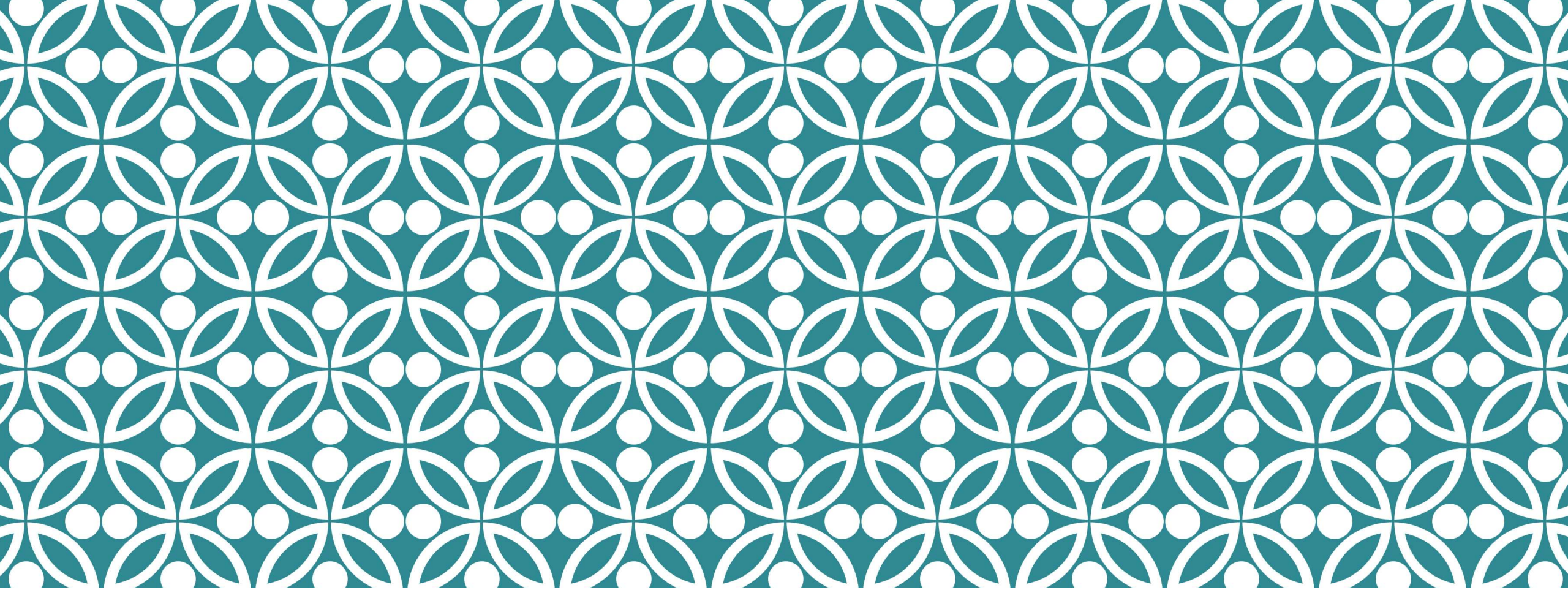


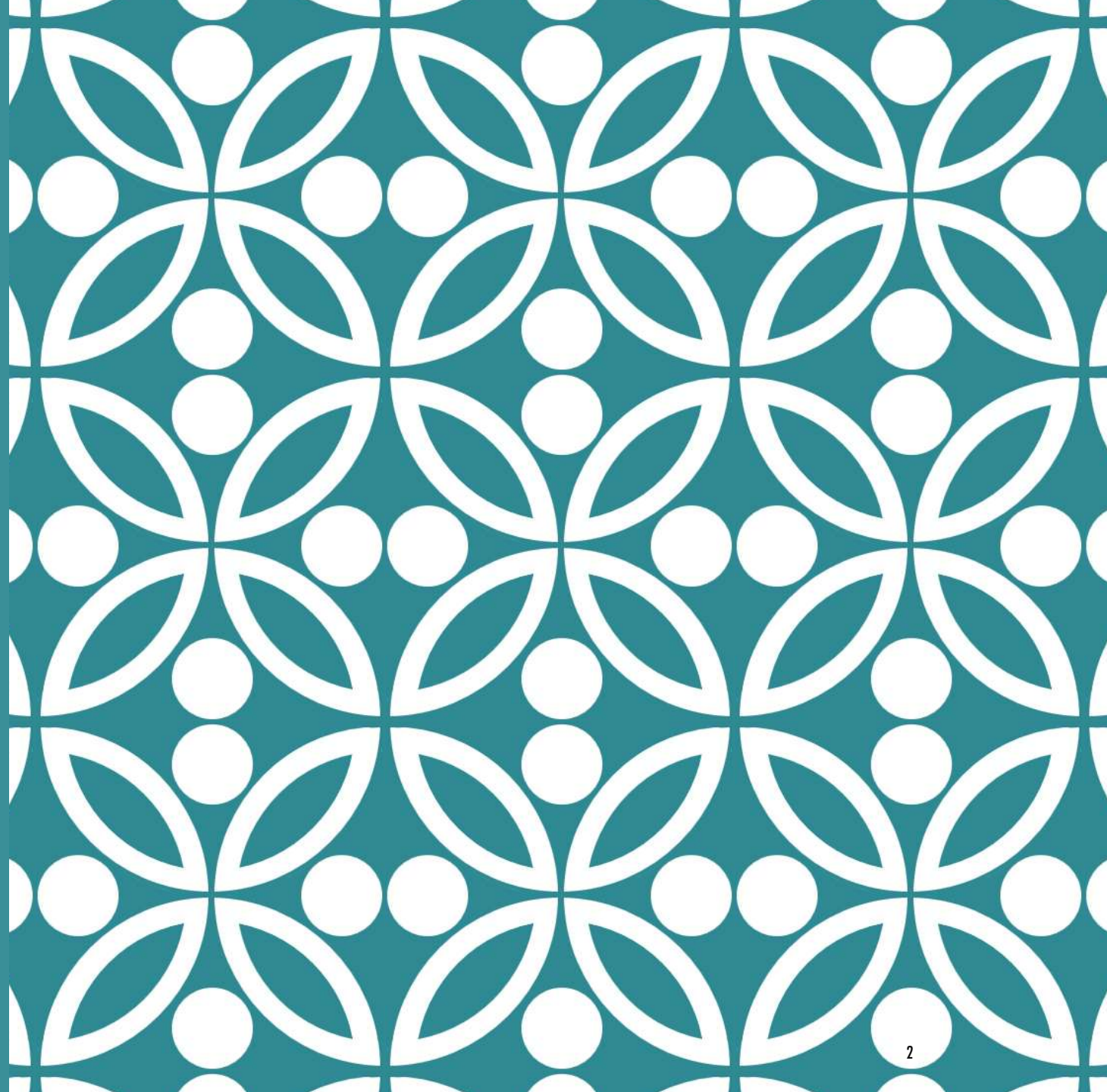
SESSION 4: TEXT AS DATA

DR. SOFIA GIL-CLAVEL

- The basics of handling text in R.
- Text Mining and Data Viz.
- Quick overview of advance topics.

1. THE BASICS OF HANDLING TEXT IN R

1.1 Basic text handling
1.2 Basic functions
1.3 Tidy text



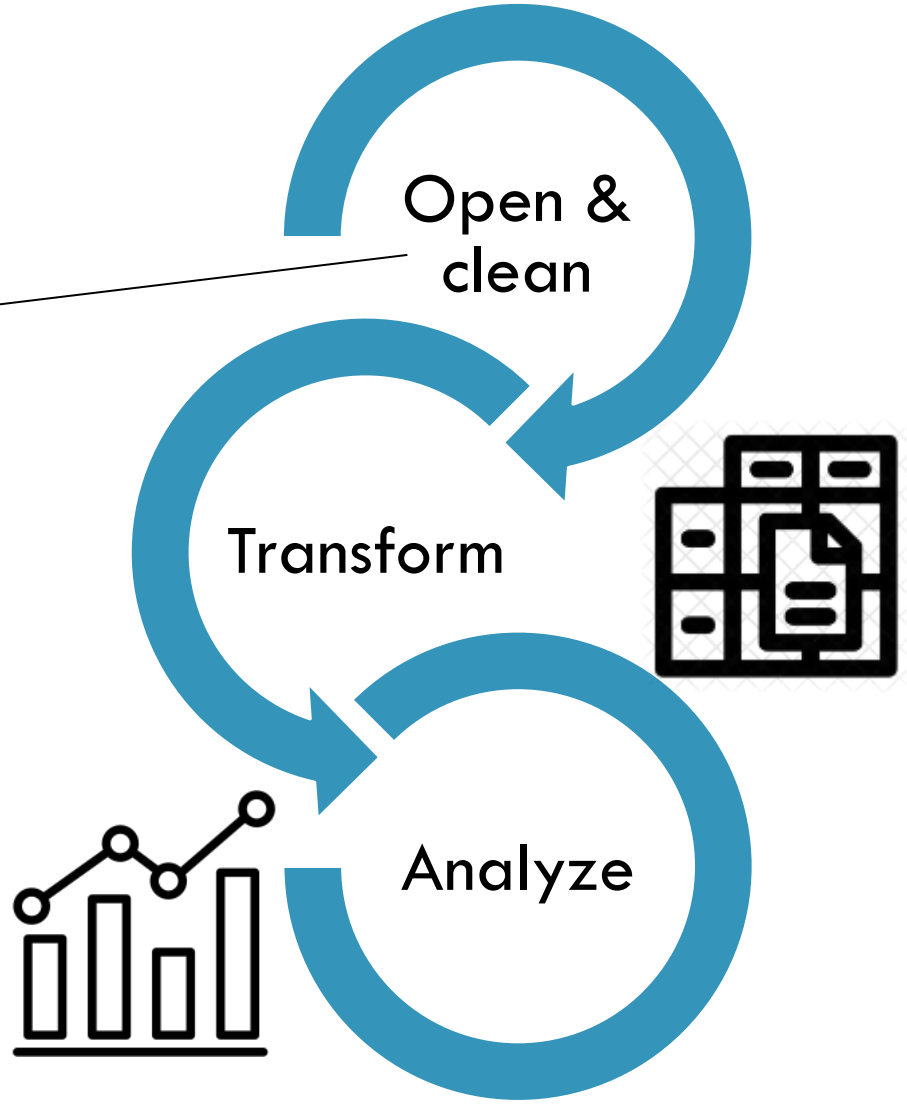
1.1 BASIC TEXT HANDLING

THE PIPELINE

Today, I went to the café in the city center to meet my friends. We sat at a table near the fenêtre, enjoying the view of the bustling streets. The ambiente inside was filled with the profumo of freshly baked croissants and the sonido of laughter. It felt so acogliente and charming, as if time had slowed down just for us. We chatted about everything from life to our latest aventures in the mountains. One of my friends, Mário, told us about his travesía through the desierto last year. We all agreed that viajar is one of the greatest joys in life.

The weather was perfect, with a brisa coming from the sea, and the sun shining brightly. We couldn't resist sitting outside on the terrace, where we could feel the fresh air on our skin. The streets around us were filled with people walking with sonrientes faces, enjoying the luz of the day. It was the kind of day where everything seemed to pause, and we embraced the moment fully.

As we sipped our cappuccini, we talked about our plans for the summer. Giulia suggested we take a trip to the montaña to escape the heat. Her idea sounded fantastico, and we all agreed to start planning it. We laughed about how we caminaríamos all day, lost in the beauty of nature. After a while, the conversation shifted to películas and funny moments from the past, and we found ourselves reminiscing about the most ridículo things that had happened over the years.

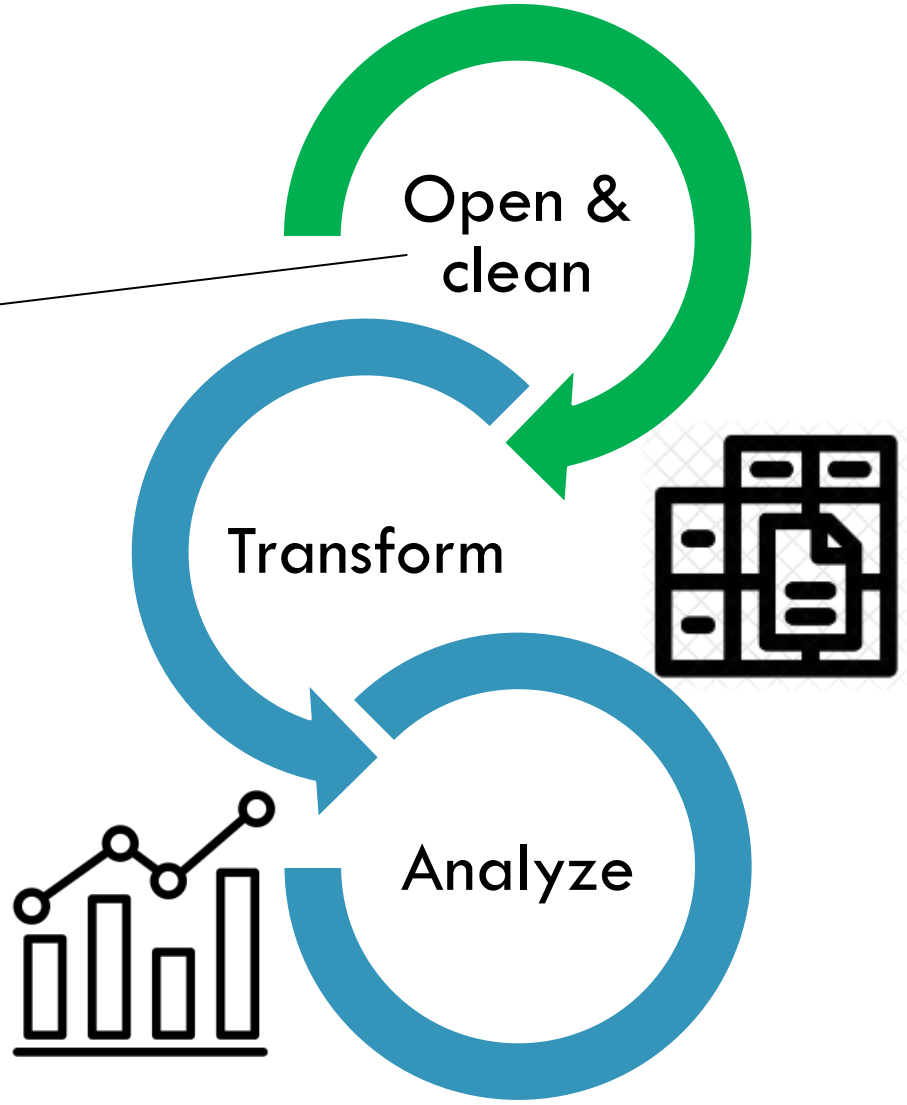


THE PIPELINE

Today, I went to the café in the city center to meet my friends. We sat at a table near the fenêtre, enjoying the view of the bustling streets. The ambiente inside was filled with the profumo of freshly baked croissants and the sonido of laughter. It felt so acogliente and charming, as if time had slowed down just for us. We chatted about everything from life to our latest aventures in the mountains. One of my friends, Mário, told us about his travesía through the desierto last year. We all agreed that viajar is one of the greatest joys in life.

The weather was perfect, with a brisa coming from the sea, and the sun shining brightly. We couldn't resist sitting outside on the terrace, where we could feel the fresh air on our skin. The streets around us were filled with people walking with sonrientes faces, enjoying the luz of the day. It was the kind of day where everything seemed to pause, and we embraced the moment fully.

As we sipped our cappuccini, we talked about our plans for the summer. Giulia suggested we take a trip to the montaña to escape the heat. Her idea sounded fantastico, and we all agreed to start planning it. We laughed about how we caminaríamos all day, lost in the beauty of nature. After a while, the conversation shifted to películas and funny moments from the past, and we found ourselves reminiscing about the most ridículo things that had happened over the years.

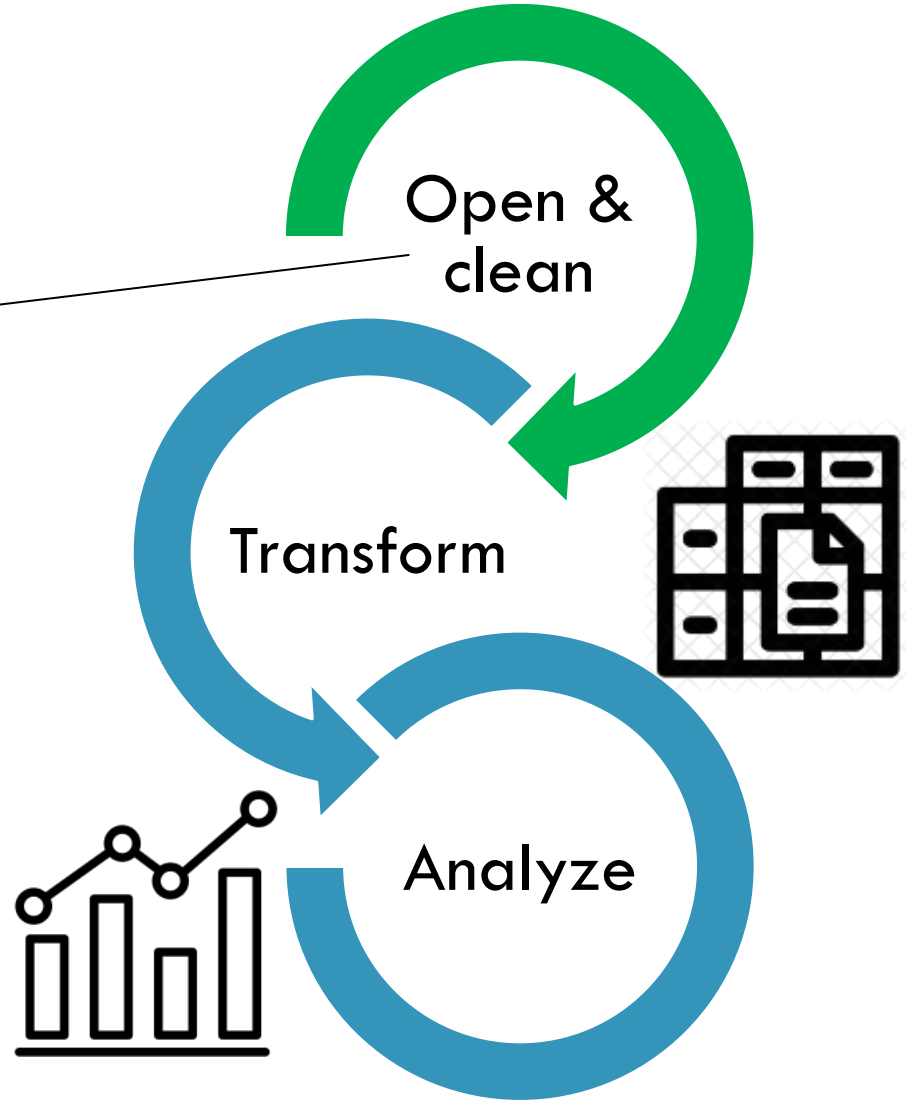


THE PIPELINE

Today, I went to the café in the city center to meet my friends. We sat at a table near the fenêtre, enjoying the view of the bustling streets. The ambiente inside was filled with the profumo of freshly baked croissants and the sonido of laughter. It felt so acogliente and charming, as if time had slowed down just for us. We chatted about everything from life to our latest aventures in the mountains. One of my friends, Mário, told us about his travesía through the desierto last year. We all agreed that viajar is one of the greatest joys in life.

The weather was perfect, with a brisa coming from the sea, and the sun shining brightly. We couldn't resist sitting outside on the terrace, where we could feel the fresh air on our skin. The streets around us were filled with people walking with sonrientes faces, enjoying the luz of the day. It was the kind of day where everything seemed to pause, and we embraced the moment fully.

As we sipped our cappuccini, we talked about our plans for the summer. Giulia suggested we take a trip to the montaña to escape the heat. Her idea sounded fantastico, and we all agreed to start planning it. We laughed about how we caminaríamos all day, lost in the beauty of nature. After a while, the conversation shifted to películas and funny moments from the past, and we found ourselves reminiscing about the most ridículo things that had happened over the years.



Let's open the text "Example_Text.txt"
What problems do you encounter?

Types of Character Data

El código ASCII

sigla en inglés de American Standard Code for Information Interchange
(Código Estadounidense Estándar para el Intercambio de Información)

Caracteres de control ASCII

DEC	HEX	Simbolo ASCII	
00	00h	NULL	(carácter nulo)
01	01h	SOH	(inicio encabezado)
02	02h	STX	(inicio texto)
03	03h	ETX	(fin de texto)
04	04h	EOT	(fin transmisión)
05	05h	ENQ	(enquiry)
06	06h	ACK	(acknowledgement)
07	07h	BEL	(timbre)
08	08h	BS	(retroceso)
09	09h	HT	(tab horizontal)
10	0Ah	LF	(salto de linea)
11	0Bh	VT	(tab vertical)
12	0Ch	FF	(form feed)
13	0Dh	CR	(retorno de carro)
14	0Eh	SO	(shift Out)
15	0Fh	SI	(shift In)
16	10h	DLE	(data link escape)
17	11h	DC1	(device control 1)
18	12h	DC2	(device control 2)
19	13h	DC3	(device control 3)
20	14h	DC4	(device control 4)
21	15h	NAK	(negative acknowle.)
22	16h	SYN	(synchronous idle)
23	17h	ETB	(end of trans. block)
24	18h	CAN	(cancel)
25	19h	EM	(end of medium)
26	1Ah	SUB	(substitute)
27	1Bh	ESC	(escape)
28	1Ch	FS	(file separator)
29	1Dh	GS	(group separator)
30	1Eh	RS	(record separator)
31	1Fh	US	(unit separator)
127	20h	DEL	(delete)

Caracteres ASCII imprimibles

DEC	HEX	Simbolo	DEC	HEX	Simbolo	DEC	HEX	Simbolo
32	20h	espacio	64	40h	@	96	60h	`
33	21h	!	65	41h	A	97	61h	a
34	22h	"	66	42h	B	98	62h	b
35	23h	#	67	43h	C	99	63h	c
36	24h	$	68	44h	D	100	64h	d
37	25h	%	69	45h	E	101	65h	e
38	26h	&	70	46h	F	102	66h	f
39	27h	'	71	47h	G	103	67h	g
40	28h	(	72	48h	H	104	68h	h
41	29h	)	73	49h	I	105	69h	i
42	2Ah	*	74	4Ah	J	106	6Ah	j
43	2Bh	+	75	4Bh	K	107	6Bh	k
44	2Ch	,	76	4Ch	L	108	6Ch	l
45	2Dh	-	77	4Dh	M	109	6Dh	m
46	2Eh	.	78	4Eh	N	110	6Eh	n
47	2Fh	/	79	4Fh	O	111	6Fh	o
48	30h	0	80	50h	P	112	70h	p
49	31h	1	81	51h	Q	113	71h	q
50	32h	2	82	52h	R	114	72h	r
51	33h	3	83	53h	S	115	73h	s
52	34h	4	84	54h	T	116	74h	t
53	35h	5	85	55h	U	117	75h	u
54	36h	6	86	56h	V	118	76h	v
55	37h	7	87	57h	W	119	77h	w
56	38h	8	88	58h	X	120	78h	x
57	39h	9	89	59h	Y	121	79h	y
58	3Ah	:	90	5Ah	Z	122	7Ah	z
59	3Bh	;	91	5Bh	[	123	7Bh	{
60	3Ch	<	92	5Ch	\	124	7Ch	|
61	3Dh	=	93	5Dh	]	125	7Dh	}
62	3Eh	>	94	5Eh	^	126	7Eh	~
63	3Fh	?	95	5Fh	_			

elCodigoASCII.com.a

ASCII extendido

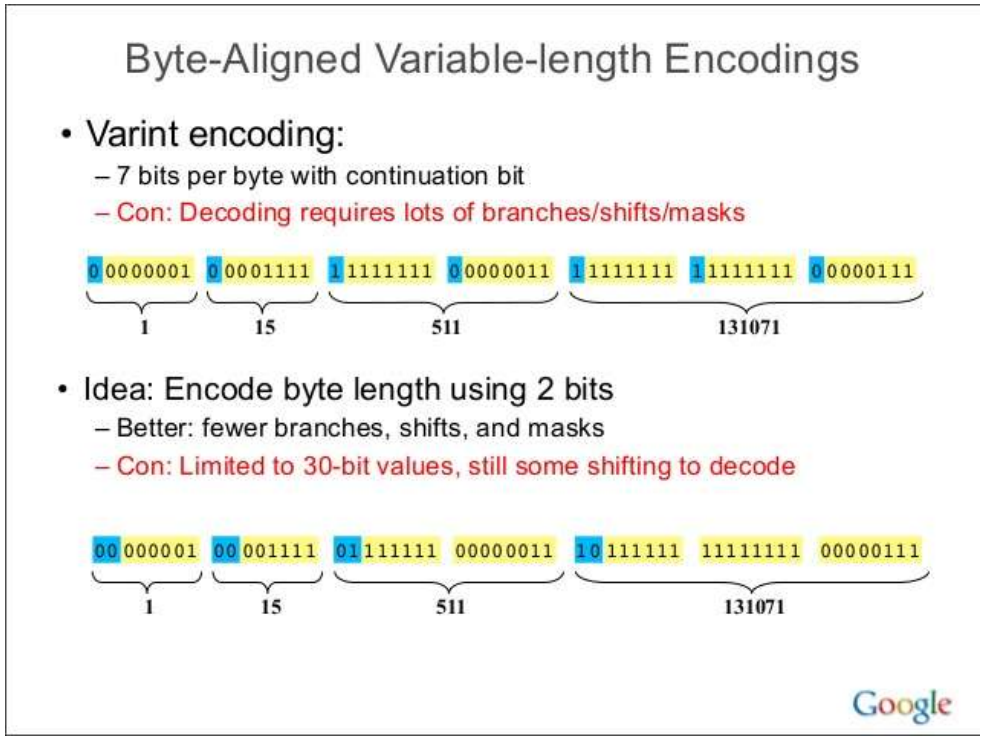
DEC	HEX	Simbolo	DEC	HEX	Simbolo	DEC	HEX	Simbolo	DEC	HEX	Simbolo
128	80h	Ç	160	A0h	á	192	C0h	└	224	E0h	Ó
129	81h	ü	161	A1h	í	193	C1h	┴	225	E1h	ß
130	82h	é	162	A2h	ó	194	C2h	┬	226	E2h	Ô
131	83h	â	163	A3h	ú	195	C3h	├	227	E3h	Ò
132	84h	ä	164	A4h	ñ	196	C4h	─	228	E4h	õ
133	85h	à	165	A5h	Ñ	197	C5h	┼	229	E5h	Õ
134	86h	å	166	A6h	ª	198	C6h	ã	230	E6h	µ
135	87h	ç	167	A7h	º	199	C7h	Ã	231	E7h	þ
136	88h	ê	168	A8h	¿	200	C8h	╚	232	E8h	Þ
137	89h	ë	169	A9h	®	201	C9h	╔	233	E9h	Ú
138	8Ah	è	170	AAh	¬	202	CAh	╩	234	EAh	Û
139	8Bh	ï	171	ABh	½	203	CBh	╦	235	EBh	Ù
140	8Ch	î	172	ACh	¼	204	CCh	╠	236	ECh	ý
141	8Dh	ì	173	ADh	¡	205	CDh	═	237	EDh	Ý
142	8Eh	Ä	174	AEh	«	206	CEh	╬	238	EEh	¯
143	8Fh	Å	175	AFh	»	207	CFh	¤	239	EFh	´
144	90h	É	176	B0h	░	208	D0h	ð	240	F0h	
145	91h	æ	177	B1h	▒	209	D1h	Ð	241	F1h	±
146	92h	Æ	178	B2h	▓	210	D2h	Ê	242	F2h	‗
147	93h	ô	179	B3h	│	211	D3h	Ë	243	F3h	¾
148	94h	ò	180	B4h	┤	212	D4h	È	244	F4h	¶
149	95h	ò	181	B5h	Á	213	D5h	ı	245	F5h	§
150	96h	û	182	B6h	Â	214	D6h	Í	246	F6h	÷
151	97h	ù	183	B7h	À	215	D7h	Î	247	F7h	¸
152	98h	ÿ	184	B8h	©	216	D8h	Ï	248	F8h	°
153	99h	Ö	185	B9h	╣	217	D9h	┘	249	F9h	¨
154	9Ah	Ü	186	BAh	║	218	DAh	┌	250	FAh	·
155	9Bh	ø	187	BBh	╗	219	DBh	█	251	FBh	¹
156	9Ch	£	188	BCh	╝	220	DCh	▄	252	FCh	³
157	9Dh	Ø	189	BDh	¢	221	DDh	¦	253	FDh	²
158	9Eh	×	190	BEh	¥	222	DEh	Ì	254	FEh	■
159	9Fh	ƒ	191	BFh	┐	223	DFh	▀	255	FFh	

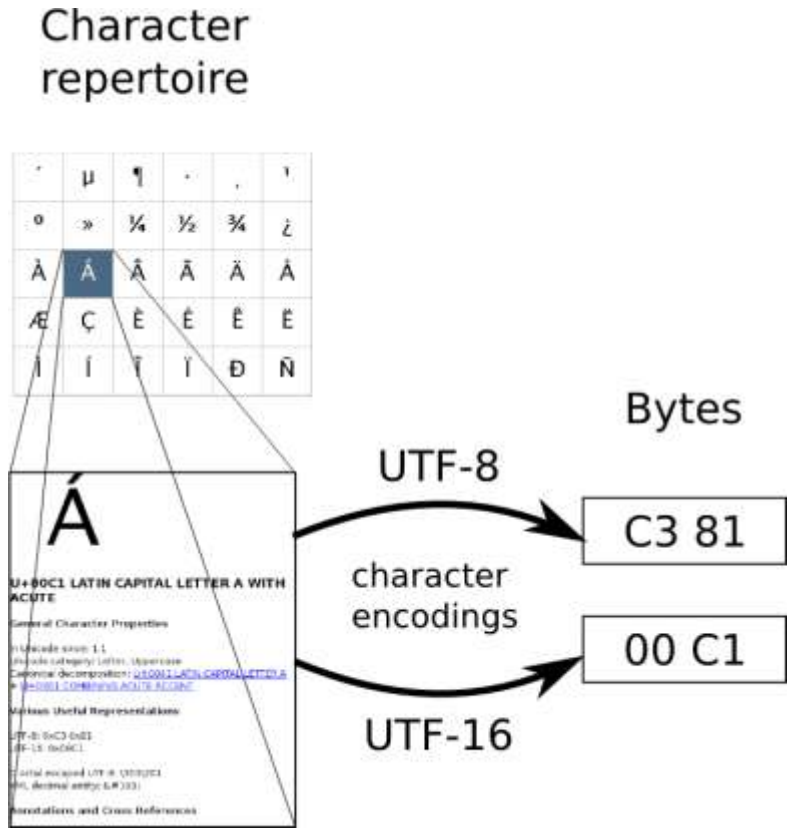
Characters that are not part of the standard ASCII and can be found in different encoding

Standard ASCII code

RELEVANCE OF USING THE RIGHT ENCODING

Strings of characters can be declared with different encodings, that is, depending on how the strings have been saved is the way in which the computer will process it.

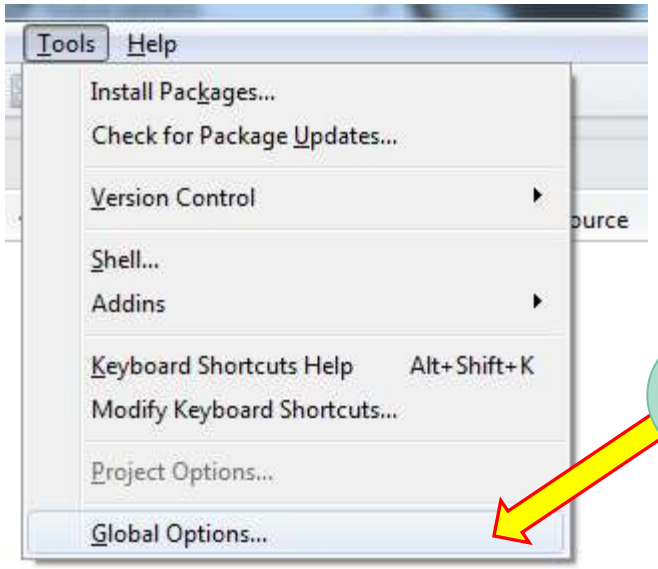




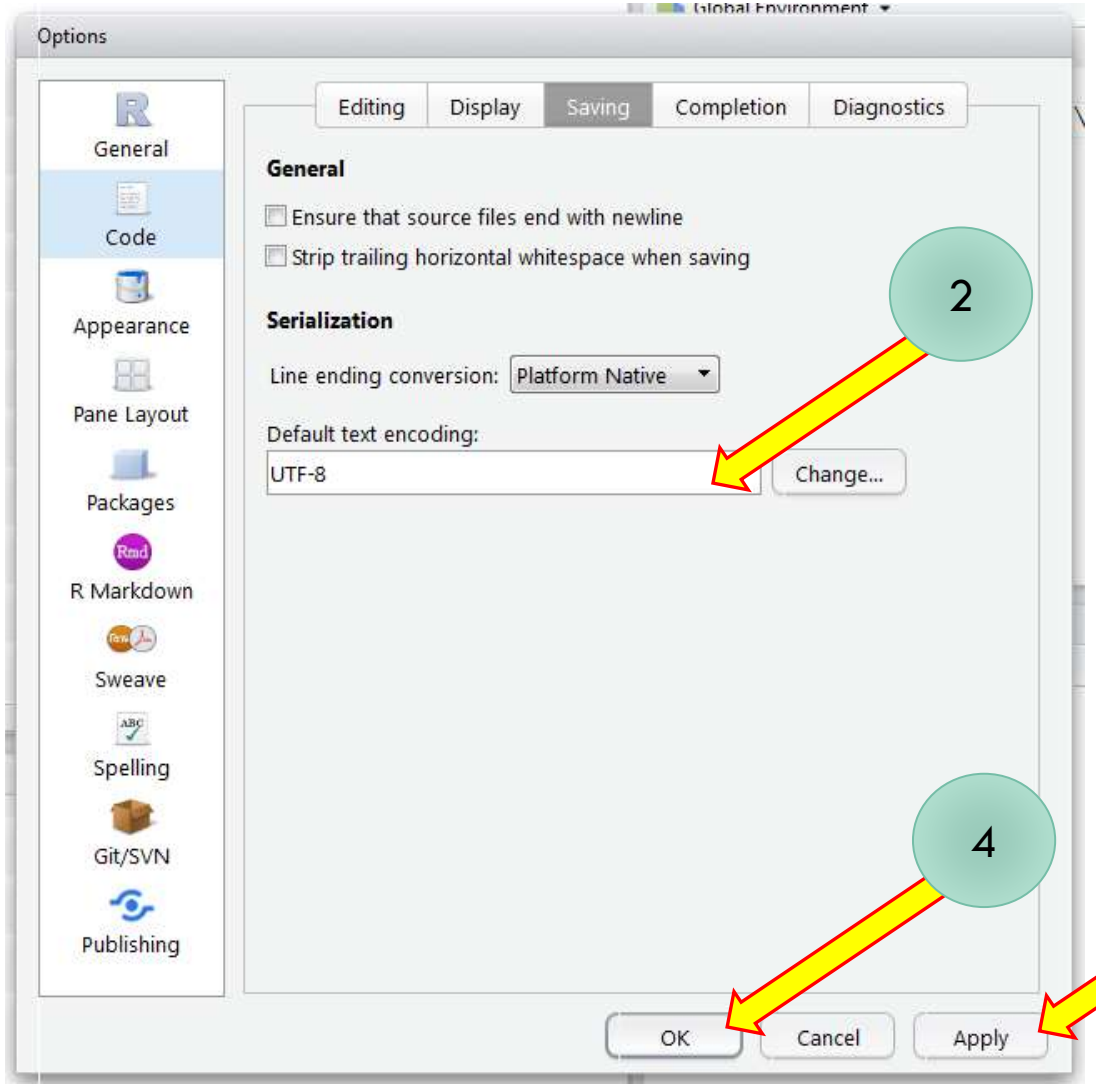
That is why it is always important that before opening a document we are sure of the type of encoding that was used to save it.

```
> x <- "fa\xE7ile"
> Encoding(x)
[1] "latin1"
> xx <- iconv(x, "latin1", "UTF-8")
> Encoding(xx)
[1] "UTF-8"
>
```

The most used type of encoding is Utf-8, which is why it is the type of encoding that will be handled here.


Tools
Help
Install Packages...
Check for Package Updates...
Version Control
Shell...
Addins
Keyboard Shortcuts Help
Alt+Shift+K
Modify Keyboard Shortcuts...
Project Options...
Global Options...

1


Options
General
Code
Appearance
Pane Layout
Packages
R Markdown
Sweave
Spelling
Git/SVN
Publishing
Editing
Display
Saving
Completion
Diagnostics
General
Ensure that source files end with newline
Strip trailing horizontal whitespace when saving
Serialization
Line ending conversion:
Platform Native
Default text encoding:
UTF-8
Change...
OK
Cancel
Apply

2

3

4

1.2 BASIC FUNCTIONS

STRINGR FOR BASIC TEXT HANDLING

R has its own functions to handle text, but it is not that convenient because the names are neither intuitive nor standardized. We will solve those problems by using the package "stringr".

Install and open the package in R.

STRINGR FUNCTIONS

stringr
`str_detect(string, pattern)`
`str_dup(string, times)`
`str_extract(string, pattern)`
`str_extract_all(string, pattern)`
`str_length(string)`
`str_locate(string, pattern)`
`str_locate_all(string, pattern)`
`str_match(string, pattern)`
`str_order(string)`
`str_replace(string, pattern, replacement)`
`str_replace_all(string, pattern, replacement)`

stringr
`str_sort(string)`
`str_split(string, pattern)`
`str_sub(string, start, end)`
`str_subset(string, pattern)`
`str_to_lower(string)`
`str_to_title(string)`
`str_to_upper(string)`
`str_trim(string)`
`str_which(string, pattern)`
`str_wrap(string)`

Source: https://cran.r-project.org/web/packages/stringr/vignettes/from-base.html

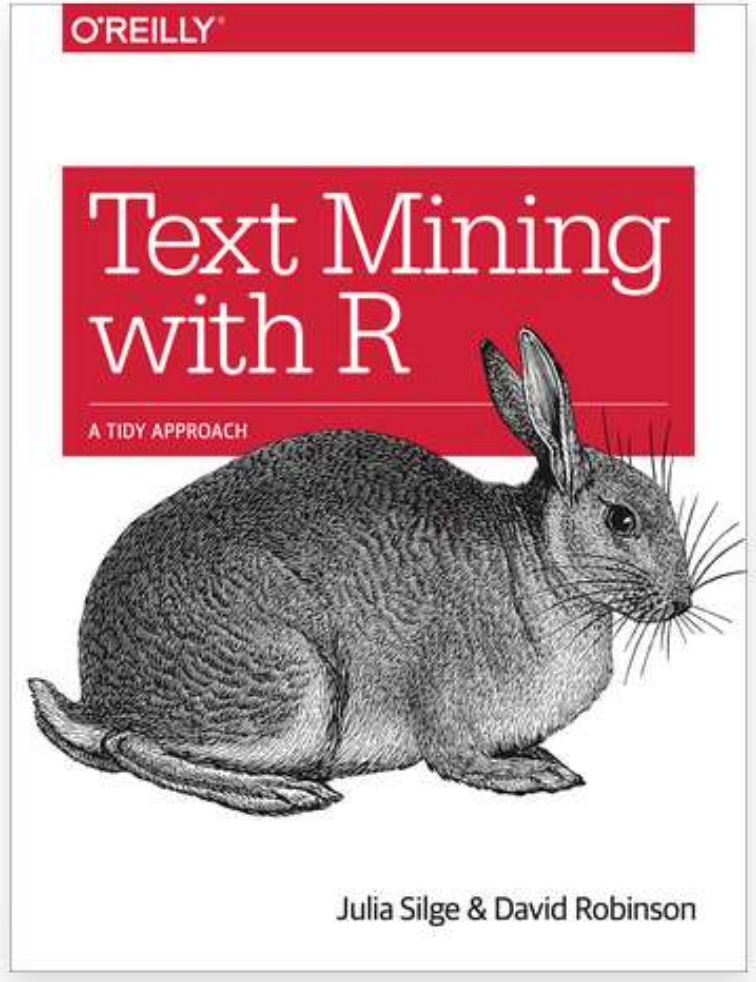
EXERCISE 1.2.A

Choose 3 stringr functions and apply them to the text document.

- What do they do?
- Think of a problem when you would use them.

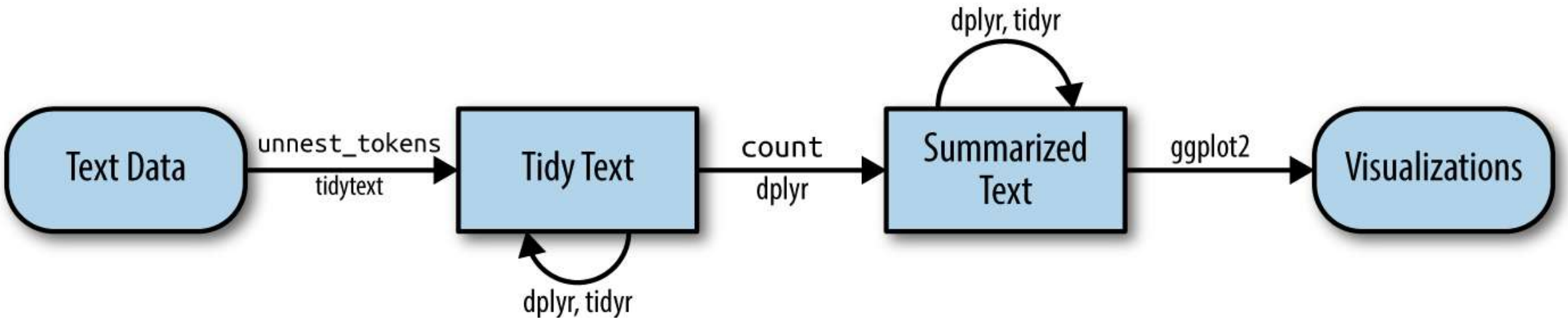
1.3 TIDY TEXT

❖ TIDYTEXT: HTTPS://WWW.TIDYTEXTMINING.COM/



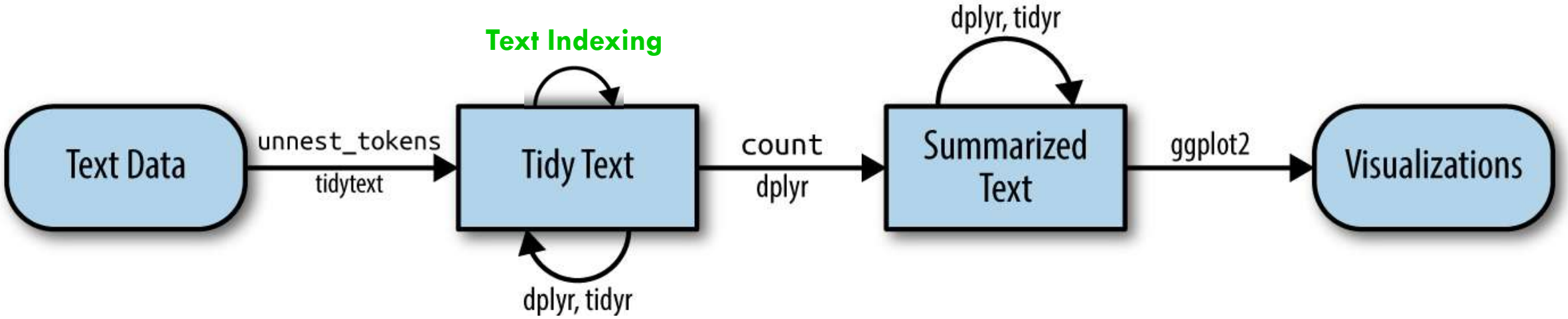
We developed the tidytext (Silge and Robinson 2016) R package because we were familiar with many methods for data wrangling and visualization, but couldn't easily apply these same methods to text. We found that using tidy data principles can make many text mining tasks easier, more effective, and consistent with tools already in wide use. Treating text as data frames of individual words allows us to manipulate, summarize, and visualize the characteristics of text easily and integrate natural language processing into effective workflows we were already using.

FLOWCHART OF A TYPICAL TEXT ANALYSIS USING TIDY DATA PRINCIPLES.



Source: https://www.tidytextmining.com/tidytext#tidytext

FLOWCHART OF A TYPICAL TEXT ANALYSIS USING TIDY DATA PRINCIPLES.



Source: https://www.tidytextmining.com/tidytext#tidytext

OPTIONAL: TEXT INDEXING

- **Stemming:** When we deal with text, often documents contain different versions of one base word, often called a *stem*. What if we aren't interested in the difference between "trees" and "tree" and we want to treat both together? That idea is at the heart of stemming, the process of identifying the base word (or stem) for a data set of words.
- **Stopwords:** The grammatical words which are called stop words and used only for grammatical functions such as "a" or "the" have no meaning, so they are usually excluded in the text preprocessing and/or text indexing.

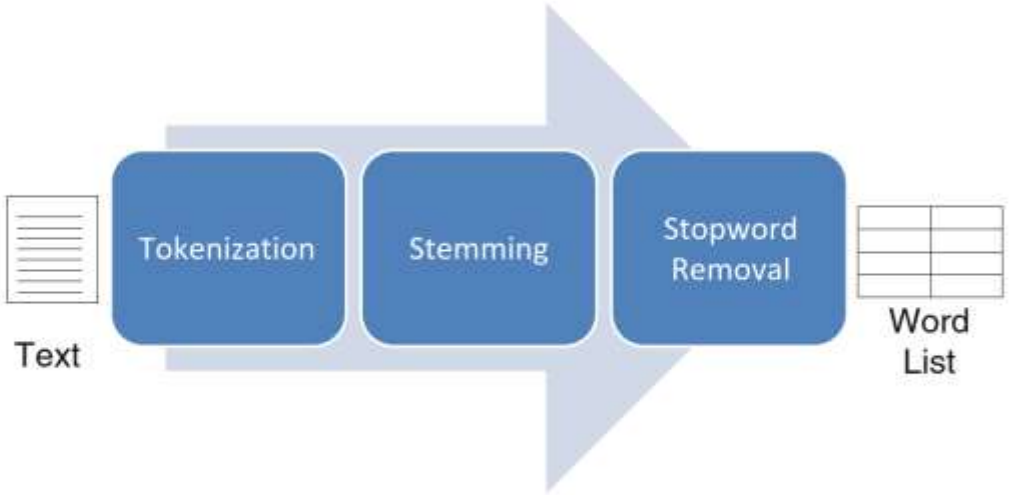


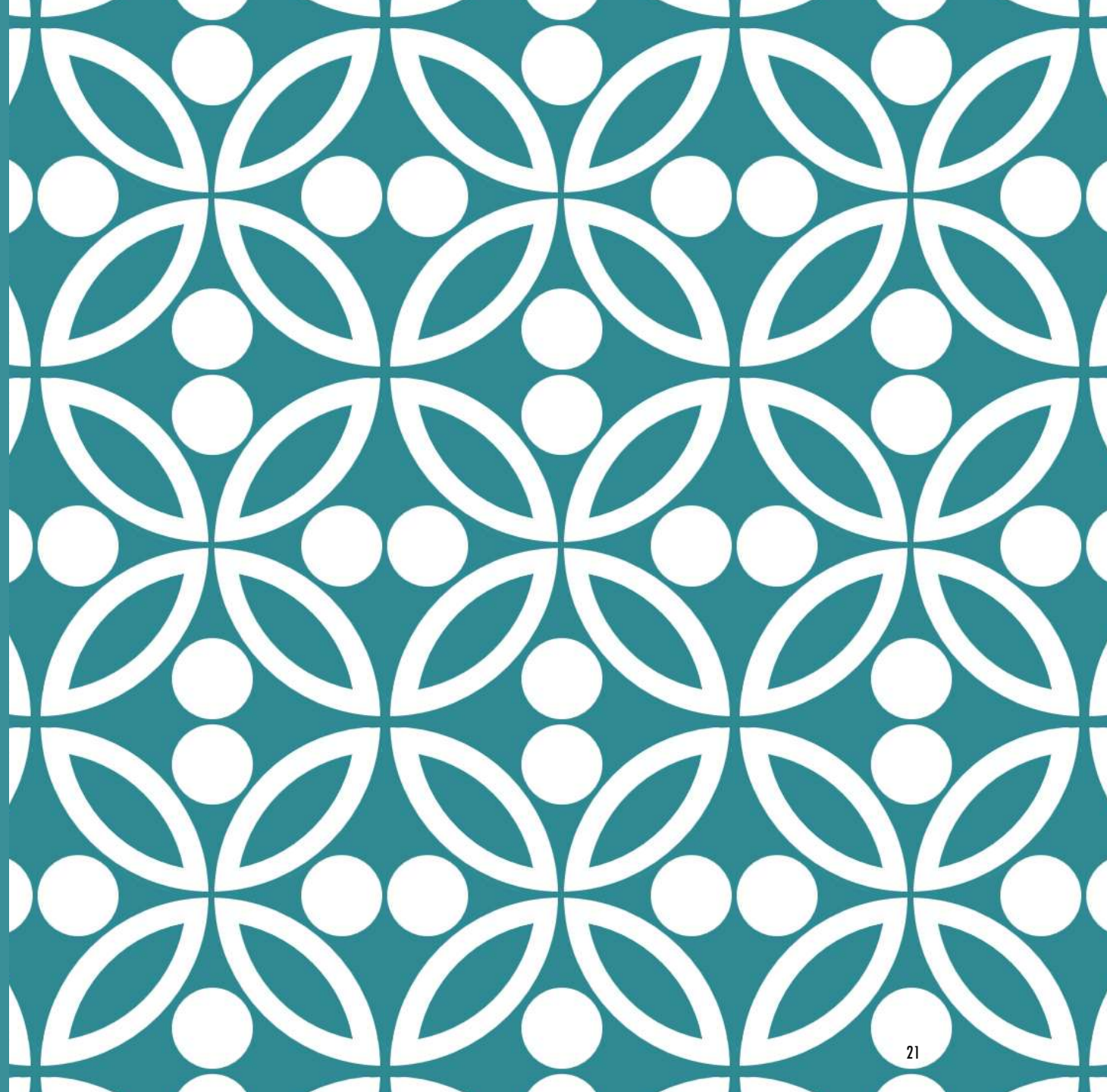
Fig. 2.1 The three steps of text indexing

Source: https://link.springer.com/book/10.1007/978-3-031-75976-5

10 MINS BREAK

2. TEXT MINING & DATAVIZ

1. What is text mining?
2. Word and document frequency
3. Relationships between words



2.1 WHAT IS TEXT MINING?

2.1 WHAT IS TEXT MINING?

Text mining, also known as text data mining, is the process of transforming unstructured text into a structured format to identify meaningful patterns and new insights.

JANE AUSTEN BOOKS

Source: https://www.amazon.nl/-/en/Jane-Austen/dp/1785995103

JANE AUSTEN BOOKS

What would be interesting to study using text mining?

Source: https://www.amazon.nl/-/en/Jane-Austen/dp/1785995103

2.2 WORD AND DOCUMENT FREQUENCY

Can we quantify what a document is about by looking at the words that make up the document?

2.2 WORD AND DOCUMENT FREQUENCY

A central question in text mining and natural language processing is **how to quantify what a document is about.**

Can we do this by looking at the words that make up the document?

Some measures of how important a word may be are:

- The ***term frequency*** **(tf),** how frequently a word occurs in a document.
- The term's **inverse document frequency (idf),** which decreases the weight for commonly used words and increases the weight for words that are not used very much in a collection of documents.
- This can be **combined** with term frequency to **calculate a term's tf-idf** (the two quantities multiplied together), the frequency of a term adjusted for how rarely it is used.

2.2 WORD AND DOCUMENT FREQUENCY

A central question in text mining and natural language processing is **how to quantify what a document is about.**

Can we do this by looking at the words that make up the document?

Some measures of how important a word may be are:

- The ***term frequency*** **(tf),** how frequently a word occurs in a document.
- The term's **inverse document frequency (idf),** which decreases the weight for commonly used words and increases the weight for words that are not used very much in a collection of documents.
- This can be **combined** with term frequency to **calculate a term's tf-idf** (the two quantities multiplied together), the frequency of a term adjusted for how rarely it is used.

2.2 TF-IDF

The idea of tf-idf is to find the important words for the content of each document by decreasing the weight for commonly used words and increasing the weight for words that are not used very much in a collection or corpus of documents, in this case, the group of Jane Austen's novels as a whole. Calculating tf-idf attempts to find the words that are important (i.e., common) in a text, but not too common.

We will use the function from tidytext:

```
bind_tf_idf(word, book, n)
```

2.3 RELATIONSHIPS BETWEEN WORDS

2.3 RELATIONSHIPS BETWEEN WORDS

Many interesting text analyses are based on the relationships between words, whether examining which words tend to follow others immediately, or that tend to co-occur within the same documents.

N-GRAMS

We can also tokenize into consecutive sequences of words, called n-grams. By seeing how often word X is followed by word Y, we can then build a model of the relationships between them.

N-GRAMS: TF-IDF

A bigram can also be treated as a term in a document in the same way that we treated individual words. For example, we can look at the tf-idf of bigrams across Austen novels. These tf-idf values can be visualized within each book, just as we did for words.

CORRELATING PAIRS OF WORDS

We may instead want to examine correlation among words, which indicates how often they appear together relative to how often they appear separately.

We'll focus on the phi coefficient, a common measure for binary correlation. The focus of the phi Φ coefficient is how much more likely it is that either both word X and Y appear, or neither do, than that one appears without the other.

	$y = 1$	$y = 0$	total
$x = 1$	n_{11}	n_{10}	$n_{1\bullet}$
$x = 0$	n_{01}	n_{00}	$n_{0\bullet}$
total	$n_{\bullet 1}$	$n_{\bullet 0}$	n

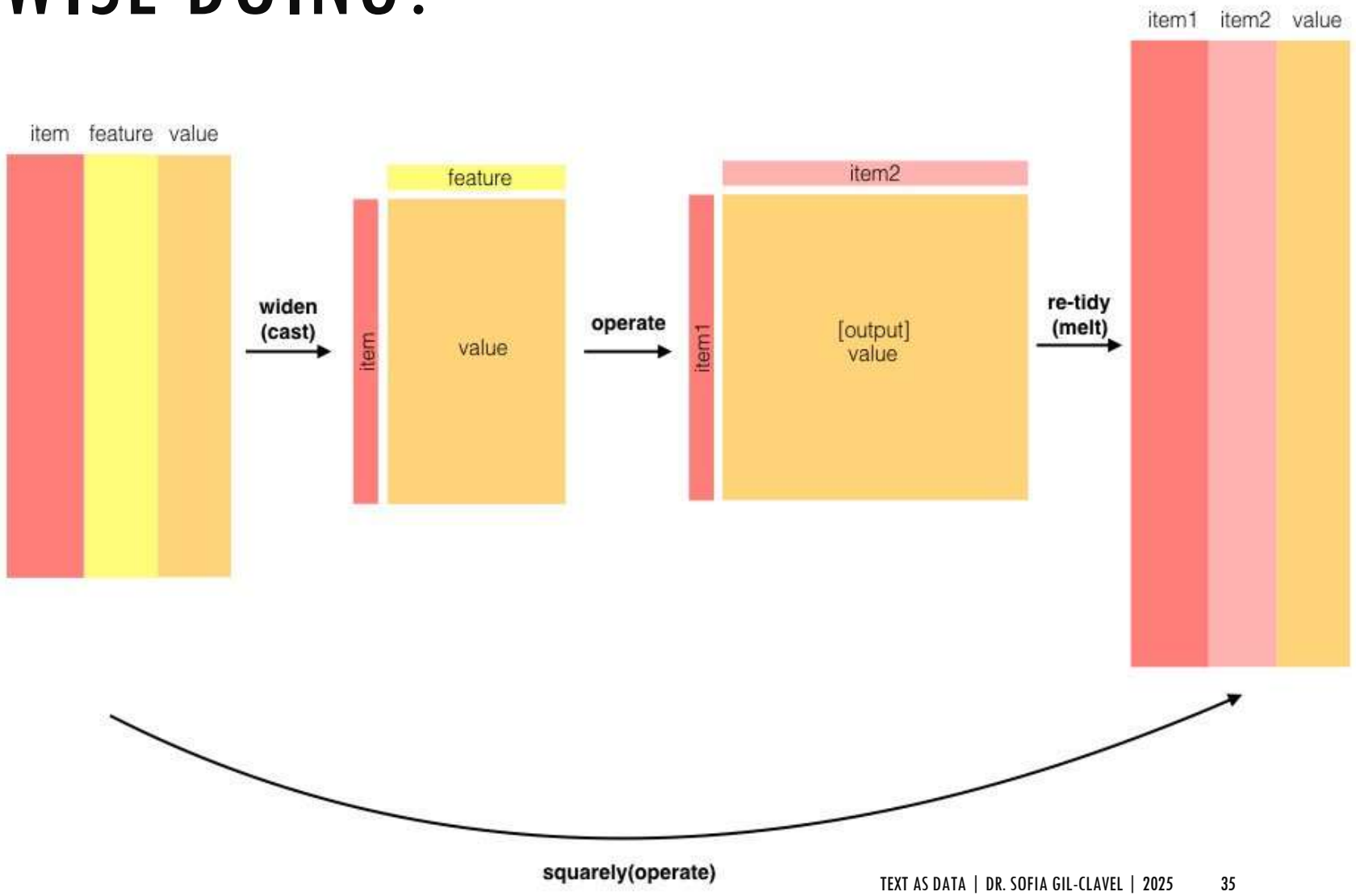
$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1\bullet}n_{0\bullet}n_{\bullet 0}n_{\bullet 1}}}$$

This is done with the package widyr and its function:

```
pairwise_cor(word, section, sort = TRUE)
```

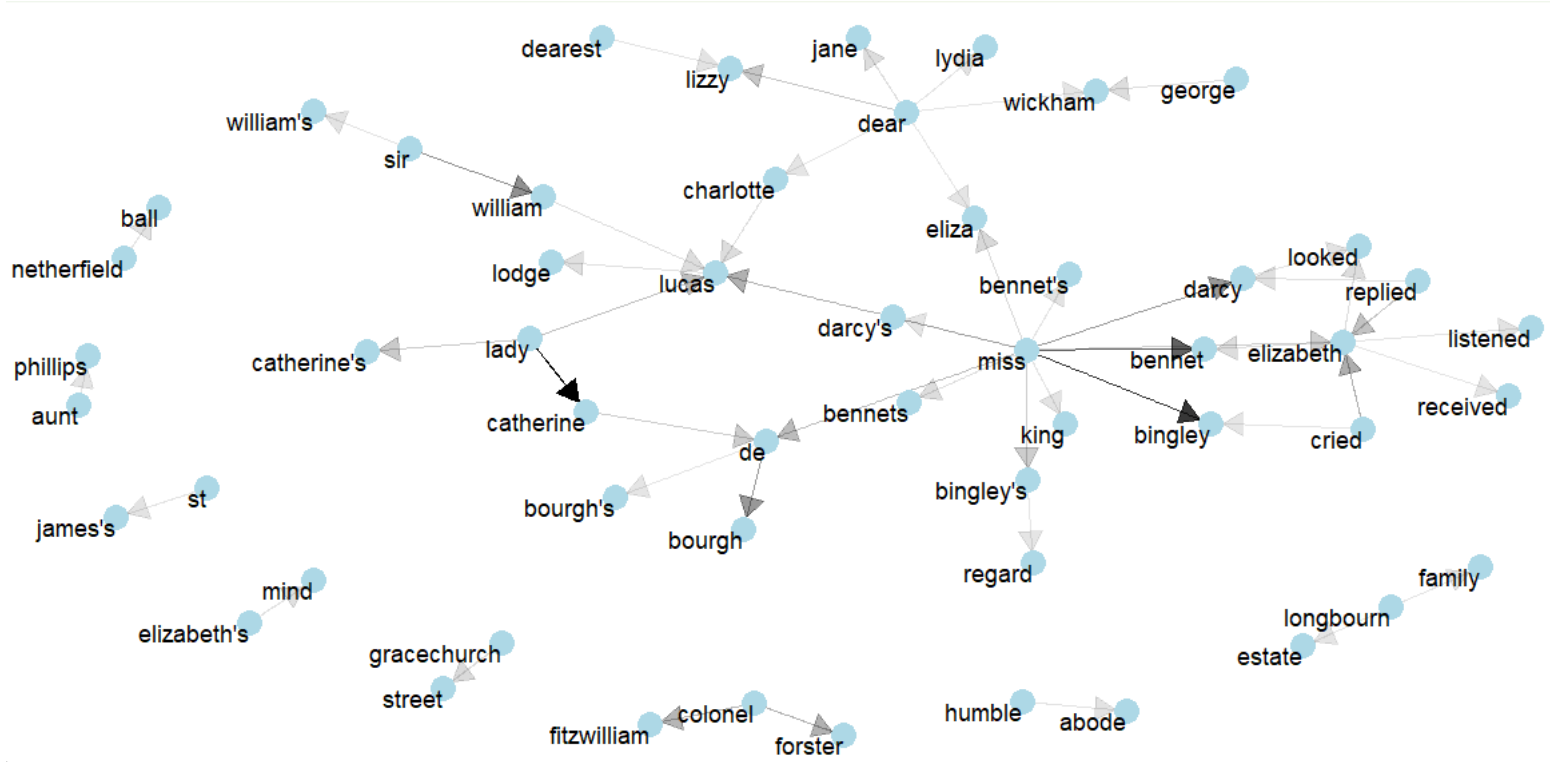
WHAT IS PAIRWISE DOING?

Counts the number of times each pair of *items* appear together within a group defined by "*feature*". In this case, it counts the number of times each pair of words appear together within a section. Later it uses this information to calculate the correlation.



NETWORK OF N-GRAMS RELATIONS

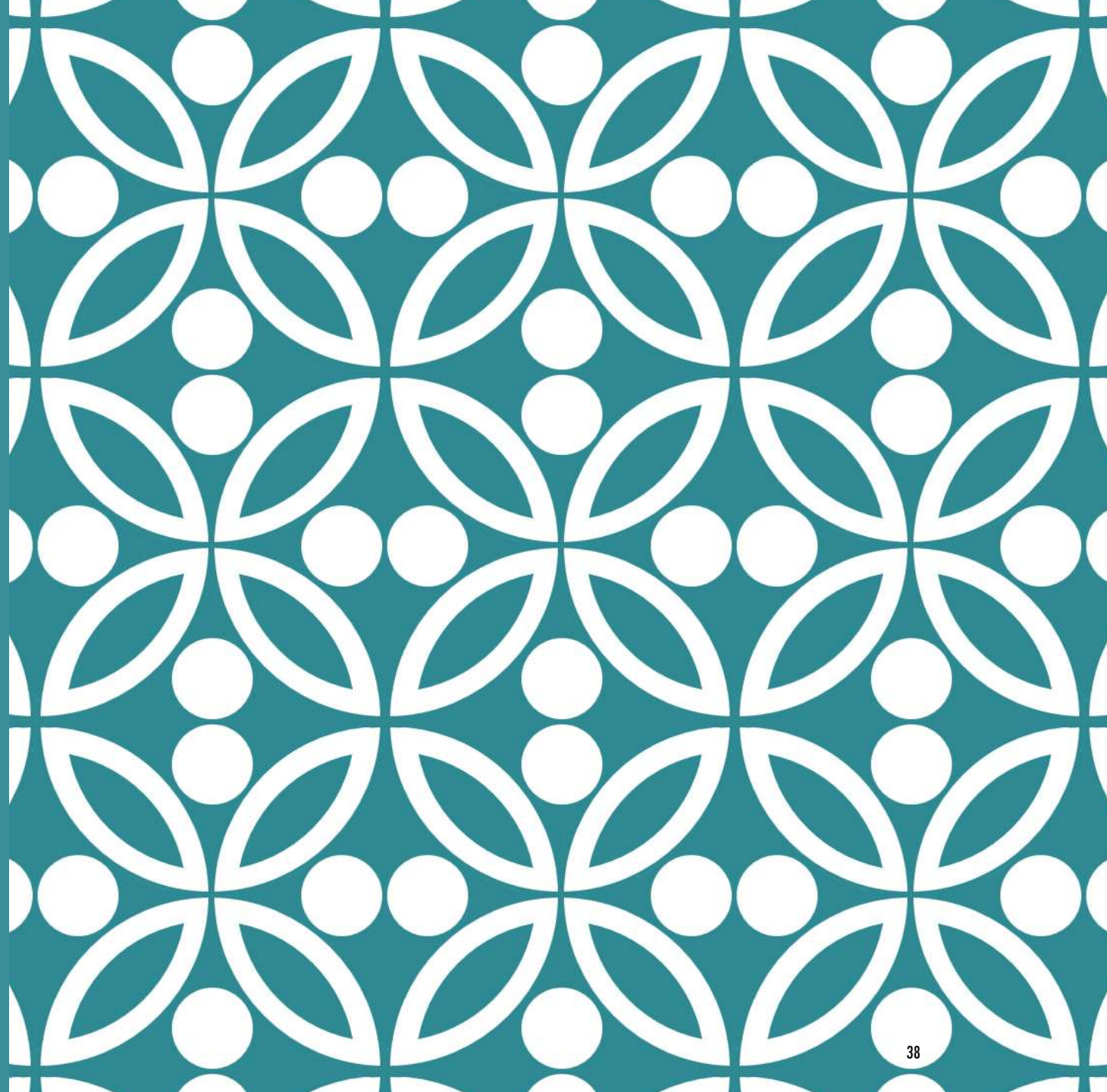
We can visualize the n-grams using the different metrics: counts, tf-idf, and correlations.



10 MINS BREAK

3. QUICK OVERVIEW OF ADVANCE TOPICS

1. Natural Language Processing (NLP)
2. Parts-Of-Speech (POS)
3. Visualizing Trees

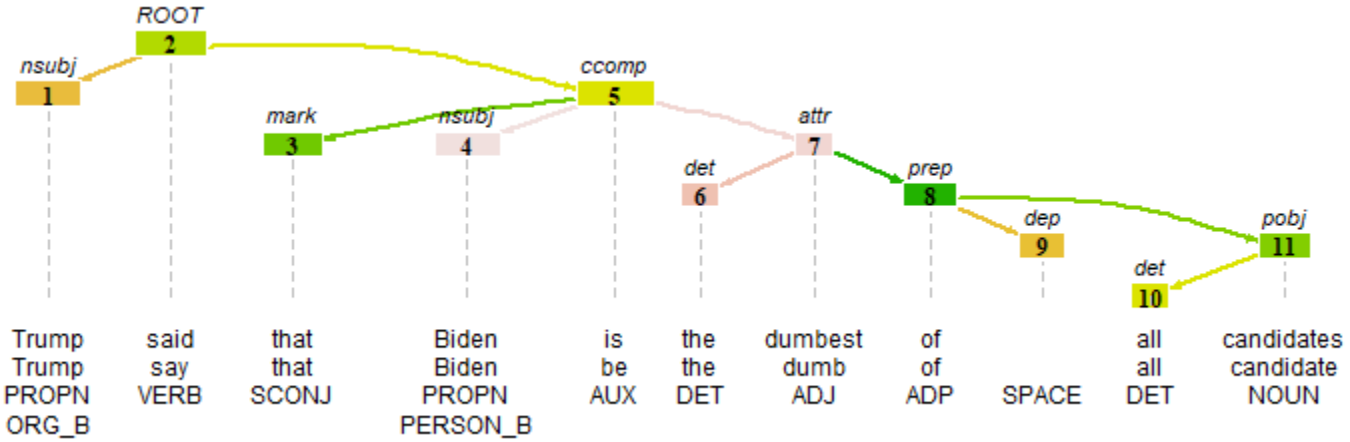


3.1 NATURAL LANGUAGE PROCESSING (NLP)

NLP enables computers and digital devices to recognize, understand, and generate text and speech by combining computational linguistics—the rule-based modeling of human language—together with statistical modeling, machine learning, and deep learning.

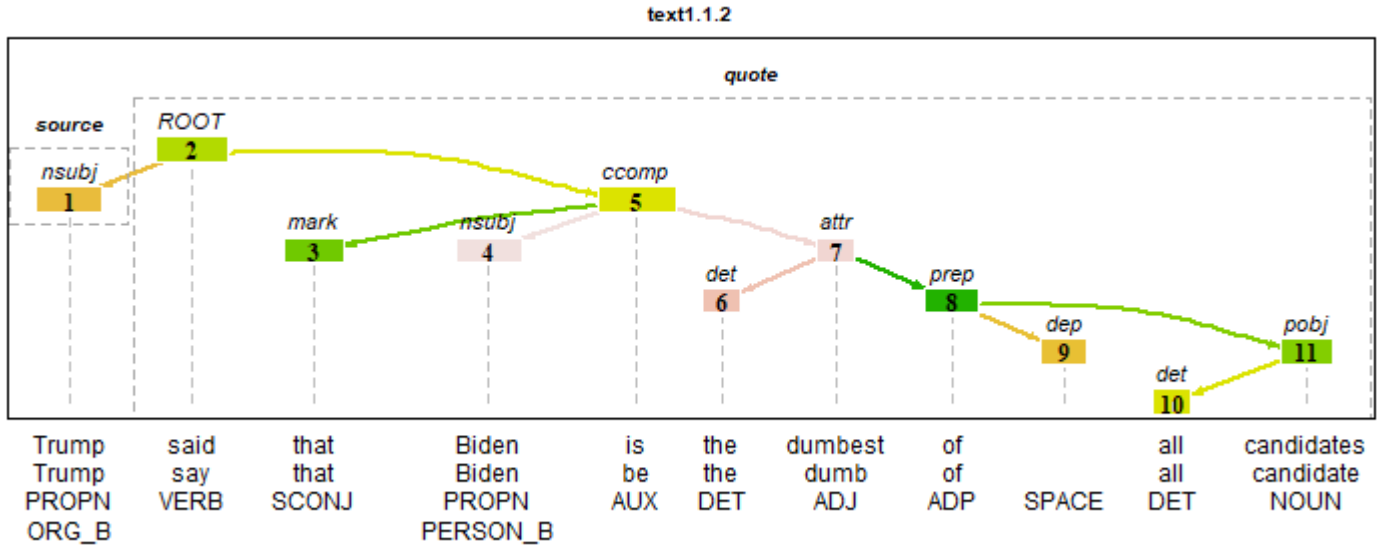
3.2 PART-OF-SPEECH (POS)

Part of speech: a class of words (as adjectives, adverbs, conjunctions, interjections, nouns, prepositions, pronouns, or verbs) identified according to the kinds of ideas they express and the way they work in a sentence.



Source: Welbers, Kasper, Wouter Van Atteveldt, and Jan Kleinnijenhuis. "Extracting Semantic Relations Using Syntax: An R Package for Querying and Reshaping Dependency Trees." *Computational Communication Research* 3, no. 2 (October 1, 2021): 1–16. https://doi.org/10.5117/CCR2021.2.003.WELB.

WHO DOES WHAT TO WHOM AND ACCORDING TO WHAT SOURCE



Source: Welbers, Kasper, Wouter Van Atteveldt, and Jan Kleinnijenhuis. "Extracting Semantic Relations Using Syntax: An R Package for Querying and Reshaping Dependency Trees." *Computational Communication Research* 3, no. 2 (October 1, 2021): 1–16. https://doi.org/10.5117/CCR2021.2.003.WELB.

[hybrid]R-Workshop: Machine Learning in R

When: April 25th between 9:30-12:30hrs

Where: Where: Online and onsite. More info will be sent via email.

To register: Register here: https://forms.office.com/e/xBY9gP3upc

Register here:

THANKS!

Dr. Sofia Gil-Clavel

https://sofiag1l.github.io/