ECPR Methods Summer School: Automated Collection of Web and Social Data

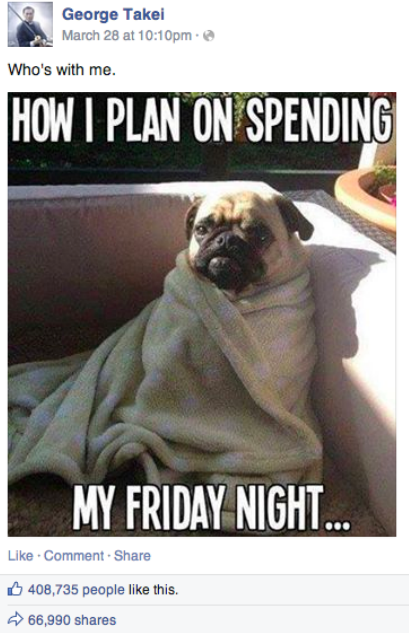
Pablo Barberá

School of International Relations
University of Southern California
pablobarbera.com

Networked Democracy Lab
www.netdem.org

Course website:

github.com/pablobarbera/ECPR-SC103



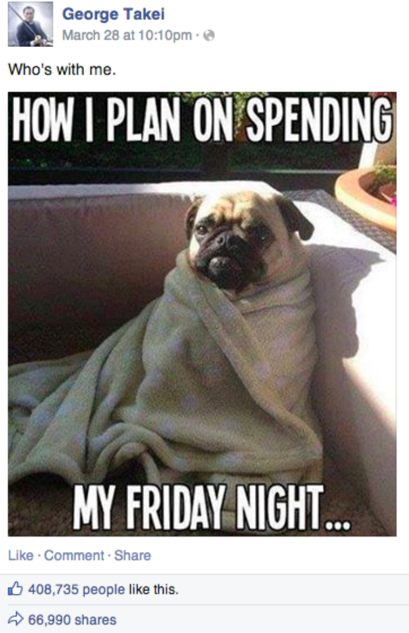
George Takei
March 28 at 10:10pm ·

Who's with me.

Like · Comment · Share

408,735 people like this.

66,990 shares



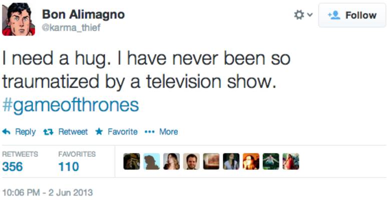
George Takei
March 28 at 10:10pm ·

Who's with me.

Like · Comment · Share

408,735 people like this.

66,990 shares

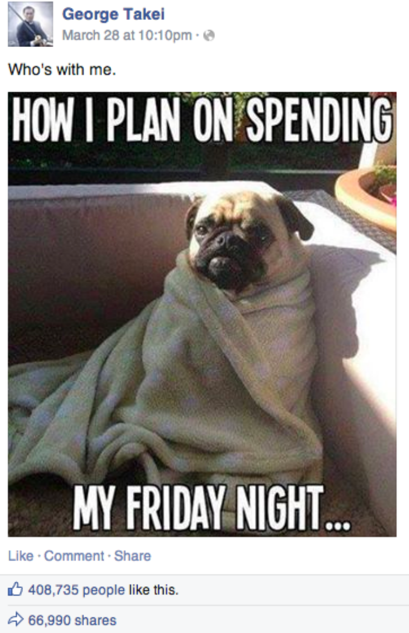


Bon Alimagno
@karma_thief

I need a hug. I have never been so traumatized by a television show. #gameofthrones

RETWEETS	FAVORITES
356	110

10:06 PM - 2 Jun 2013

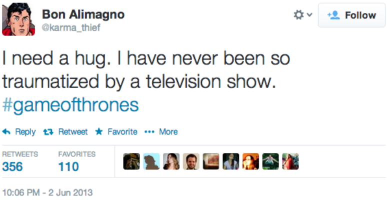


George Takei
March 28 at 10:10pm ·

Who's with me.

408,735 people like this.

66,990 shares

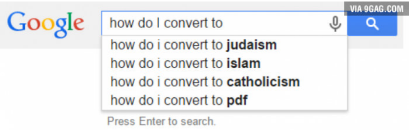


Bon Alimagno
@karma_thief

I need a hug. I have never been so traumatized by a television show. #gameofthrones

RETWEETS 356
FAVORITES 110

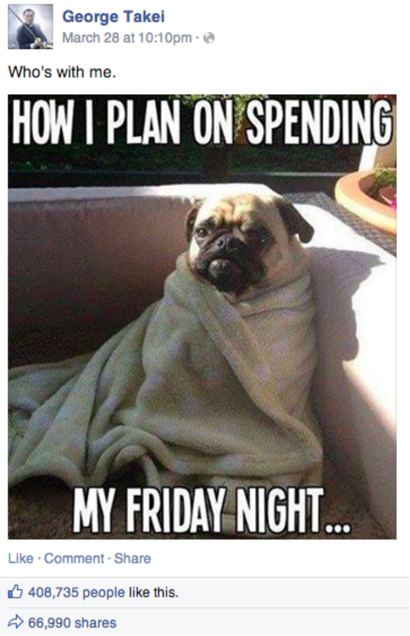
10:06 PM - 2 Jun 2013



how do I convert to

- how do i convert to **judaism**
- how do i convert to **islam**
- how do i convert to **catholicism**
- how do i convert to **pdf**

Press Enter to search.



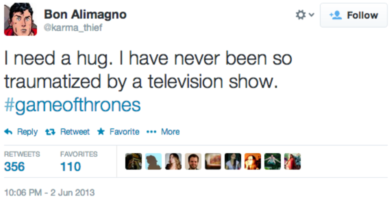
George Takei
March 28 at 10:10pm ·

Who's with me.

Like · Comment · Share

408,735 people like this.

66,990 shares

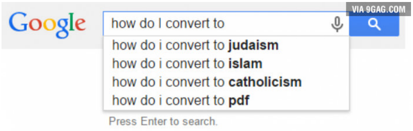


Bon Alimagno
@karma_thief

I need a hug. I have never been so traumatized by a television show. #gameofthrones

RETWEETS	FAVORITES
356	110

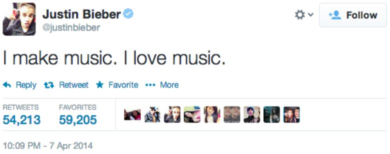
10:06 PM - 2 Jun 2013



how do I convert to

how do i convert to **judaism**
how do i convert to **islam**
how do i convert to **catholicism**
how do i convert to **pdf**

Press Enter to search.

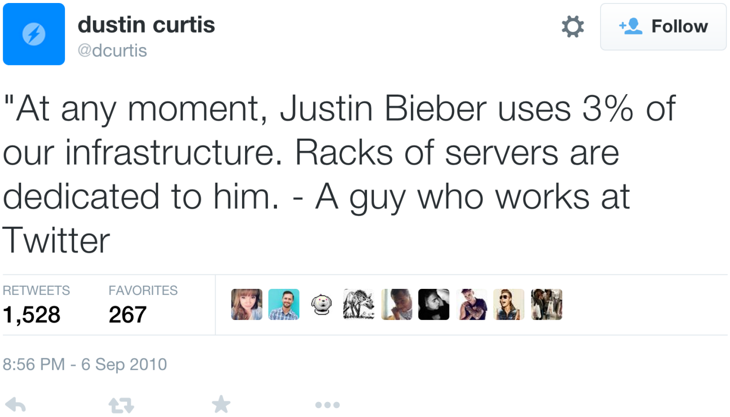


Justin Bieber
@justinbieber

I make music. I love music.

RETWEETS	FAVORITES
54,213	59,205

10:09 PM - 7 Apr 2014



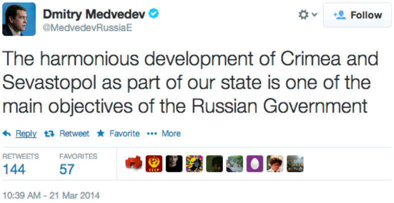
dustin curtis
@dcurtis

"At any moment, Justin Bieber uses 3% of our infrastructure. Racks of servers are dedicated to him. - A guy who works at Twitter

RETWEETS
1,528

FAVORITES
267

8:56 PM - 6 Sep 2010

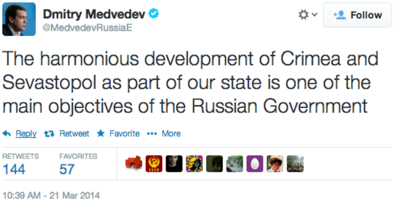


Dmitry Medvedev
@MedvedevRussiaE

The harmonious development of Crimea and Sevastopol as part of our state is one of the main objectives of the Russian Government

RETWEETS 144 FAVORITES 57

10:39 AM - 21 Mar 2014

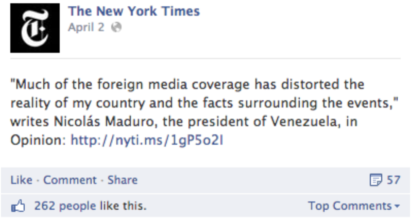


Dmitry Medvedev
@MedvedevRussiaE

The harmonious development of Crimea and Sevastopol as part of our state is one of the main objectives of the Russian Government

RETWEETS 144 FAVORITES 57

10:39 AM - 21 Mar 2014

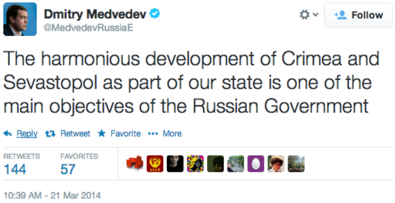


The New York Times
April 2

"Much of the foreign media coverage has distorted the reality of my country and the facts surrounding the events," writes Nicolás Maduro, the president of Venezuela, in Opinion: http://nyti.ms/1gP5o2I

57

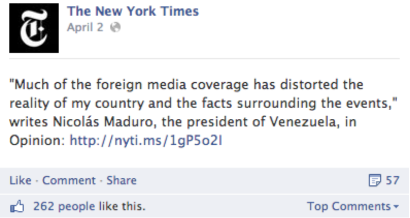
262 people like this.


Dmitry Medvedev
@MedvedevRussiaE

The harmonious development of Crimea and Sevastopol as part of our state is one of the main objectives of the Russian Government

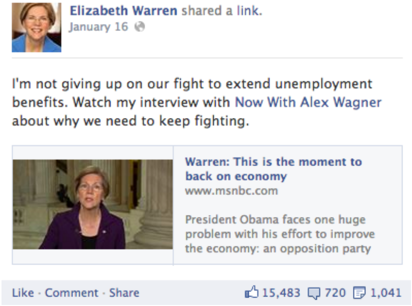
RETWEETS 144 FAVORITES 57

10:39 AM - 21 Mar 2014


The New York Times
April 2

"Much of the foreign media coverage has distorted the reality of my country and the facts surrounding the events," writes Nicolás Maduro, the president of Venezuela, in Opinion: http://nyti.ms/1gP5o2I

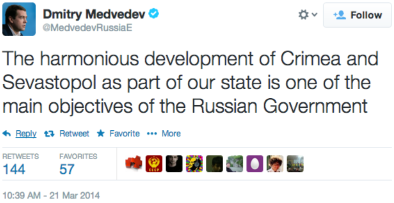
262 people like this.


Elizabeth Warren shared a link.
January 16

I'm not giving up on our fight to extend unemployment benefits. Watch my interview with Now With Alex Wagner about why we need to keep fighting.

Warren: This is the moment to back on economy
www.msnbc.com

President Obama faces one huge problem with his effort to improve the economy: an opposition party



Dmitry Medvedev ✓
@MedvedevRussiaE

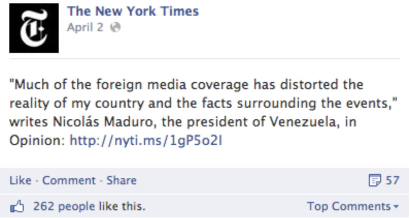
Follow

The harmonious development of Crimea and Sevastopol as part of our state is one of the main objectives of the Russian Government

Reply · Retweet · ★ Favorite · ••• More

RETWEETS	FAVORITES
144	57

10:39 AM - 21 Mar 2014

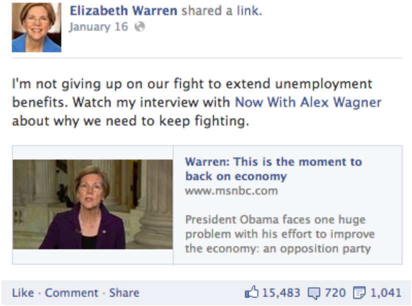


The New York Times
April 2

"Much of the foreign media coverage has distorted the reality of my country and the facts surrounding the events," writes Nicolás Maduro, the president of Venezuela, in Opinion: http://nyti.ms/1gP5o2I

Like · Comment · Share 57

262 people like this. Top Comments ▾



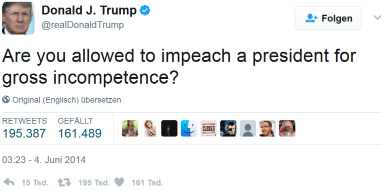
Elizabeth Warren shared a link.
January 16

I'm not giving up on our fight to extend unemployment benefits. Watch my interview with Now With Alex Wagner about why we need to keep fighting.

Warren: This is the moment to back on economy
www.msnbc.com

President Obama faces one huge problem with his effort to improve the economy: an opposition party

Like · Comment · Share 15,483 720 1,041



Donald J. Trump ✓
@realDonaldTrump

Folgen

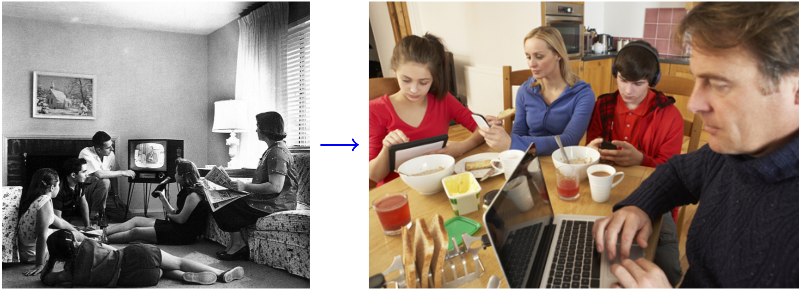
Are you allowed to impeach a president for gross incompetence?

Original (Englisch) übersetzen

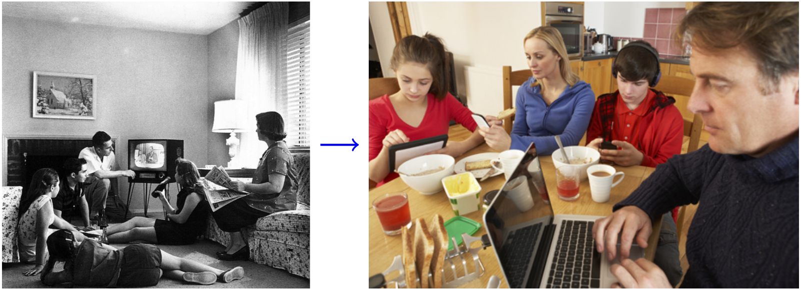
RETWEETS	GEFÄLLT
195.387	161.489

03:23 - 4. Juni 2014

15 Tsd. 195 Tsd. 161 Tsd.



Shift in communication patterns



Shift in communication patterns

Digital footprints of human behavior

This course

1. Collecting web and social data
   - Web scraping: tables, unstructured data, information behind web forms...
   - Through Application Programming Interfaces (APIs)
   - Social media: Twitter, Facebook
2. Manipulating web and social data
   - Network and text data
   - Extracting data from PDF files
   - Cleaning, merging, and reshaping data

Hello!

About me

- Assistant Professor in Computational Social Science at the London School of Economics as of January 2018

About me

- Assistant Professor in Computational Social Science at the London School of Economics as of January 2018
- Currently Assistant Professor at University of Southern California

About me

- Assistant Professor in Computational Social Science at the London School of Economics as of January 2018
- Currently Assistant Professor at University of Southern California
- PhD in Politics, New York University (2015)

About me

- Assistant Professor in Computational Social Science at the London School of Economics as of January 2018
- Currently Assistant Professor at University of Southern California
- PhD in Politics, New York University (2015)
- Data Science Fellow at NYU, 2015–2016

About me

- Assistant Professor in Computational Social Science at the London School of Economics as of January 2018
- Currently Assistant Professor at University of Southern California
- PhD in Politics, New York University (2015)
- Data Science Fellow at NYU, 2015–2016
- My research:

About me

- Assistant Professor in Computational Social Science at the London School of Economics as of January 2018
- Currently Assistant Professor at University of Southern California
- PhD in Politics, New York University (2015)
- Data Science Fellow at NYU, 2015–2016
- My research:
  - Social media and politics, comparative electoral behavior, corruption and accountability

About me

- Assistant Professor in Computational Social Science at the London School of Economics as of January 2018
- Currently Assistant Professor at University of Southern California
- PhD in Politics, New York University (2015)
- Data Science Fellow at NYU, 2015–2016
- My research:
    - Social media and politics, comparative electoral behavior, corruption and accountability
    - Social network analysis, Bayesian statistics, text as data methods

About me

- Assistant Professor in Computational Social Science at the London School of Economics as of January 2018
- Currently Assistant Professor at University of Southern California
- PhD in Politics, New York University (2015)
- Data Science Fellow at NYU, 2015–2016
- My research:
    - Social media and politics, comparative electoral behavior, corruption and accountability
    - Social network analysis, Bayesian statistics, text as data methods
    - Author of R packages to analyze data from social media

About me

- Assistant Professor in Computational Social Science at the London School of Economics as of January 2018
- Currently Assistant Professor at University of Southern California
- PhD in Politics, New York University (2015)
- Data Science Fellow at NYU, 2015–2016
- My research:
  - Social media and politics, comparative electoral behavior, corruption and accountability
  - Social network analysis, Bayesian statistics, text as data methods
  - Author of R packages to analyze data from social media
- Contact:

About me

- Assistant Professor in Computational Social Science at the London School of Economics as of January 2018
- Currently Assistant Professor at University of Southern California
- PhD in Politics, New York University (2015)
- Data Science Fellow at NYU, 2015–2016
- My research:
    - Social media and politics, comparative electoral behavior, corruption and accountability
    - Social network analysis, Bayesian statistics, text as data methods
    - Author of R packages to analyze data from social media
- Contact:
    - `pbarbera@usc.edu`

About me

- Assistant Professor in Computational Social Science at the London School of Economics as of January 2018
- Currently Assistant Professor at University of Southern California
- PhD in Politics, New York University (2015)
- Data Science Fellow at NYU, 2015–2016
- My research:
  - Social media and politics, comparative electoral behavior, corruption and accountability
  - Social network analysis, Bayesian statistics, text as data methods
  - Author of R packages to analyze data from social media
- Contact:
  - `pbarbera@usc.edu`
  - `www.pablobarbera.com`

Scraping the web

Scraping the web: what? why?

An increasing amount of data is available on the web:

- Speeches, sentences, biographical information...

Scraping the web: what? why?

An increasing amount of data is available on the web:

- Speeches, sentences, biographical information...
- Social media data, newspaper articles, press releases...

Scraping the web: what? why?

An increasing amount of data is available on the web:

- Speeches, sentences, biographical information...
- Social media data, newspaper articles, press releases...
- Geographic information, conflict data...

Scraping the web: what? why?

An increasing amount of data is available on the web:

- Speeches, sentences, biographical information...
- Social media data, newspaper articles, press releases...
- Geographic information, conflict data...

These datasets are often provided in an unstructured format.

Scraping the web: what? why?

An increasing amount of data is available on the web:

- Speeches, sentences, biographical information...
- Social media data, newspaper articles, press releases...
- Geographic information, conflict data...

These datasets are often provided in an unstructured format.

Web scraping is the process of extracting this information automatically and transforming it into a structured dataset.

Scraping the web: two approaches

Two different approaches:

1. Screen scraping: extract data from source code of website, with html parser and/or regular expressions

Scraping the web: two approaches

Two different approaches:

1. Screen scraping: extract data from source code of website, with html parser and/or regular expressions
    - `rvest` package in R

Scraping the web: two approaches

Two different approaches:

1. Screen scraping: extract data from source code of website, with html parser and/or regular expressions
   - `rvest` package in R
2. Web APIs (application programming interfaces): a set of structured http requests that return JSON or XML data

Scraping the web: two approaches

Two different approaches:

1. Screen scraping: extract data from source code of website, with html parser and/or regular expressions
   - `rvest` package in R
2. Web APIs (application programming interfaces): a set of structured http requests that return JSON or XML data
   - `httr` package to construct API requests

Scraping the web: two approaches

Two different approaches:

1. Screen scraping: extract data from source code of website, with html parser and/or regular expressions
   - `rvest` package in R
2. Web APIs (application programming interfaces): a set of structured http requests that return JSON or XML data
   - `httr` package to construct API requests
   - Packages specific to each API: weatherData, WDI, Rfacebook... Check CRAN Task View on Web Technologies and Services for more examples

The rules of the game

1. Respect the hosting site's wishes:

The rules of the game

1. Respect the hosting site's wishes:
    - First, check if an API exists or if data are available for download

The rules of the game

1. Respect the hosting site's wishes:
    - First, check if an API exists or if data are available for download
    - Some websites *disallow* scrapers on their `robots.txt` files

The rules of the game

1. Respect the hosting site's wishes:
   - First, check if an API exists or if data are available for download
   - Some websites *disallow* scrapers on their `robots.txt` files

2. Limit your bandwidth use:

The rules of the game

1. Respect the hosting site's wishes:
    - First, check if an API exists or if data are available for download
    - Some websites *disallow* scrapers on their `robots.txt` files

2. Limit your bandwidth use:
    - Wait one or two seconds after each hit

The rules of the game

1. Respect the hosting site's wishes:
   - First, check if an API exists or if data are available for download
   - Some websites *disallow* scrapers on their `robots.txt` files
2. Limit your bandwidth use:
   - Wait one or two seconds after each hit
   - Scrape only what you need, and just once (e.g. store the html file in disk, and then parse it)

The rules of the game

1. Respect the hosting site's wishes:
   - First, check if an API exists or if data are available for download
   - Some websites *disallow* scrapers on their `robots.txt` files
2. Limit your bandwidth use:
   - Wait one or two seconds after each hit
   - Scrape only what you need, and just once (e.g. store the html file in disk, and then parse it)
3. When using APIs, read documentation

The rules of the game

1. Respect the hosting site's wishes:
   - First, check if an API exists or if data are available for download
   - Some websites *disallow* scrapers on their `robots.txt` files

2. Limit your bandwidth use:
   - Wait one or two seconds after each hit
   - Scrape only what you need, and just once (e.g. store the html file in disk, and then parse it)

3. When using APIs, read documentation
   - Is there a batch download option?

The rules of the game

1. Respect the hosting site's wishes:
    - First, check if an API exists or if data are available for download
    - Some websites *disallow* scrapers on their `robots.txt` files
2. Limit your bandwidth use:
    - Wait one or two seconds after each hit
    - Scrape only what you need, and just once (e.g. store the html file in disk, and then parse it)
3. When using APIs, read documentation
    - Is there a batch download option?
    - Are there any rate limits?

The rules of the game

1. Respect the hosting site's wishes:
    - First, check if an API exists or if data are available for download
    - Some websites *disallow* scrapers on their `robots.txt` files
2. Limit your bandwidth use:
    - Wait one or two seconds after each hit
    - Scrape only what you need, and just once (e.g. store the html file in disk, and then parse it)
3. When using APIs, read documentation
    - Is there a batch download option?
    - Are there any rate limits?
    - Can you share the data?

The art of web scraping

Workflow:

1. Learn about structure of website

The art of web scraping

Workflow:

1. Learn about structure of website
2. Build prototype code

The art of web scraping

Workflow:

1. Learn about structure of website
2. Build prototype code
3. Generalize: functions, loops, debugging

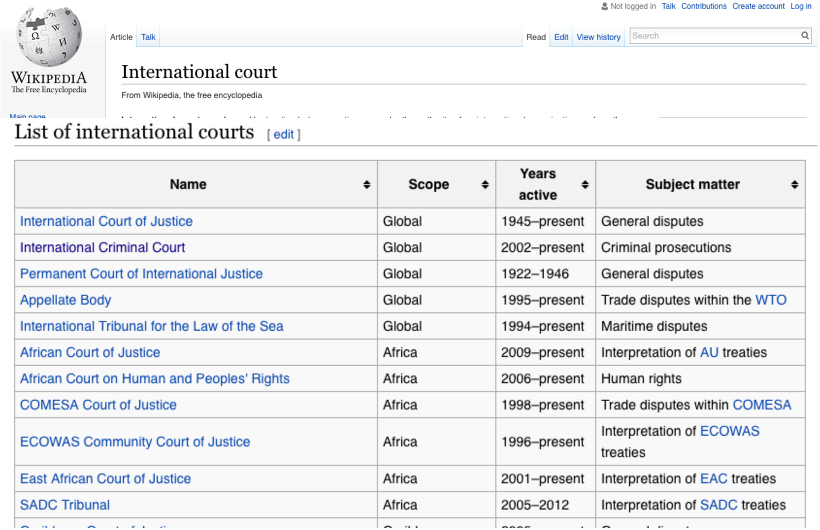
The art of web scraping

Workflow:

1. Learn about structure of website
2. Build prototype code
3. Generalize: functions, loops, debugging
4. Data cleaning

Three main scenarios

1. Data in table format



International court

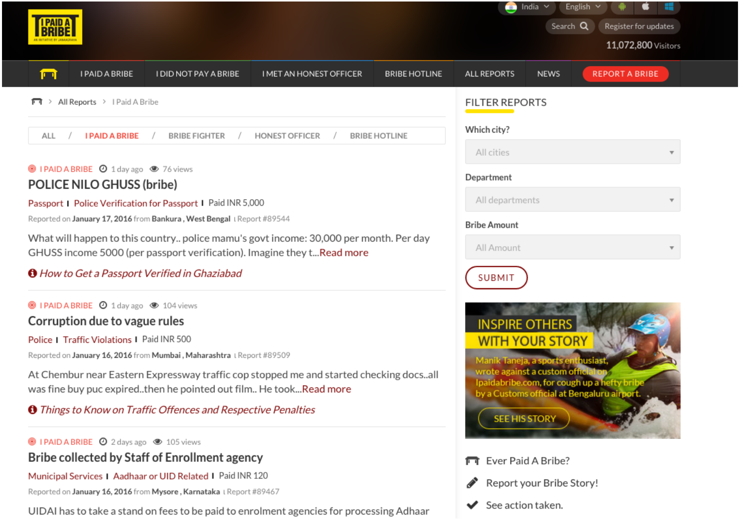
From Wikipedia, the free encyclopedia

List of international courts [edit]

Name	Scope	Years active	Subject matter
International Court of Justice	Global	1945–present	General disputes
International Criminal Court	Global	2002–present	Criminal prosecutions
Permanent Court of International Justice	Global	1922–1946	General disputes
Appellate Body	Global	1995–present	Trade disputes within the WTO
International Tribunal for the Law of the Sea	Global	1994–present	Maritime disputes
African Court of Justice	Africa	2009–present	Interpretation of AU treaties
African Court on Human and Peoples' Rights	Africa	2006–present	Human rights
COMESA Court of Justice	Africa	1998–present	Trade disputes within COMESA
ECOWAS Community Court of Justice	Africa	1996–present	Interpretation of ECOWAS treaties
East African Court of Justice	Africa	2001–present	Interpretation of EAC treaties
SADC Tribunal	Africa	2005–2012	Interpretation of SADC treaties

Three main scenarios

2. Data in unstructured format



11,072,800 Visitors

All Reports > I Paid A Bribe

ALL / I PAID A BRIBE / BRIBE FIGHTER / HONEST OFFICER / BRIBE HOTLINE

I PAID A BRIBE 1 day ago 76 views

POLICE NILO GHUSS (bribe)

Passport | Police Verification for Passport | Paid INR 5,000

Reported on **January 17, 2016** from **Bankura , West Bengal** | Report #89544

What will happen to this country.. police mamu's govt income: 30,000 per month. Per day GHUSS income 5000 (per passport verification). Imagine they t...Read more

How to Get a Passport Verified in Ghaziabad

I PAID A BRIBE 1 day ago 104 views

Corruption due to vague rules

Police | Traffic Violations | Paid INR 500

Reported on **January 16, 2016** from **Mumbai , Maharashtra** | Report #89509

At Chembur near Eastern Expressway traffic cop stopped me and started checking docs..all was fine buy puc expired..then he pointed out film.. He took...Read more

Things to Know on Traffic Offences and Respective Penalties

I PAID A BRIBE 2 days ago 105 views

Bribe collected by Staff of Enrollment agency

Municipal Services | Aadhaar or UID Related | Paid INR 120

Reported on **January 16, 2016** from **Mysore , Karnataka** | Report #89467

UIDAI has to take a stand on fees to be paid to enrolment agencies for processing Adhaar

FILTER REPORTS

Which city?

All cities

Department

All departments

Bribe Amount

All Amount

SUBMIT

INSPIRE OTHERS WITH YOUR STORY

Manik Taneja, a sports enthusiast, wrote against a custom official on Ipaidabribe.com, for cough up a hefty bribe by a Customs official at Bengaluru airport.

SEE HIS STORY

Ever Paid A Bribe?

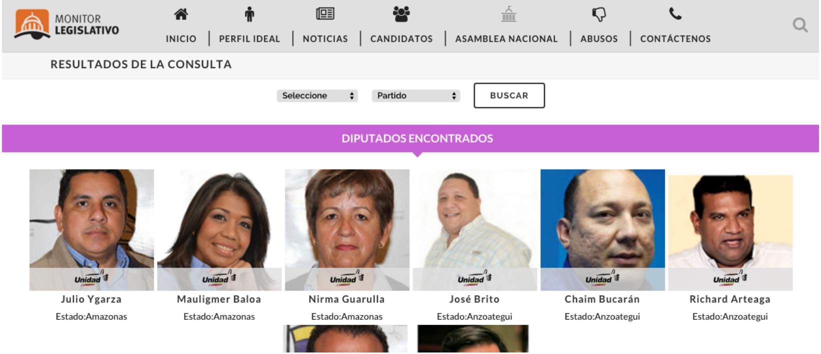
Report your Bribe Story!

See action taken.

www.ipaidabribe.com/reports/paid

Three main scenarios

3. Data hidden behind web forms



RESULTADOS DE LA CONSULTA

DIPUTADOS ENCONTRADOS

Julio Ygarza
Estado:Amazonas

Mauligmer Baloa
Estado:Amazonas

Nirma Guarulla
Estado:Amazonas

José Brito
Estado:Anzoategui

Chaim Bucarán
Estado:Anzoategui

Richard Arteaga
Estado:Anzoategui

Candidates on 2015 Venezuelan parliamentary election

Three main scenarios

1. Data in table format

Three main scenarios

1. Data in table format
    - Automatic extraction with `rvest`

Three main scenarios

1. Data in table format
    - Automatic extraction with `rvest`
2. Data in unstructured format

Three main scenarios

1. Data in table format
   - Automatic extraction with `rvest`
2. Data in unstructured format
   - Element identification with `selectorGadget`

Three main scenarios

1. Data in table format
    - Automatic extraction with `rvest`
2. Data in unstructured format
    - Element identification with `selectorGadget`
    - Automatic extraction with `rvest`

Three main scenarios

1. Data in table format
   - Automatic extraction with `rvest`
2. Data in unstructured format
   - Element identification with `selectorGadget`
   - Automatic extraction with `rvest`
3. Data hidden behind web forms

Three main scenarios

1. Data in table format
   - Automatic extraction with `rvest`
2. Data in unstructured format
   - Element identification with `selectorGadget`
   - Automatic extraction with `rvest`
3. Data hidden behind web forms
   - Automation of web browser behavior with `selenium`