http://v.gd/webscraping

Video

Web Scraping

Gathering Data from Websites

John Little

Data Visualization Services

Duke University Libraries

DVS: Data & Visualization

Contact Us: askdata@duke.edu

Location: Bostock Library, the /Edge (1st floor)

Hours: Walk-in, M-F / 8-5

OR, **Anytime** the Library is open (with ID)

Walk-in, Attendant Hours:

http://library.duke.edu/data/about/schedule.html

Workshops:

http://library.duke.edu/data/news

Services:

- Data Sources
- Data Management
- Data Cleaning
- Data Analysis
- Mapping & GIS
- Data Visualization

Web Scraping - What We'll Cover

1. **Build** a data **corpus** of congressional press releases
2. **APIs** and **gather** latitude and longitude -- using **JSON** formatted **data**
3. A brief hands-on introduction into **HTML parsing**
4. **APIs** and **Documentation** (FTP) -- OpenSecrets.org
5. Discussion of **APIs** and **Social Media** data gathering
6. A brief discussion on the **ethics** of scraping

This is **not** a programming workshop, but...

1. We will discuss **Python** and BeautifulSoup
2. We will **not learn** or use **Python** in the workshop
3. However, **some automation** tools are used in this workshop
4. Web Scraping is about deconstructing websites. Effective scraping requires learning about technical infrastructure as well as subject content
5. Not a workshop on **Text Analysis** (tools that calculate or correlate your data)
6. Not a workshop on **data cleaning**

Technical Definitions

Deconstruction v Construction

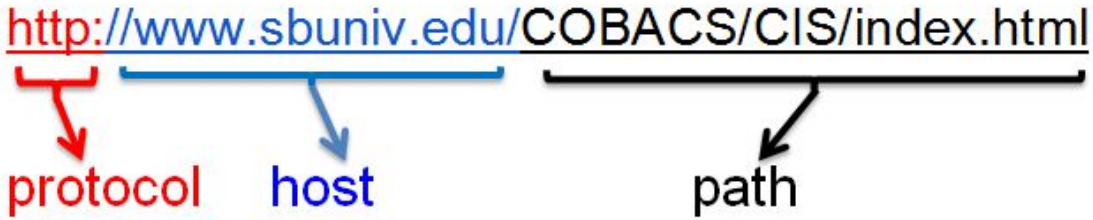
Definitions

- **Scraping**
  Using tools to gather data you can see on a webpage

  A wide range of web scraping techniques and tools exist. These can be as simple as copy/paste and increase in complexity to automation tools, HTML parsing, APIs and programming

Scraping propolis from the sides of the bee box
Image by Abalg~commonswiki

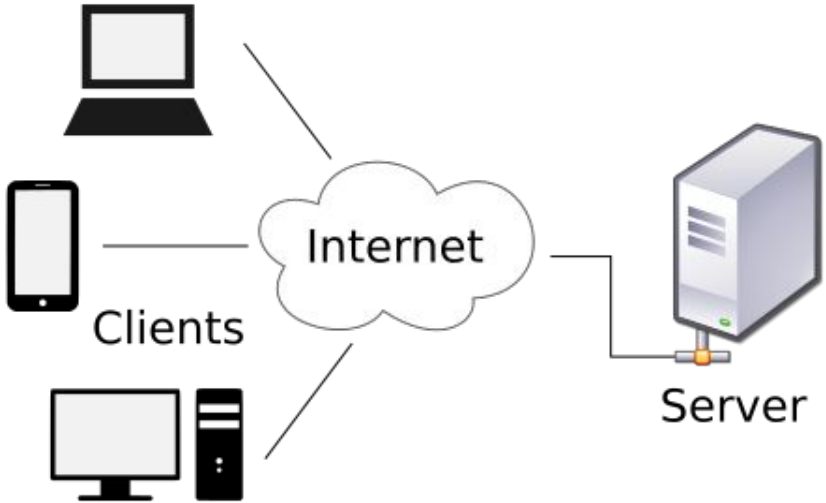
Definitions



- Scraping
- **HTTP**
  HyperText Transfer Protocol

  Machine interchange information transported over the Internet to enable multi-media data exchange, aka WWW. The protocol defines aspects of authentication, requests, status codes, persistent connections, client/server request/response. etc.

  Access a server on port 80; the declarative Document Type Definition (HTML, XML, JSON, etc.)



Images from commons.WikiMedia.org

Definitions

- Scraping
- HTTP
- **HTML**
  HyperText Markup Language

  The standard markup language on the Web

  As the web evolves so does the proliferation of technical wrappers surrounding the visible content of websites (text and data)

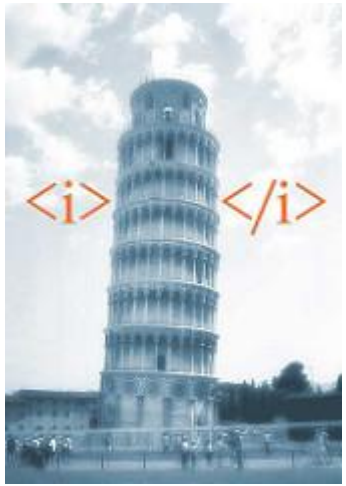


Image by Jesper Rønn-Jensen

```
<html>
  <head>
    <title>N&aacute;</title>
  </head>
  <body>
    Text stránky
  </body>
</html>
```

Image by Michaelbrabec

Definitions

- Scraping
- HTTP
- HTML
- **Parsing**
  The act of analyzing the strings and symbols to reveal only the data you need

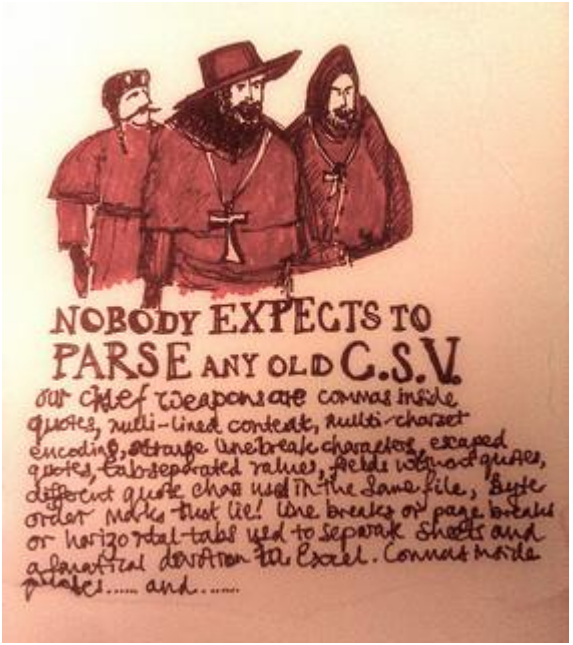


Image by Paul Downey

Definitions

- Scraping
- HTTP
- HTML
- Parsing
- **Crawling**
  Moving across or through a website in an attempt to gather data from more than one URL or page

Image by Dave Gingrich

Definitions

- Scraping
- HTTP
- HTML
- Parsing
- Crawling
- **JSON**
  Javascript Open Notation

  Readable text used to transmit data objects consisting of attribute-value pairs

  -- Wikipedia

```
{
  "firstName": "John",
  "lastName": "Smith",
  "isAlive": true,
  "age": 25,
  "address": {
    "streetAddress": "21 2nd Street",
    "city": "New York",
    "state": "NY",
    "postalCode": "10021-3100"
  },
  "phoneNumbers": [
    {
      "type": "home",
      "number": "212 555-1234"
    },
    {
      "type": "office",
      "number": "646 555-4567"
    }
  ],
  "children": [],
  "spouse": null
}
```

Definitions

- Scraping
- HTTP
- HTML
- Parsing
- JSON
- Crawling
- **API**
  Application Programming Interface

A set of rules and protocols used to build a software application. In the context of Web Scraping an API is a method used to gather clean data from a website (i.e. data that is not wrapped in HTML, Javascript, bound in HTTP, etc.)

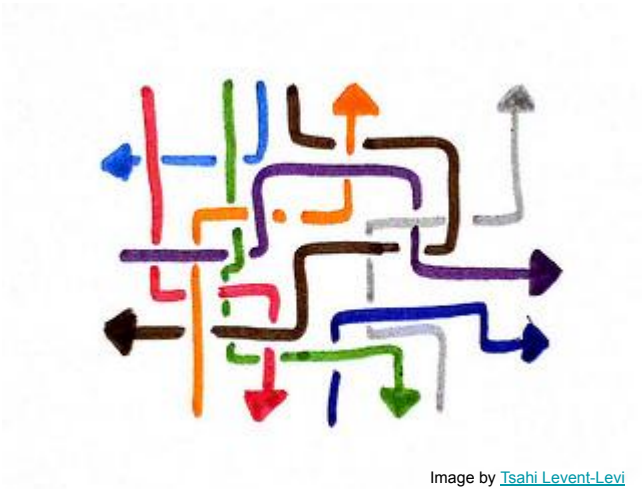
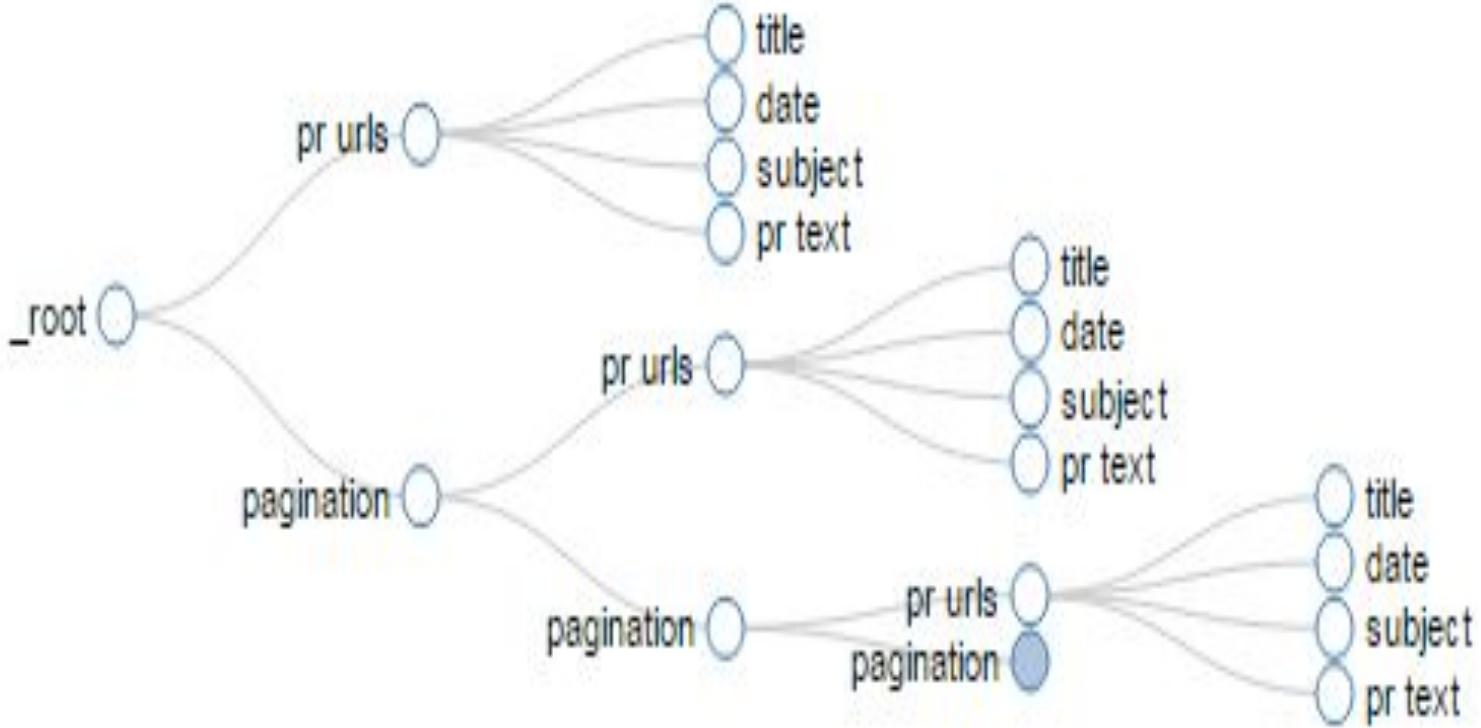


Image by Tsahi Levent-Levi

Webscraper.io
Demonstration & Hands-on

Demo: Scraping Congressional Press Releases

- Representative Nancy Pelosi's Press Releases
  - **CONTENT**
    - Structure of the Press Release subsection of the site
      - Pagination
      - Links to each release
      - Common elements of release
  - **TOOLS**
    - Webscraper.io tool works inside of Chrome
      - Tutorials
      - Documentation
      - Community
      - Free or, alternatively, Fee for Service


_root
pr urls
title
date
subject
pr text
pagination
pr urls
title
date
subject
pr text
pagination
pr urls
pagination
title
date
subject
pr text

Now You Try It

1. http://v.gd/webscraping1111
2. Follow the download & installation instructions: step 9a
3. Find some congressional press release sites: step 9b
4. Follow the instructions: steps 1-8

Google Sheets

Demonstration

ImportHTML
ImportXML

Google Sheets -- ImportHTML

- Example Site: http://www.boxofficemojo.com/
- IMPORTHTML(url, query, index)
- In Practice:

```
=IMPORTHTML("http://www.boxofficemojo.com/","table",3)
```

http://www.boxofficemojo.com/movies/?page=intl&id=annie2014.htm

```
=IMPORTHTML(A6,"table",6)
```

Help Docs: https://support.google.com/docs/answer/3093339?hl=en

Google Sheets -- ImportXML

- Example Site: http://www.nytimes.com/
- IMPORTXML(url, xpath_query)
- In Practice:

```
=IMPORTXML("http://nytimes.com/", "//*/p[contains(@class, 'summary')]")
```

Help Docs: https://support.google.com/docs/answer/3093342?hl=en

Resources

YouTube Videos

- Web Scraping with Google Sheets
- Importing Data 4 ImportXML
- Web scraping using Google Docs - Xpath

Other Resources

- **CSS Tutorial**
- XPath
- XPath Language defined by W3C

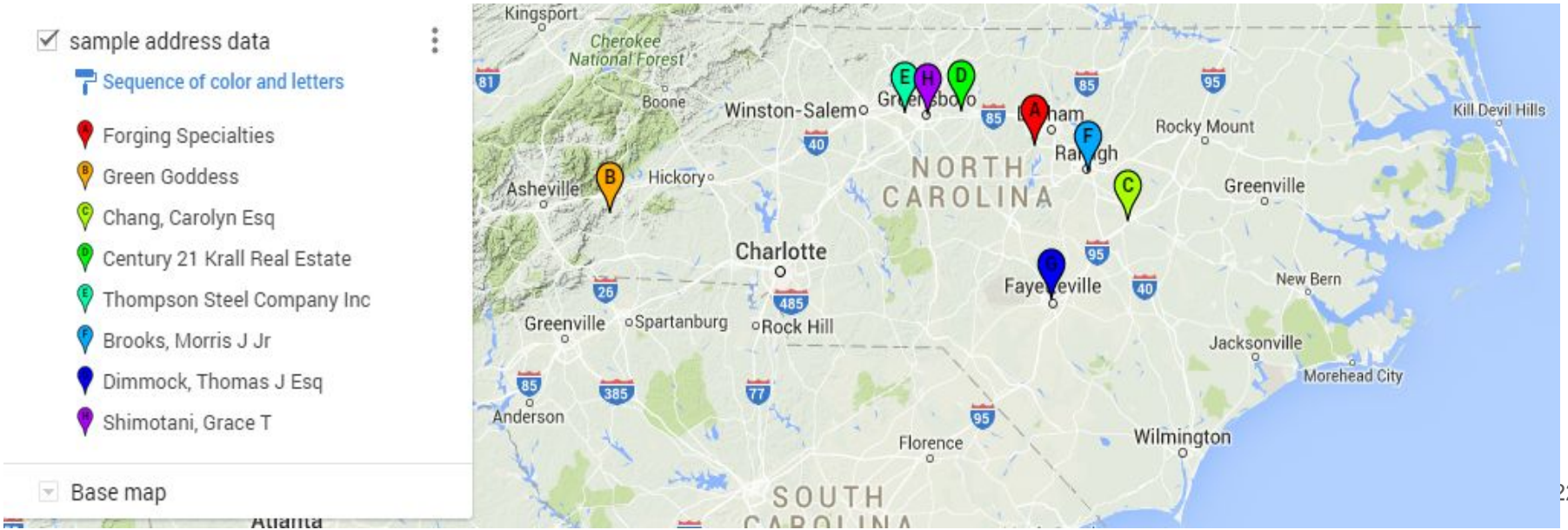
Your Turn:

- https://github.com/data-and-visualization/Rfun

APIs, Parsing, & JSON

OpenRefine & JSON file format

- Demonstration (http://v.gd/parsing3333)
  - A step-by-step guide using OpenRefine to gather JSON data via Google Map's API; then parse the JSON for latitude & longitude



Parsing HTML

OpenRefine & ParseHtml

- BeautifulSoup Libraries
  - Refine uses the Jython Libraries and has Jsoup
  - jSoup is a Java library built on BeautifulSoup -- a tool for HTML Extraction
- Resources (OpenRefine)
  - Step-by-step example documented in the demonstration above
- Documentation
  - Refine's documentation on HTML Parsing
  - jSoup Documentation
- Now You Try it -- http://v.gd/parsing2222

Case Study: OpenSecrets and documentation

OpenSecrets API and FTP

- OpenSecrets tracks the effects of money and lobbying in elections and politics
- OpenSecrets has an API
- OpenSecrets API Documentation
- OpenSecrets Bulk Data downloader
  - Login
  - Lobby.zip

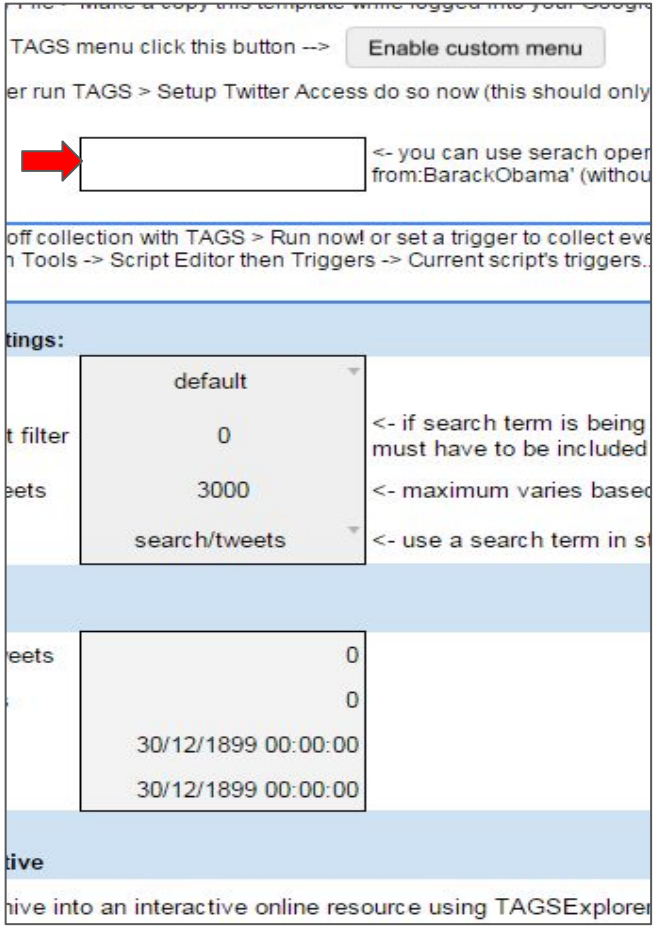
Social Media

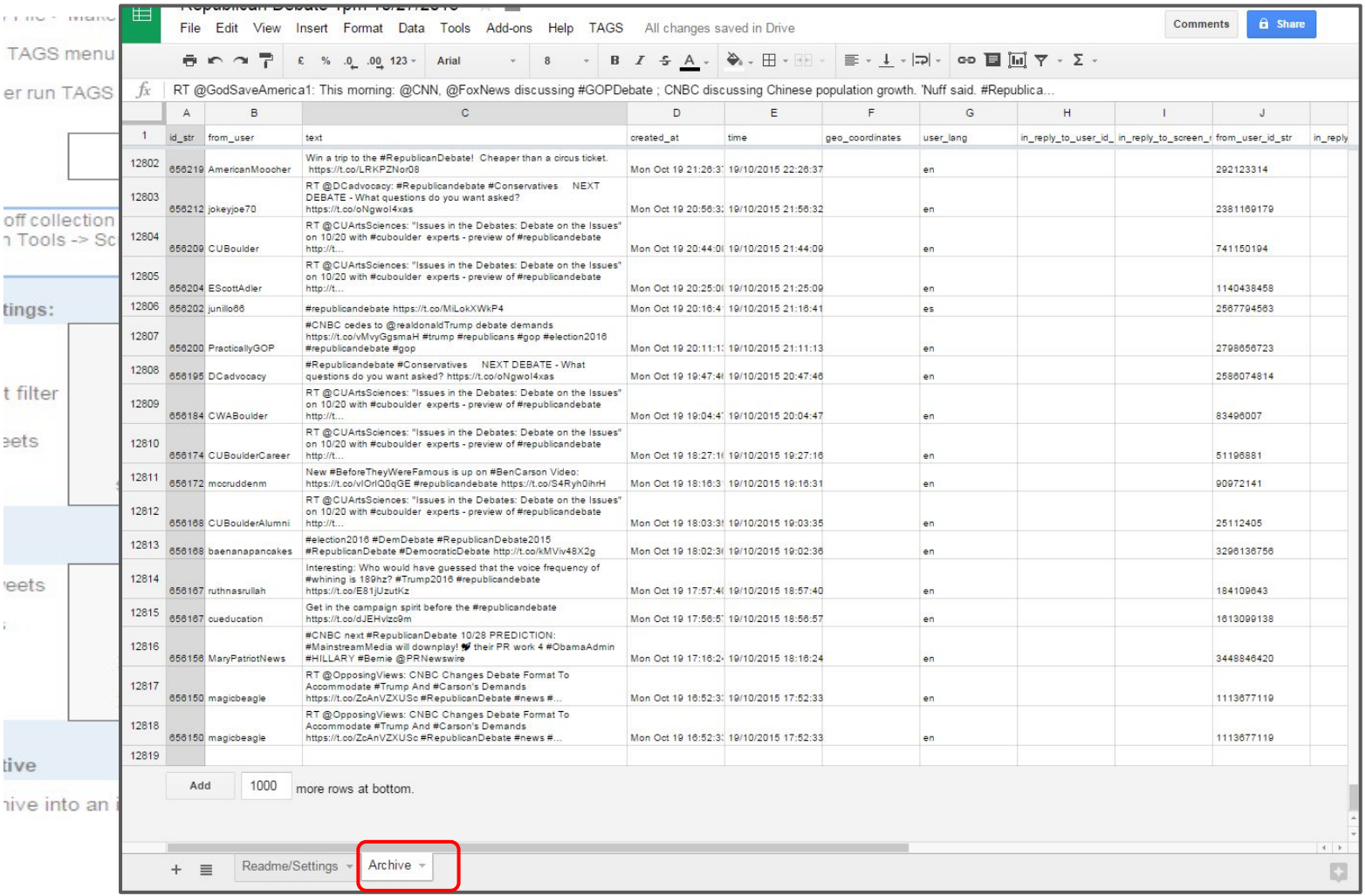
Social Media

1. Many ways to gather social media data
    a. IFTTT where you compose rules to connect sites and can deposit data in a spreadsheet
    b. APIs - often requires registered keys
    c. Buy your data from a service such as GNIP
2. After you download it you may want to perform analysis
    a. Sentiment Analysis, Word Frequency, Correlation, etc.
    b. Text Analysis tools (from Digital Humanities LibGuide)
    c. Digital Studio's program on working with Texts: Comparing and choosing texts analysis tools

TAGS: a tool for collecting Twitter streams

- TAGS (“New Sheets”; Version 6.0ns) - https://tags.hawksey.info/
  - Form driven (not command line)
  - Minimal setup
  - Data are collected in Google Sheets
  - Gather twitter stream data by type
    - screen-name stream data
    - screen-name status updates
    - twitter user favorited tweets
    - Search term for last 7 days: hashtag stream, username, boolean logic
    - Limit by date
    - Schedule to run hourly - set your interval, or run once.
    - 3 minute setup-video; easy to use - https://youtu.be/Vm0kjAvH5HM
    - Outputs: raw CSV structured data, plus default social graph visualizations


TAGS menu click this button -->
Enable custom menu
er run TAGS > Setup Twitter Access do so now (this should only
<- you can use serach oper
from:BarackObama' (withou
off collection with TAGS > Run now! or set a trigger to collect eve
h Tools -> Script Editor then Triggers -> Current script's triggers..
tings:
default
t filter
0
<- if search term is being
must have to be included
eets
3000
<- maximum varies based
search/tweets
<- use a search term in s
eets
0
0
30/12/1899 00:00:00
30/12/1899 00:00:00
tive
hive into an interactive online resource using TAGSExplorer



	id_str	from_user	text	created_at	time	geo_coordinates	user_lang	in_reply_to_user_id_	in_reply_to_screen_	from_user_id_str	in_reply
12802	656219	AmericanMoocher	Win a trip to the #RepublicanDebate! Cheaper than a circus ticket. https://t.co/LRKPZNor08	Mon Oct 19 21:26:3	19/10/2015 22:26:37		en			292123314	
12803	656212	jokeyjoe70	RT @DCadvocacy: #Republicandebate #Conservatives NEXT DEBATE - What questions do you want asked? https://t.co/oNgwol4xas	Mon Oct 19 20:56:3	19/10/2015 21:56:32		en			2381169179	
12804	656209	CUBoulder	RT @CUArtsSciences: "Issues in the Debates: Debate on the Issues" on 10/20 with #cuboulder experts - preview of #republicandebate http://t...	Mon Oct 19 20:44:0	19/10/2015 21:44:09		en			741150194	
12805	656204	EScottAdler	RT @CUArtsSciences: "Issues in the Debates: Debate on the Issues" on 10/20 with #cuboulder experts - preview of #republicandebate http://t...	Mon Oct 19 20:25:0	19/10/2015 21:25:09		en			1140438458	
12806	656202	junillo66	#republicandebate https://t.co/MiLokXWkP4	Mon Oct 19 20:16:4	19/10/2015 21:16:41		es			2567794563	
12807	656200	PracticallyGOP	#CNBC cedes to @realdonaldTrump debate demands https://t.co/vMvyGgsmaH #trump #republicans #gop #election2016 #republicandebate #gop	Mon Oct 19 20:11:1	19/10/2015 21:11:13		en			2798656723	
12808	656195	DCadvocacy	#Republicandebate #Conservatives NEXT DEBATE - What questions do you want asked? https://t.co/oNgwol4xas	Mon Oct 19 19:47:4	19/10/2015 20:47:46		en			2586074814	
12809	656184	CWABoulder	RT @CUArtsSciences: "Issues in the Debates: Debate on the Issues" on 10/20 with #cuboulder experts - preview of #republicandebate http://t...	Mon Oct 19 19:04:4	19/10/2015 20:04:47		en			83496007	
12810	656174	CUBoulderCareer	RT @CUArtsSciences: "Issues in the Debates: Debate on the Issues" on 10/20 with #cuboulder experts - preview of #republicandebate http://t...	Mon Oct 19 18:27:1	19/10/2015 19:27:16		en			51196881	
12811	656172	mccruddenm	New #BeforeTheyWereFamous is up on #BenCarson Video: https://t.co/vIOrlQ0qGE #republicandebate https://t.co/S4Ryh0lhrH	Mon Oct 19 18:16:3	19/10/2015 19:16:31		en			90972141	
12812	656168	CUBoulderAlumni	RT @CUArtsSciences: "Issues in the Debates: Debate on the Issues" on 10/20 with #cuboulder experts - preview of #republicandebate http://t...	Mon Oct 19 18:03:3	19/10/2015 19:03:35		en			25112405	
12813	656168	baenanapancakes	#election2016 #DemDebate #RepublicanDebate2015 #RepublicanDebate #DemocraticDebate http://t.co/kMViv48X2g	Mon Oct 19 18:02:3	19/10/2015 19:02:36		en			3296136756	
12814	656167	ruthnasrullah	Interesting: Who would have guessed that the voice frequency of #whining is 189hz? #Trump2016 #republicandebate https://t.co/E81jUzutKz	Mon Oct 19 17:57:4	19/10/2015 18:57:40		en			184109643	
12815	656167	cueducation	Get in the campaign spirit before the #republicandebate https://t.co/dJEHvlzc9m	Mon Oct 19 17:56:5	19/10/2015 18:56:57		en			1613099138	
12816	656156	MaryPatriotNews	#CNBC next #RepublicanDebate 10/28 PREDICTION: #MainstreamMedia will downplay! their PR work 4 #ObamaAdmin #HILLARY #Bernie @PRNewswire	Mon Oct 19 17:16:2	19/10/2015 18:16:24		en			3448846420	
12817	656150	magicbeagle	RT @OpposingViews: CNBC Changes Debate Format To Accommodate #Trump And #Carson's Demands https://t.co/ZcAnVZXUSc #RepublicanDebate #news #...	Mon Oct 19 16:52:3	19/10/2015 17:52:33		en			1113677119	
12818	656150	magicbeagle	RT @OpposingViews: CNBC Changes Debate Format To Accommodate #Trump And #Carson's Demands https://t.co/ZcAnVZXUSc #RepublicanDebate #news #...	Mon Oct 19 16:52:3	19/10/2015 17:52:33		en			1113677119	
12819											

TAGS menu click this button --> Enable custom menu

er run TAGS > Setup Twitter Access do so now (this should only

<- you can use serach oper
from:BarackObama' (withou

off collection with TAGS > Run now! or set a trigger to collect eve
n Tools -> Script Editor then Triggers -> Current scripts triggers..

tings:

	default
t filter	0
eets	3000
	search/tweets

<- if search term is being
must have to be included
<- maximum varies based
<- use a search term in st

eets	0
	0
	30/12/1899 00:00:00
	30/12/1899 00:00:00

tive

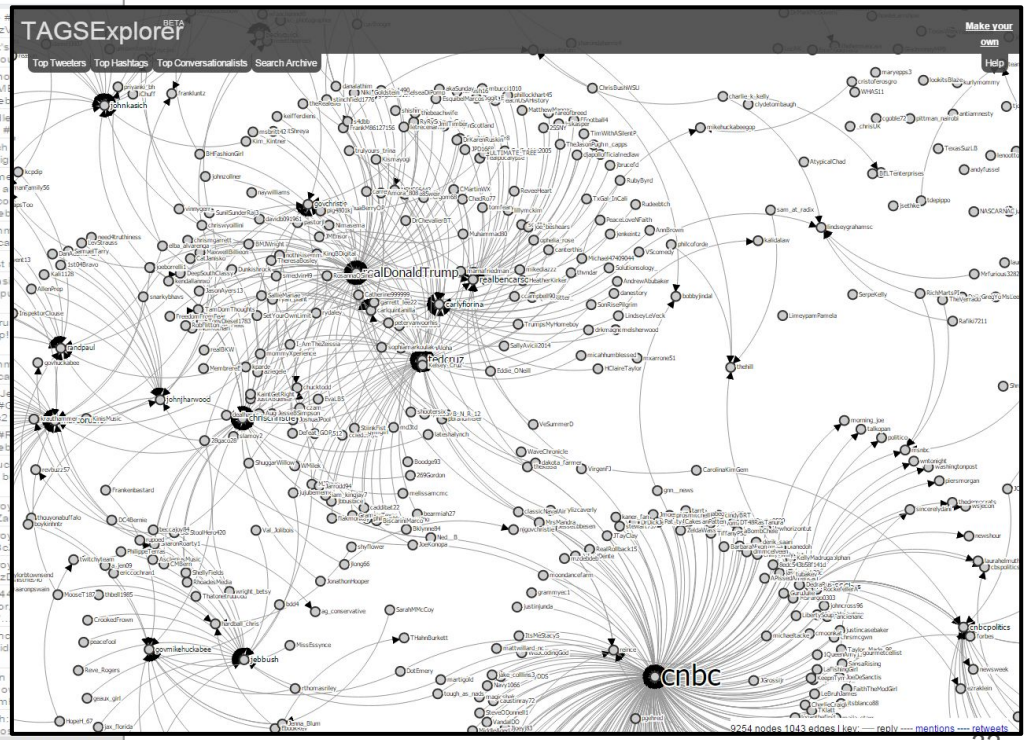
hive into an interactive online resource using TAGSExplorer

File Edit View Insert Format Data Tools

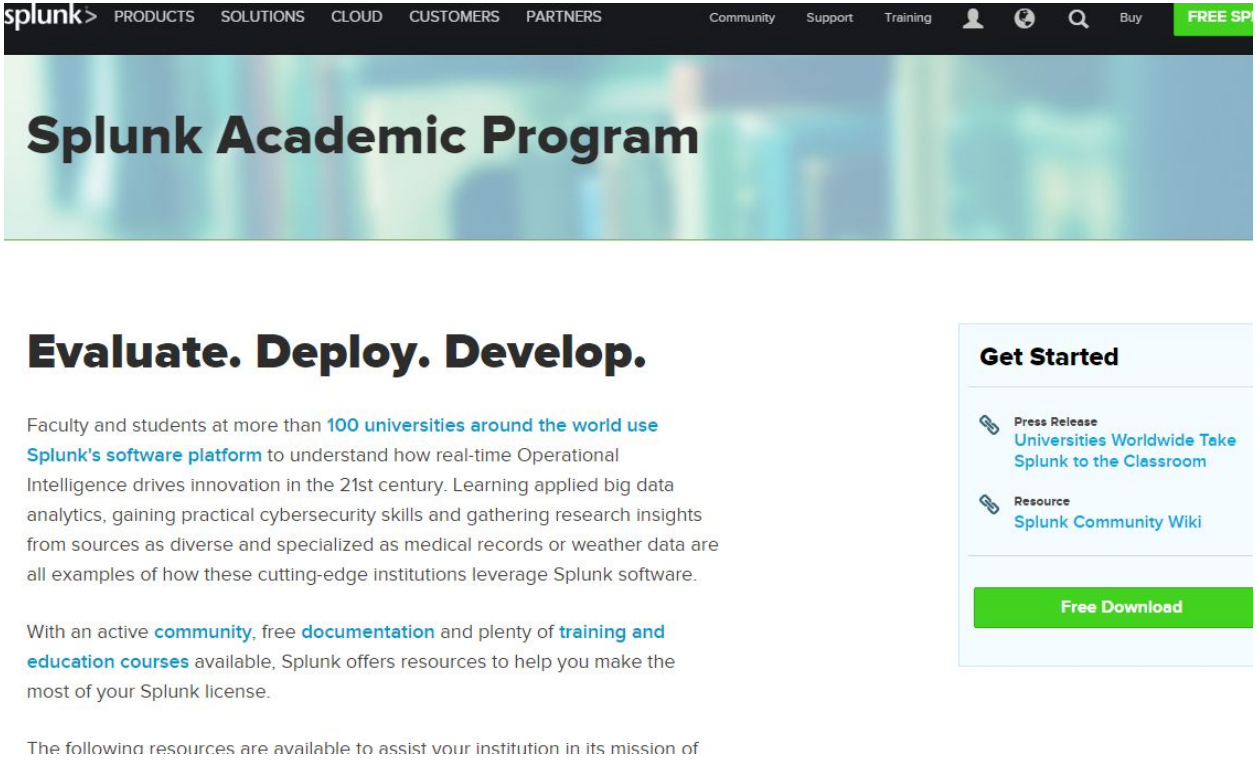
Arial

	A	B	C
	id_str	from_user	text
659751		chuckbalsamo	My take on the s https://t.co/UJ4z
659751		maryfshort	Obfuscation - it's https://t.co/mHo
659751		JamesHowdenIII	RT @kurtbraun PLEASE LET M #RepublicanDeb
659750		BillKempin	Jeb Bush handle Johnny Cueto. #
659750		iilleK	Stoked to watch to miss it last nig
659750		CallmeDownsie	Attention all Ame Just because I a I suggest you co #RepublicanDeb
659750		ArvelMauldin	RT @estheramn way. #Republica
659750		elderied5	The debate last
659750		StarDonovan	Fuming that bas watch the #Rep rights.
659750		B_N_R_12	@realDonaldTru night Mr. Trump #Trump2016
659750		Political_Bill	RT @estheramn way. #Republica
659749		Gilliganista	#JohnKasich #J #Republicans # https://t.co/LPc2
659749		RevMattPurkey	Last night the # #RepublicanDeb
659749		cristinapalumbo	@GovMikeHuc @goldencorral b trots.
659749		Jacquie_Kubin	Ted Cruz destro https://t.co/k4jZa
659749		WTCPolitics1	Ted Cruz destro https://t.co/Xu3c
659749		CommDigiNews	Ted Cruz destro https://t.co/f1iZz
659749		darcrich	RT @MWM444 helping the poor more tax outs."
659749		t1nydancaa	Biggest differenc laughs at candid responses.
659749		uBeenBitten	The Republican https://t.co/dp1o https://t.co/UXm
			RT @AsbillJosh They Live so clo

Readme/Settings Archive



Splunk Academic Program



https://www.splunk.com/en_us/solutions/industries/higher-education/academic-licenses.html

Ethics

Ethics & Technical Challenges of Web Scraping

Like copyright, there are rules, protocols, and best practices

1. Do you identify yourself?
2. Do you document your downloading steps?
3. Do you document your analysis steps?
4. Is your research reproducible?
5. Are your methods and processes confidential or proprietary?
6. How do you protect and document your sources?
7. How do you give credit to your sources?
8. Articles:
    a. To Scrape or Not to Scrape: Technical and Ethical Challenges of Collecting Data off the Web
    b. On the ethics of web scraping from a data journalism perspective

Resources

- Catalog of Web Scraping tools
  https://docs.google.com/spreadsheets/d/1A_9wBEmc8VP6HUm2R-7jdPU9FWY8dSfNsfljQ6cz638/edit#gid=0

Tools used or demonstrated in this Workshop

- Webscraper.io - http://webscraper.io/ (script and crawl a website)
- OpenRefine - http://openrefine.org (scripting, cleaning, parsing, utility data tool)
  - Regular Expressions
  - GREL
  - Jsoup
- TAGS - https://tags.hawksey.info/ (Twitter Stream collector)
- Splunk - Academic Program (Twitter Stream collector / Search and Discover)

Ethics

- To Scrape or Not to Scrape: Technical and Ethical Challenges of Collecting Data off the Web
- http://gijn.org/2015/08/12/on-the-ethics-of-web-scraping-and-data-journalism/

Thank You!

Please complete feedback forms

John Little
http://v.gd/webscraping

Data & Visualization Services
Duke University Libraries

This **presentation** contains links to directions and sample data. Each section of the workshop can be completed at your own pace.