GRADUATE SCHOOL OF DECISION SCIENCES

Universität Konstanz

Web Data Collection with R

Session 1: Introduction

Sascha Göbel
sascha.goebel@uni-konstanz.de

November 4th, 2020

Welcome!

Today: Introductory session

- what this course is about and learning goals
- course schedule
- prerequisites and requirements for receiving a grade and credits
- course registration

Before we start: My background

- PhD Candidate in PolSci at the GSDS
- learned R 5 years ago in a course on Web Data Collection with R
- interested in political participation, representation, public opinion, and information technology and politics

Welcome!

Today: Introductory session

- what this course is about and learning goals
- course schedule
- prerequisites and requirements for receiving a grade and credits
- course registration

Before we start: My background

- PhD Candidate in PolSci at the GSDS
- learned R 5 years ago in a course on Web Data Collection with R
- interested in political participation, representation, public opinion, and information technology and politics

Before we start: Important!

- BA and MA seminar for Politics and Public Administration and SEDS students (prioritized)
- students enrolled in other programs have to check if they can receive credits for this course
- same for PhD students
- I'm not allowed to grade PhD students at the GSDS

Before we start: How to Zoom?

- please use your real name
- please turn on your camera
- please mute your microphone and unmute only when speaking
- if you have a question or a comment, click "Raise Hand"

Before we start: Important!

- BA and MA seminar for Politics and Public Administration and SEDS students (prioritized)
- students enrolled in other programs have to check if they can receive credits for this course
- same for PhD students
- I'm not allowed to grade PhD students at the GSDS

Before we start: How to Zoom?

- please use your real name
- please turn on your camera
- please mute your microphone and unmute only when speaking
- if you have a question or a comment, click "Raise Hand"

Motivation

Whats your Motivation?

- Let’s conduct a small poll!

Course goals

- acquire basic knowledge of web technologies
- be able to identify the feasibility and appropriate strategy of web data collection
- be able to extract data from static and dynamic webpages and APIs using diverse tools
- be able to process data gathered on the web for subsequent analyses using R

Motivation

Whats your Motivation?

- Let's conduct a small poll!

Course goals

- acquire basic knowledge of web technologies
- be able to identify the feasibility and appropriate strategy of web data collection
- be able to extract data from static and dynamic webpages and APIs using diverse tools
- be able to process data gathered on the web for subsequent analyses using R

Why is this useful?

– Let's have a look at some research using web data:

Political Advertising on the Wikipedia Marketplace of Information (Göbel and Munzert 2018)

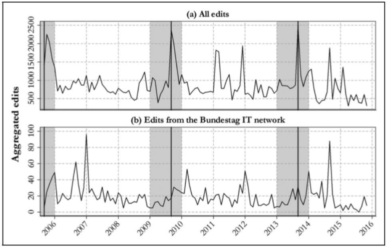


Figure 3. Edits on members of parliament (MPs') Wikipedia entries over time. For (a) $N = 108{,}775$; for (b) $N = 2{,}365$. Edits only cover MPs with a seat in the Bundestag during the respective period. Reference lines denote Bundestag elections, shaded areas the election years.

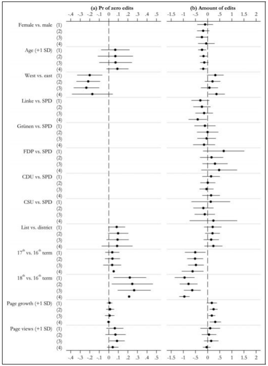
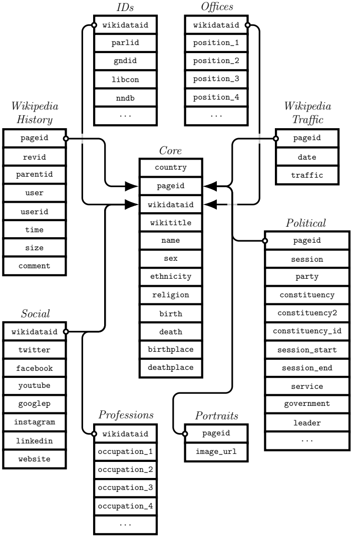


Figure 5. First differences in factors associated with edits from the parliament's IP range. Based on zero-inflated negative binomial models in Table 2. Computed holding each of the other predictors not at their means but at the observed values for each case in the sample. Solid lines denote 95% confidence intervals.

The Comparative Legislators Database (Göbel and Munzert 2020)



IDs
- wikidataid
- parlid
- gndid
- libcon
- nndb
- ...

Offices
- wikidataid
- position_1
- position_2
- position_3
- position_4
- ...

Wikipedia History
- pageid
- revid
- parentid
- user
- userid
- time
- size
- comment

Wikipedia Traffic
- pageid
- date
- traffic

Core
- country
- pageid
- wikidataid
- wikititle
- name
- sex
- ethnicity
- religion
- birth
- death
- birthplace
- deathplace

Political
- pageid
- session
- party
- constituency
- constituency2
- constituency_id
- session_start
- session_end
- service
- government
- leader
- ...

Social
- wikidataid
- twitter
- facebook
- youtube
- googlep
- instagram
- linkedin
- website

Professions
- wikidataid
- occupation_1
- occupation_2
- occupation_3
- occupation_4
- ...

Portraits
- pageid
- image_url

Voting and Social Media-Based Political Participation (Göbel 2020)

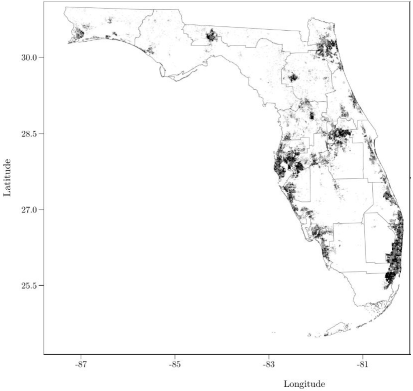
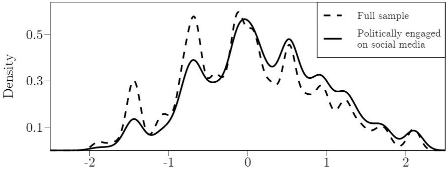


Figure 4: Voting propensities in the sample and among politically involved on social media.



Does Communication Signal Voter Preferences and Participation? (Göbel and Roth 2020)

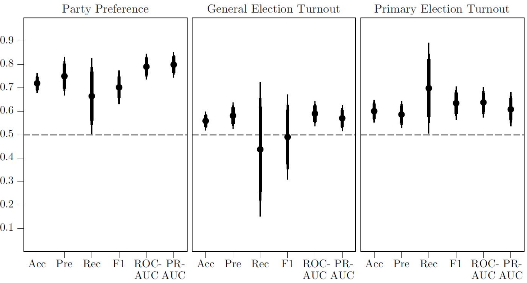
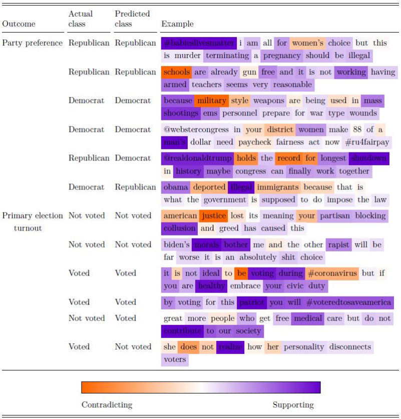


Table 3: Examples of model explanations (colors indicate support for the predicted class).



Outcome	Actual class	Predicted class	Example
Party preference	Republican	Republican	#babieslivesmatter i am all for women's choice but this is murder terminating a pregnancy should be illegal
	Republican	Republican	schools are already gun free and it is not working having armed teachers seems very reasonable
	Democrat	Democrat	because military style weapons are being used in mass shootings ems personnel prepare for war type wounds
	Democrat	Democrat	@webstercongress in your district women make 88 of a man's dollar need paycheck fairness act now #ru4fairpay
	Republican	Democrat	@realdonaldtrump holds the record for longest shutdown in history maybe congress can finally work together
	Democrat	Republican	obama deported illegal immigrants because that is what the government is supposed to do impose the law
Primary election turnout	Not voted	Not voted	american justice lost its meaning your partisan blocking collusion and greed has caused this
	Not voted	Not voted	biden's morals bother me and the other rapist will be far worse it is an absolutely shit choice
	Voted	Voted	it is not ideal to be voting during #coronavirus but if you are healthy embrace your civic duty
	Voted	Voted	by voting for this patriot you will #voteredtosaveamerica
	Not voted	Voted	great more people who get free medical care but do not contribute to our society
	Voted	Not voted	she does not realize how her personality disconnects voters

Contradicting Supporting

Why is this useful?

- get your hands on tons of original data
- no more copy-paste, automatize and save resources
- hard skills offer extended career choices
- an opportunity to apply what you have learned in prior courses
- reproducible and updatable data collection (in principle)

But there are also downsides

- data cleaning can be cumbersome
- R can be frustrating
- web data collection is not new, need to think harder about RQs
- legal and ethical challenges
- the web can change, code and data access can break

Course Schedule

4 Nov. 2020: Introduction

- Freelon (2018): Sections "Introduction" and "Web Scraping"
- Monroe (2013)
- Nyhuis(2017): Section "Introduction" and "Conceptual Challenges"

11 Nov. 2020: R Primer

- Tutorial script

18 Nov. 2020: Basic Web Technologies

- Munzert et al. (2015): Chapter 2 "HTML" and chapter 3 "XML and JSON"

25 Nov. 2020: Legal and Ethical Considerations

- Salganik (2018): Chapter 6 “Ethics”
- Freelon (2018): Section “Terms of Service”
- Steinert-Threkeld (2018): Chapter 6.2 “Ethics”
- Munzert et al. (2015): Chapter 9.3 “Web Scraping: Good Practice”

2 Dec. 2020: Collecting Data from Static Websites: XPath

- Munzert et al. (2015): Chapter 4 “XPath” and Chapter 9.2.2 “XPath”
- Nyhuis(2017): Section “Web Scraping”

9 Dec. 2020: Collecting Data from Static Websites: XPath practice session

16 Dec. 2020: Regular Expressions and Data Cleaning

- Munzert et al. (2015): Chapter 8 “Regular expressions and essential string functions” and Chapter 9.2.1 “Regular Expressions”

23 Dec. 2020: Collecting Data from APIs I

- Nyhuis(2017): Section “Application Programming Interfaces (APIs)”
- Munzert et al. (2015): Chapter 5 “HTTP”, Chapter 9.1.10 “Retrieving Data from APIs”, Chapter 9.1.11 “Authentication with OAuth”, and Chapter 9.2.3 “Application Programming Interfaces”

24 Dec. 2020 – 6 Jan. 2021: Christmas break

13 Jan. 2020: Collecting Data from APIs II/Collecting Data from Dynamic Websites

20 Jan. 2020: Collecting Data from Dynamic Websites

- Munzert et al. (2015): Chapter 6 “AJAX” and Chapter 9.1.9 “Scraping Data from AJAX-Enriched Webpages with Selenium/Rwebdriver”

27 Jan. 2020: Presentations in class

3 Feb. 2020: Presentations in class

10 Feb. 2020: Presentations in class

Prerequisites

Do I need to know R?

- no, but prior knowledge of R will be helpful
- you should be interested in programming and learning a programming language
- expect a higher workload if you have no prior background in R

Assignments

Preparation before class

- read the weekly literature
- from session 3: bring one new idea (RQ and potential web data source) to class

Grade and Credits

- presentation of planned project (max. 10 min)
- term paper and corresponding code (approx. 10 pages)
  - political science-related RQ
  - describes appropriate data source
  - describes implementation of web data collection
  - presents first exploration of the data with reference to RQ

Assignments

Preparation before class

- read the weekly literature
- from session 3: bring one new idea (RQ and potential web data source) to class

Grade and Credits

- presentation of planned project (max. 10 min)
- term paper and corresponding code (approx. 10 pages)
  - political science-related RQ
  - describes appropriate data source
  - describes implementation of web data collection
  - presents first exploration of the data with reference to RQ

Registration

Registration is conditional on the commitment to obtain a grade and credits!

If you are still interested

- send an email to sascha.goebel@uni-konstanz.de including your program of study
- until November 9 at 2 p.m.
- I will add you to the course on ILIAS
- on ILIAS you find the Zoom link and password for the other sessions
- if the number of participants exceeds 20 a lottery will take place

Next week

R Primer

- Please have a look at the r-primer.html on ILIAS
- Please have R and RStudio installed