

Class 6: Data Science Foundations

MA5953: Web Scraping and Text Mining

Dr. Miriam Sorace

www.miriamsorace.net

21 January 2022

Outline of Today's Class

Some Housekeeping ...

R Revision + Key Definitions

Data Principles and Ethics

Markup Languages

Some Housekeeping ...

Support/Contact

- ▶ For technical/content questions use the [Discussion Forums](#)
 - ▶ But also: exchange contact info with each other and study together! There is no better way to learn than in teams!
- ▶ For extenuating circumstances/mitigation/extensions contact CEMS Student Support: cemssupport@kent.ac.uk
- ▶ [Book Office Hour Slot](#)
- ▶ If all of the above did not work, you can email me at: m.sorace@kent.ac.uk

Course Details & Materials

- ▶ [Course GitHub Page](#)

Spring Term Learning Outcomes

1. Collect web data independently
 - ▶ Building Web Scrapers via R
2. Carry out text analysis via R

MAST5953: Web Scraping & Text Mining

Spring Term Structure

1. Class 6: Data Science Foundations
2. Class 7: Web Scraping and Regular Expressions
3. Class 8: Scraping Social Media Data
4. Class 8: Text Mining I: Pre-Processing and the Document-Term Matrix
5. Class 10: Text Mining II: Sentiment Analysis
6. Class 11: Text Mining III: Topic Models

MAST5953: Web Scraping & Text Mining

Class Structure

1. Introductory lecture

2. Computer lab

- ▶ 2 Groups
 - ▶ Group 1: Fridays from 9am
 - ▶ Group 2: Fridays from 12pm
- ▶ In the labs we will work together through the problem sets/R code.
- ▶ Being present in the computer lab will massively help to pre-test all relevant lines of code which will be used in the assessment!

First Assessment - Feedback

Common Mistakes

- ▶ Not following the RMarkdown set-up instructions as presented in class - and/or not troubleshooting the R error correctly (StackOverflow is a life-saver!), as explained in Class 1
 - ▶ We will review RMarkdown today
- ▶ Failing to respond to the assessment brief: many took the assessment as a way to test a research hypothesis, when it was actually about carrying quality checks on an original survey item!
 - ▶ Read the assessment prompt carefully, there is a lot of detail on the [Course GitHub Page](#)!
- ▶ Many uncritically used/failed to apply the R code to the specific requirements of their own study. The aim is not to see whether you can copy/paste all my lines of code in the exact order I present them in class! Show you've understood what each line of code does and can apply it when it suits.
 - ▶ **Study the code and the notes I provide on it in the .Rmd files.**
 - ▶ Practice every week what we've learnt in class. Do not leave it to the week your assignment is due!

First Assessment - Feedback

- ▶ Your grade is on Kent Vision.
- ▶ I have also provided written feedback on each of your scripts on Moodle: please read it carefully as it can help for your next assessment!

Spring Term Assessment

- ▶ 1,000 words Web Scraping / Text Mining Task
 - ▶ Due: 11th March 2021 - **4pm!**
 - ▶ Try to submit in advance as experiencing computer issues close to the deadline can mean you have to request a formal extension to CEMS, Moodle won't allow submissions after the deadline.
 - ▶ If glitches arise, submit to the late submission box then immediately alert CEMS to explain the late submission.
 - ▶ For issues with submission, and to request extensions/mitigation contact CEMS Student Support: cemssupport@kent.ac.uk

Spring Term Assessment

The Assessment Brief

In this task you will have to **scrape tweets from 2 politicians** of your choice and carry out **either a sentiment analysis or a topic model analysis**, on the basis of a research question of your choosing. The 1,000 words report will include (a) a section **describing/justifying the choice of research question** and **describing the text data scraped**; (b) a section where the **text mining method is presented**, and (c) a section where the **results from the comparison are presented (with visualisations and/or numerical summaries)**. The appendix and tables/figures and R code do not count towards the word limit.

PLEASE include a word count at the top of your report.

R Markdown will need to be used to generate the reports. The snippets of R code will need to be visible and will not count towards the word limit. The appendix will also not count towards the word limit. +/- 10% of the word limit is allowed.

Your Feedback to Me

And how it will shape the course

- ▶ Things that went well:
 - ▶ I am good at explaining/clarifying things (81%) - so do attend my classes, they are beneficial :) :)
 - ▶ The applied examples from research were particularly helpful to put stuff in context
 - ▶ A lot of helpful resources in the course GitHub page
 - ▶ Students felt included and felt they had tons of opportunities to ask questions
- ▶ Things that can be improved:
 - ▶ Some felt the need of a computer class/seminar
 - ▶ We'll have tons of this in this term, I will consider it for the survey bit in the future :)
 - ▶ More time devoted to clarify the assessment criteria/give feedback:
 - ▶ I usually discuss it in the first class of each term, and then leave space in the last class of term to allow for questions.
 - ▶ Hopefully the information above will also prove helpful for this!
 - ▶ The computer lab format will also help with this I believe.

R Revision + Key Definitions

Course Requirements: R

Particularly for the Spring Term Sessions!

- ▶ Prior knowledge of R is a must, make sure you understand the language basics (packages, objects/vectors, core functions and vector + data management operations), and that you know how to trouble-shoot errors and install packages
 - ▶ Great R Intro Resources:
 - ▶ Adler, Joseph. 2009. *R in a Nutshell. A Desktop Quick Reference*. O'Reilly
 - ▶ Teetor, Paul. 2011. *R Cookbook*. O'Reilly.
 - ▶ I recommend the following website for revisions:
<https://stats.idre.ucla.edu/r/>

Course Requirements: R

- ▶ Make sure you have installed R and RStudio and everything is up-to-date.
 - ▶ R:
 - ▶ Newest Version: 4.1.2
 - ▶ Type `rversions::r_release()` in R to check
 - ▶ Follow the instructions to download R here: [R Installer](#)
 - ▶ RStudio:
 - ▶ Updates needed: Go to «Help» - «Check for Updates» - & follow instructions in the pop-up
 - ▶ To download RStudio: [RStudio Installer](#)

R debugging

Examples of some common errors

- ▶ If it says: **there is no package called 'xxx'** you just need to type:
 - ▶ `install.packages("xxx")`
- ▶ If it says: **could not find function 'xxx'**, just call the library by typing:
 - ▶ `require(xxx)` or `library(xxx)`
- ▶ If it says: **cannot open the connection**, it means you have not specified the correct working directory (i.e. folder) where the file is located
 - ▶ `setwd()`
- ▶ Great Resources to Trouble-Shoot Errors:
 - ▶ StackOverflow
 - ▶ Stack Exchange
 - ▶ **just Google it, it's not cheating :)**



John B. Holbein
@JohnHolbein1



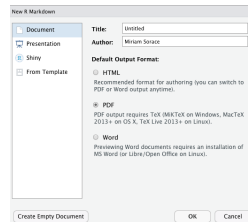
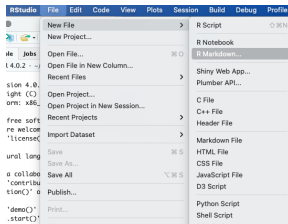
Students often don't realize that being good at coding really just means being good at Googling.

3:03 PM · Jan 11, 2021 · Twitter Web App

32 Retweets **18** Quote Tweets **554** Likes

Creating an RMarkdown File

► From scratch:



- Opening an existing file/template: «File» - «Open File» - Navigate to the folder where you saved the .Rmd file & select the file.
- Revise the [RMarkdown Cheat Sheet](#)

Problems with Knitting?

- ▶ If all troubleshooting fails, try to select the 'Word' or 'HTML' options instead of the pdf option when creating your RMarkdown
 - ▶ E.g. instead of output: `pdf_document`, use output: `html_document`
 - ▶ You can then convert to a pdf later (online pages or via 'Save as' in your laptop)!

Exercise 1: RMarkdown

- ▶ Create a new .Rmd file
- ▶ Knit the template R provides
- ▶ Create another .Rmd file but select the 'html' option instead

Key Terms: What is Data?

- ▶ Data = coded information for statistical processing
 - ▶ letters from an historical archive
 - ▶ legislative speeches
 - ▶ survey answers
 - ▶ tweets
 - ▶ Instagram pictures
- ▶ Data can be *primary* (collected directly by the researcher - e.g. original survey) or *secondary* (taken from an existing source - e.g. web database)

Key Terms: Data Science

- ▶ Collecting, organising and analysing 'big data'
- ▶ Advent of Web means large amounts of data are online
- ▶ Programming skills (R features prominently) are a must

Key Terms: Web Scraping

- ▶ Building a computer program (*scraper*) to grab specific content from webpages and convert it into usable datasets
- ▶ \neq Spidering/Web Crawling: grabbing entire webpage and links in an unstructured way

Disclaimer

you will quickly learn that there is no 'universal' web scraper: web pages are never the same and the same web page can change over time! Finding the perfect scraper to your specific data collection needs requires *customised* programs!

Web Scraping: When Should I Bother?

- ▶ data collection task is repeated (e.g. database update)
- ▶ data collection task is complex
 - ▶ large data
 - ▶ expensive (time or money) to do manually
- ▶ Web scraping is less error-prone, more time-efficient, and fully reproducible!

Web Scraping Example

How Facebook Started












Check out this short video from the movie:

[The Social Media - the 'hacking/data scraping' scene](#)

Web Scraping Example

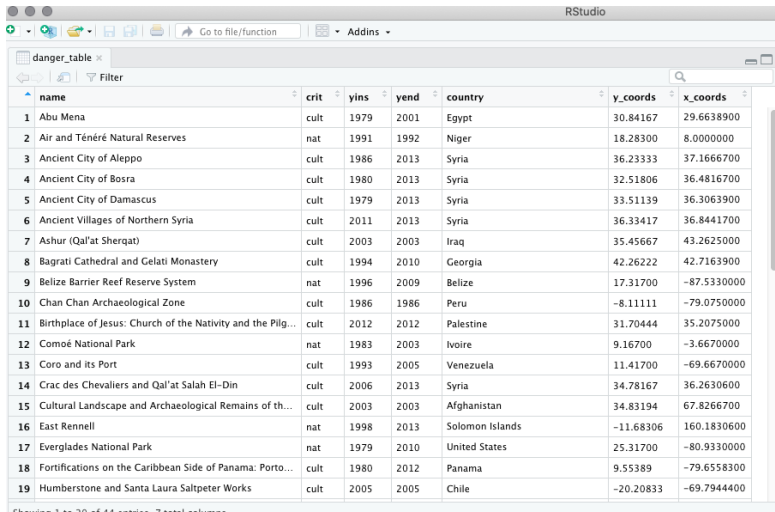
Heritage Sites in Danger

Which sites are threatened and where are they located?

Name	Image	Location	Criteria	Area ha (acre)	Year (WHIS)	Endangered	Reason
Abu Mena		Abu Mena, Egypt ↳ 30°50'30"N 29°39'50"E	Cultural: (iv)	182 (450)	1979	2001–	Cave-ins in the area caused by the clay at the surface, which becomes semi-liquid when met with "excess water"
Air and Ténéré Natural Reserves		Adia Department, Niger ↳ 16°17'N 8°0'E	Natural: (vi), (ix), (x)	7,736,000 (19,130,000)	1991	1992–	Military conflict and civil disturbance in the region as well as a reduction of wildlife population and degradation of the vegetation cover
Ancient City of Aleppo		Aleppo Governorate, Syria ↳ 36°14'N 37°10'E	Cultural: (ii)(iv)	350 (860)	1986	2013–	Syrian Civil War, currently held by the government. Bombings continue threatening the site.
Ancient City of Bosra		Damascus Governorate, Syria ↳ 32°31'5"N 36°28'54"E	Cultural: (ii)(iv)	—	1980	2013–	Syrian Civil War, held by the government.
Ancient City of Damascus		Damascus Governorate, Syria ↳ 33°30'41"N 36°18'23"E	Cultural: (ii)(iii)(iv) (vi)(iv)	86 (210)	1979	2013–	Syrian Civil War, rebel gunfire and mortar shelling, mainly from adjacent Jisr suburb endangers foundations.
Ancient Villages of Northern Syria		Syria ↳ 36°20'3"N 36°50'39"E	Cultural: (ii)(iv)(vi)	12,290 (30,400)	2011	2013–	Syrian Civil War, some held by rebels. Reports of looting and demolitions by Islamist groups.
Archaeological Site of Cyrene		Jebel Akhdar, Libya ↳ 32°49'30"N 21°51'30"E	Cultural: (ii), (iv), (vi)	—	1982	2016–	Libyan Civil War, presence of armed groups, already incurred and potential further damage.
Archaeological Site of Leptis Magna		Khoms, Libya ↳ 32°38'18"N 14°17'35"E	Cultural: (i), (iv), (vi)	—	1982	2016–	Libyan Civil War, presence of armed groups, already incurred and potential further damage.
Archaeological Site of Sabratha		Sabratha, Libya ↳ 32°48'19"N 12°29'9"E	Cultural: (iv)	—	1982	2016–	Libyan Civil War, presence of armed groups, already incurred and potential further damage.
Ashtur (Ga'rat Shergat)		Salah ad Din, Iraq ↳ 32°27'24"N 43°15'40"E	Cultural: (ii), (iv)	70 (170)	2003	2003–	A planned reservoir that would have partially flooded the site was suspended in the wake of the Iraq War by the new administration; lack of adequate protection.
Chan Chan Archaeological Zone		La Libertad, Peru ↳ 8°0'40"S 79°4'30"W	Cultural: (i), (iv)	600 (1,500)	1986	1986–	Natural erosion

Web Scraping Example

Heritage Sites in Danger - Example Code in GitHub page



The screenshot shows the RStudio interface with a data table named 'danger_table' loaded. The table contains 19 rows of data, each representing a heritage site in danger. The columns are: name, crit, yins, yend, country, y_coords, and x_coords. The data is sorted by 'crit' (cultural or natural) and then by 'yins' (year inscribed).

	name	crit	yins	yend	country	y_coords	x_coords
1	Abu Mena	cult	1979	2001	Egypt	30.84167	29.6638900
2	Air and Ténéré Natural Reserves	nat	1991	1992	Niger	18.28300	8.0000000
3	Ancient City of Aleppo	cult	1986	2013	Syria	36.23333	37.1666700
4	Ancient City of Bosra	cult	1980	2013	Syria	32.51806	36.4816700
5	Ancient City of Damascus	cult	1979	2013	Syria	33.51139	36.3063900
6	Ancient Villages of Northern Syria	cult	2011	2013	Syria	36.33417	36.8441700
7	Ashur (Qal'at Sherqat)	cult	2003	2003	Iraq	35.45667	43.2625000
8	Bagrati Cathedral and Gelati Monastery	cult	1994	2010	Georgia	42.26222	42.7163900
9	Belize Barrier Reef Reserve System	nat	1996	2009	Belize	17.31700	-87.5330000
10	Chan Chan Archaeological Zone	cult	1986	1986	Peru	-8.11111	-79.0750000
11	Birthplace of Jesus: Church of the Nativity and the Pilg...	cult	2012	2012	Palestine	31.70444	35.2075000
12	Comoé National Park	nat	1983	2003	Ivoire	9.16700	-3.6670000
13	Coro and its Port	cult	1993	2005	Venezuela	11.41700	-69.6670000
14	Crac des Chevaliers and Qal'at Salah El-Din	cult	2006	2013	Syria	34.78167	36.2630600
15	Cultural Landscape and Archaeological Remains of th...	cult	2003	2003	Afghanistan	34.83194	67.8266700
16	East Rennell	nat	1998	2013	Solomon Islands	-11.68306	160.1830600
17	Everglades National Park	nat	1979	2010	United States	25.31700	-80.9330000
18	Fortifications on the Caribbean Side of Panama: Porto...	cult	1980	2012	Panama	9.55389	-79.6558300
19	Humberstone and Santa Laura Saltpeter Works	cult	2005	2005	Chile	-20.20833	-69.7944400

Heritage Sites in Danger



Key Terms: Text Mining

- ▶ Automatic categorization of text data
- ▶ Usually done on the basis of document word frequency similarity
- ▶ Useful when dealing with textual 'big data', to impose a structure

Data Principles and Ethics

Data Collection Principles

Data Quality

- ▶ Is the data suited to answer my research question? Theory before data!
- ▶ Is the data source accurate? Cross-validate!
 - ▶ Wikipedia is not a qualitative gold standard ...

Data Collection Principles

Data Cleaning

- ▶ Is the data complete?
- ▶ Is the data consistently measured?
 - ▶ Use consistent variable names
 - ▶ Use consistent file names
- ▶ Data should be organised as a single rectangle
 - ▶ units in row, variables in columns
 - ▶ first row should contain variable names
 - ▶ in naming, avoid spaces and special characters. Use hyphens/underscores
 - ▶ only 1 piece of observation per cell
 - ▶ avoid numbering missing values. Use blank space, "NA", or "."

Data Collection Principles

Example Dataset

StataCleanedMetaData																
Search Sheet																
General																
Conditional Formatting																
Format as Table																
Cell Styles																
Insert																
Delete																
Format																
AutoSum																
Fill																
Clear																
Sort & Filter																
A1																
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
		TITLE	ProcedureType	Migr_topic	EUROVOC	Date	Sub_Topic	URL	TYPE	AMENOMEN	SUBJECTS	SUMMARY	PROC_URL	COM_ACTIO	COUNC_ACT	EP_ACTIONS
1																
2	31990J0364	Council Direc Consultation	0	"health insu	1990		1 https://eur-l		FALSE	"Internal market - Principles",		FALSE	https://eur-l	7	2	8 DG II Ji Industry
3	31990J0269	Council Direc Consultation	0	"occupation	1990		2 https://eur-l		FALSE	"Social provisions", "Safety at		TRUE	https://eur-l	12	2	4 Directorate-General for Employment, Social
4	31990J0679	Council Direc Consultation	0	"occupation	1990		2 https://eur-l		FALSE	"Social provisions", "Safety at		FALSE	https://eur-l	7	2	6 Directorate-General for Employment, Social
5	31990J0641	Council Direc Consultation	0	"occupation	1990		2 https://eur-l		FALSE	"Safety at work and elsewhere		FALSE	https://eur-l	6	1	1 Directorate-General for Environment
6	31990K1360	Council Regu Consultation	0	"vocational	1990		1 https://eur-l	R	FALSE	"Provisions governing the Insti		FALSE		NA	NA	NA
7	31990J0270	Council Direc Consultation	0	"occupation	1990		2 https://eur-l		FALSE	"Safety at work and elsewhere		TRUE	https://eur-l	11	3	6 Directorate-General for Employment, Social
8	31990J0394	Council Direc Consultation	0	"occupation	1990		2 https://eur-l		FALSE	"public health", "Safety at wor		FALSE	https://eur-l	11	3	6 Directorate-General for Employment, Social
9	31990J0365	Council Direc Consultation	0	"insurance"	1990		1 https://eur-l		FALSE	"Internal market - Principles",		FALSE	https://eur-l	7	2	4 DG II Ji Industry
10	31991J0383	Council Direc Consultation	0	"worker info	1991		2 https://eur-l		FALSE	"Social provisions", "Approxim		TRUE	https://eur-l	9	3	6 Directorate-General for Employment, Social
11	31991J0533	Council Direc Consultation	0	"worker info	1991		1 https://eur-l		FALSE	"Social provisions", "Approxim		TRUE	https://eur-l	7	2	5 Directorate-General for Employment, Social
12	31991J0368	Council Direc Consultation	0	"occupation	1991		2 https://eur-l		TRUE	"Internal market - Principles",		FALSE	https://eur-l	9	2	6 DG II Ji Industry
13	31991J0534	Council Direc Consultation	0	"agriculture	1991		1 https://eur-l		TRUE	"Social provisions", "Informati		FALSE	https://eur-l	2	1	0 Legal service
14	31991J0382	Council Direc Consultation	0	"occupation	1991		2 https://eur-l		TRUE	"Safety at work and elsewhere		FALSE	https://eur-l	7	3	5 Directorate-General for Employment, Social
15	31992R2079	Council Regu Consultation	0	"early retire	1992		4 https://eur-l	R	TRUE	"Social provisions", "Agricultu		FALSE	https://eur-l	4	2	4 Directorate-General for Agriculture and Rura
16	31992J0057	Council Direc Consultation	0	"occupation	1992		2 https://eur-l		FALSE	"Internal market - Principles",		TRUE	https://eur-l	10	3	6 Directorate-General for Employment, Social
17	31992J0058	Council Direc Consultation	0	"occupation	1992		2 https://eur-l		TRUE	"Internal market - Principles",		TRUE	https://eur-l	10	3	4 Directorate-General for Employment, Social
18	31992R2434	Council Regu Consultation	0	"free move	1992		1 https://eur-l	R	TRUE	"Free movement of workers"]		FALSE	https://eur-l	7	2	4 Directorate-General for Employment, Social
19	31992J0056	Council Direc Consultation	0	"worker info	1992		1 https://eur-l		TRUE	"Social provisions", "Approxim		FALSE	https://eur-l	6	2	4 Directorate-General for Employment, Social
20	31992R1247	Council Regu Consultation	1	"disabled pe	1992		1 https://eur-l	R	TRUE	"Social provisions for migrant wo		FALSE	https://eur-l	4	2	1 Directorate-General for Employment, Social
21	31992R3654	Council Regu Consultation	0	"discount se	1992		4 https://eur-l	R	TRUE	"Public products"]		FALSE	https://eur-l	2	1	0 Directorate-General for Agriculture and Rura
22	31992J0104	Council Direc Consultation	0	"occupation	1992		2 https://eur-l		FALSE	"Approximation of laws", "Saf		TRUE	https://eur-l	11	2	6 Directorate-General for Employment, Social
23	31992J0091	Council Direc Consultation	0	"humanisa	1992		2 https://eur-l		FALSE	"Safety at work and elsewhere		TRUE	https://eur-l	12	2	3 Directorate-General for Employment, Social
24	31992J0029	Council Direc Consultation	0	"occupation	1992		2 https://eur-l		FALSE	"Transport", "Safety at work a		TRUE	https://eur-l	10	3	7 Directorate-General for Employment, Social
25	31992J0085	Council Direc Consultation	0	"humanisa	1992		2 https://eur-l		FALSE	"Safety at work and elsewhere		TRUE	https://eur-l	11	3	8 Directorate-General for Employment, Social
26	31992J0104	Council Direc Consultation	0	"humanisa	1993		2 https://eur-l		FALSE	"Internal market - Principles",		FALSE	https://eur-l	9	3	5 Directorate-General for Employment, Social
27	31992R1347	Council Regu Consultation	0	"board of di	1993		4 https://eur-l	R	TRUE	"Provisions under Article 235 E		FALSE	https://eur-l	5	1	3 Directorate-General for Budget
28	31993J0086	Council Direc Consultation	0	"occupation	1993		2 https://eur-l		FALSE	"public health", "Safety at wor		FALSE	https://eur-l	7	3	4 Directorate-General for Employment, Social
29	31993J0103	Council Direc Consultation	0	"non-service	1993		2 https://eur-l		FALSE	"Approximation of laws", "Fiel		TRUE	https://eur-l	9	3	4 Directorate-General for Employment, Social
30	31994J0045	Council Direc Consultation	0	"group of cc	1994		1 https://eur-l		FALSE	"Approximation of laws", "Soc		FALSE	https://eur-l	8	3	4 Directorate-General for Employment, Social
31	31994J0033	Council Direc Consultation	0	"labour star	1994		2 https://eur-l		FALSE	"Approximation of laws", "Soc		TRUE	https://eur-l	8	3	7 Directorate-General for Employment, Social
32	31994R2063	Council Regu Consultation	0	"vocational	1994		1 https://eur-l	R	TRUE	"Provisions governing the Insti		FALSE	https://eur-l	3	1	2 Task Force for the Accession Negotiations
33	31994R3096	Council Regu Consultation	0	"discount se	1994		4 https://eur-l	R	TRUE	"European Agricultural Guidan		FALSE	https://eur-l	2	1	0 Directorate-General for Agriculture and Rura
34	31994J0058	Council Direc Consultation	0	"vocational	1994		2 https://eur-l		FALSE	"Internal market - Principles",		FALSE	https://eur-l	6	3	2 DG VII Ji Transport

Data Collection Principles

Data Transparency & Reproducibility

- ▶ Store the data in open formats (best: “.csv”), avoid software-specific formats
- ▶ Document everything!
 - ▶ Variable codebook
 - ▶ R scripts / do files
- ▶ Have a master dataset that you **never** overwrite
 - ▶ data you can always go back to
- ▶ backup your files! (Dropbox / Google Drive)

Data Collection Principles

Example Variable Codebook

General Questions on European Integration

CHES 2019 Codebook

EU_POSITION = overall orientation of the party leadership towards European integration in 2019.

- 1 = Strongly opposed
- 2 = Opposed
- 3 = Somewhat opposed
- 4 = Neutral
- 5 = Somewhat in favor
- 6 = In favor
- 7 = Strongly in favor

EU_POSITION_SD = standard deviation of expert placement of overall orientation of the party leadership towards European integration in 2019.

EU_SALIENCE = relative salience of European integration in the party's public stance in 2019.

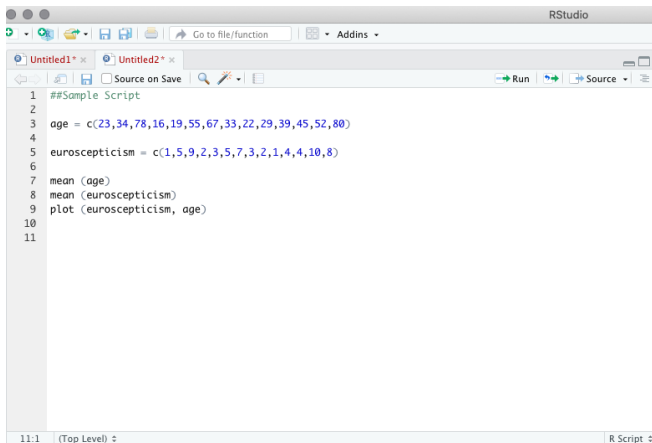
- 0 = European Integration is of no importance.
- :
- 10 = European Integration is of great importance.

EU_DISSENT = degree of dissent on European integration in 2019.

- 0 = Party was completely united.
- :
- 10 = Party was extremely divided.

Data Collection Principles

Example R Script



The screenshot shows the RStudio interface with two untitled files open. The active file, 'Untitled2', contains an R script. The script defines two vectors, 'age' and 'euroscepticism', and then calculates their means and creates a plot. The status bar at the bottom indicates the cursor is at line 11:1 at the top level of the script.

```
1 ##Sample Script
2
3 age = c(23,34,78,16,19,55,67,33,22,29,39,45,52,80)
4
5 euroscepticism = c(1,5,9,2,3,5,7,3,2,1,4,4,10,8)
6
7 mean (age)
8 mean (euroscepticism)
9 plot (euroscepticism, age)
10
11
```

That's why using RMarkdown is great!

Ethical issues

- ▶ do not violate copyrights/terms of use: check if you have permission
- ▶ cite/give appropriate credit for data
- ▶ Respect privacy: anonymise!
- ▶ Do not coerce: make sure you have consent

Markup Languages

Web page Structures

- ▶ **HTML** *Hypertext Markup Language*: series of symbols that define the structure and presentation of a web page. Invaluable to learn in web scraping as this is the standard language.
- ▶ **XML** *Extensible Markup Language*: almost identical to HTML but more flexible since tags can be user-defined.
- ▶ **JSON** *JavaScript Object Notation*: more lightweight formatting language, no start/end tags, just key/value pairs and curly or square brackets to express hierarchies.

HTML



Figure 2.1 Browser view of a simple HTML document



Figure 2.2 Source view of a simple HTML document

Some Housekeeping ...
oooooooooooo

Revision + Key Definitions
oooooooooooooooooooo

Data Principles and Ethics
oooooooooooo

Markup Languages
ooo●oooooooooooo

HTML: Element Inspector

This screenshot shows the Wikipedia page for the University of Kent. The Element Inspector is open, displaying the HTML structure of the page. The main content area shows the university's name, its location, and a brief history. The Element Inspector highlights the 'h1' tag for 'University of Kent'.

Wikipedia, the free encyclopedia

Not logged in - Talk Contributions Create account Log in

Article Talk Feed Edit View history Search Wikipedia

University of Kent

From Wikipedia, the free encyclopedia

Not to be confused with Kent State University.

The **University of Kent** (formerly the **University of Kent at Canterbury**, abbreviated as **UKC**) is a semi-collegiate public research university based in Kent, United Kingdom. It was founded in 1965 and is a plate glass university. The University was granted its Royal Charter on 4 January 1969 and the following year Princess Marina, Duchess of Kent was formally installed as the first Chancellor.^[a]

The university has its main campus north of Canterbury situated within 300 acres (1.2 km²) of park land, housing over 6,000 students, as well as campuses in Maidstone and Tonbridge in Kent and European postgraduate centres in Brussels, Athens, Rome and Paris.^[b] The University is international, with students from 156 different nationalities and 41% of its academic and research staff being from outside the United Kingdom.^[a] It is a member of the Santander Network of European universities encouraging social and economic development.^[c]

Canterbury [d]

History

- 1.1 Origins
- 1.2 1965 to 2000
- 1.3 2000 to present

Campus

- 2.1 Canterbury campus
- 2.1.1 Facilities
- 2.1.2 Transport and access
- 2.2 Maidstone campus
- 2.3 Tonbridge campus

Organisation and administration

- 2.1 Facilities, departments and schools
- 2.2 Colleges

Coat of Arms of the University of Kent

Further facts

- University of Kent at Canterbury
- Latin: Our service reaches out

Media

In English

Literal translation: "Whom to serve is to reign"

Book of Common Prayer translation: "Whose service is perfect freedom"

Type

Established

4 January 1965

Endowment

US \$58 million (as of 31 July 2016)^[e]

Other projects

This screenshot shows the same Wikipedia page for the University of Kent, but with the Element Inspector open to the 'Code' tab. This view displays the raw HTML markup for the page, including the title, the main text, and the various templates and categories used. The 'h1' tag for 'University of Kent' is highlighted.

Wikipedia, the free encyclopedia

Not logged in - Talk Contributions Create account Log in

Article Talk Feed Edit View history Search Wikipedia

University of Kent

From Wikipedia, the free encyclopedia

Not to be confused with Kent State University.

The **University of Kent** (formerly the **University of Kent at Canterbury**, abbreviated as **UKC**) is a semi-collegiate public research university based in Kent, United Kingdom. It was founded in 1965 and is a plate glass university. The University was granted its Royal Charter on 4 January 1969 and the following year Princess Marina, Duchess of Kent was formally installed as the first Chancellor.^[a]

The university has its main campus north of Canterbury situated within 300 acres (1.2 km²) of park land, housing over 6,000 students, as well as campuses in Maidstone and Tonbridge in Kent and European postgraduate centres in Brussels, Athens, Rome and Paris.^[b] The University is international, with students from 156 different nationalities and 41% of its academic and research staff being from outside the United Kingdom.^[a] It is a member of the Santander Network of European universities encouraging social and economic development.^[c]

Canterbury [d]

History

- 1.1 Origins
- 1.2 1965 to 2000
- 1.3 2000 to present

Campus

- 2.1 Canterbury campus
- 2.1.1 Facilities
- 2.1.2 Transport and access
- 2.2 Maidstone campus
- 2.3 Tonbridge campus

Organisation and administration

- 2.1 Facilities, departments and schools
- 2.2 Colleges

Coat of Arms of the University of Kent

Further facts

- University of Kent at Canterbury
- Latin: Our service reaches out

Media

In English

Literal translation: "Whom to serve is to reign"

Book of Common Prayer translation: "Whose service is perfect freedom"

Type

Established

4 January 1965

Endowment

US \$58 million (as of 31 July 2016)^[e]

Other projects

HTML: Core Elements

- ▶ **head**: metadata element, usually containing document characteristics
- ▶ **title**: document title, used by search engines
- ▶ **body**: contains the webpage contents
- ▶ **header**: defines section headers
- ▶ **div**: defines a section of the html document
- ▶ **link**: defines links to external resources. **Attributes** like `rel=""` and `href=""` define, respectively, type of link and url of the external resource

HTML: Formatting Tags

- ▶ **p**: defines separate paragraphs
- ▶ **br**: defines a line break
- ▶ **b**: bold text
- ▶ **i**: text in italic
- ▶ **ul**: unordered lists
- ▶ **ol**: ordered lists
- ▶ **h1 ... h6**: headline size
- ▶ **span**: usually combined with attribute `class=""` or `style=""` marks up parts of text to change colors and style
- ▶ **table**: used to begin a table **tr** used to begin a row, **td** for cells, and **th** for header cells.

Always preceded and followed starting and closing brackets. For more information see <https://www.w3schools.com/html/default.asp>.

XML & JSON: Examples

```
1 <?xml version="1.0" encoding="ISO-8859-1"?>
2 <bond_movies>
3   <movie id="1">
4     <name>Dr. No</name>
5     <year>1962</year>
6     <actors bond="Sean Connery" villain="Joseph Wiseman"/>
7     <budget>1.1M</budget>
8     <boxoffice>59.5M</boxoffice>
9   </movie>
10  <movie id="2">
11    <name>Live and Let Die</name>
12    <year>1973</year>
13    <actors bond="Roger Moore" villain="Yaphet Kotto"/>
14    <budget>7M</budget>
15    <boxoffice>126.4M</boxoffice>
16  </movie>
17  <movie id="3">
18    <name>Skyfall</name>
19    <year>2012</year>
20    <actors bond="Daniel Craig" villain="Javier Bardem"/>
21    <budget>175M</budget>
22    <boxoffice>1108.6M</boxoffice>
23  </movie>
24 </bond_movies>
```

Figure 3.1 An XML code example: James Bond movies

```
1 {"indy movies" : {
2   {
3     "name" : "Raiders of the Lost Ark",
4     "year" : 1981,
5     "actors" : {
6       "Indiana Jones" : "Harrison Ford",
7       "Dr. René Belloq" : "Paul Freeman"
8     },
9     "producers" : ["Frank Marshall", "George Lucas", "Howard Kazanjian"],
10    "budget" : 18000000,
11    "academy_award_ve" : true
12  },
13  {
14    "name" : "Indiana Jones and the Temple of Doom",
15    "year" : 1984,
16    "actors" : {
17      "Indiana Jones" : "Harrison Ford",
18      "Mola Ram" : "Amish Puri"
19    },
20    "producers" : ["Robert Watts"],
21    "budget" : 28170000,
22    "academy_award_ve" : true
23  },
24  {
25    "name" : "Indiana Jones and the Last Crusade",
26    "year" : 1989,
27    "actors" : {
28      "Indiana Jones" : "Harrison Ford",
29      "Walter Donovan" : "Julian Glover"
30    },
31    "producers" : ["Robert Watts", "George Lucas"],
32    "budget" : 48000000,
33    "academy_award_ve" : false
34  }
35 }
```

Figure 3.9 JSON code example: Indiana Jones movies

Steps of Web Scraping

1. **Inspect** background structure of the web page
2. **Read** HTML/XML/JSON structure into R
3. **Parse** = extract information from relevant webpage elements only & convert to usable format (data frame)

Exercise 2: Inspecting Web-Pages

- ▶ Open a webpage (e.g. newspaper or Wikipedia).
- ▶ Have a look at the source code (right-click on the webpage with your mouse and select the command that mentions 'page source' or 'inspect element').
- ▶ Check out the structure of the code underlying the webpage: can you find some familiar element?

Exercise 3: Create a Web-Page

1. Write up a Word Document with a section head in bold and a series of paragraphs and an ordered list (you can also copy/paste content from a website)
2. Save the word document as an html file (scroll down the 'file format' field, after clicking 'save as').
3. Open the html file and explore it with the 'Inspect Element' function in your browser (right-clicking). What does the HTML structure of your document look like?

Exercise 4: Re-create a Web-Page Structure

1. Go to: [w3schools.com/html/tryit](https://www.w3schools.com/html/tryit)
2. Play with HTML editor and try to re-create the main section of this webpage (doesn't have to be an identical/perfect result, try to approximate it as best you can!): e.g. [EU Treaties](#)

The w3schools website also has suggestions on useful tags you can use, with examples, see:

<https://www.w3schools.com/html/default.asp>)