Regular Expressions 101

Adela Sobotkova

History and Classical Studies

29th August 2022

Pre-Carpentry survey

At the start and end of every carpentries workshop, I poll participants.
`http://bit.ly/DigitalMethodsSurvey`

Figure 1: HASS pre-survey

Code of Conduct

This class is using a great deal of material from The Carpentries. All interactions related to this class, inside and outside, abide by The Carpentries Code of Conduct.
Report code of conduct violations to Adela, or student representative.
`https://docs.carpentries.org/topic_folders/policies/code-of-conduct.html`
In summary, I want to emphasise:

- Use welcoming and inclusive language
- Be respectful of different viewpoints and experiences
- Gracefully accept constructive criticism
- Focus on what is best for the community
- Show courtesy and respect towards other community members

Sticky notes

We use sticky notes during our in person hands-on sessions to indicate progress or needs for assistance.
We also use them as minute cards for feedback and the end of each session.

Starting the hands-on

https://librarycarpentry.org/lc-data-intro/

https://datacarpentry.org/socialsci-workshop/

https://datacarpentry.org/spreadsheets-socialsci/setup.html

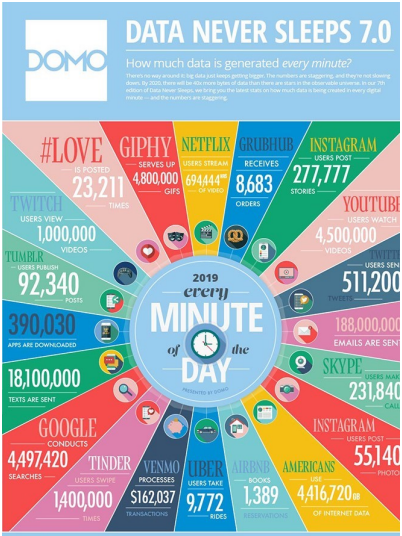
https://datacarpentry.org/openrefine-socialsci/setup.html

https://swcarpentry.github.io/shell-novice/

https://swcarpentry.github.io/git-novice/

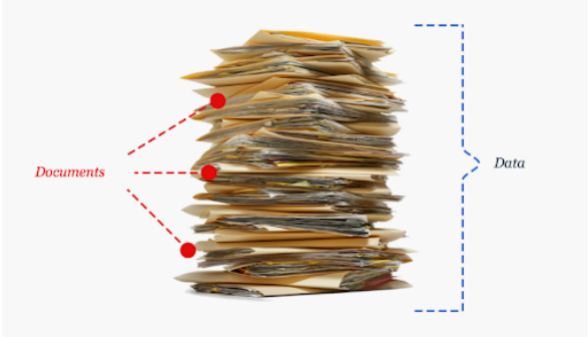
https://datacarpentry.org/r-socialsci/setup.html

Data never sleeps



Data

... much of the data out there is text-based.



Computers are stupid

... but they don't get tired.

People interpret. Machines don't. Computer only does what you tell it to. If it throws up an error it is often not your fault, rather in most cases the computer has failed to interpret what you mean because it can only work with what it knows (ergo, it is bad at interpreting).

Regular Expressions

Regular Expressions (regexes) are a powerful way to handle a multitude of different types of data. They can be used to find patterns in text and make sophisticated replacements. Think of them as find and replace on steroids. It's good to learn what they can do and how to apply them to your research.
Good training sites are:

`www.regex101.com`

`www.regexr.com`

How can you use Regular Expressions in your work?

Figure 3: www.xkcd.com/208/ CC BY NC

Regular Expressions - Carpentry exercises

`https://librarycarpentry.org/lc-data-intro/`

What will the regular expression Fr[ea]nc[eh]$ match?

How do you match the whole words "colour" and "color"?

How would you match the date format dd-MM-yyyy or dd-MM-yy at the end of a line only?

Regular Expressions - exercise 1

Avian report `http://bit.ly/regexexercise1`

- Find literal words: avian, AVIAN, Avian
- Find only capitalised words

Regular Expressions - exercise 2

American history dates `http://bit.ly/regexexercise2`

- Match all the dates
- Convert them to YYYY-MM-DD format

...

Juan Ponce de León sights Florida for the first time, on 3.27, 1513 Giovanni da Verrazzano explored the Atlantic coast of North America under French employ, on 4.17.1524 The Roanoke Colony was found deserted, on 8/15/1590 John Smith founded the Jamestown settlement, on 5/14, 1607 The Dutch laid claim to the territories of New Netherland, on 11.11.1614 The Massachusetts Bay Colony founded, on 3-4-1629

Regular Expressions - exercises 3 and 4 : stopword lists

Make your own stop-word lists for R and for Voyant!

...

A stop-word list for R needs to be formatted as a block of comma-separated words, enclosed in quotations. It should look like the text in exercise 4, ie.: "a", "an", "the", "of",...

...

Contrariwise, stop-word list for Voyant should have exactly one word on each line without any quotations or punctuation. (Click here to see Voyant software)

...

Take the Voyant list below and convert it to an R stopword list:

— `http://bit.ly/regexexercise3`

Take the R stopword list below and convert it back to a Voyant stopword list (each word on a new line and without any quotations):

— `http://bit.ly/regexexercise4`

Regular Expressions - exercise 5 : Dis Manibus invocations

Dis Manibus is the full version of the invocation of the underworld gods, which typically occurs on funerary inscriptions. Depending on the status of the individual it appears in these canonic versions:

D M

D M S

Dis Manibus

Diis Manibus

Dis Manibus Sacrum

Diis Manibus Sacrum

Can you write a regular expression that would help you find all the records in the EDH_tiny inscriptions (n=81476) that contain some invocation of the underworld gods? How many are there?

Examples from (large-scale) historical research

AUCTORITATES Project by & Vojtech Kase

Extracting 70K+ biblical references from an extensive corpus of ancient Christian texts (e.g. "1 Cor. 11:17-34"); code: `https://bit.ly/2R8NoI1`; data example: `https://bit.ly/2GQ961D`

stoa0162.stoa004.opp-lat3. Jerome	Epistulae, 120-154	cf.	Act. 9, 15 "Rom.	Acts 9, 15	a	acts_9_15	single	chapter_verse	Acts	9	15		
stoa0162.stoa004.opp-lat3. Jerome	Epistulae, 120-154		9 Ps. 142, 2 13	Psa. 142, 2	c	psa_142_2	single	chapter_verse	Psa.	142	2		
stoa0162.stoa004.opp-lat3. Jerome	Epistulae, 120-154	cf.	I Reg. 16, 7 21	1Kgs. 16, 7	a	1kgs_16_7	single	chapter_verse	1Kgs.	16	7		
stoa0162.stoa004.opp-lat3. Jerome	Epistulae, 120-154	cf.	Ps. 1, 2 1	Psa. 1, 2	a	psa_1_2	single	chapter_verse	Psa.	1	2		
stoa0162.stoa004.opp-lat3. Jerome	Epistulae, 120-154	cf.	Matth. 13, 10-17. Mar	Matt. 13, 10-17.	a	matt_13_10to17	ranged	chapter_verse	Matt.	13	10-17	10	17
stoa0162.stoa004.opp-lat3. Jerome	Epistulae, 120-154	10-17.	Marc. 4, 10-12. Luc._	Mark 4, 10-12.	c	mark_4_10to12	ranged	chapter_verse	Mark	4	10-12	10	12
stoa0162.stoa004.opp-lat3. Jerome	Epistulae, 120-154	cf.	Hier. 51, 7 13	Jer. 51, 7	a	jer_51_7	single	chapter_verse	Jer.	51	7		
stoa0162.stoa004.opp-lat3. Jerome	Epistulae, 120-154	cf.	II Tim. 3, 6-7 25	2Tim. 3, 6-7	a	2tim_3_6to7	ranged	chapter_verse	2Tim.	3	6-7	6	7
stoa0162.stoa004.opp-lat3. Jerome	Epistulae, 120-154	cf.	II Tim. 4, 3 27	2Tim. 4, 3	a	2tim_4_3	single	chapter_verse	2Tim.	4	3		
stoa0162.stoa004.opp-lat3. Jerome	Epistulae, 120-154	cf.	Ezech. 13, 10 1	Eze. 13, 10	a	eze_13_10	single	chapter_verse	Eze.	13	10		
stoa0162.stoa004.opp-lat3. Jerome	Epistulae, 120-154	cf.	Matth. 15, 14. Luc._6,	Matt. 15, 14.	a	matt_15_14	single	chapter_verse	Matt.	15	14		
stoa0162.stoa004.opp-lat3. Jerome	Epistulae, 120-154	cf.	Act. 9, 15 28	Acts 9, 15	a	acts_9_15	single	chapter_verse	Acts	9	15		
stoa0162.stoa004.opp-lat3. Jerome	Epistulae, 120-154		28 I Cor. 4, 7 2	1Cor. 4, 7	c	1cor_4_7	single	chapter_verse	1Cor.	4	7		
stoa0162.stoa004.opp-lat3. Jerome	Epistulae, 120-154		8 Phil. 2, 13 9	Phil. 2, 13	c	phil_2_13	single	chapter_verse	Phil.	2	13		
stoa0162.stoa004.opp-lat3. Jerome	Epistulae, 120-154		9 Ioh. 5, 17 15	John 5, 17	c	john_5_17	single	chapter_verse	John	5	17		
stoa0162.stoa004.opp-lat3. Jerome	Epistulae, 120-154		15 Ps. 33, 9 18	Psa. 33, 9	c	psa_33_9	single	chapter_verse	Psa.	33	9		
stoa0162.stoa004.opp-lat3. Jerome	Epistulae, 120-154	cf.	I Ioh. 4, 18 21	1John 4, 18	a	1john_4_18	single	chapter_verse	1John	4	18		
stoa0162.stoa004.opp-lat3. Jerome	Epistulae, 120-154	cf.	I Cor. 13, 9—12 26	1Cor. 13, 9—12	a	1cor_13_9	single	chapter_verse	1Cor.	13	9		
stoa0162.stoa004.opp-lat3. Jerome	Epistulae, 120-154	cf.	Phil. 3, 12-13 2	Phil. 3, 12-13	a	phil_3_12to13	ranged	chapter_verse	Phil.	3	12-13	12	13
stoa0162.stoa004.opp-lat3. Jerome	Epistulae, 120-154		13 Ps. 16, 4 15	Psa. 16, 4	c	psa_16_4	single	chapter_verse	Psa.	16	4		
stoa0162.stoa004.opp-lat3. Jerome	Epistulae, 120-154		15 Matth. 7, 13 Matth.	Matt. 7, 13	c	matt_7_13	single	chapter_verse	Matt.	7	13		
stoa0162.stoa004.opp-lat3. Jerome	Epistulae, 120-154		13 Matth. 5, 44 etc.	Matt. 5, 44	c	matt_5_44	single	chapter_verse	Matt.	5	44		
stoa0162.stoa004.opp-lat3. Jerome	Epistulae, 120-154	cf.	Act. 9, 15 17	Acts 9, 15	a	acts_9_15	single	chapter_verse	Acts	9	15		
stoa0162.stoa004.opp-lat3. Jerome	Epistulae, 120-154		21 Gal. 2, 16 24	Gal. 2, 16	c	gal_2_16	single	chapter_verse	Gal.	2	16		
stoa0162.stoa004.opp-lat3. Jerome	Epistulae, 120-154	et	Rom. 9, 13 11	Rom. 9, 13	c	rom_9_13	single	chapter_verse	Rom.	9	13		
stoa0162.stoa004.opp-lat3. Jerome	Epistulae, 120-154	cf.	Ios. c. 7 12	Josh. c. 7	a	josh_7_12	single	chapter_verse	Josh.	7	12		
stoa0162.stoa004.opp-lat3. Jerome	Epistulae, 120-154	cf.	I Reg. capp. 2-4 14	1Kgs. capp. 2-4	a	1kgs_.2to4	ranged	no_chapter_vers	1Kgs.		2-4	2	4
stoa0162.stoa004.opp-lat3. Jerome	Epistulae, 120-154	cf.	Rom. c. 9 1	Rom. c. 9	a	rom_9_1	single	chapter_verse	Rom.	9	1		
stoa0162.stoa004.opp-lat3. Jerome	Epistulae, 120-154		8 Ps. 31, 1 8	Psa. 31, 1	c	psa_31_1	single	chapter_verse	Psa.	31	1		
stoa0162.stoa004.opp-lat3. Jerome	Epistulae, 120-154	cf.	Hebr. 5, 12 1	Heb. 5, 12	a	heb_5_12	single	chapter_verse	Heb.	5	12		
stoa0162.stoa004.opp-lat3. Jerome	Epistulae, 120-154		3 I Cor. 1, 19 4	1Cor. 1, 19	c	1cor_1_19	single	chapter_verse	1Cor.	1	19		
stoa0162.stoa004.opp-lat3. Jerome	Epistulae, 120-154	cf.	Ioh. 19, 24 17	John 19, 24	a	john_19_24	single	chapter_verse	John	19	24		

Examples from (large-scale) historical research

Dating PHI Project Extracting dates from 200K+ ancient Greek inscriptions (e.g. "early 2nd c. BC"); code: `https://bit.ly/32h1XiM`

PHI_ID	tildeinfo	raw_date	not_before	not_after	certainty	or	date_tags	source
262001	IStr 427	IStr 427					['unknown']	
266001	Pont. — Amasia — Rom. Imp. period	Rom. Imp. period	-31	410			['range', 'period']	PeriodO
231001	Att. — Athens: Agora — 220/19	220/19	-220	-219			['range']	
79501	Thasos	Thasos					['unknown']	
218501	Eg. — el-Boueib	el-Boueib					['unknown']	
210001	Korinthia — Korinthos — ca. 100-150 AD	ca. 100-150 AD	95	155			['range', 'phase', 'ca']	
179501	Korinthia — Korinthos — 267-668 AD	267-668 AD	267	668			['range']	
183001	N. Black Sea — Pantikapaion (Kerch) — 1st c. BC — IosPE IV 253	1st c. BC	-100	-1			['range', 'cent']	
183501	N. Black Sea — Pantikapaion (Kerch) — Roman period? — IosPE IV 348	Roman period?	-146	324	?		['range', 'period']	
184001	N. Black Sea — Tanais — 155 AD — IosPE II 438	155 AD	155	155			['exact']	
278501	Bith. — Prusa ad Olympum (Bursa) — 199 AD	199 AD	199	199			['exact']	
298501	Dacia Sup. — Tibiscum (Jupa) — 2nd/3rd c. AD	2nd/3rd c. AD	101	300			['range', 'cent', 'morece']	
288501	Mys.: Aisepos — Zeus Olbios bei Buğdaylı? — JHS 25,1905,56 Nr.4	JHS 25,1905,56 Nr.4					['unknown']	
149501	Makedonia (Bottiaia) — Beroia — ca. 100-150 AD — SEG 35.714	ca. 100-150 AD	95	155			['range', 'phase', 'ca']	
150001	Makedonia (Bottiaia) — Beroia — Rachi — 1st c. BC — ABSA 39 (1938/39) 95, 4	1st c. BC	-100	-1			['range', 'cent']	
63001	Delos — ca. 300-250 BC	ca. 300-250 BC	-305	-245			['range', 'phase', 'ca']	
63501	Delos — beg. 2nd c. BC	beg. 2nd c. BC	-200	-190			['range', 'cent', 'phase', 'beg']	
233001	Att. — Athens: Agora — stoich. 29 — 301/0-295/4 a. — *Hesp. 13.1944.242,7 — *SEG 24.119	301/0-295/4 a.	-301	-294			['range']	

Feedback time

On your green sticky, write one thing you liked in this session.
On your red sticky, write one thing I could improve upon for next time. Be specific.