

WEEK 1: WEB DATA COLLECTION 1

SECU0057

BENNETT KLEINBERG

16 JAN 2020

Applied Data Science

What is this? and Why do we need it?

AAAHH: SO WE'RE TALKING BIG DATA!



PROBLEMS WITH “BIG DATA”

- what is “big”?
- data = data?
- complexity of data?
- sexiness of small data

THINGS YOU'LL DO

- Accessing news article data and academic literature through APIs
- Building proper web-scrappers to collect data on the FBI's most wanted terrorists
- Analysing patterns of language use in text data
- Utilising Natural Language Processing to understand controversial news coverage
- Identifying the semantic relationships of concepts through automated methods
- Building fake news and sentiment classification systems using machine learning
- Utilising clustering to identify patterns in YouTube videos
- Building neural networks for prediction tasks
- your own project

LEARNING HOURS

Component	Amount	Duration	Total hours
Lectures	10	1h	10h
Tutorials/practicals	10	1h	10h
Assessment: class test	1	1.5h	1.5h
Assessment: project	1	40.5h	40.5h
Homework/revision/self-study	11	8h	88h
TOTAL	-	-	150

≈ 14 hours per week

BECOMING A REAL PROBLEM-SOLVER

- Principle 1: There's no magic in Data Science
- Principle 2: Data never come in a spreadsheet
- Principle 3: Data are hidden in front of you
- Principle 4: Programming can only ever be the vehicle

~~It all starts with the data.~~

It all starts with a problem.

WEEK 1: WEB DATA COLLECTION 1

TODAY

- Types of webscraping
- Using APIs
- “Real” webscraping: basics of a webpage

WHAT IS WEBSRAPING ANYWAY?

THE GAME CHANGER!

- direct broadcasting of ideas
- “unfiltered” and “uncensored” (?)
- location-enabled
- and: *en masse*

TYPES OF WEBSCRAPING

	Data shared	Data not shared
Ready-made table	Download	<i>closed source</i>
Not ready-made	API	Real webscraping

APPLICATION PROGRAMMING INTERFACES (APIs)

API: BASICS

Goal:

- helping developers interact with the platform
- facilitating interaction in an automatable manner
- analogous to the GUI
- part of it: enabling data access
- contains precise documentation

WHAT AN API DOES NOT DO:

- give you all the data
- be free forever
- give you full control

There's no free lunch!

CORE ELEMENTS OF AN API:

- GET requests
- POST requests

Implementable in different ways...

CLASSES OF APIs

1. Web APIs

- send requests through the browser
- add URL parameters

`https://data.police.uk/api/crimes-at-location`

2. Libraries/packages for APIs

- depending on the API: python, js, php, ruby
- = frameworks to access the API
- = methods implemented in different languages

3. API wrappers

- R packages that use the API

USEFUL WEBSITES THAT HAVE AN API

- Twitter (`twitterR, rtweet`)
- YouTube (`tuber`)
- Instagram
- Facebook
- Reddit
- The Guardian

Recommended: try online tutorials

API: NEWSPAPER COVERAGE

Accessing The Guardian data using `GuardianR`

```
library(GuardianR)

get_guardian("SOME SEARCH TERM",
             from.date="START_DATE",
             to.date="END_DATE",
             api.key="YOUR_ACCESS_KEY")
```

HOW TO OBTAIN THE ACCESS KEY?

<https://open-platform.theguardian.com/access/>

Get Started

Choose the level of access you need



How it works

The Open Platform is a public web service for accessing all the content the Guardian creates, categorised by tags and section. To get started, You need an key to successfully authenticate against the API.

Choose the level of access required for your application:

Developer

This key is for any non-commercial usage of the content, such as student dissertations, hackathons, nonprofit app developers.

- Up to 12 calls per second
- Up to 5,000 calls per day
- Access to article text
- Access to over 1,900,000 pieces of content
- **Free for non-commercial usage**

Commercial

This key is for any commercial enterprises and developers wanting to monetise our content.

- Custom throttle limit
- Custom quota limit
- Access to article text, images, audio and videos
- Access to entire Guardian content store
- **Price dependent on usage**

[Register developer key](#)

[Request commercial key](#)

STEP 1: SENDING API REQUESTS

Aim: search for keyword “crime” in Aug 2019

```
request.crime = get_guardian("crime",
                             from.date="2019-08-01",
                             to.date="2019-08-31",
                             api.key="ea765...")
```

Note: save these requests in an R object (variable)

STEP 2: UNDERSTANDING THE DATA

Dimensions (rows, columns) of the `request.crime` object

```
dim(request.crime)
```

```
## [1] 430 27
```

COLUMN NAMES

```
names(request.crime)
```

```
## [ 1 ] "id"                                "sectionId"                "sectionName"
## [ 4 ] "webPublicationDate"                 "webTitle"                  "webUrl"
## [ 7 ] "apiUrl"                            "newspaperPageNumber"     "trailText"
## [10 ] "headline"                           "showInRelatedContent"   "lastModified"
## [13 ] "hasStoryPackage"                   "score"                    "standfirst"
## [16 ] "shortUrl"                          "wordcount"                "commentable"
## [19 ] "allowUgc"                           "isPremoderated"         "byline"
## [22 ] "publication"                      "newspaperEditionDate"   "shouldHideAdverts"
## [25 ] "liveBloggingNow"                   "commentCloseDate"       "body"
```

DISPLAYING THE DATA

id	sectionName	webTitle	headline
books/2019/aug/07/top-10-true-books-duncan-campbell	Books	Top 10 true crime books	Top 10 true crime books
books/2019/aug/23/best-recent-crime-and-thrillers-review-roundup	Books	The best recent crime and thrillers â€“ review roundup	The best recent crime and thrillers â€“ review roundup
politics/2019/aug/12/boris-johnson-crime-crackdown-policing-prisons	Politics	Johnson's crackdown on offenders will only entrench crime	Johnson's crackdown on offenders will only entrench crime

id	sectionName	webTitle	headline
commentisfree/2019/aug/12/boris-johnson-crime-fighting-crusade-posturing-brexit	Opinion	Boris Johnsonâ€™s â€˜crime fightingâ€™ crusade is mere posturing Simon Jenkins	Boris Johnsonâ€™s â€˜crime fightingâ€™ crusade is mere posturing
uk-news/2019/aug/05/rural-crime-in-britain-hits-seven-year-high	UK news	Rural crime in Britain hits seven-year high	Rural crime in Britain hits seven-year high

DISPLAYING THE DATA

	headline	wordcount	byline	newspaperEditionDate
11	Surge in violent crime in Barcelona prompts calls for legal reform	451	Stephen Burgen in Barcelona	2019-08-24T00:00:00Z
12	Monica Lewinsky to produce American Crime Story drama about Clinton scandal	505	Gwilym Mumford	2019-08-08T00:00:00Z
13	Eric Cantona speech: humans 'will become eternal' â€“ unless crime or war intervene	397	Guardian sport	NA
14	87 bird crime incidents last year and just one conviction, says RSPB	388	Patrick Barkham	2019-08-29T00:00:00Z
15	Video of trans women forced from LA bar prompts hate crime	994	Sam Levin in Los Angeles	NA

headline

wordcount

byline

newspaperEditionDate

THE ARTICLE BODY

```
## [1] "<p>The police minister, Kit Malthouse, has insisted that a new wa
```

STEP 3: USING THE DATA

Note: not all sections are relevant (e.g., book reviews vs politics vs uk news)

Option 1: sending a new API request R package docs

“UK NEWS” CRIME COVERAGE

```
dim(request.crime.uknews)
```

```
## [1] 32 27
```

```
knitr:::kable(request.crime.uknews[1:5, c(1,3,5,10)])
```

id	sectionName	webTitle	headline
uk-news/2019/aug/05/rural-crime-in-britain-hits-seven-year-high	UK news	Rural crime in Britain hits seven-year high	Rural crime in Britain hits seven-year high
uk-news/2019/aug/02/east-london-volunteers-set-up-knife-amnesty-bin	UK news	'There's not enough police': East London volunteers take on knife crime	'There's not enough police': East London volunteers take on knife crime
uk-news/2019/aug/08/female-knife-possession-crimes-in-england-rise-by-73	UK news	Female knife possession crimes in England rise by 73%	Female knife possession crimes in England rise by 73%

id	sectionName	webTitle	headline
uk-news/2019/aug/14/nca-freezes-more-than-100m-after-court-order-on-imported-money	UK news	NCA freezes Â£100m suspected to be from corruption overseas	NCA freezes Â£100m suspected to be from corruption overseas
uk-news/2019/aug/07/alesha-macphail-killer-appeals-against-27-year-sentence	UK news	Alesha MacPhail killer appeals against 27-year minimum sentence	Alesha MacPhail killer appeals against 27-year minimum sentence

OPTION 2: SUBSETTING THE DATA

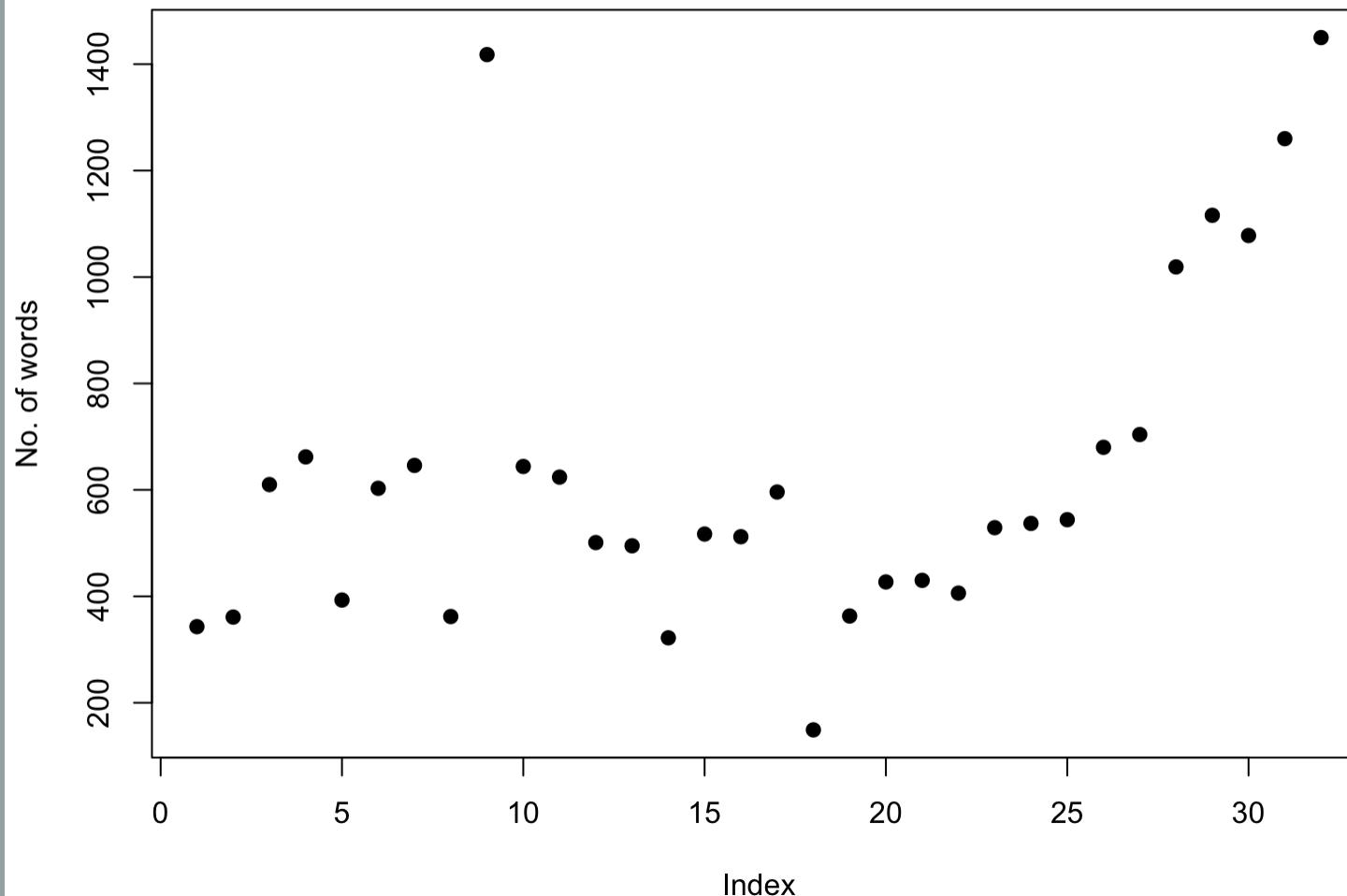
```
crime_data_uknews = request.crime[request.crime$sectionId == "uk-news",  
dim(crime_data_uknews)]
```

```
## [1] 32 27
```

STEP 3: USING THE DATA

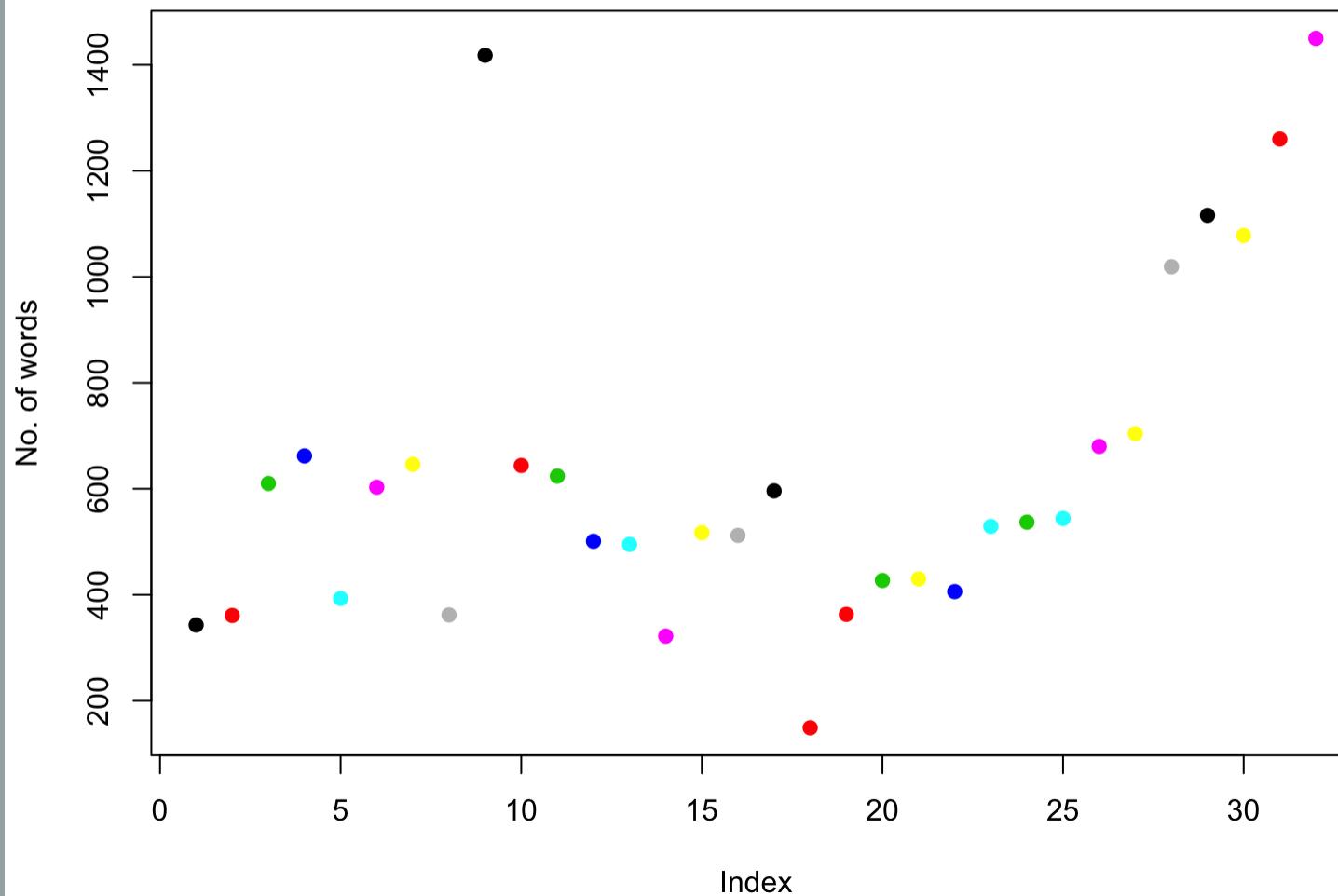
Aim: investigate article length

Crime article lengths (UK News section)



... BY AUTHOR

Crime article lengths (UK News section)



NUMERICALLY

```
tapply(X = as.numeric(as.character(request.crime.uknews$wordcount))  
, INDEX = request.crime.uknews$byline  
, FUN = mean)
```

```
##                                     Frances Perraudin  
##                                         343.0  
##                                     Amy Walker  
##                                         362.0  
## Jamie Grierson Home affairs correspondent  
##                                         573.5  
## Rupert Neate Wealth correspondent  
##                                         662.0  
## Libby Brooks Scotland correspondent  
##                                         393.0  
## Helen Pidd North of England editor  
##                                         1026.5  
## Vikram Dodd Police and crime correspondent  
##                                         862.0  
##                                     Letters  
##                                         362.0  
## Duncan Campbell  
##                                         1418.0
```

APIs: PROS & CONS

Pro

- easy to access
- nicely documentation
- **works even if website changes**

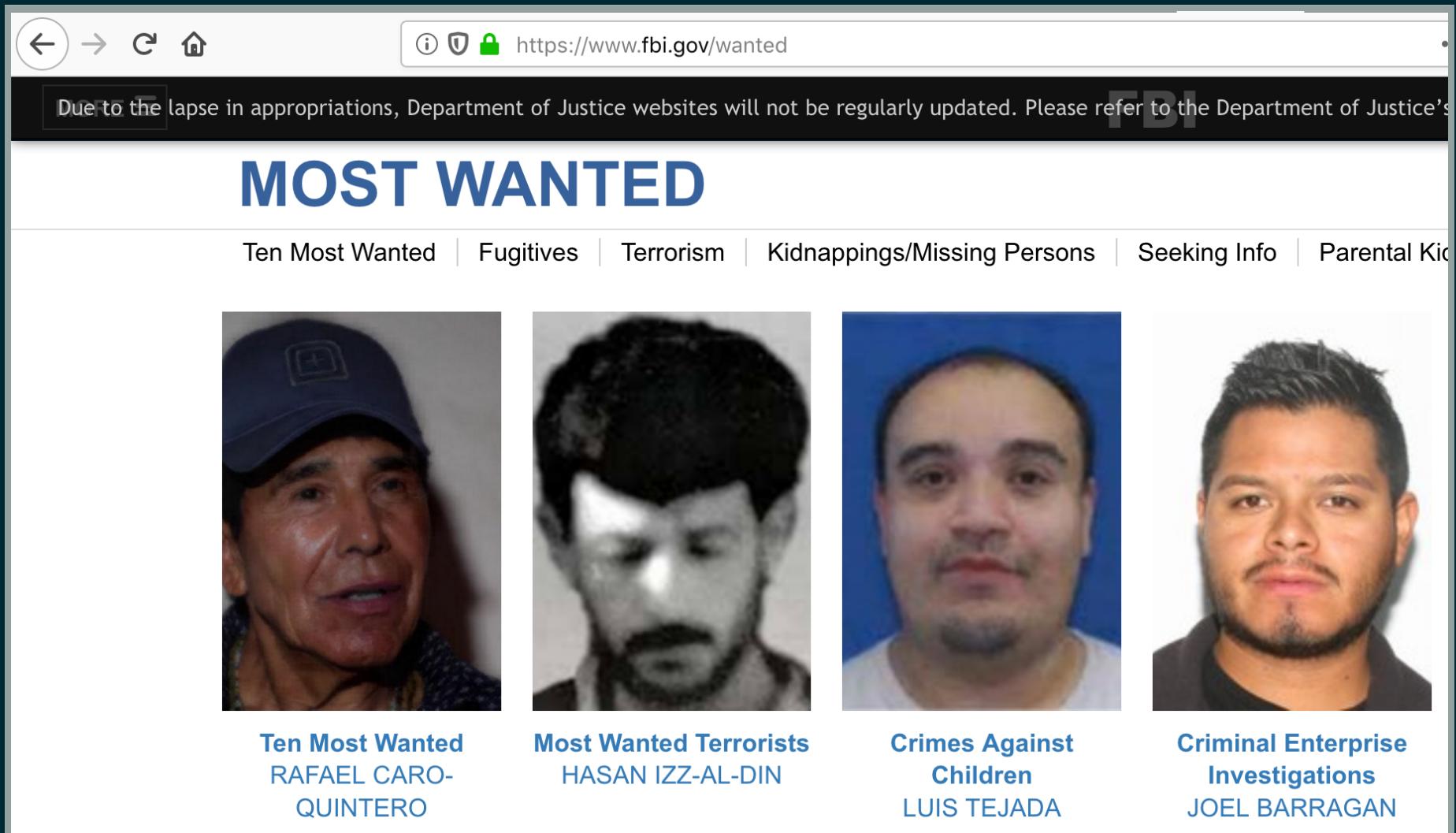
Cons

- quota limits (\$ \$ \$)
- under the platforms' control
- only for few platforms

DON'T LET THE DATA DETERMINE YOUR RESEARCH!

BEYOND APIs

What about:



A screenshot of a web browser displaying the FBI's "Most Wanted" page at <https://www.fbi.gov/wanted>. The page header includes standard navigation icons and a URL bar. A prominent message at the top states: "Due to the lapse in appropriations, Department of Justice websites will not be regularly updated. Please refer to the Department of Justice's". Below this, the word "FBI" is displayed in its signature font. The main section features a large blue header "MOST WANTED". Below it is a horizontal menu with links: "Ten Most Wanted" (highlighted), "Fugitives", "Terrorism", "Kidnapping/Missing Persons", "Seeking Info", and "Parental Kid...". The page then displays four mugshots of men, each with a caption below it. From left to right: 1. Rafael Caro-Quintero, labeled "Ten Most Wanted RAFAEL CARO-QUINTERO". 2. Hasan Izz-Al-Din, labeled "Most Wanted Terrorists HASAN IZZ-AL-DIN". 3. Luis Tejada, labeled "Crimes Against Children LUIS TEJADA". 4. Joel Barragan, labeled "Criminal Enterprise Investigations JOEL BARRAGAN".

← → ⌛ 🏠 https://www.fbi.gov/wanted

Due to the lapse in appropriations, Department of Justice websites will not be regularly updated. Please refer to the Department of Justice's

FBI

MOST WANTED

Ten Most Wanted | Fugitives | Terrorism | Kidnapping/Missing Persons | Seeking Info | Parental Kid...



Ten Most Wanted
RAFAEL CARO-
QUINTERO

Most Wanted Terrorists
HASAN IZZ-AL-DIN

**Crimes Against
Children**
LUIS TEJADA

**Criminal Enterprise
Investigations**
JOEL BARRAGAN



National Crime Agency (GB)

<https://www.missingpersons.police.uk/en-gb/case->



Bureau Reference:	18-009282
Location	London
Gender	Male
Date found	21 September 2018
Age	30 - 40
Ethnicity	White European

[View more details](#)

[View case details](#)



Bureau Reference:	18-006195
Location	Dunsden
Gender	Male
Date found	15 May 2018
Age	16 - 100
Ethnicity	White European

[View case details](#)

NO APIs

- incels.me
- Stormfront
- 4chan
- APIs are restrictive!

... WHAT ABOUT:
YOUR RESEARCH PROJECT -> NO API?



Main problem:
Really ‘juicy’ data of the Internet vs APIs

“REAL” WEBSRAPING: BASICS OF A WEBPAGE

THREE ELEMENTS OF A WEBPAGE

1. Structure
2. Behaviour
3. Style

THREE ELEMENTS OF A WEBPAGE

1. Structure

2. Behaviour

- JavaScript (\neq Java)
- user interaction
- examples: alerts, popups, server-interaction

3. Style

THREE ELEMENTS OF A WEBPAGE

1. Structure
2. Behaviour
3. Style

- CSS (Cascading Style Sheets)
- formatting, design, responsiveness
- examples: submit buttons, app interfaces

THREE ELEMENTS OF A WEBPAGE

1. Structure

- HTML (hypertext markup language)
- structured with <tags>
- contains the pure content of the webpage

2. Behaviour

3. Style

FOR NOW: HTML

The very basics of HTML:

Raw architecture of a webpage

```
<!DOCTYPE html>
<html>
<body>

HERE COMES THE VISIBLE PART !!

</body>
</html>
```

Note: Every tag `< >` is closed `< />`. Content is contained within the tag.

HTML BASICS

Ways to put content in the `<body> . . . </body>` tag:

- headings: `<h1>I'm a heading at level 1</h1>`



CONTENT IN THE BODY TAG

- paragraphs: <p>This is a paragraph</p>

The screenshot shows a web browser window with a dark blue header bar. In the header, there are icons for back, forward, refresh, and home. To the right of the icons is a URL bar containing the text: "file:///Users/bennettkleinberg/GitHub/ucl_aca_20182019/slides/html_example.html". The main content area of the browser displays the following text:

```
I'm a heading at level 1
I'm the first paragraph
I'm the second paragraph
I'm the third paragraph
```

The text "I'm a heading at level 1" is displayed in a large, bold, black font. The other three lines of text ("I'm the first paragraph", "I'm the second paragraph", "I'm the third paragraph") are in a smaller, regular black font.

WEB SCRAPING LOGIC

If all webpages are built in this structure...

...then we could access this structure programmatically.

BUT WHERE DO I FIND THAT STRUCTURE?

Is it just “there”?

YES!!

HOW TO SEE THE HTML STRUCTURE?

← → ⌂ ⌂ file:///Users/bennettkleinberg/GitHub/ucl_aca_20182019/slides/html_example.html

I'm a heading at level 1

I'm the first paragraph

I'm the second paragraph

I'm the third paragraph

A photograph of the British Museum's neoclassical portico and dome at dusk. The building is illuminated from within, with warm light glowing through the columns and windows. The dome is topped with a flag. The sky is a clear, pale blue.



I'm a heading at level 1

I'm the first paragraph

I'm the second paragraph

I'm the third paragraph



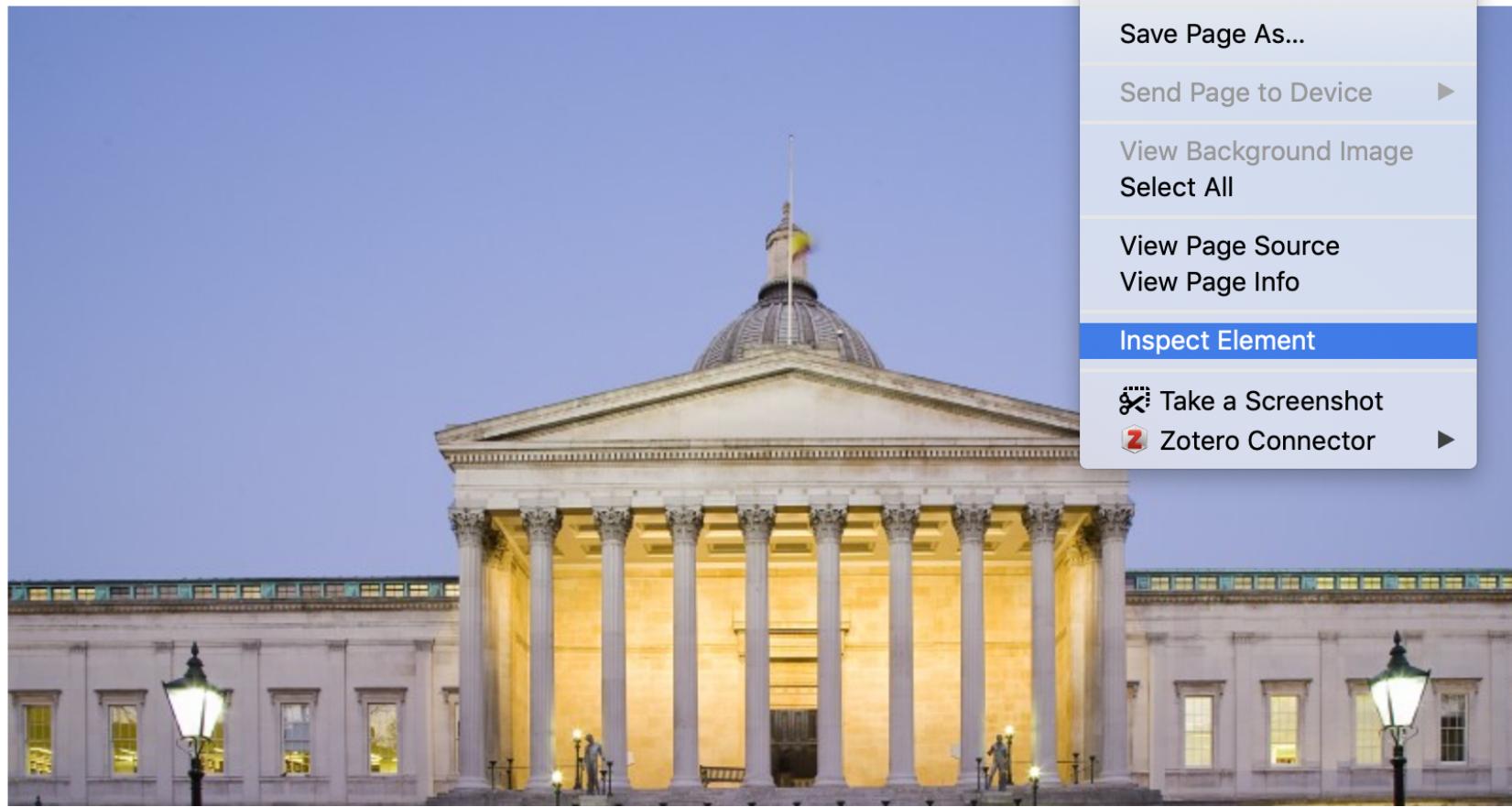


I'm a heading at level 1

I'm the first paragraph

I'm the second paragraph

I'm the third paragraph





Inspector Console Debugger Style Editor Performance Memory ... X

+ Search HTML

```
<!DOCTYPE html>
<html class="gr__"> event
<head></head>
<body data-gr-c-s-loaded="true">
  <h1>I'm a heading at level 1</h1>
  <p id="paragraph1" class="paragraph_class">I'm the first paragraph</p>
  <p class="paragraph_class">I'm the second paragraph</p>
  <p class="paragraph_class">I'm the third paragraph</p>
  
  <br>
  <br>
  <a href="https://www.ucl.ac.uk/">Click here to go to UCL's website</a>
  <br>
  <br>
  ▶ <table>[...]</table>
  ▶ <ul>[...]</ul>
</body>
</html>
```

html.gr__ > body

```
▼<table>
  ▼<tbody>
    ▼<tr>
      <th>Departments</th>
      <th>Location</th>
    </tr>
    ▼<tr>
      <td>Dept. of Security and Crime Science</td>
      <td>Division of Psychology and Language Sciences</td>
    </tr>
    ► <tr>...</tr>
  </tbody>
</table>
```

WEBSCRAPING IN A NUTSHELL

1. understand the structure of a webpage
2. exploit that structure for web-scraping

WHAT'S NEXT?

- Today's tutorial: - Today's tutorial: building your own API scraper, using the data, some HTML
- Homework: arxiv scraping, Twitter access, project start

Next week: Web scraping 2 (more HTML, browser simulation)