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

Pablo Barberá

London School of Economics
pablobarbera.com

Course website:
pablobarbera.com/ECPR-SC104

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

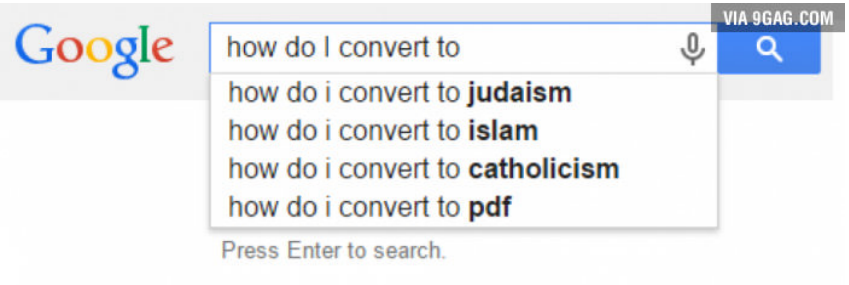
Follow

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

Reply Retweet Favorite ••• More

RETWEETS 356 FAVORITES 110

10:06 PM - 2 Jun 2013



Justin Bieber
@justinbieber

Follow

I make music. I love music.

Reply Retweet Favorite ••• More

RETWEETS 54,213 FAVORITES 59,205

10:09 PM - 7 Apr 2014

dustin curtis
@dcurtis

Follow

"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

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 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

Like · Comment · Share 57

262 people like this. Top Comments

Total seats won

650 of 650 seats declared at 9 Jun 21.06 BST

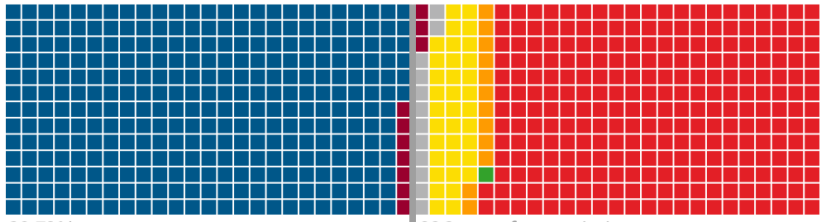
Con 318 (-13)

Lab 262 (+32)

LD 12 (+3)

SNP 35 (-19)

DUP 10 (+2) | UKIP 0 (0) | Green 1 (0) | Others 12



68.73% turnout

326 seats for a majority

Donald J. Trump
@realDonaldTrump

Folgen

Are you allowed to impeach a president for gross incompetence?

Original (Englisch) übersetzen

RETWEETS 195.387 GEFÄLLT 161.489

03:23 - 4. Juni 2014

15 Tsd. 195 Tsd. 161 Tsd.

How can we collect web and social data to answer social science questions?

Course outline

1&2 Scraping data from the web
- Key tools for webscraping
- Scraping tables
- Scraping web data in unstructured format
- Parsing RSS feeds

3 Working with APIs
- How to build an http request
- Interacting with newspapers' APIs

4 Collecting social media data
- Twitter's Streaming API
- Twitter's REST API

5 Advanced topics
- Parsing data in PDF format
- Text encoding

Hello!

About me: Pablo Barberá

- Assistant Professor of Computational Social Science at the London School of Economics
  - Previously Assistant Prof. at Univ. 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
  - Text as data methods, social network analysis, Bayesian statistics
  - Author of R packages to analyze data from social media
- Contact:
  - `P.Barbera@lse.ac.uk`
  - `www.pablobarbera.com`
  - `@p_barbera`

About me: Tom Paskhalis

- PhD candidate in Social Research Methods at the London School of Economics
- My research:
  - Interest groups and political parties
  - Text as data, record linkage, Bayesian statistics
  - Author/contributor to R packages to scrape websites and PDF documents
- Contact:
  - `T.G.Paskhalis@lse.ac.uk`
  - `tom.paskhal.is`
  - `@tpaskhalis`

About me: Alberto Stefanelli

- Prospective Phd candidate at KU Leuven
  - Previously Master Student at Central European University
  - Vice president of the Populism Research Group at Central European University and member of the survey and experimental teams of Team Populism
  - External Consultant and data analyst for the ECPR Methods Schools and the Intellectual Theme Initiative project Text Analysis across Disciplines

- My research:
  - Electoral behavior, public opinion, political communication, party finance
  - Graphical causal models, machine learning, text analysis, and big data

- Contact:
  - `alberto.stefanelli.main@gmail.com`
  - `alberto-stefanelli.netlify.com`
  - `@sergsagara`

Your turn!

1. Name?
2. Affiliation?
3. Research interests?
4. Previous experience with R?
5. Why are you interested in this course?

Course philosophy

How to learn the techniques in this course?

- Lecture approach: not ideal for learning how to code
- You can only learn by doing.

→ We will cover each concept three times during each session

1. Introduction to the topic (20-30 minutes)
2. Guided coding session (30-40 minutes)
3. Coding challenges (30 minutes)

- You're encouraged to continue working on the coding challenges after class. Solutions will be posted the following day.
- Additional questions? We can arrange one-on-one meetings after class

Course logistics

ECTS credits:

- Attendance: 2 credits (pass/fail grade)
- Submission of at least 3 coding challenges: +1 credit
  - Due before beginning of following class via email to Tom or Alberto
  - Only applies to challenge 2 of the day
  - Graded on a 100-point scale
- Submission of class project: +1 credit
  - Due by August 20th
  - Goal: collect and analyze data from the web or social media
  - 5 pages max (including code) in Rmarkdown format
  - Graded on a 100-point scale

If you wish to obtain more than 2 credits, please indicate so in the attendance sheet

Social event

Save the date:

Wednesday Aug. 1st, 6pm

Location TBA

Why we're using R

- Becoming *lingua franca* of statistical analysis in academia
- What employers in private sector demand
- It's free and open-source
- Flexible and extensible through *packages* (over 10,000 and counting!)
- Powerful tool to conduct automated text analysis, social network analysis, and data visualization, with packages such as quanteda, igraph or ggplot2.
- Command-line interface and scripts favors reproducibility.
- Excellent documentation and online help resources.

R is also a full programming language; once you understand how to use it, you can learn other languages too.

RStudio Server

RStudio

File Edit Code View Project Workspace Plots Tools Help

Go to file/function Project: (None)

diamondPricing.R* | formatPlot.R | diamonds

Source on Save Source

```
1  library(ggplot2)
2  source("plots/formatPlot.R")
3
4  View(diamonds)
5  summary(diamonds)
6
7  summary(diamonds$price)
8  aveSize <- round(mean(diamonds$carat), 4)
9  clarity <- levels(diamonds$clarity)
10
11 p <- qplot(carat, price,
12            data=diamonds, color=clarity,
13            xlab="Carat", ylab="Price",
14            main="Diamond Pricing")
15
```

15:1 (Top Level) R Script

Console ~/

```
       x                y                z
 Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
 1st Qu.: 4.710   1st Qu.: 4.720   1st Qu.: 2.910
 Median : 5.700   Median : 5.710   Median : 3.530
 Mean   : 5.731   Mean   : 5.735   Mean   : 3.539
 3rd Qu.: 6.540   3rd Qu.: 6.540   3rd Qu.: 4.040
 Max.   :10.740   Max.   :58.900   Max.   :31.800
> summary(diamonds$price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    326     950    2401    3933    5324   18820
> aveSize <- round(mean(diamonds$carat), 4)
> clarity <- levels(diamonds$clarity)
> p <- qplot(carat, price,
+            data=diamonds, color=clarity,
+            xlab="Carat", ylab="Price",
+            main="Diamond Pricing")
>
> format.plot(p, size=24)
>
```

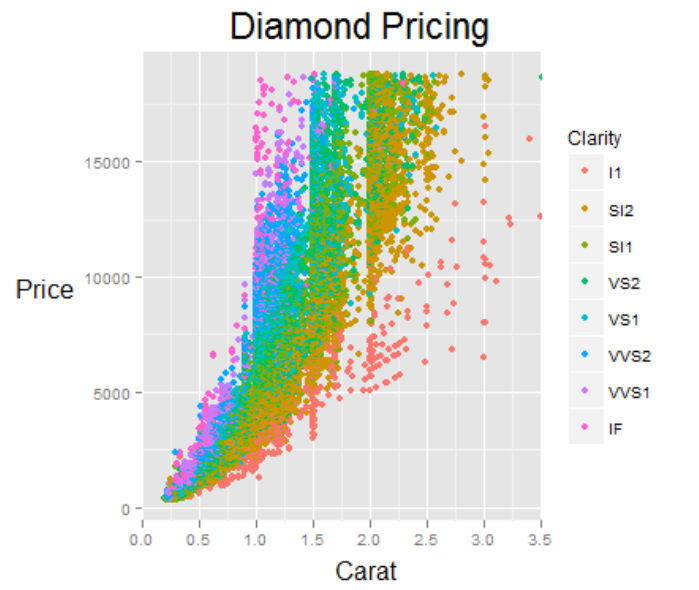
Workspace | History

Load Save Import Dataset Clear All

Data	
diamonds	53940 obs. of 10 variables
Values	
aveSize	0.7979
clarity	character[8]
p	ggplot[8]
Functions	
format.plot(plot, size)	

Files | Plots | Packages | Help

Zoom Export Clear All



Course website

pablobarbera.com/ECPR-SC104