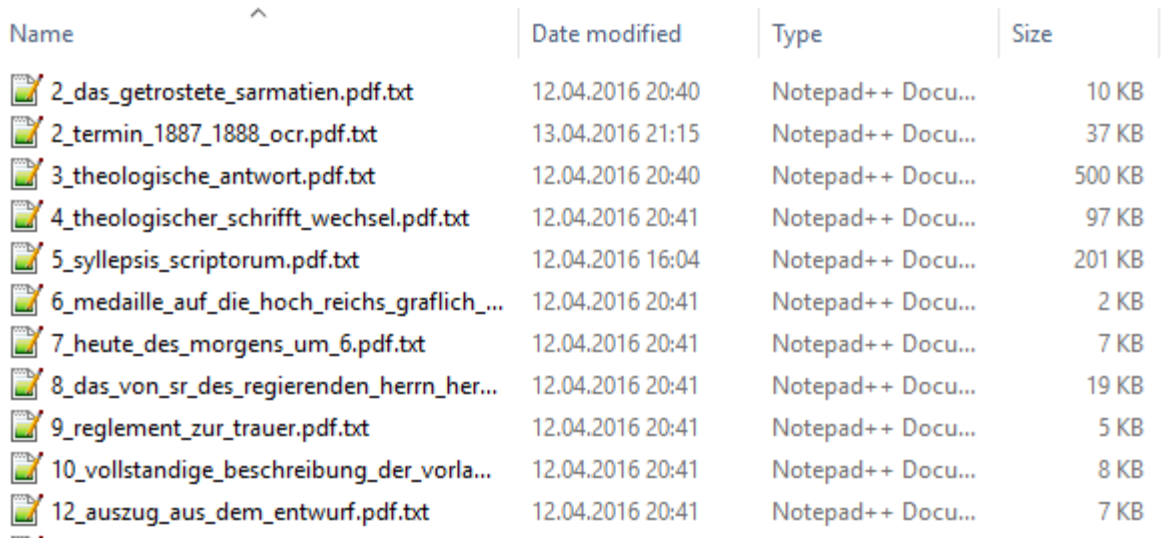
Workshop on digital humanities

SIMPLE TEXT ANALYSES, 05.09.2017

PEETER TINITS

Texts on computers and workflow

Simple texts can be read in R



Name	Date modified	Type	Size
2_das_getrostete_sarmatien.pdf.txt	12.04.2016 20:40	Notepad++ Docu...	10 KB
2_termin_1887_1888_ocr.pdf.txt	13.04.2016 21:15	Notepad++ Docu...	37 KB
3_theologische_antwort.pdf.txt	12.04.2016 20:40	Notepad++ Docu...	500 KB
4_theologischer_schrifft_wechsel.pdf.txt	12.04.2016 20:41	Notepad++ Docu...	97 KB
5_syllepsis_scriptorum.pdf.txt	12.04.2016 16:04	Notepad++ Docu...	201 KB
6_medaille_auf_die_hoch_reichs_graflich_...	12.04.2016 20:41	Notepad++ Docu...	2 KB
7_heute_des_morgens_um_6.pdf.txt	12.04.2016 20:41	Notepad++ Docu...	7 KB
8_das_von_sr_des_regierenden_herrn_her...	12.04.2016 20:41	Notepad++ Docu...	19 KB
9_reglement_zur_trauer.pdf.txt	12.04.2016 20:41	Notepad++ Docu...	5 KB
10_vollstandige_beschreibung_der_vorla...	12.04.2016 20:41	Notepad++ Docu...	8 KB
12_auszug_aus_dem_entwurf.pdf.txt	12.04.2016 20:41	Notepad++ Docu...	7 KB

In this workshop we will use the predownloaded texts in the library.

Get the data

Main page:

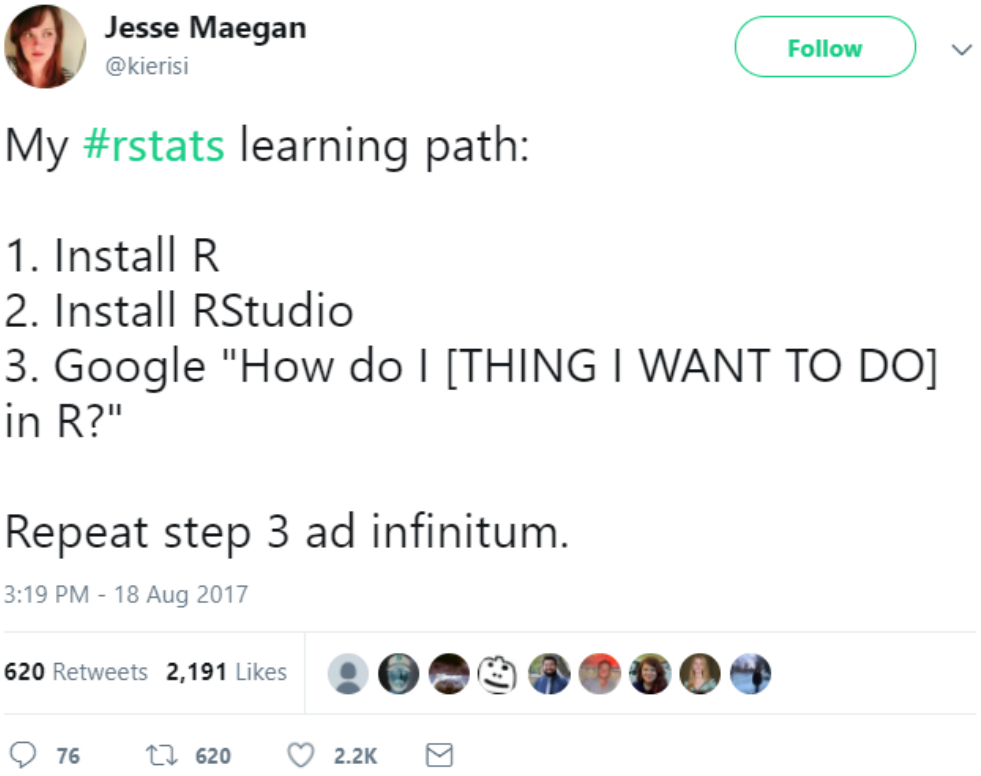
https://github.com/peeter-t2/DH-workshop-BAIE17

1. Navigate to the folder where you want to keep the files, e.g. the downloads folder.
2. Run git
3. type "git clone https://github.com/peeter-t2/DH-workshop-BAIE17.git"

Alternatively:

1. Click on the download .zip link in the top right green button, or follow this link: https://github.com/peeter-t2/DH-workshop-BAIE17/archive/master.zip
2. Unpack the files where you want them.

How to use R.



Jesse Maegan
@kierisi

Follow

My #rstats learning path:

1. Install R
2. Install RStudio
3. Google "How do I [THING I WANT TO DO] in R?"

Repeat step 3 ad infinitum.

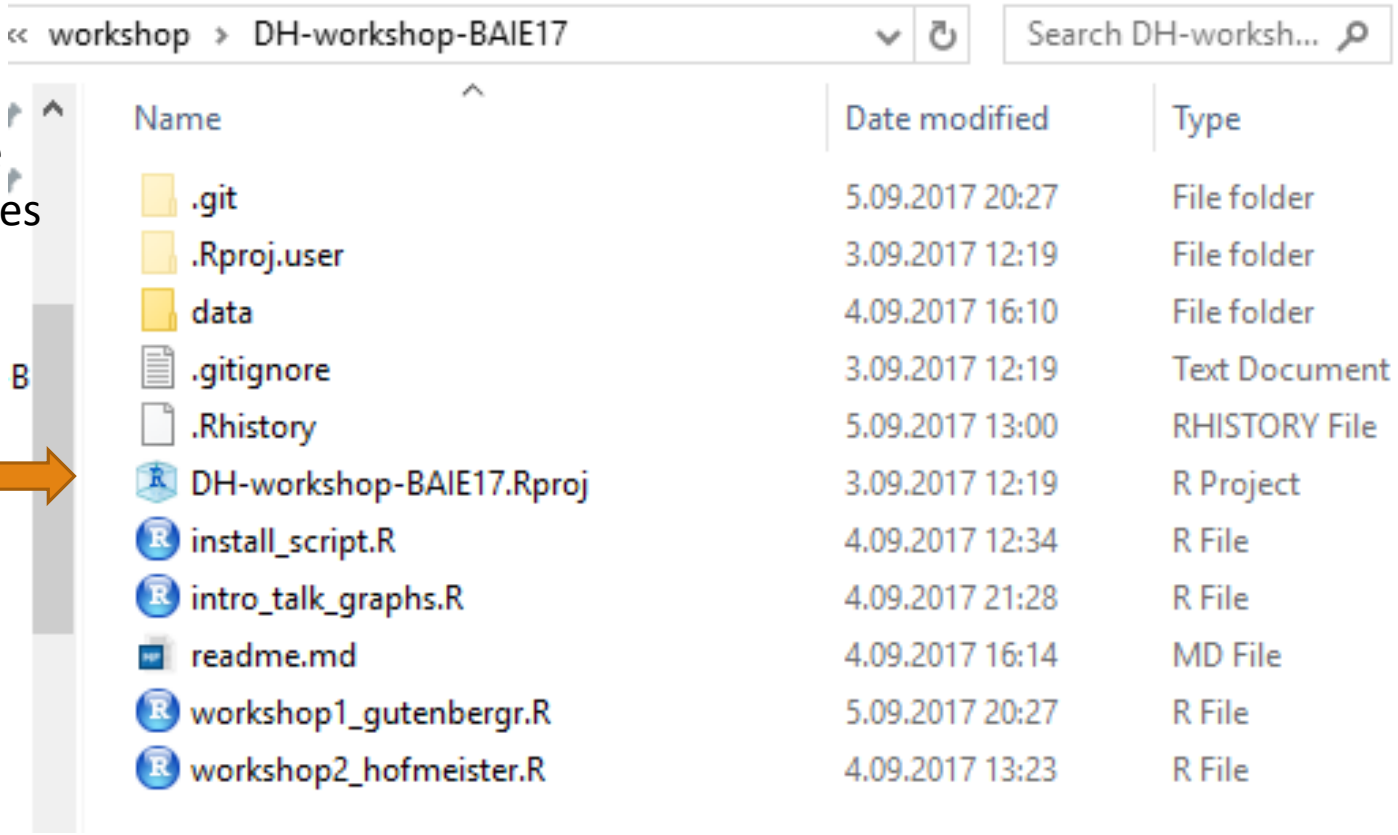
3:19 PM - 18 Aug 2017

620 Retweets 2,191 Likes

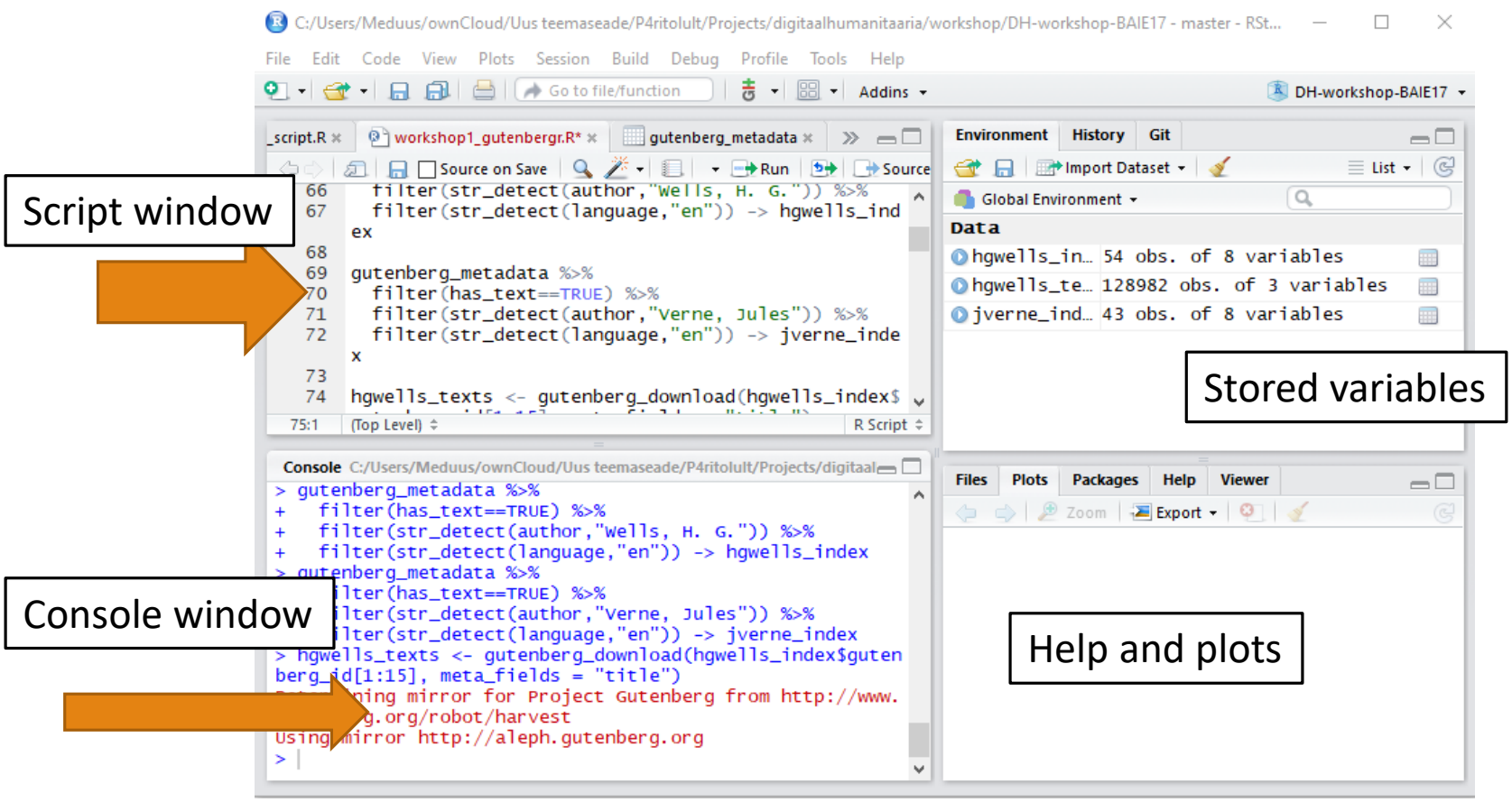
76 620 2.2K

To get started

To get started, run the Rproj file. This simplifies things in Rstudio, sets the working directory and remembers your actions.



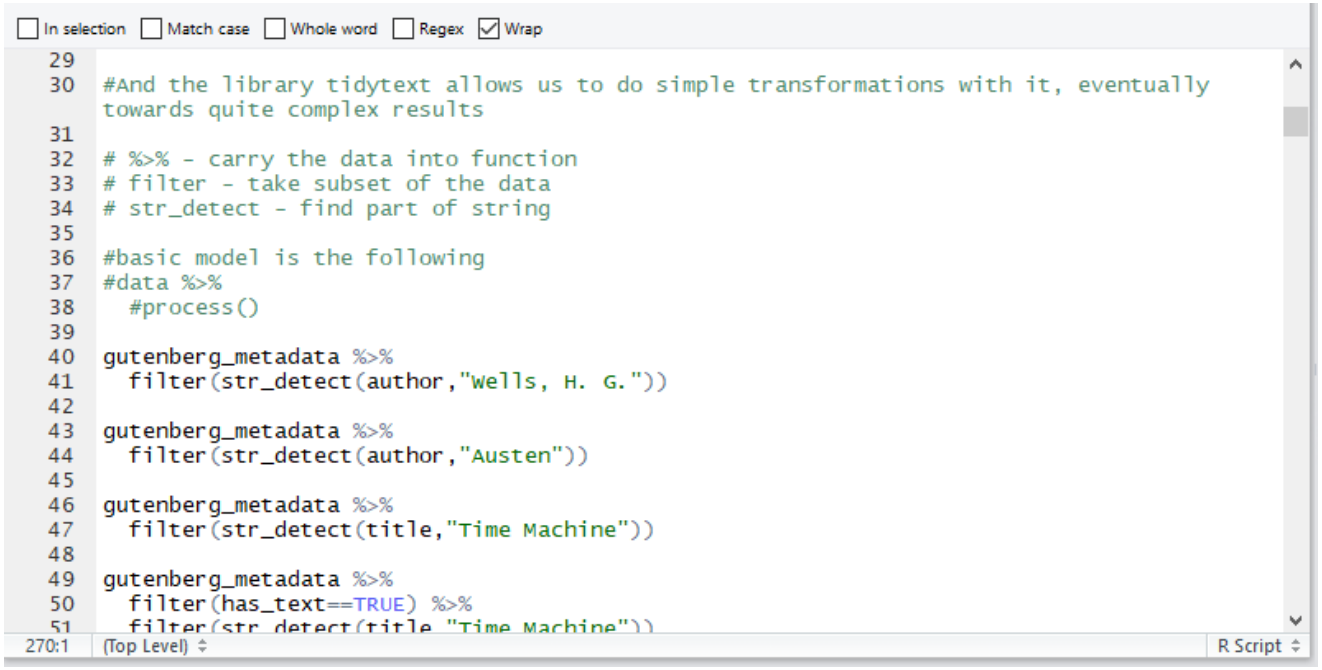
RStudio view



Script files

Green = comments, (and text strings – e.g. „Wells, H. G.“)

black,blue, etc = code

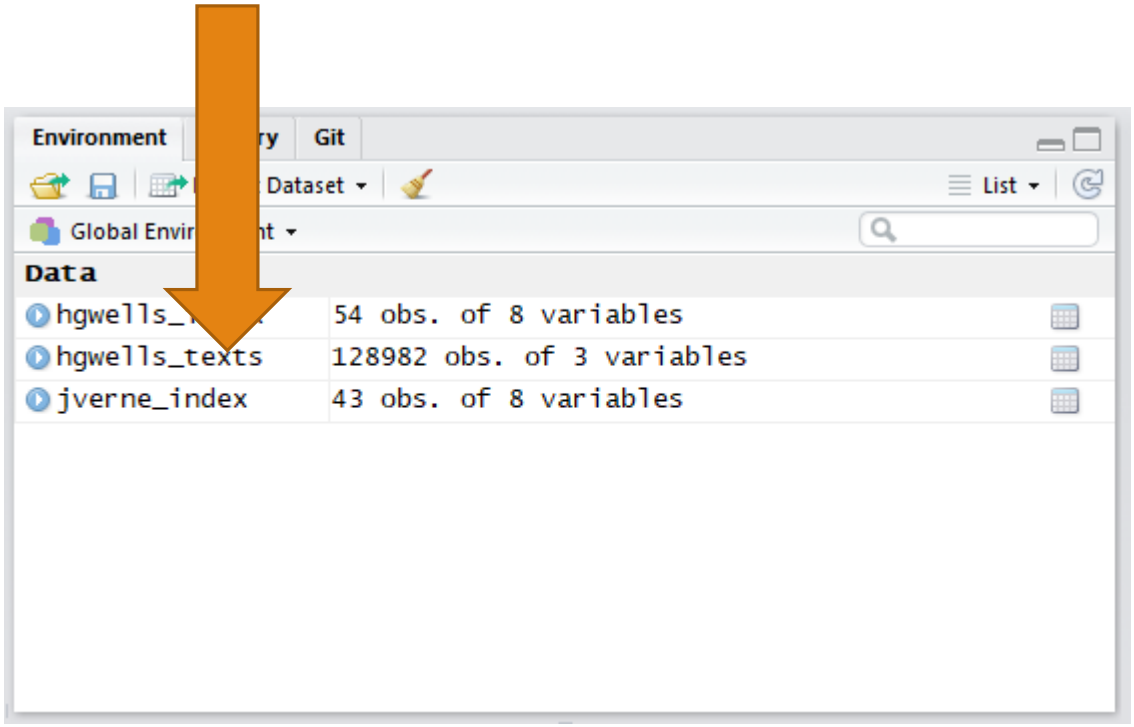


To run code

Pick a line and click run

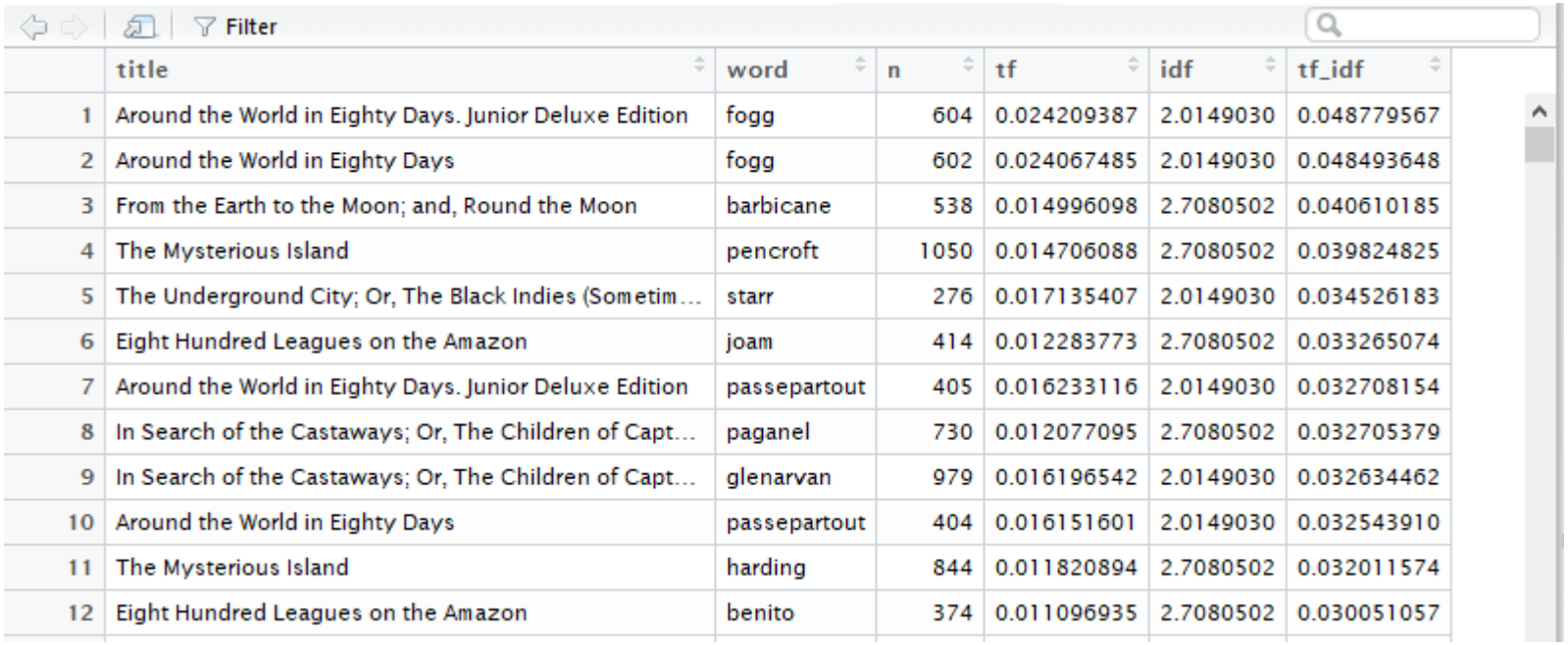
To look at the data

To view a dataframe



How data looks like

Just a table really ☺



Filter

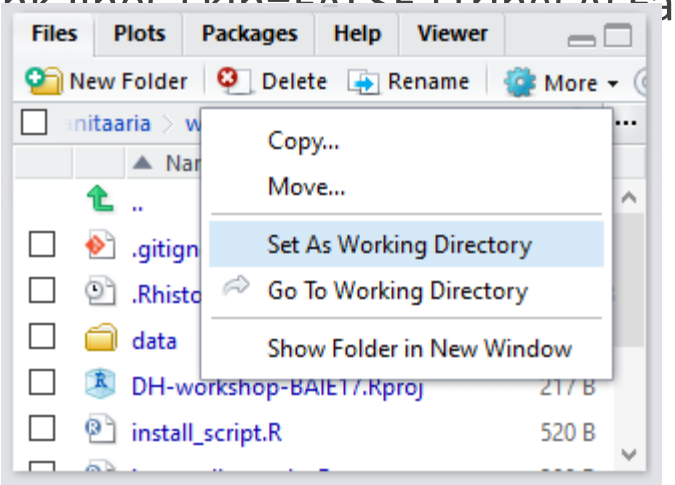
	title	word	n	tf	idf	tf_idf
1	Around the World in Eighty Days. Junior Deluxe Edition	fogg	604	0.024209387	2.0149030	0.048779567
2	Around the World in Eighty Days	fogg	602	0.024067485	2.0149030	0.048493648
3	From the Earth to the Moon; and, Round the Moon	barbicane	538	0.014996098	2.7080502	0.040610185
4	The Mysterious Island	pencroft	1050	0.014706088	2.7080502	0.039824825
5	The Underground City; Or, The Black Indies (Sometim…	starr	276	0.017135407	2.0149030	0.034526183
6	Eight Hundred Leagues on the Amazon	joam	414	0.012283773	2.7080502	0.033265074
7	Around the World in Eighty Days. Junior Deluxe Edition	passepartout	405	0.016233116	2.0149030	0.032708154
8	In Search of the Castaways; Or, The Children of Capt…	paganel	730	0.012077095	2.7080502	0.032705379
9	In Search of the Castaways; Or, The Children of Capt…	glenarvan	979	0.016196542	2.0149030	0.032634462
10	Around the World in Eighty Days	passepartout	404	0.016151601	2.0149030	0.032543910
11	The Mysterious Island	harding	844	0.011820894	2.7080502	0.032011574
12	Eight Hundred Leagues on the Amazon	benito	374	0.011096935	2.7080502	0.030051057

To read files

You need a right working directory (RProject does this automatically)

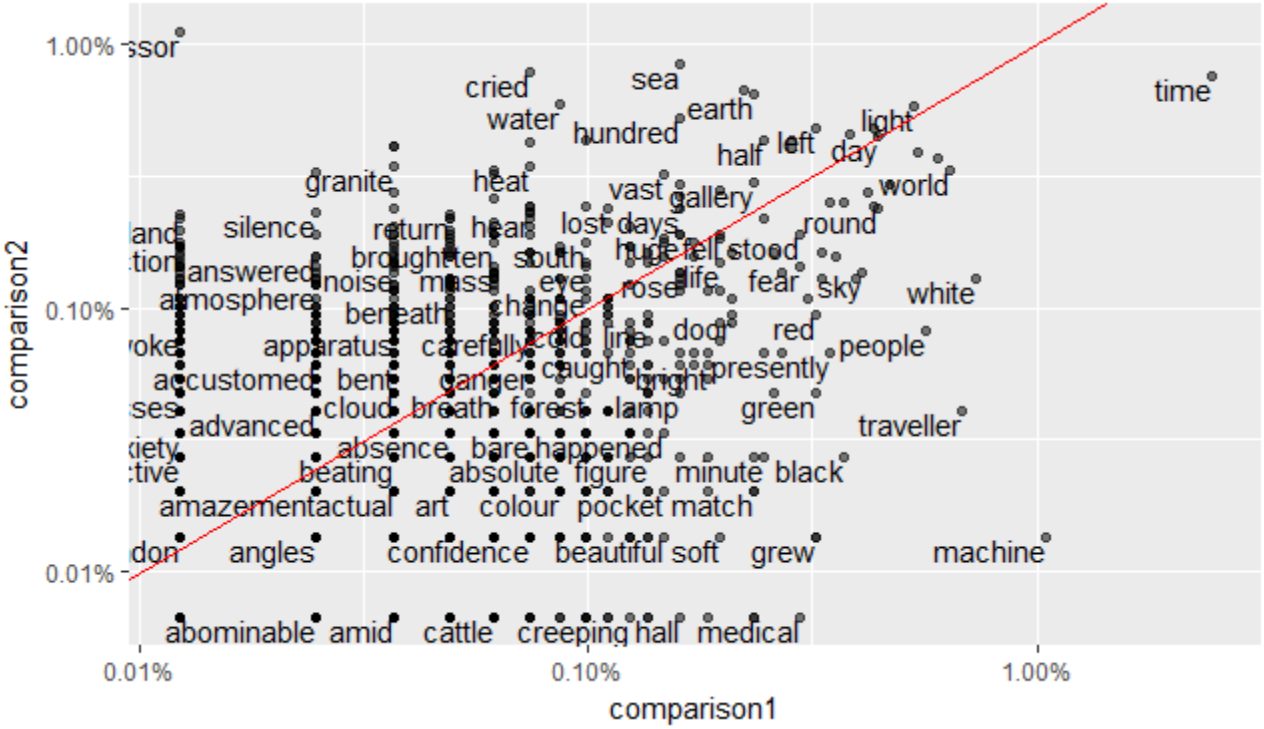
It looks for „data“ folder in the working directory.

hofmeister <- read.table("data/7hfms10.txt", sep="\t",quote = "",header=FALSE,blank.lines.skip=FALSE,stringsAsFactors=F)



Comparing word frequencies

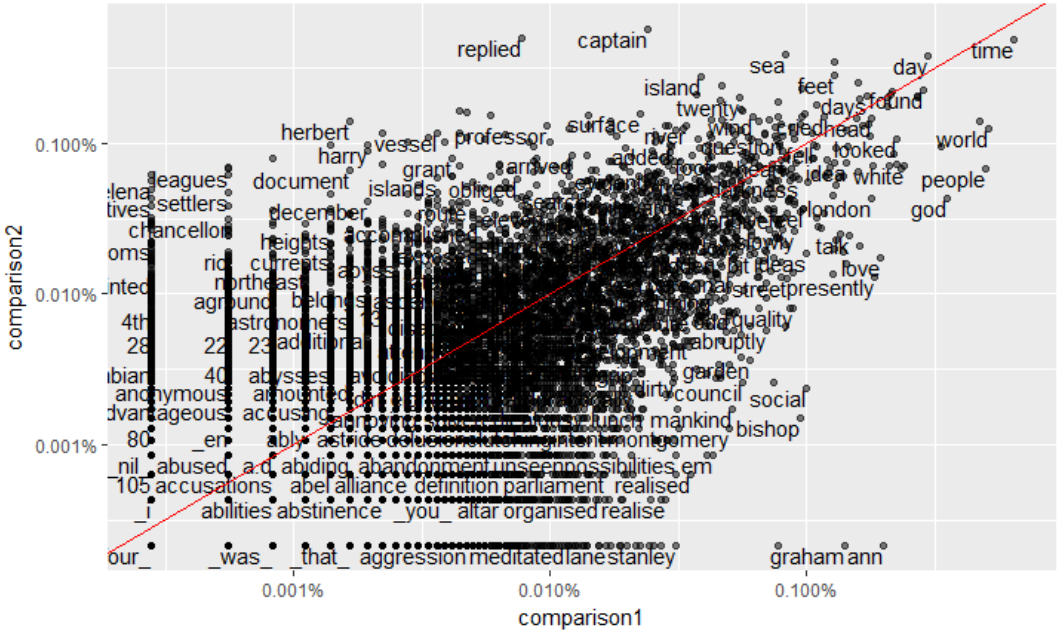
H.G. Wells Time machine (bottom)
vs
J. Verne Journey to middle of earth



Comparison of many texts

HG Wells (bottom)

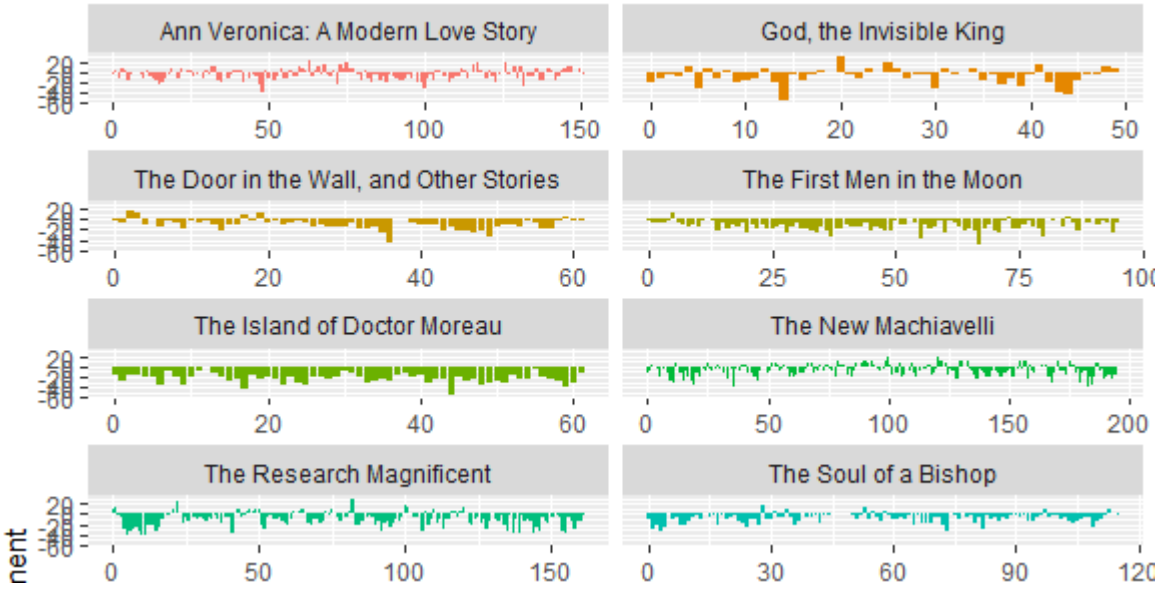
Jules Verne (left)



Sentiment analysis in Lord of the Rings.

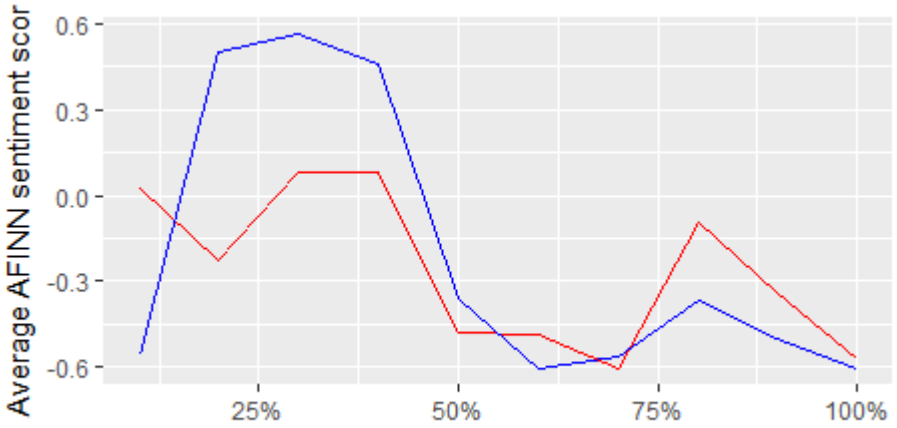
HG Wells sentiment analysis

Counting words with sentiments and their locations within text

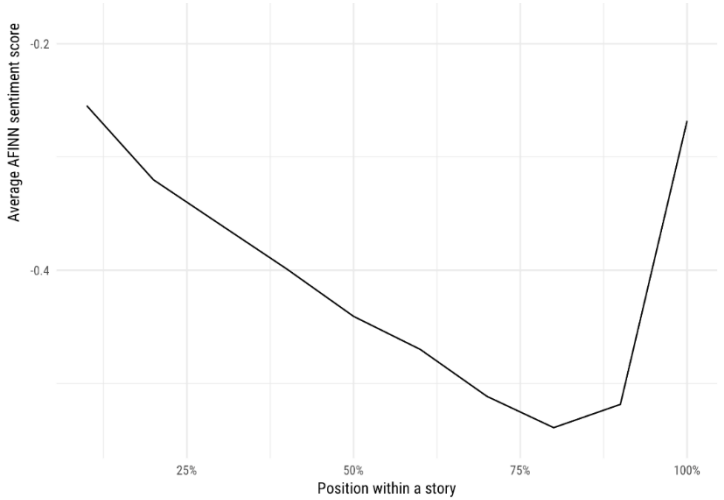


Sentiment averages in text

Verne – blue, Wells – red

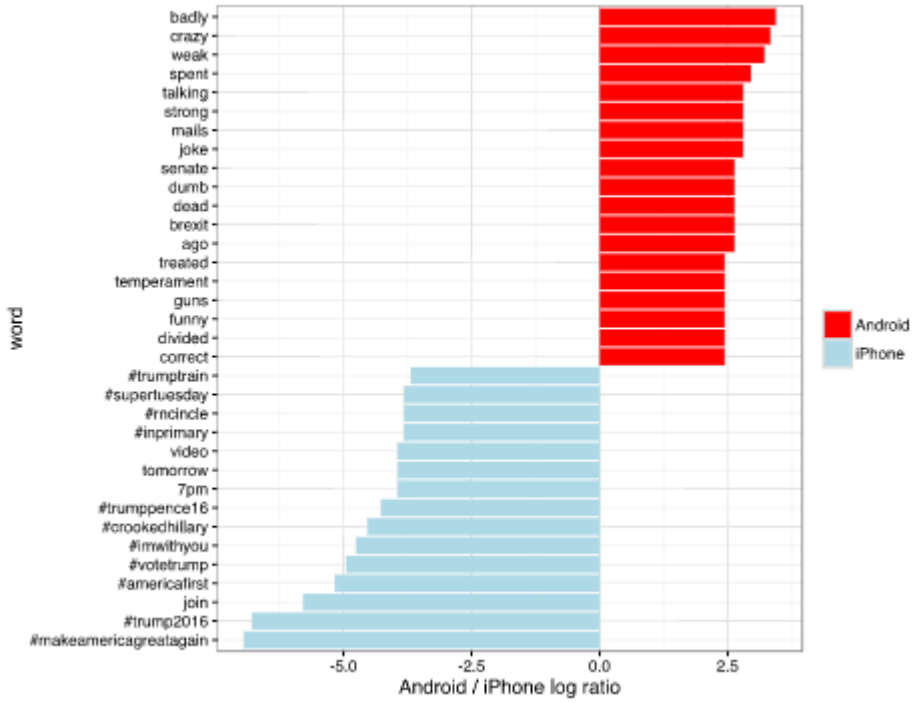


Compare with 100,000 plot descriptions



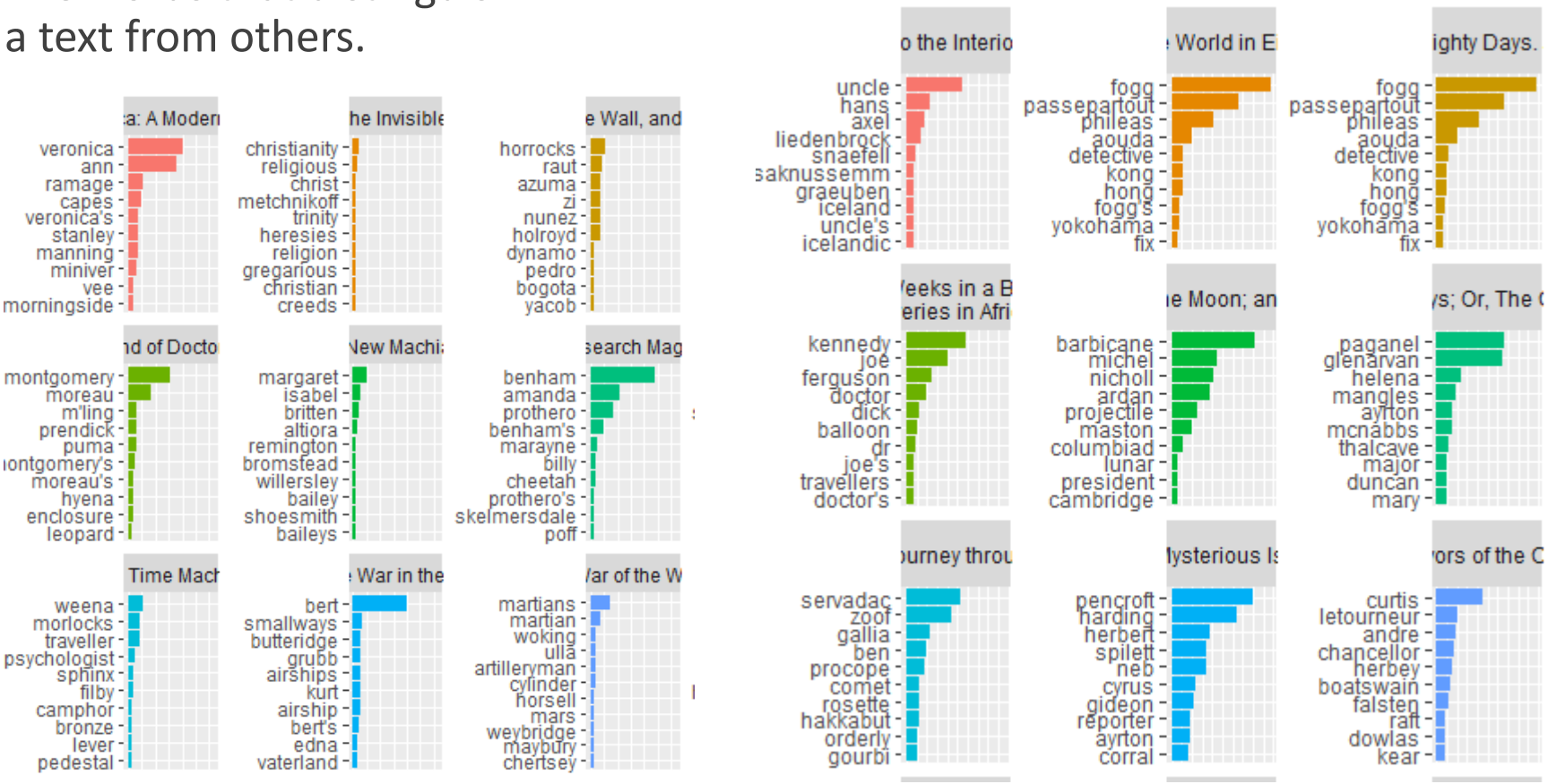
Keywords in 2 groups of texts

E.g. Trump twitter Android & IPhone



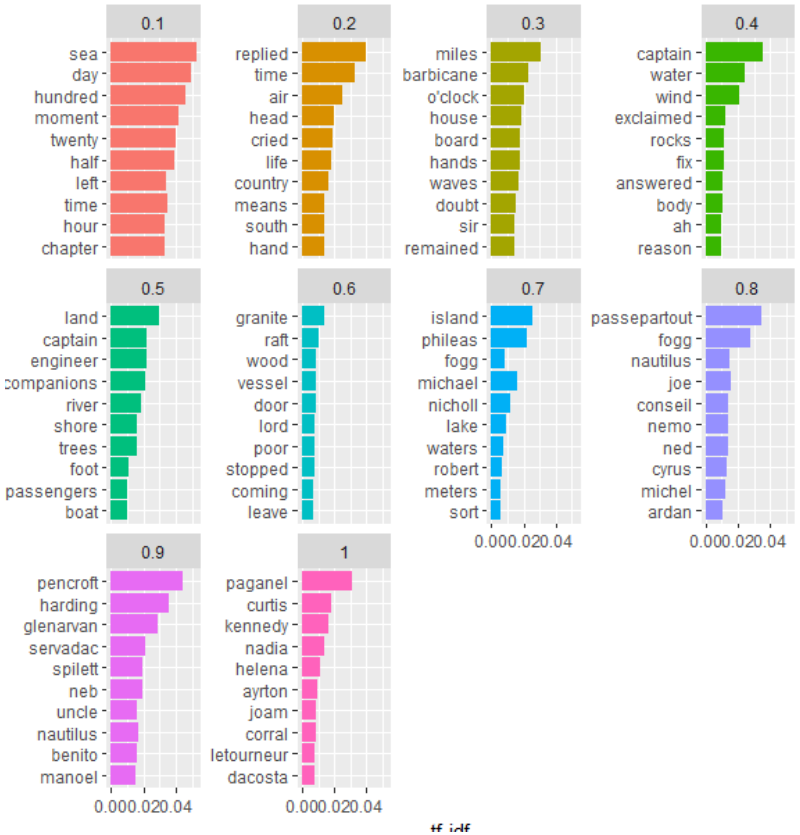
Keyword analysis

The words that distinguish a text from others.



Keywords by position in story

All of downloaded Jules Verne



Basic operations on texts

%>% - pushes the result to be processed on the next line

data %>%

process()

operations:

unnest_tokens(words,line) – make lines into words

mutate(new_var = operation(old_var) – make/change column

filter(var==what_you_want) – get only the rows with right value

count(var) – count unique values

arrange(column) – sort by column

arrange(desc(column) - reverse

- str_detect(where, „what“) – check if it contains „what“
- group_by(what)
- ungroup()

- anti_join(with_what, „var“) – remove matching values
- inner_join(with_what, „var“) – keep only matching values
- left_join(with_what, „var“) – just add where possible

- rename() – just when needed
- summarize() – based on group_by

- bind_df_idf(of, by, value) – get key words

Thanks for attending!