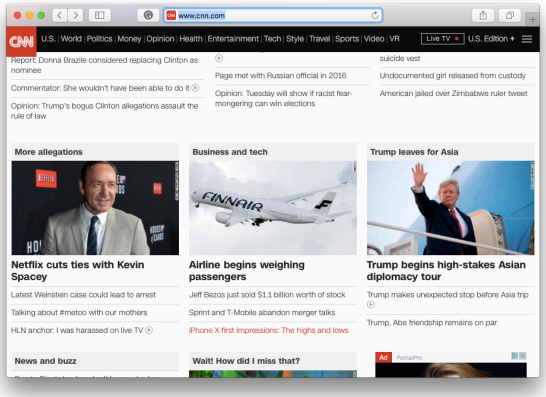
What happen when I surf the internet?

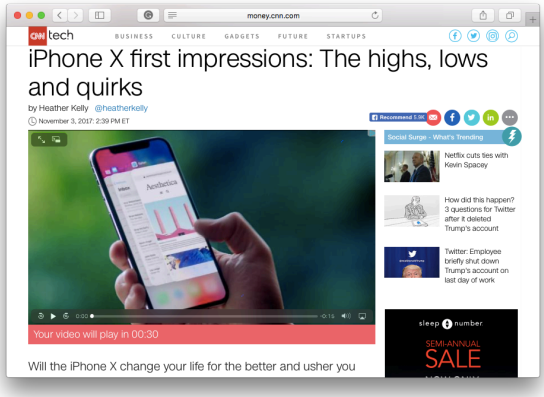



```
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET https://www.google.com HTTP/1.1" 200 96433 ” "
"Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/
16.0.912.77 Safari/535.7”

0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433
”https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7
(KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7”
```

What happen when I surf the internet?








```
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET https://www.google.com HTTP/1.1" 200 96433 ” "
"Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/
16.0.912.77 Safari/535.7”

0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433
”https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7
(KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7”

0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://money.cnn.com/2017/10/31/technology/
gadgets/iphone-x-first-impressions/index.html HTTP/1.1" 200 96433 ” http://www.cnn.com"
"Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/
16.0.912.77 Safari/535.7”
```

What else do you think is recorded in log files?

```
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET https://www.google.com HTTP/1.1" 200 96433 ” " "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like
Gecko) Chrome/16.0.912.77 Safari/535.7”

0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433 ”https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/
535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7”

0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://money.cnn.com/2017/10/31/technology/gadgets/iphone-x-first-impressions/index.html HTTP/1.1" 200 96433 ” http://
www.cnn.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7”
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET https://www.google.com HTTP/1.1" 200 96433 ” " "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like
Gecko) Chrome/16.0.912.77 Safari/535.7”

0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433 ”https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/
535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7”

0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://money.cnn.com/2017/10/31/technology/gadgets/iphone-x-first-impressions/index.html HTTP/1.1" 200 96433 ” http://
www.cnn.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7”
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET https://www.google.com HTTP/1.1" 200 96433 ” " "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like
Gecko) Chrome/16.0.912.77 Safari/535.7”

0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433 ”https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/
535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7”

0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://money.cnn.com/2017/10/31/technology/gadgets/iphone-x-first-impressions/index.html HTTP/1.1" 200 96433 ” http://
www.cnn.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7”
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET https://www.google.com HTTP/1.1" 200 96433 ” " "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like
Gecko) Chrome/16.0.912.77 Safari/535.7”

0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433 ”https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/
535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7”

0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://money.cnn.com/2017/10/31/technology/gadgets/iphone-x-first-impressions/index.html HTTP/1.1" 200 96433 ” http://
www.cnn.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7”
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET https://www.google.com HTTP/1.1" 200 96433 ” " "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like
Gecko) Chrome/16.0.912.77 Safari/535.7”

0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433 ”https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/
535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7”

0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://money.cnn.com/2017/10/31/technology/gadgets/iphone-x-first-impressions/index.html HTTP/1.1" 200 96433 ” http://
www.cnn.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7”


0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://money.cnn.com/2017/10/31/technology/gadgets/iphone-x-first-impressions/index.html HTTP/1.1" 200 96433 ” http://
www.cnn.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7”
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET https://www.google.com HTTP/1.1" 200 96433 ” " "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like
Gecko) Chrome/16.0.912.77 Safari/535.7”

0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433 ”https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/
535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7”

0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://money.cnn.com/2017/10/31/technology/gadgets/iphone-x-first-impressions/index.html HTTP/1.1" 200 96433 ” http://
www.cnn.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7”
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET https://www.google.com HTTP/1.1" 200 96433 ” " "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like
Gecko) Chrome/16.0.912.77 Safari/535.7”

0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433 ”https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/
535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7”

0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://money.cnn.com/2017/10/31/technology/gadgets/iphone-x-first-impressions/index.html HTTP/1.1" 200 96433 ” http://
www.cnn.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7”
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET https://www.google.com HTTP/1.1" 200 96433 ” " "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like
Gecko) Chrome/16.0.912.77 Safari/535.7”

0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433 ”https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/
535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7”

0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://money.cnn.com/2017/10/31/technology/gadgets/iphone-x-first-impressions/index.html HTTP/1.1" 200 96433 ” http://
www.cnn.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7”
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET https://www.google.com HTTP/1.1" 200 96433 ” " "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like
Gecko) Chrome/16.0.912.77 Safari/535.7”
```

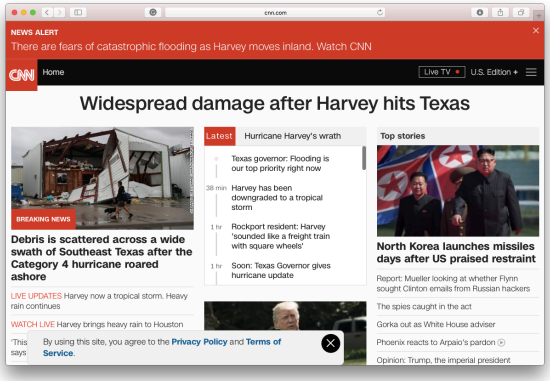
Web usage data
Server log files

- Each time a client requests a resource the server of that resource may record the following in its log files:
  - Client IP
  - Access time
  - HTTP request method
  - URI
  - Protocol
  - HTTP status code
  - Bytes sent
  - Referring URI
  - User-Agent

Server log files

```
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433 ”https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7"
```

- Client IP: 0.247.222.86

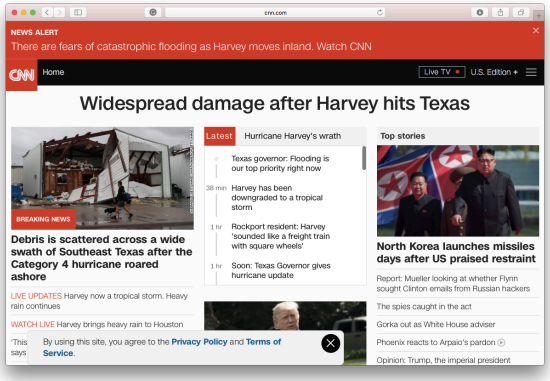


http://www.cnn.com

Server log files

```
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/
1.1" 200 96433 ”https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac
OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77
Safari/535.7"
```

- Client IP: 0.247.222.86
- Access time: 0/Sep/2017:07+0000

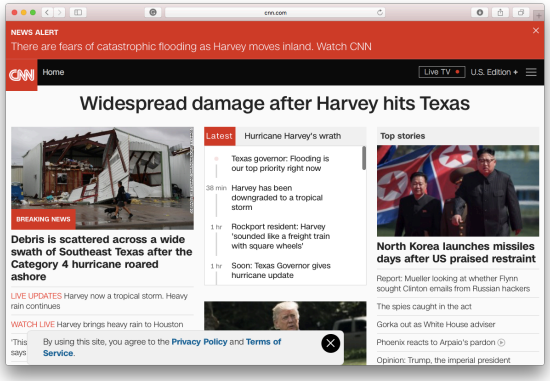


http://www.cnn.com

Server log files

```
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/
1.1" 200 96433 ”https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac
OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77
Safari/535.7"
```

- Client IP: 0.247.222.86
- Access time: 0/Sep/2017:07+0000
- HTTP request method: GET

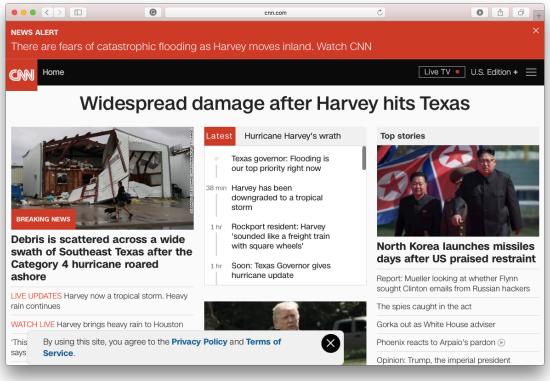


http://www.cnn.com

Server log files

```
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/
1.1" 200 96433 "https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac
OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77
Safari/535.7"
```

- Client IP: 0.247.222.86
- Access time: 0/Sep/2017:07+0000
- HTTP request method: GET
- URI: http://www.cnn.com

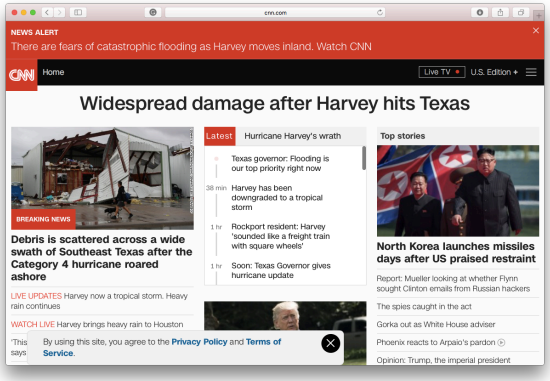


http://www.cnn.com

Server log files

```
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433 ”https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7"
```

- Client IP: 0.247.222.86
- Access time: 0/Sep/2017:07+0000
- HTTP request method: GET
- URI: http://www.cnn.com
- Protocol: HTTP/1.1

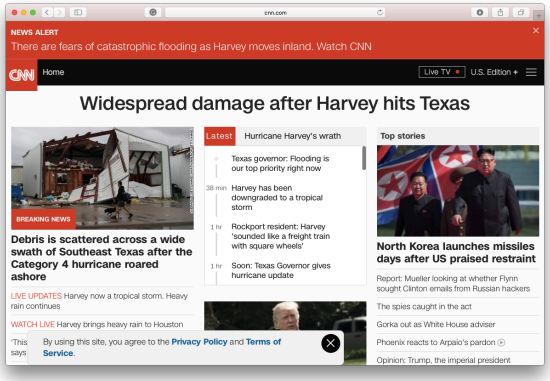


http://www.cnn.com

Server log files

```
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433 ”https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7"
```

- Client IP: 0.247.222.86
- Access time: 0/Sep/2017:07+0000
- HTTP request method: GET
- URI: http://www.cnn.com
- Protocol: HTTP/1.1
- HTTP status code: 200

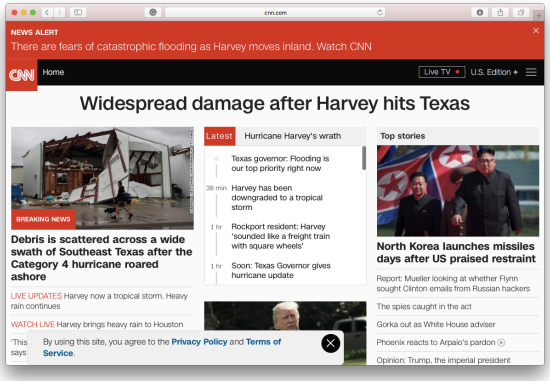


http://www.cnn.com

Server log files

```
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433 ”https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7"
```

- Client IP: 0.247.222.86
- Access time: 0/Sep/2017:07+0000
- HTTP request method: GET
- URI: http://www.cnn.com
- Protocol: HTTP/1.1
- HTTP status code: 200
- Bytes sent: 96433

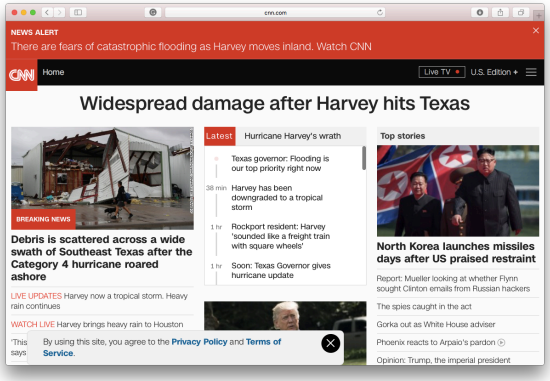


http://www.cnn.com

Server log files

```
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433 "https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7"
```

- Client IP: 0.247.222.86
- Access time: 0/Sep/2017:07+0000
- HTTP request method: GET
- URI: http://www.cnn.com
- Protocol: HTTP/1.1
- HTTP status code: 200
- Bytes sent: 96433
- Referring URI: https://www.google.com

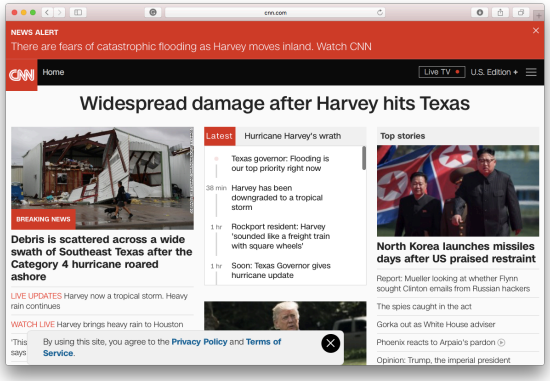


http://www.cnn.com

Server log files

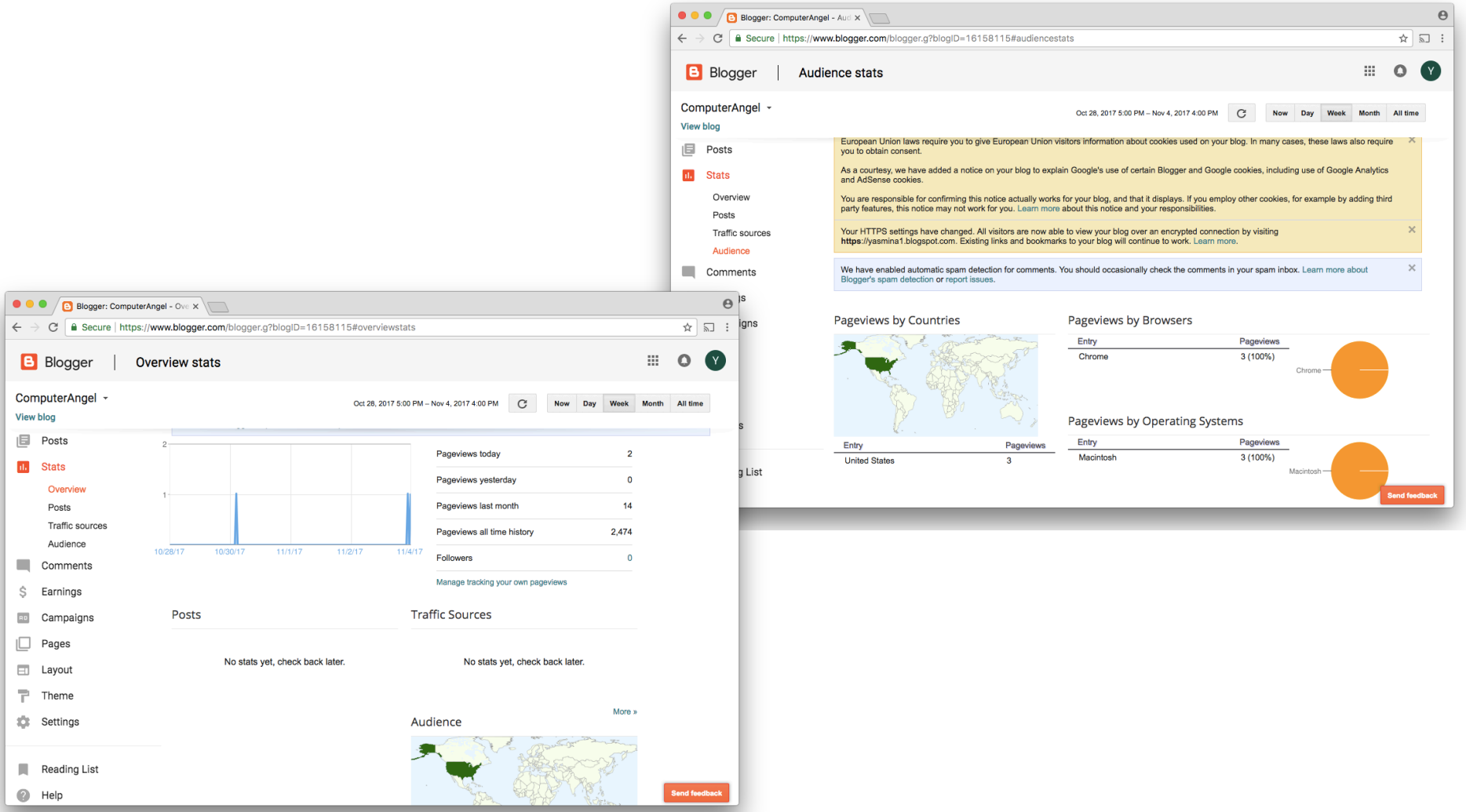
```
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433 "https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7"
```

- Client IP: 0.247.222.86
- Access time: 0/Sep/2017:07+0000
- HTTP request method: GET
- URI: http://www.cnn.com
- Protocol: HTTP/1.1
- HTTP status code: 200
- Bytes sent: 96433
- Referring URI: https://www.google.com
- User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7

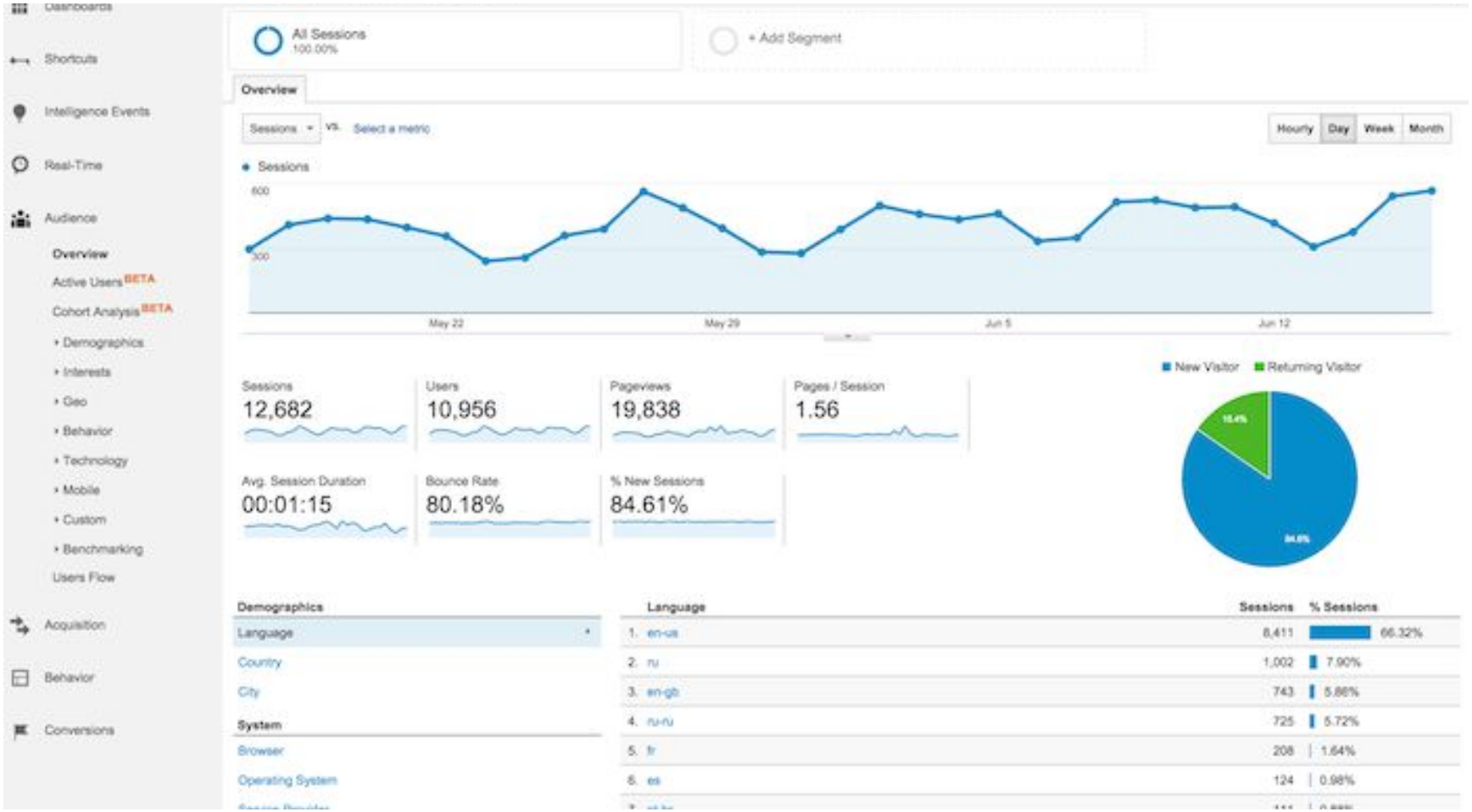


http://www.cnn.com

What we get from the log
Web usage data processing



Google analytics



Source: https://moz.com/blog/absolute-beginners-guide-to-google-analytics

Analyzing Web usage data

- Data preprocessing
- Data cleaning
- User identification
- Session identification
- Pattern detection

Data preprocessing

- Parse log files

```
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/
1.1" 200 96433 ”https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac
OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77
Safari/535.7"
```

Client IP	Access time	HTTP request method	URI	Protocol	HTTP status code	Bytes sent	Referring URI	User-Agent
0.247.222.86	0/Sep/ 2017:07+0000	GET	http://www.cnn.com	HTTP/1.1	200	96433	https://www.google.com	Mozilla/5.0 (Macintosh; …

Python library for parsing logs: https://github.com/rory/apache-log-parser

Data cleaning

Filter our irrelevant requests

0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET https://www.google.com HTTP/1.1" 200 96433 " " "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7"

0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433 "https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7"

0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://money.cnn.com/2017/10/31/technology/gadgets/iphone-x-first-impressions/index.html HTTP/1.1" 200 96433 " http://www.cnn.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7"

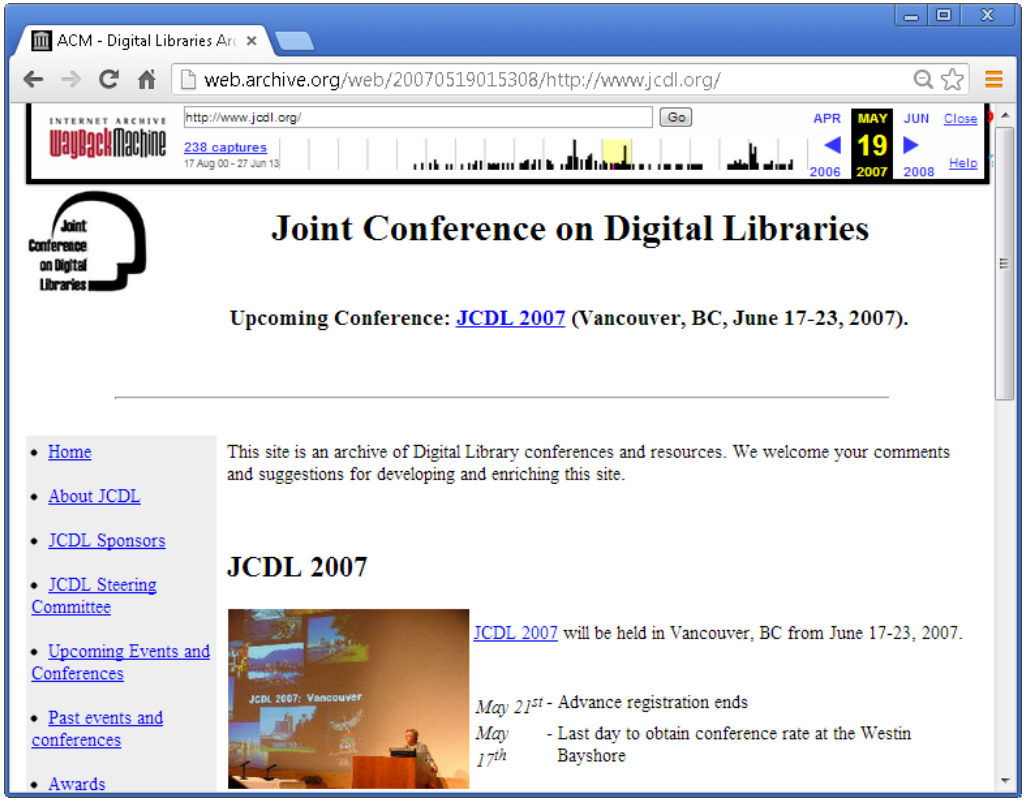
But I viewed only two cnn pages = 2 requests

~400 requests

Data cleaning

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET
http://web.archive.org/web/20070519015308/
http://www.jcdl.org/  HTTP/1.1" 200 2137 "-"
"Mozilla/5.0"

0.11.160.135 [02/Feb/2012:00:01:03] "GET
http://web.archive.org/web/20070519015308im_/
http://www.jcdl.org/images/jcdl2007-edie.jpg
HTTP/1.1" 200 2137 "-" "Mozilla/5.0"

0.11.160.135 [02/Feb/2012:00:01:03] "GET
http://staticweb.archive.org/images/toolbar/
wayback-toolbar-logo.png HTTP/1.1" 200 3700
"-" "Mozilla/5.0"

0.151.147.108 [02/Feb/2012:00:01:03] "GET
http://web.archive.org/web/20100102003557/
about:blank HTTP/1.1" 302 0 "www.xx.com"
"Mozilla/4.0"

0.26.129.146 - - [02/Feb/2012:00:01:54] "GET
http://web.archive.org/web/20140004100000/
http://www.jcdl.org/ HTTP/1.1" 302 0 "-"
"Mozilla/5.0"
```

http://web.archive.org/web/20070519015308/
http://www.jcdl.org/



Embedded resources

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET
http://web.archive.org/web/20070519015308/
http://www.jcdl.org/  HTTP/1.1" 200 2137 "-"
"Mozilla/5.0"

0.11.160.135 [02/Feb/2012:00:01:03] "GET
http://web.archive.org/web/20070519015308im_/
http://www.jcdl.org/images/jcdl2007-edie.jpg
HTTP/1.1" 200 2137 "-" "Mozilla/5.0"

0.11.160.135 [02/Feb/2012:00:01:03] "GET
http://staticweb.archive.org/images/toolbar/
wayback-toolbar-logo.png HTTP/1.1" 200 3700
"-" "Mozilla/5.0"

0.151.147.108 [02/Feb/2012:00:01:03] "GET
http://web.archive.org/web/20100102003557/
about:blank HTTP/1.1" 302 0 "www.xx.com"
"Mozilla/4.0"

0.26.129.146 - - [02/Feb/2012:00:01:54] "GET
http://web.archive.org/web/20140004100000/
http://www.jcdl.org/ HTTP/1.1" 302 0 "-"
"Mozilla/5.0"
```

http://web.archive.org/web/20070519015308/
http://www.jcdl.org/

Embedded resources

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET
http://web.archive.org/web/20070519015308/
http://www.jcdl.org/  HTTP/1.1" 200 2137 "-"
"Mozilla/5.0"

0.11.160.135 [02/Feb/2012:00:01:03] "GET
http://web.archive.org/web/20070519015308im_/
http://www.jcdl.org/jcdl2007-edie.jpg
HTTP/1.1" 200 2137 "-" "Mozilla/5.0"

0.11.160.135 [02/Feb/2012:00:01:03] "GET
http://staticweb.archive.org/images/toolbar/
wayback-toolbar-logo.png HTTP/1.1" 200 3700
"-" "Mozilla/5.0"

0.151.147.108 [02/Feb/2012:00:01:03] "GET
http://web.archive.org/web/20100102003557/
about:blank HTTP/1.1" 302 0 "www.xx.com"
"Mozilla/4.0"

0.26.129.146 - - [02/Feb/2012:00:01:54] "GET
http://web.archive.org/web/20140004100000/
http://www.jcdl.org/ HTTP/1.1" 302 0 "-"
"Mozilla/5.0"
```

http://web.archive.org/web/20070519015308/
http://www.jcdl.org/

Static resources

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET
http://web.archive.org/web/20070519015308/
http://www.jcdl.org/  HTTP/1.1" 200 2137 "-"
"Mozilla/5.0"

0.11.160.135 [02/Feb/2012:00:01:03] "GET
http://web.archive.org/web/20070519015308im_/
http://www.jcdl.org/jcdl2007-edie.jpg
HTTP/1.1" 200 2137 "-" "Mozilla/5.0"

0.11.160.135 [02/Feb/2012:00:01:03] "GET
http://staticweb.archive.org/images/toolbar/
wayback-toolbar-logo.png HTTP/1.1" 200 3700
"-" "Mozilla/5.0"

0.151.147.108 [02/Feb/2012:00:01:03] "GET
http://web.archive.org/web/20100102003557/
about:blank HTTP/1.1" 302 0 "www.xx.com"
"Mozilla/4.0"

0.26.129.146 - - [02/Feb/2012:00:01:54] "GET
http://web.archive.org/web/20140004100000/
http://www.jcdl.org/ HTTP/1.1" 302 0 "-"
"Mozilla/5.0"
```

http://web.archive.org/web/20070519015308/
http://www.jcdl.org/

Static resources

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET
http://web.archive.org/web/20070519015308/
http://www.jcdl.org/  HTTP/1.1" 200 2137 "-"
"Mozilla/5.0"

0.11.160.135 [02/Feb/2012:00:01:03] "GET
http://web.archive.org/web/20070519015308im_/
http://www.jcdl.org/jcdl2007-edie.jpg
HTTP/1.1" 200 2137 "-" "Mozilla/5.0"

0.11.160.135 [02/Feb/2012:00:01:03] "GET
http://staticweb.archive.org/images/toolbar/
wayback-toolbar-logo.png HTTP/1.1" 200 3700
"-" "Mozilla/5.0"

0.151.147.108 [02/Feb/2012:00:01:03] "GET
http://web.archive.org/web/20100102003557/
about:blank HTTP/1.1" 302 0 "www.xx.com"
"Mozilla/4.0"

0.26.129.146 - - [02/Feb/2012:00:01:54] "GET
http://web.archive.org/web/20140004100000/
http://www.jcdl.org/ HTTP/1.1" 302 0 "-"
"Mozilla/5.0"
```

http://web.archive.org/web/20070519015308/
http://www.jcdl.org/

User identification

- Grouping: based on the IP and User-Agent

```
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET https://www.google.com HTTP/1.1" 200 96433 ” "
"Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/
16.0.912.77 Safari/535.7”

0.247.222.86 - - [0/Sep/2017:07:05:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433
”https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7
(KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7”

0.247.222.86 - - [0/Sep/2017:07:12:46 +0000] "GET http://money.cnn.com/2017/10/31/technology/
gadgets/iphone-x-first-impressions/index.html HTTP/1.1" 200 96433 ” http://www.cnn.com"
"Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/
16.0.912.77 Safari/535.7”
```

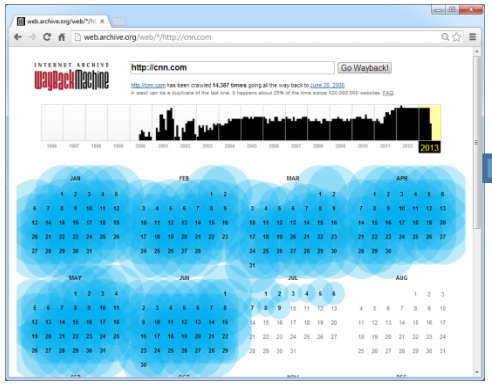
Same IP

Session identification

- Segment the user activity record of each user into sessions, each representing a single visit to the site
- Threshold timeout: 10 minutes Liu et al. 2007, Spiliopoulou et al. 2003

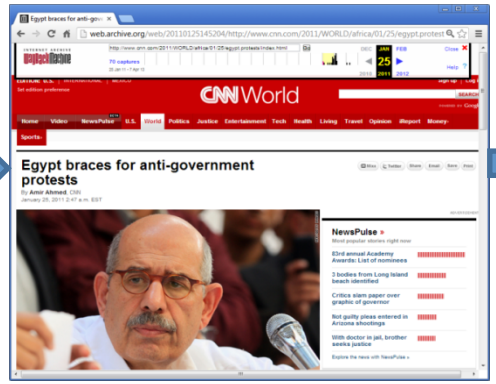
```
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET https://www.google.com HTTP/1.1" 200 96433 ” "
"Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/
16.0.912.77 Safari/535.7”

0.247.222.86 - - [0/Sep/2017:07:05:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433
”https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7
(KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7”

0.247.222.86 - - [0/Sep/2017:07:12:46 +0000] "GET http://money.cnn.com/2017/10/31/technology/
gadgets/iphone-x-first-impressions/index.html HTTP/1.1" 200 96433 ” http://www.cnn.com"
"Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/
16.0.912.77 Safari/535.7”
```

Session: set of Web pages requested by a particular user in a single visit



p1

1 mins



p2

4 mins

p3

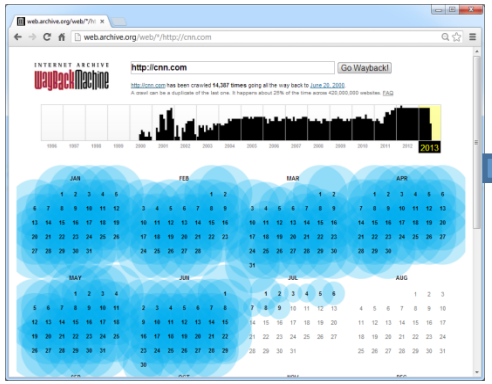
3 mins

p4

9 mins

p5

Session: set of Web pages requested by a particular user in a single visit



p1

1 mins

4 mins

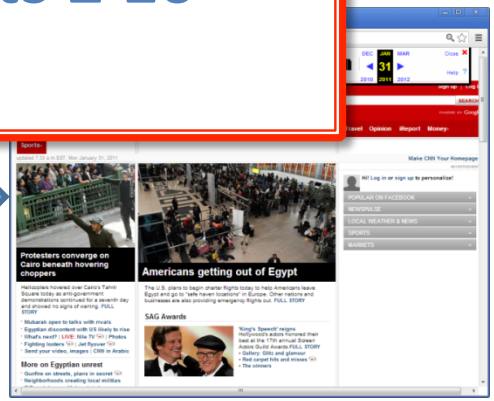
p3

Time between two requests ≤ 10

3 mins

p4

9 mins



p5

User 1

User 2

Session 1

Session 2

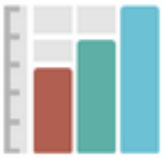
Session 3

Session 4

Session 1

Session 2

Session 2



page views

users

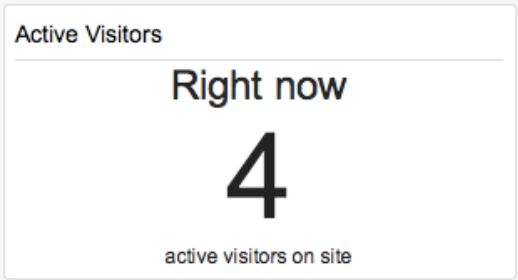
visits

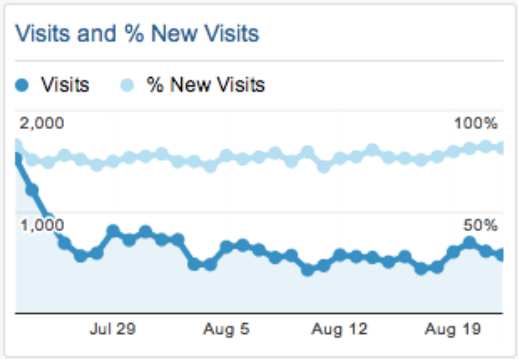
time spent on the site

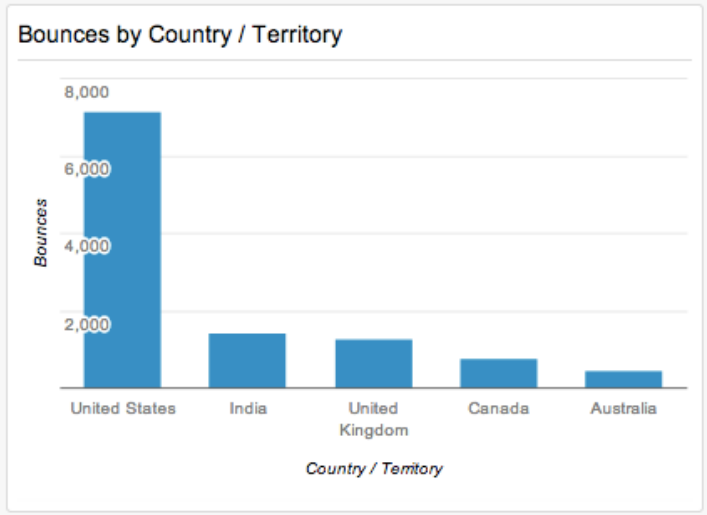
where people come from

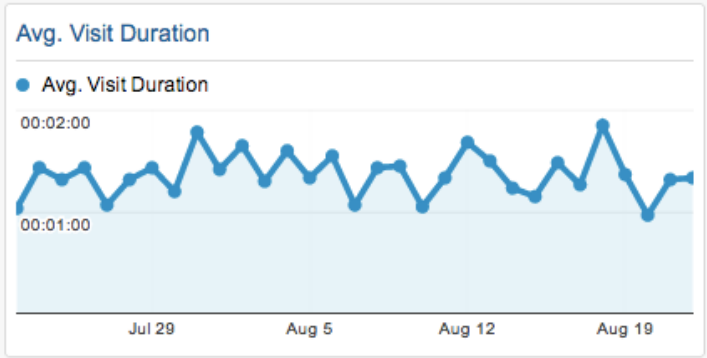
Log file → Users → sessions → patterns

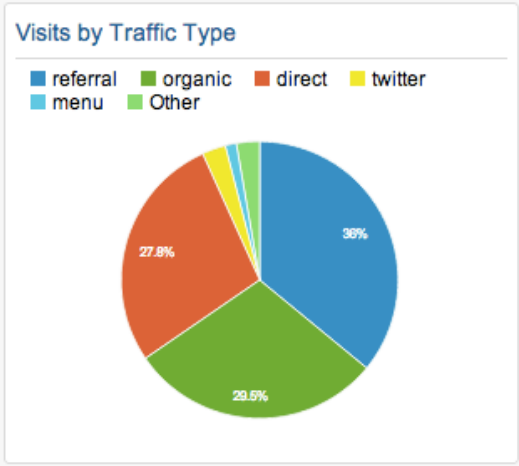
Google analytics

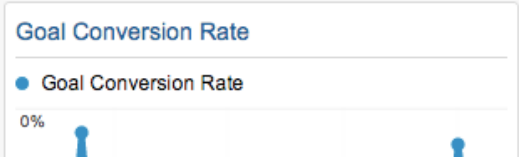




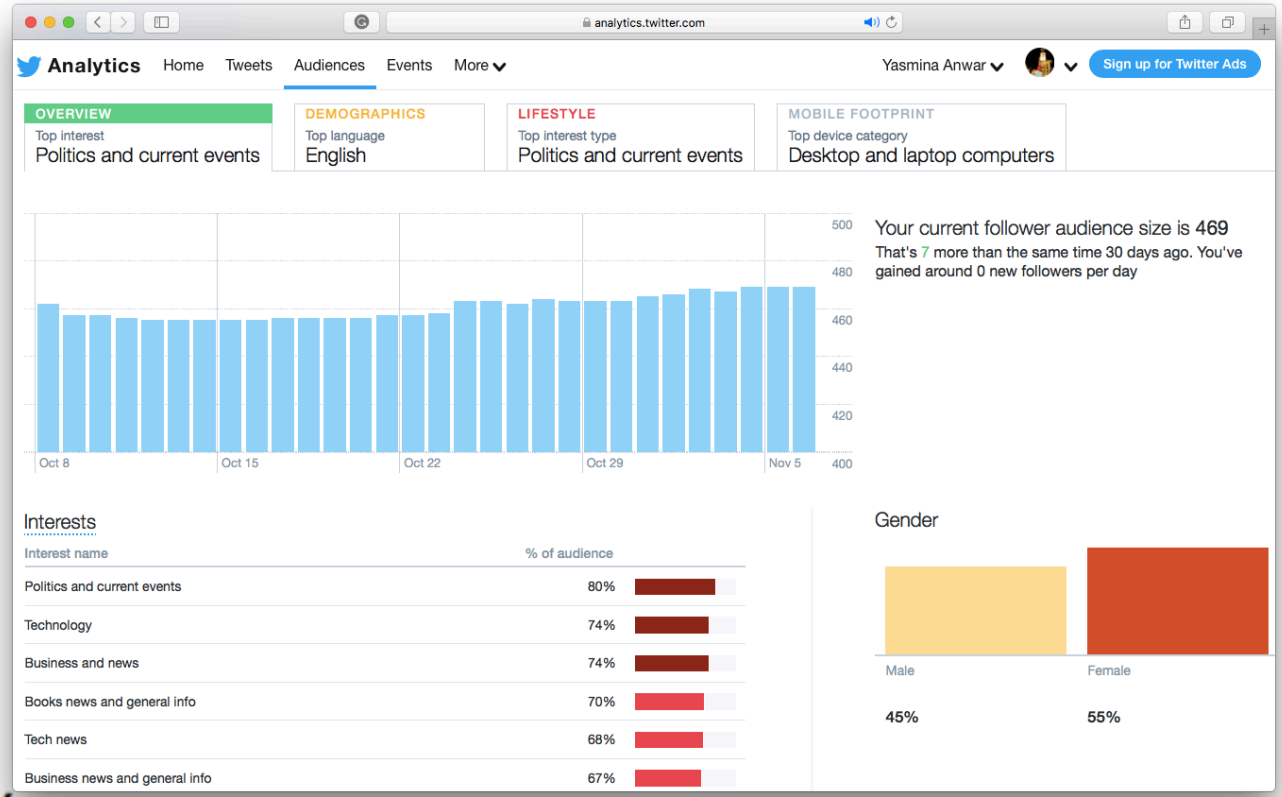


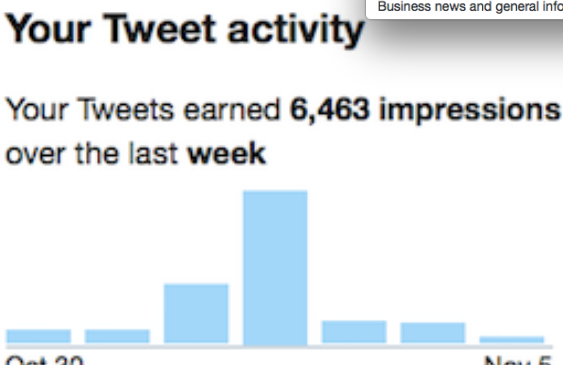






Twitter activity





Resources

- Chapter 12: Web Usage Mining by B. Mobasher (pdf)
- Access Patterns for Robots and Humans in Web Archives (pdf)

Next

- How to generate good visualizations