Computational Text Analysis and the Politics of Education

University of Marburg

Milan Thies

28.06.2025

Day 2

Seminar Overview

Day 1 Theoretical foundations

I. Texts as Traces of Political Conflict and Change

II. The Politics of Education

III. Qualitative Text Analysis

Day 2 Practical Foundations

IV. **A Newcomer's Guide to Computational Text Analysis**

V. **Working with R: Introduction to R & Quanteda**

VI. **Workflow & Descriptive Analyses**

Today is the most important day to avoid frustration in the future – let's get the workflow sorted out!

Seminar Overview

Day 3 Text-as-Data Methods

VI.   **Word Counts and Dictionaries**

VII.   **Supervised- and Unsupervised Methods**

Day 4 Application

VIII.   **In-Class Project: Applied CTA (and the Politics of Education)**

IX.   **Presentation, Reflection and Outlook**

Credits

I want to thank Theresa Gessler for first introducing me to the world of computational text analysis but – most importantly for this seminar – her course material on CTA. Some of the following material draws on hers. Her material can be accessed via this link.

A Newcomer's Guide to Computational Text Analysis

A Newcomer's Guide to Computational Text Analysis

1. Why computations text analysis?
2. What is computational text analysis and what are its basic assumptions?
3. Workflow
4. Methods of Computational Text Analysis

Co-funded by the
Erasmus+ Programme
of the European Union

Why Computational Text Analysis?

- Political conflict and change manifests in text (see slides of yesterday)
- Increasing amount of accessible text allows for new insights into politics and policy
- …but also makes it impossible to be processed manually
- Computational text analysis allows us to collect, analyse and systematize large amounts of data, learn about political processes, policy development and the context of concepts that are used in policy-making

What is Computational Text Analysis

- "A variant of content analysis that is expressly quantititative, not just in terms of representing textual content numerically but also in analyzing it, typically using computers" Kenneth Benoit

- Also known as "text-as-data"

- Rapidly growing field within several disciplines including political science

Basic Assumptions

- Text represents an observable implication of some underlying characteristic of interest (the topic of the text, some attribute of the author like e.g. a political position)

- Text can be represented by extracting theirs feature (usually words)

- We can analyse the frequency of features with quantitative methods to measure the characteristics underlying a text

The Bag of Words Representation

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1

Document → **Tokens** → **Document-Feature-Matrix**

How does this work? Creating a bag of words

- We need a corpus, which is a collection of texts.
- We tokenize the corpus, so each word represents a feature

```
toks <- tokens(c("This is a corpus for our seminar on computational text analysis and the politics of education",
"This is the second document of this example corpus"))
toks
```

```
## Tokens consisting of 2 documents.
## text1 :
##  [1] "This"          "is"           "a"            "corpus"
##  [5] "for"           "our"          "seminar"      "on"
##  [9] "computational" "text"         "analysis"     "and"
## [ ... and 4 more ]
##
## text2 :
## [1] "This"     "is"       "the"      "second"   "document" "of"       "this"
## [8] "example"  "corpus"
```

Getting word frequencies per document: The document feature matrix

```
dfm(toks)
```

```
## Document-feature matrix of: 2 documents, 19 features (36.84% sparse) and 0 docvars.
##        features
## docs     this is a corpus for our seminar on computational text
##    text1    1  1 1      1   1   1       1  1             1    1
##    text2    2  1 0      1   0   0       0  0             0    0
## [ reached max_nfeat ... 9 more features ]
```

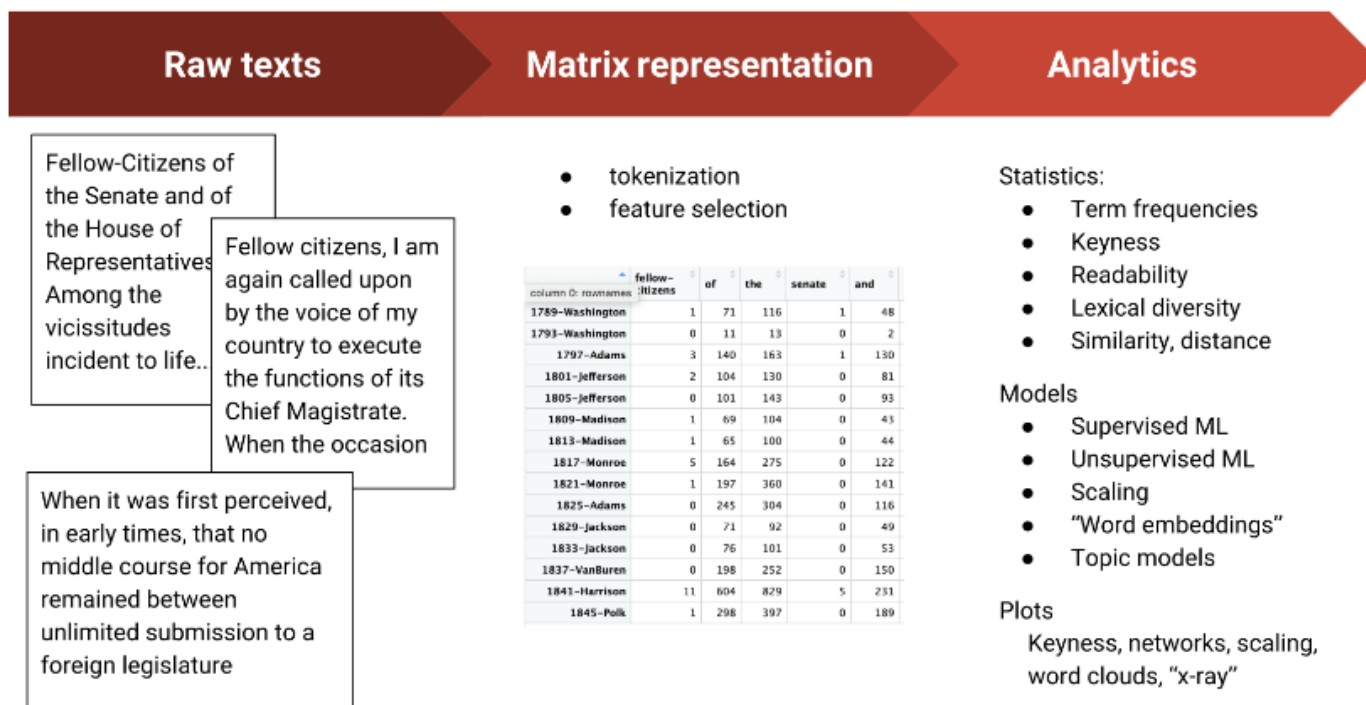# Getting the sorted frequency of features
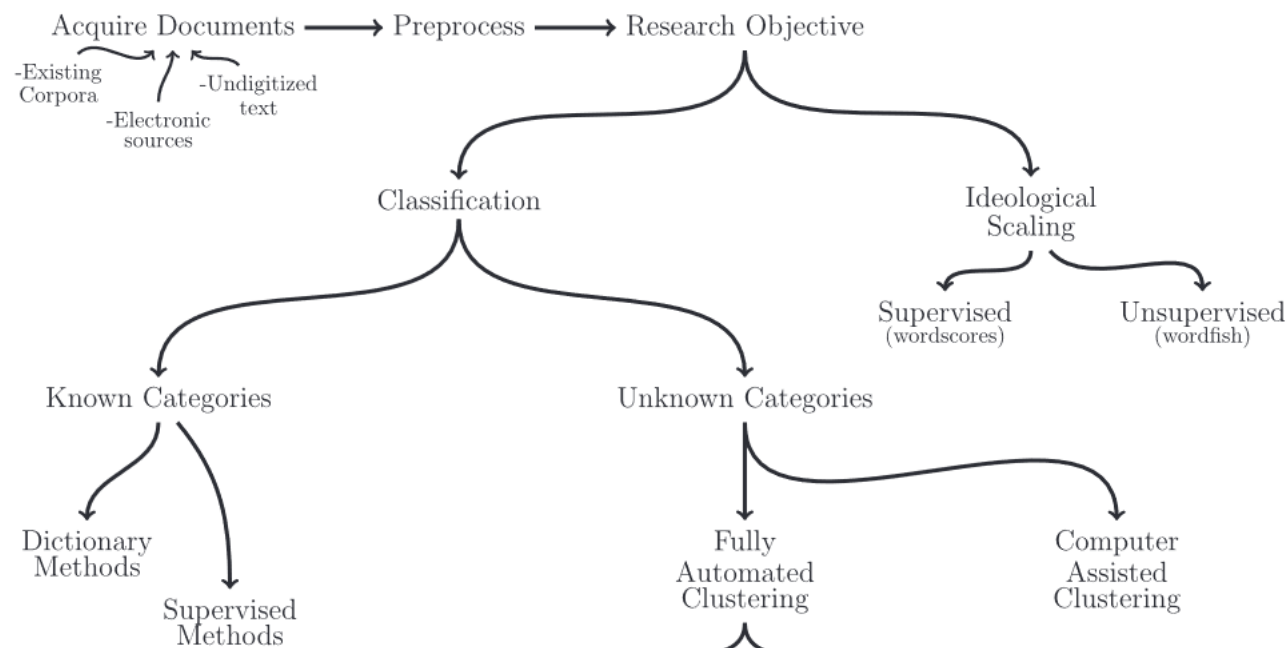
```
dfm(toks) %>% topfeatures()
```

```
##    this      is corpus     the      of       a     for     our seminar      on
##       3       2       2       2       2       1       1       1       1       1
```

The workflow (simple version), demystified by Kenneth Benoit (Quanteda Initiative)

column 0: rownames	fellow-citizens	of	the	senate	and
1789-Washington	1	71	116	1	48
1793-Washington	0	11	13	0	2
1797-Adams	3	140	163	1	130
1801-Jefferson	2	104	130	0	81
1805-Jefferson	0	101	143	0	93
1809-Madison	1	69	104	0	43
1813-Madison	1	65	100	0	44
1817-Monroe	5	164	275	0	122
1821-Monroe	1	197	360	0	141
1825-Adams	0	245	304	0	116
1829-Jackson	0	71	92	0	49
1833-Jackson	0	76	101	0	53
1837-VanBuren	0	198	252	0	150
1841-Harrison	11	604	829	5	231
1845-Polk	1	298	397	0	189

Raw texts

Fellow-Citizens of the Senate and of the House of Representatives Among the vicissitudes incident to life..

Fellow citizens, I am again called upon by the voice of my country to execute the functions of its Chief Magistrate. When the occasion

When it was first perceived, in early times, that no middle course for America remained between unlimited submission to a foreign legislature

Matrix representation

- tokenization
- feature selection

Analytics

Statistics:
- Term frequencies
- Keyness
- Readability
- Lexical diversity
- Similarity, distance

Models
- Supervised ML
- Unsupervised ML
- Scaling
- "Word embeddings"
- Topic models

Plots
Keyness, networks, scaling, word clouds, "x-ray"

Computational Text Analysis Methods

Acquire Documents → Preprocess → Research Objective

-Existing Corpora
-Electronic sources
-Undigitized text

Classification

Ideological Scaling

Supervised (wordscores)

Unsupervised (wordfish)

Known Categories

Unknown Categories

Dictionary Methods

Supervised Methods

Fully Automated Clustering

Computer Assisted Clustering

Example:
- Research objective: Classify parliamentary speeches
- We want to know if politicians often mention social equality in their speeches on education
- We can choose dictionary methods or supervised methods
- Dictionary methods: How often does a specified number of words appear in a document?
- Supervised methods, like seeded topic modelling

Working with R: Introduction to R & Quanteda

A very brief introduction to R

- R is a statistical programming language in which you can perform a large variety of operations

- We will focus on those connected to the analysis of text

- If you do not understand everything while we go through the slides: Do not panic! You will learn it step by step today and I am happy to go through it with you afterwards

- Recommendation: Do not work through the R scripts while I present the steps to you. I know it is tempting but you might miss something important. We will have enough time just for working in the script.

A typical R Session

Environment – objects & values. You can refer to them in your script

Script – tells R what to do. Always safe this. Most important window. Here is where you write code.

Console – output. You can find the output of your code here.

Other information – packages, files etc.

What can we do in Rstudio?
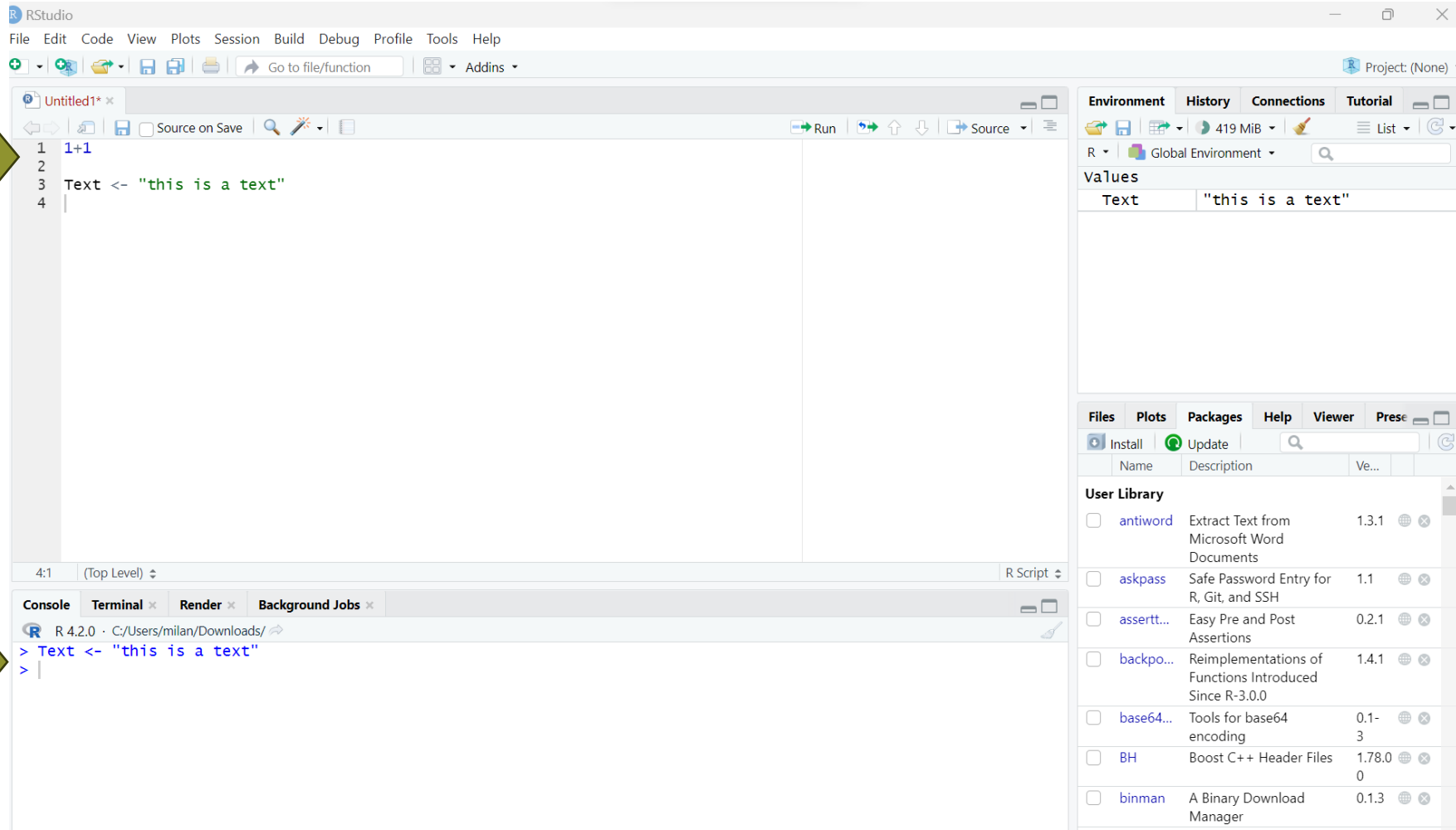
Input – code (= 1+1)

Output – result (= 2)

www.eui.eu
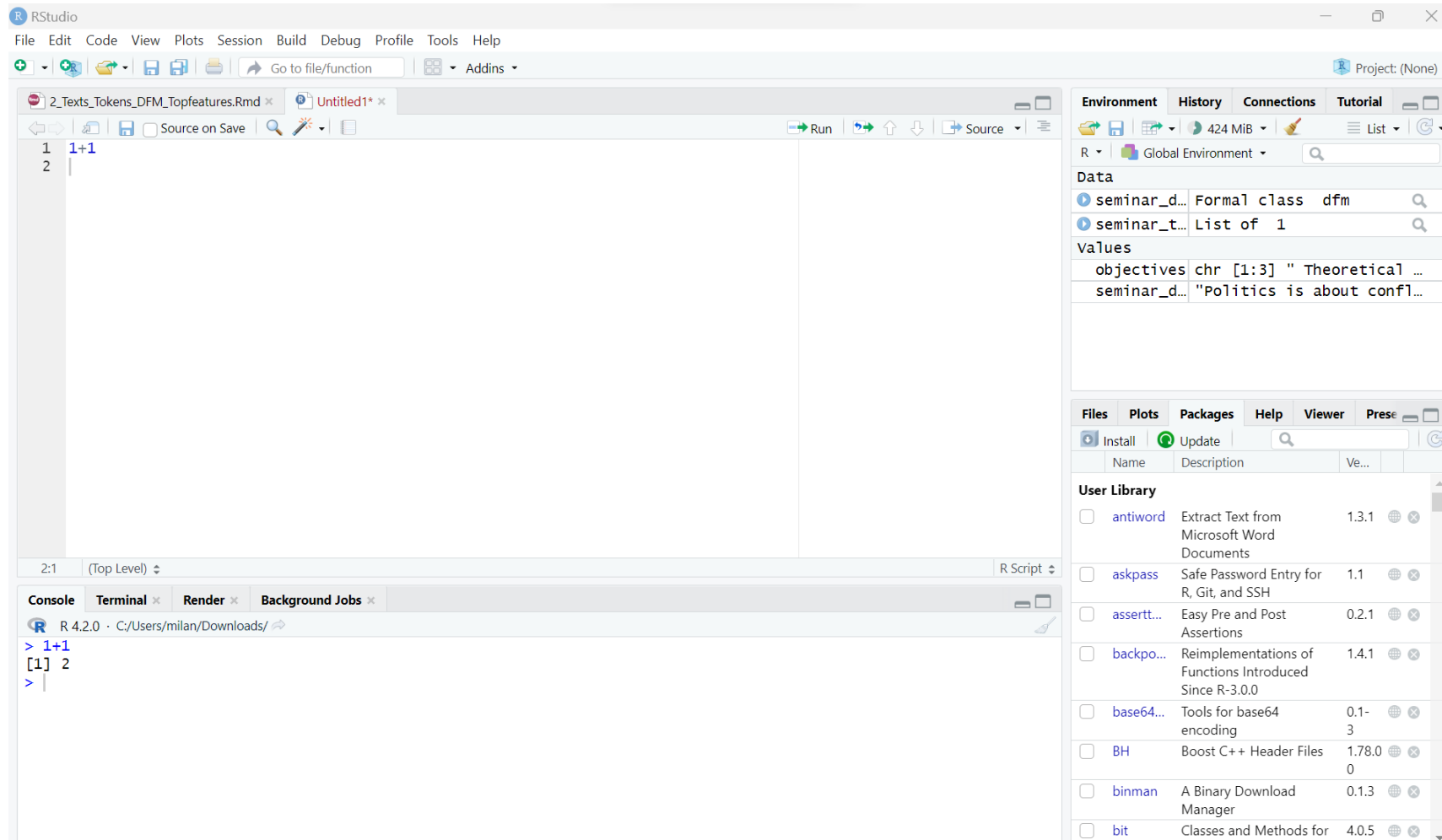
What can we do in Rstudio?

Input – code (name of an object <- value of the object

Output – operation required by code
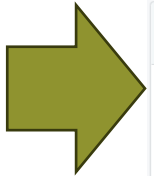
What can we do in Rstudio?

- We can do a great variety of things, mostly statistical programming.

- Most importantly for this seminar is that we can analyse text.

- In this seminar we create an object consisting of text documents called **corpus** and then perform operations on the corpus using the package "**quanteda**".

- In R the operations we can run depend on the packages installed and loaded.

R Markdown

- Instead of writing a normal R script, we will work in R Markdown. (I know, even more coding… sorry!)

- Markdown is a simple formatting syntax creating HTML, PDF, and MS Word documents including code written for R.

- Why do we work in Markdown? It allows you to see exactly which step of your code produced which output. I highly recommend working in Markdown when you are doing computational text analysis. In the next days you will see why.

- Further information: http://rmarkdown.rstudio.com

R Markdown

- I will provide Markdown scripts during the seminar, but this is how you create one in RStudio:
  - File
  - New File
  - R Markdown
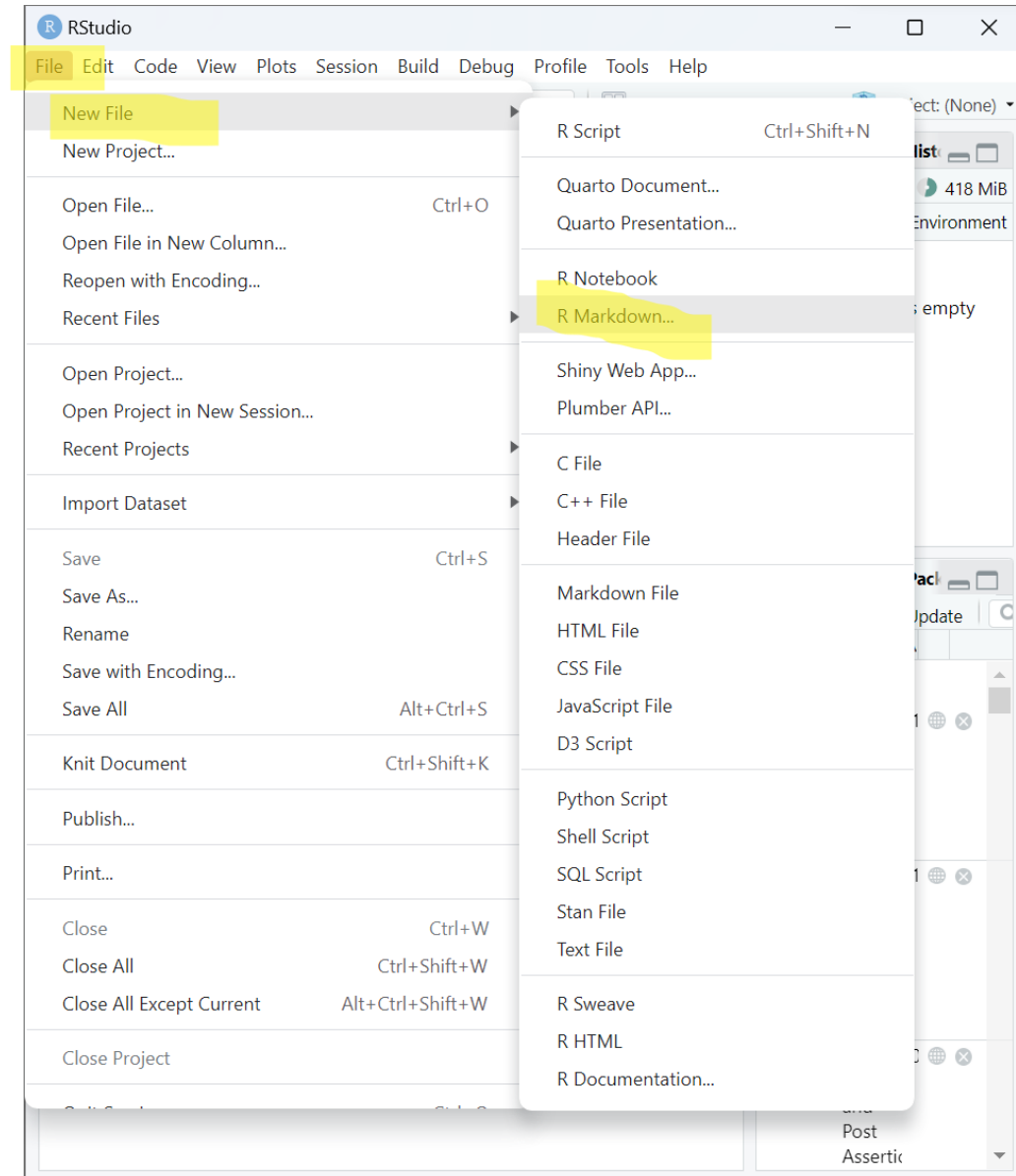
R Markdown

Content – Will appear in the document

Code – Start with ```` ```{r} ```` and end with ```` ``` ````

Output – Will appear in document

All in the script window

R RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function   Addins

1_RMarkdown.Rmd

Knit on Save   Knit   Run

Source Visual   Outline
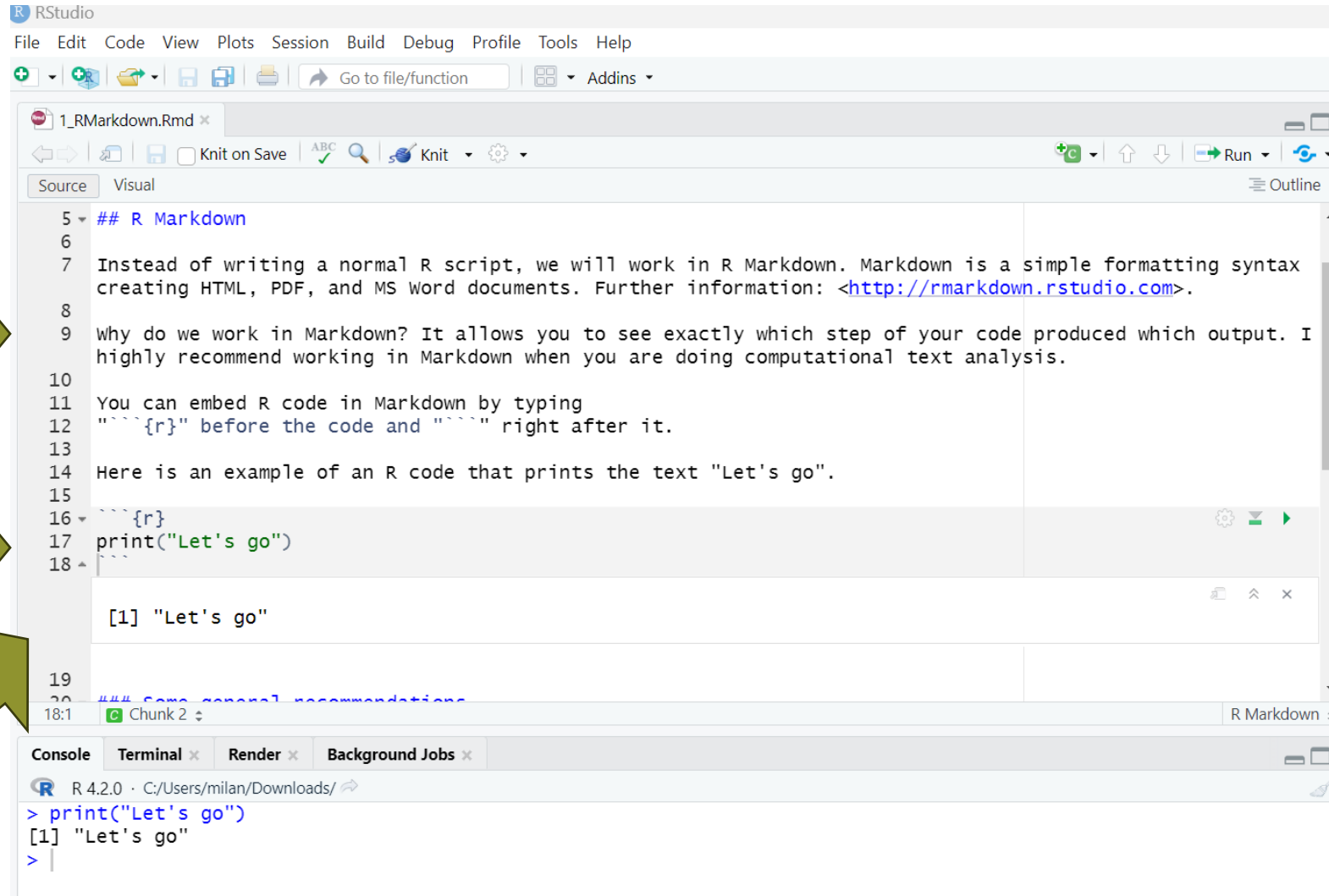
```
 5  ## R Markdown
 6
 7  Instead of writing a normal R script, we will work in R Markdown. Markdown is a simple formatting syntax
    creating HTML, PDF, and MS Word documents. Further information: <http://rmarkdown.rstudio.com>.
 8
 9  Why do we work in Markdown? It allows you to see exactly which step of your code produced which output. I
    highly recommend working in Markdown when you are doing computational text analysis.
10
11  You can embed R code in Markdown by typing
12  "```{r}" before the code and "```" right after it.
13
14  Here is an example of an R code that prints the text "Let's go".
15
16  ```{r}
17  print("Let's go")
18  ```

    [1] "Let's go"

19
20  ### Some general recommendations
```

18:1   C Chunk 2   R Markdown

Console   Terminal   Render   Background Jobs

R R 4.2.0 · C:/Users/milan/Downloads/

```
> print("Let's go")
[1] "Let's go"
>
```

Co-funded by the Erasmus+ Programme of the European Union

Script 1 – R Markdown

- Open document 1_Rmarkdown (find it on ILIAS / my website), understand how the code I wrote for you works.

- Create a new chunk of code including any operation you want, for example a simple math operation (6-2) or assign a value to an object.

- If you have any questions, let me know.

The ABC of Computational Text Analysis – Text, Tokens, DFM

- Who still remembers what tokens are and what a document-feature-matrix is?

- Let's apply this pipeline in R

Step 1: Assigning text to an object

RStudio interface showing:

Tabs: 1_RMarkdown.Rmd, 2_Texts_Tokens_DFM_Topfeatures.Rmd

Source code editor:

```
14
15  Step 1: We first need a text.
16  {r}
17  seminar_description <- "Politics is about conflict and cooperation between societal groups. Harold
    Lasswell once defined it as "who gets what, when, how?". Actors in the political arena, be it
    politicians, lobbyists, or mass movements, engage in (distributional) conflicts through language - be it
    spoken out loud in speeches, dialogues, and protest chants, or written down in position papers,
    protocols, and regulations. Throughout the seminar, students will learn how to use computational methods
    to explore the traces of conflict and cooperation preceding political change.We will focus on a policy
    field that not only affects every university student personally but also one that has been at the
    forefront of global change: Education. Over the past century, education has expanded from being a
    privilege of the few to an almost universal and global experience. In Europe, this trend is now
    increasingly leading to educational upgrading and a massive expansion of higher education. But how has
    this quiet revolution come about? And what are the central conflicts of educational governance today? The
    seminar is organized in three stages of the learning process. First, students will learn about the
    foundations of text-as-data in political science, as well as central theories of education politics.
    Further, they will study how to manually analyze texts without computational assistance. Next, we will
    focus on the methods of computational text analysis. Students will learn about the theoretical
    underpinnings of computational text analysis as well as the use of text-as-data methods through hands-on
    exercises using the R statistical programming language. We will also discuss recent examples of empirical
    research using computational text analysis. Last but not least, students will apply the learned methods
    in an in-class project. Small groups of students will be provided with textual data on a policy process
```

Console output:
```
his quiet revolution come about? And what are the central conflicts of educational governance today? The semina
r is organized in three stages of the learning process. First, students will learn about the foundations of tex
t-as-data in political science, as well as central theories of education politics. Further, they will study how
to manually analyze texts without computational assistance. Next, we will focus on the methods of computational
text analysis. Students will learn about the theoretical underpinnings of computational text analysis as well a
s the use of text-as-data methods through hands-on exercises using the R statistical programming language. We w
ill also discuss recent examples of empirical research using computational text analysis. Last but not least, s
tudents will apply the learned methods in an in-class project. Small groups of students will be provided with t
extual data on a policy process and subsequently implement one or several of the techniques learned throughout
the seminar"
>
```

Environment pane:
Values
seminar_d... "Politics is about confl...

Packages (User Library):
- antiword — Extract Text from Microsoft Word Documents — 1.3.1
- askpass — Safe Password Entry for R, Git, and SSH — 1.1
- assertt... — Easy Pre and Post Assertions — 0.2.1
- backpo... — Reimplementations of Functions Introduced Since R-3.0.0 — 1.4.1
- base64... — Tools for base64 encoding — 0.1-3
- BH — Boost C++ Header Files — 1.78.0-0
- binman — A Binary Download Manager — 0.1.3
- bit — Classes and Methods for ... — 4.0.5

Step 2: Tokenization

This is a screenshot of RStudio showing the tokenization step.

The code editor shows:

```
22
23
24  Step 2: We need to separate the text into its features. To safe the list of features (tokens) we can
    assign it an own name so we can work with it in further steps.
25 - ```{r}
26
27  seminar_toks <- tokens(seminar_description)
28
29  tokens(seminar_description)
30
31 - ```
```

Output panel:

```
Tokens consisting of 1 document.
text1 :
  [1] "Politics"    "is"        "about"      "conflict"   "and"        "cooperation"
  [7] "between"     "societal"  "groups"     "."          "Harold"     "Lasswell"
[ ... and 329 more ]
```

Annotations on the slide:

Two ways to do it:

1. **Creating a new object and then tokenizing**

2. **Creating a direct output**

Console panel:

```
> seminar_toks <- tokens(seminar_description)
> tokens(seminar_description)
Tokens consisting of 1 document.
text1 :
  [1] "Politics"    "is"        "about"      "conflict"   "and"        "cooperation"
  [7] "between"     "societal"  "groups"     "."          "Harold"     "Lasswell"
[ ... and 329 more ]
```

- For many operations you will need a document-feature-matrix

- The dfm tells you the frequency per feature for every feature of each document

- Again: 2 ways: creating an object you can work with later (no direct output, option 1) or creating a direct output but no object (option 2)

```
34  Step 3: We can organise the features in a document-feature-matrix (dfm) which will also give us the
    frequency of each feature.
35 ▾ ```{r}
36  seminar_dfm <- dfm(seminar_toks)
37
38  dfm(seminar_toks)
39 ▴
```

```
Document-feature matrix of: 1 document, 176 features (0.00% sparse) and 0 docvars.
        features
docs     politics is about conflict and cooperation between societal groups  .
  text1         2  3     4        2   8           2       1        1      2 13
[ reached max_nfeat ... 166 more features ]
```

Step 4: Topfeatures

- You can also sort the dfm by frequency with the topfeatures function
- Topfeatures are a great way of understanding the differences between texts and what texts are about
- How to depends on whether you have created a dfm object before (option 1). If not you can use %>% (also called pipe) as a way of combining two arguments. Here "create dfm" AND "organise by frequency" (option 2)

```
42  Step 4: We can learn about the most used features using the **topfeatures** function drawing on the
    document-feature-matrix we have created.
43 ▾  ```{r}
44  topfeatures(seminar_dfm)
45 ▴  ```
```

	,	the	of	.	will	and	in	as	students	to
	19	17	14	13	9	8	8	5	5	5

```
46  Alternative step 4: We can also combine several steps by using the "%>%" connector. Here we create a dfm
    from the tokens we have created (list of features [= mostly words]) and then we also directly ask it to
    print the topfeatures (= the most frequent features).
47 ▾  ```{r}
48  dfm(seminar_toks) %>% topfeatures()
49
50 ▴  ```
```

	,	the	of	.	will	and	in	as	students	to
	19	17	14	13	9	8	8	5	5	5

Script 2 – ABC of Text Analysis

- Open document 2_Texts_Tokens_DFM_Topfeatures (find it on ILIAS/ my website)

- Work through the script

- If you have any questions, let me know

Almost ready to start analysing text

- Results not really interesting because we find that the most frequent features are not really interesting. What we still need to learn in how to preprocess the data.

- Next section is about your setup and the workflow for text analysis including the preprocessing

Workflow & Descriptive Analyses

The real workflow – including pre-prosessing

- Install & load the packages you will use
- Acquire documents and import them into RStudio
- Pre-Processing
  - Tokenize
  - Create document-feature-matrix
  - Learn about data at hand
  - Remove non-text (punctuation, symbols, numbers)
  - Remove meaningless words (stopwords, potentially trim)
  - Stem
- Choose method for text analysis
- Plot and visualise

Working Directory

- Before we start: find or set the working directory

- WD is a folder in which RStudio stores output and saves your script

- You can set a new working directory with the setwd() function
  - How? You copy the path to a folder you like into the brackets in quotation marks. Make sure to change all \ to /.
  - Example: setwd("C:/Users/.../Data & Scripts")

Installing and loading packages

- Next you install the packages you might need and load those you use (through the library function)

```{r}
packages <- c("tidyverse","quanteda","quanteda.textstats",
"quanteda.textplots","quanteda.textmodels","readtext","rmarkdown","knitr")
install.packages(packages)

library(tidyverse)
library(quanteda)
library(quanteda.textmodels)
library(quanteda.textplots)
library(quanteda.textstats)
library(readtext)
```

Acquire documents and import them into RStudio: Eurydice

- Governments describe their own education systems at length

- Data set I scraped last year

- No longitudinal data, but great standardized cross-sectional data

- Issues: Contains words in original language sometimes

- Nicely organised into table for you to work with

Export PDF

▌ **Key features of the Education System**

According to the Austrian Federal Constitutional Law Article 14 - as amended (Bundesverfassungs-Gesetz, B-VG, Art. 14) democracy, humanity, solidarity, peace and justice, openness and tolerance towards everyone regardless of race, social status and financial background are fundamental principles of education in Austria.

▌ **Key features related to governance**

Concerning **kindergartens and crèches** the provinces (Bundesländer) are responsible for legislation and implementation and maintained to high degree by municipalities (Gemeinden). However, there is also a large **private sector**.

Concerning **schools** responsibilities for legislation and its implementation are divided between the federation (Bund) and the provinces (Bundesländer) where it is executed by the parliaments of the provinces (Landtage) and the offices of the provincial governments (Ämter der Landesregierungen). In specific matters enumerated in the Constitution, the federation sets the framework, while detailed legislation is implemented by the parliaments of the provinces. The federation has overwhelming responsibility for the education system, including virtually all areas of school organisation, the organisation of school instruction, private schools as well as the remuneration and retirement law governing education staff.

Legislation and execution of all matters pertaining to **universities** and **higher education** is a

Table of Contents   In other countries

Overview

> 1. Political, social and economic background and trends
> 2. Organisation and governance
> 3. Funding in education
> 4. Early childhood education and care
> 5. Primary education
> 6. Secondary and post-secondary non-tertiary education
> 7. Higher education
> 8. Adult education and training
> 9. Teachers and education staff

Acquire documents and import them into RStudio: Eurydice

- Data set available for download in ILIAS/ on my website
- Find out the path to the document. Does everybody know how to find the path?
- Steps:
  - Copy the path to the RDS document "eurydice".
  - Use the readRDS() function to import it into the environment.
  - Create a new object (systems)
  - Create a corpus that only includes the features of the RDS imported.

```r
systems <- readRDS("C:/Users/                        Data & Scripts/eurydice.RDS")

systems_corp <- corpus(systems)
```

Eurydice in R

	country	chapter	text	url
1	Albania	Political,Social and Economic Background and Trends	This chapter discusses political, social and economic backgr...	https://eacea.ec.europa
2	Albania	Historical Development	Albanian territories were inhabited since 100,000 years ago....	https://eacea.ec.europa
3	Albania	Main Executive and Legislative Bodies	The President and the Assembly The President of the Repub...	https://eacea.ec.europa
4	Albania	Population: Demographic Situation, Languages and Religion...	Population and Demographic Situation The Republic of Alba...	https://eacea.ec.europa
5	Albania	Political and Economic Situation	Albanian is classified as a middle-income country since 200...	https://eacea.ec.europa
6	Albania	Organisation and Governance	This chapter is divided into six sub-sections the first discuss...	https://eacea.ec.europa
7	Albania	Fundamental Principles and National Policies	Fundamental principles of national education system in the ...	https://eacea.ec.europa
8	Albania	Lifelong learning strategy	Lifelong learning has been a cross-cutting issue, addressed ...	https://eacea.ec.europa
9	Albania	Organisation of the education system and of its structure	Pre-university education system. Organization. Based on law...	https://eacea.ec.europa
10	Albania	Organisation of private education	Private pre-university education Likewise with public Pre-uni...	https://eacea.ec.europa
11	Albania	National qualifications framework	Albanian Qualifications Framework was adopted by Law No...	https://eacea.ec.europa
12	Albania	Administration and governance at central and/or regional le...	Pre university education According to the pre-university edu...	https://eacea.ec.europa
13	Albania	Administration and governance at local and/or institutional ...	Pre- University education The Law on pre-university educati...	https://eacea.ec.europa

Pre-Processing: Tokenize

- At word, sentence or character level. We will tokenize at word level and store tokenized text in new object.

- Why not at character level?

```r
50   ```{r}
51   tokens(systems_corp,"word")
52   tokens(systems_corp ,"sentence")
53   tokens(systems_corp,"character")
54   systems_toks <- tokens(systems_corp)
55
56   ```
```

46:16   # Tokenization   R Markdown

Console   **Terminal**   **Background Jobs**

R R 4.2.0 · ~/

```
Tokens consisting of 4,465 documents and 3 docvars.
text1 :
 [1] "This"       "chapter"    "discusses"  "political"  ","          "social"     "and"
 [8] "economic"   "background" "and"        "trends"     "of"
[ ... and 59 more ]

text2 :
 [1] "Albanian"   "territories" "were"       "inhabited"  "since"      "100,000"
 [7] "years"      "ago"         "."          "Over"       "the"        "centuries"
[ ... and 409 more ]

text3 :
```

Pre-Processing: Document-Feature-Matrix

- You know the drill: Use the dfm() function and store in new object

```{r}
systems_dfm <- dfm(systems_toks)
systems_dfm
```

```
Document-feature matrix of: 4,465 documents, 71,392 features (99.49% sparse) and 3 docvars.
       features
docs     this chapter discusses political  , social and economic background trends
   text1    1       2         1         2  1     1   5        2          1      1
   text2    1       0         0         1 20     4  17        3          0      0
   text3    0       0         0         0  4     0  22        0          0      0
   text4    2       0         0         0 34     0  14        0          0      0
   text5    0       0         0         0  5     0   5        2          0      0
   text6    1       2         3         0  0     0   2        0          0      0
[ reached max_ndoc ... 4,459 more documents, reached max_nfeat ... 71,382 more features ]
```

Get to know the data: Keywords in context

- Get window of word before and after a keyword

- How?

```{r}
kwic(systems_toks, "education",window=4)
```

- What do we see?

<chr>	<chr>		<chr>		<chr>
[text3, 354]	. The Ministry of	|	Education	|	, Sports and Youth
[text3, 375]	in charge for drafting	|	education	|	policies in the country
[text3, 386]	the area of pre-univeristy	|	education	|	it has full competences
[text3, 397]	the area of higher	|	education	|	it shares competences with
[text3, 403]	shares competences with higher	|	education	|	institutions in compliance with
[text5, 140]	%. Expenses for	|	education	|	in 2019 as percentage
[text5, 164]	INSTAT. Expenses for	|	education	|	in 2020 account to
[text6, 20]	policies that rule the	|	education	|	system in Albania.

Co-funded by the
Erasmus+ Programme
of the European Union

Get to know the data:

- We can analyze the number of tokens (=words) per document

- The number of features or documents in the dfm

- The most frequent features before starting to clean the data

```
1. Number of tokens per document

```{r}
ntoken(systems_toks) %>% head()
```

 text1 text2 text3 text4 text5 text6
    71   421   410   455   175   79

2. The number of feature in our dfm

```{r}
nfeat(systems_dfm)
```

 [1] 60998
```

3 Steps of Pre-Processing:

1. Remove non-text

2. Remove uninformative text

3. Unite features

- Removing non-text means removing, punctuation, numbers and symbols

- In process of tokenization, we can ask to remove each feature: remove_punct = T means that it is true (T) that we want to remove punctuation.

- We need to create dfm from clean tokens again

```
121  systems_toks <- tokens(systems_corp,  remove_punct=T,
122    remove_numbers=T, remove_symbols=T)
123  systems_dfm <- dfm(systems_toks)
```

3 Steps of Pre-Processing:
Step 2: Remove uninformative text

- Texts often contain words that are not interesting or even harmful for our analysis. See for example the topfeatures brefore pre-processing:

```
      the        of       and        in        to education       for         a        is       are
   408922    251546    218485    138468    117839    105486     99107     86702     58311     47690
```

- These words are with one exception stopwords. Let's remove them!

```
systems_dfm <- dfm_remove(systems_dfm,stopwords("en"))
```

- There are lists of stopwords for many languages, include all that matter.

- We can also remove words using the dfm_trim() function. Example: words that appear too often to be relevant or words that appear to few times to matter.

3 Steps of Pre-Processing:
Step 3: Unite features

- If we reduce the features to their wordstem, we unite words/features with the same stem, often sharing the same meaning. E.g. "systems" and "system" or "programm" and "programme"

```
systems_dfm <- dfm_wordstem(systems_dfm,"en")
```

Now you: Script 3

- Work on Script "3_Pre-Processing & Descriptives"
- **Work at your own pace, ask me or your peers for help**
- Do not start with the "Advanced Transformations" and the "Descriptive Analysis" section yet

How did it go? How far did you get?

Use existing document variables

- Our initial data set had variables that describe the text: Information about the text. Which country does the text describe? Which part of the education system does the text describe?

- We want to use that information in the analysis but lost it by creating a corpus.

- Let's re-attach the docvars

	country	chapter	text	url
1	Albania	Political,Social and Economic Background and Trends	This chapter discusses political, social and economic backgr...	https://eacea.ec.europa
2	Albania	Historical Development	Albanian territories were inhabited since 100,000 years ago...	https://eacea.ec.europa
3	Albania	Main Executive and Legislative Bodies	The President and the Assembly The President of the Repub...	https://eacea.ec.europa
4	Albania	Population: Demographic Situation, Languages and Religion...	Population and Demographic Situation The Republic of Alba...	https://eacea.ec.europa
5	Albania	Political and Economic Situation	Albanian is classified as a middle-income country since 200...	https://eacea.ec.europa
6	Albania	Organisation and Governance	This chapter is divided into six sub-sections the first discuss...	https://eacea.ec.europa
7	Albania	Fundamental Principles and National Policies	Fundamental principles of national education system in the ...	https://eacea.ec.europa
8	Albania	Lifelong learning strategy	Lifelong learning has been a cross-cutting issue, addressed ...	https://eacea.ec.europa
9	Albania	Organisation of the education system and of its structure	Pre-university education system. Organization. Based on law...	https://eacea.ec.europa
10	Albania	Organisation of private education	Private pre-university education Likewise with public Pre-uni...	https://eacea.ec.europa
11	Albania	National qualifications framework	Albanian Qualifications Framework was adopted by Law No....	https://eacea.ec.europa
12	Albania	Administration and governance at central and/or regional le...	Pre university education According to the pre-university edu...	https://eacea.ec.europa
13	Albania	Administration and governance at local and/or institutional ...	Pre- University education The Law on pre-university educati...	https://eacea.ec.europa

```
docvars(systems_dfm, "country") <- systems$country
docvars(systems_dfm, "chapter") <- systems$chapter
```

```
systems_... Large dfm (220995175...
    ..@ docvars :'data.frame':...
    .. ..$ docname_: chr [1:4465].
    .. ..$ docid_  : Factor w/ 44.
    .. ..$ segid_  : int [1:4465].
    .. ..$ country : chr [1:4465].
    .. ..$ chapter : chr [1:4465].
```

Co-funded by the Erasmus+ Programme of the European Union

Advanced Transformations: Group documents

- After attaching the country and chapter variable we can analyze text per country or chapter by using the dfm_group function

- This helps if we want to analyze the education system of one country or compare one sector of the education system across countries

```{r}
country_system_dfm <- dfm_group(systems_dfm,country)
chapter_system_dfm <- dfm_group(systems_dfm,chapter)
```

```
country_system_d…  Large dfm (2128285 elements,  5.8 MB)
    ..@ docvars :'data.frame':      43 obs. of  4 variables:
    .. ..$ docname_: chr [1:43] "Albania" "Austria" "Belgium - Fle
    .. ..$ docid_  : Factor w/ 43 levels "Albania","Austria",..: 
    .. ..$ segid_  : int [1:43] 1 1 1 1 1 1 1 1 1 1 ...
    .. ..$ country : chr [1:43] "Albania" "Austria" "Belgium - Fle
```

Advanced Transformations: Create a subset

- Sometimes we want to analyse only some part of the data

- Therefore, we need to create a subset of the data

- Let's create a subset in which we exclude Germany and one in which we only keep data on Germany with dfm_subset()

```
nogermany_system_dfm <- dfm_subset(systems_dfm,country!="Germany")
Germany_dfm <- dfm_subset(systems_dfm, country == "Germany")
```

Descriptive Analyses

What we will cover

- How to conduct and visualise simple text-as-data analyses

- Analyses based on absolute word frequencies: Topfeatures and Wordclouds

- Analysis based on relative word frequencies (keyness / keyword analysis)
  - How frequent are words compared to the overall corpus?

- What we will not cover today, but next weekend:
  - Frequencies of selected words
  - Dictionaries
  - Topic Models
  - In-depth discussion on best practices and reading (Martin, 2018)

Topfeatures

- We have seen the mechanics of how to get topfeatures earlier today. Let's compare how useful this is to understand what a text is about before and after pre-processing.

- Before:

```
      the         of        and         in         to education        for          a         is        are
   408922     251546     218485     138468     117839     105486      99107      86702      58311      47690
```

- After:

```
      educ     school    student   institut      train   programm
    125054      71368      27918      27828      25322      23814
   teacher     higher       year     provid
     23326      23166      22670      19144
```
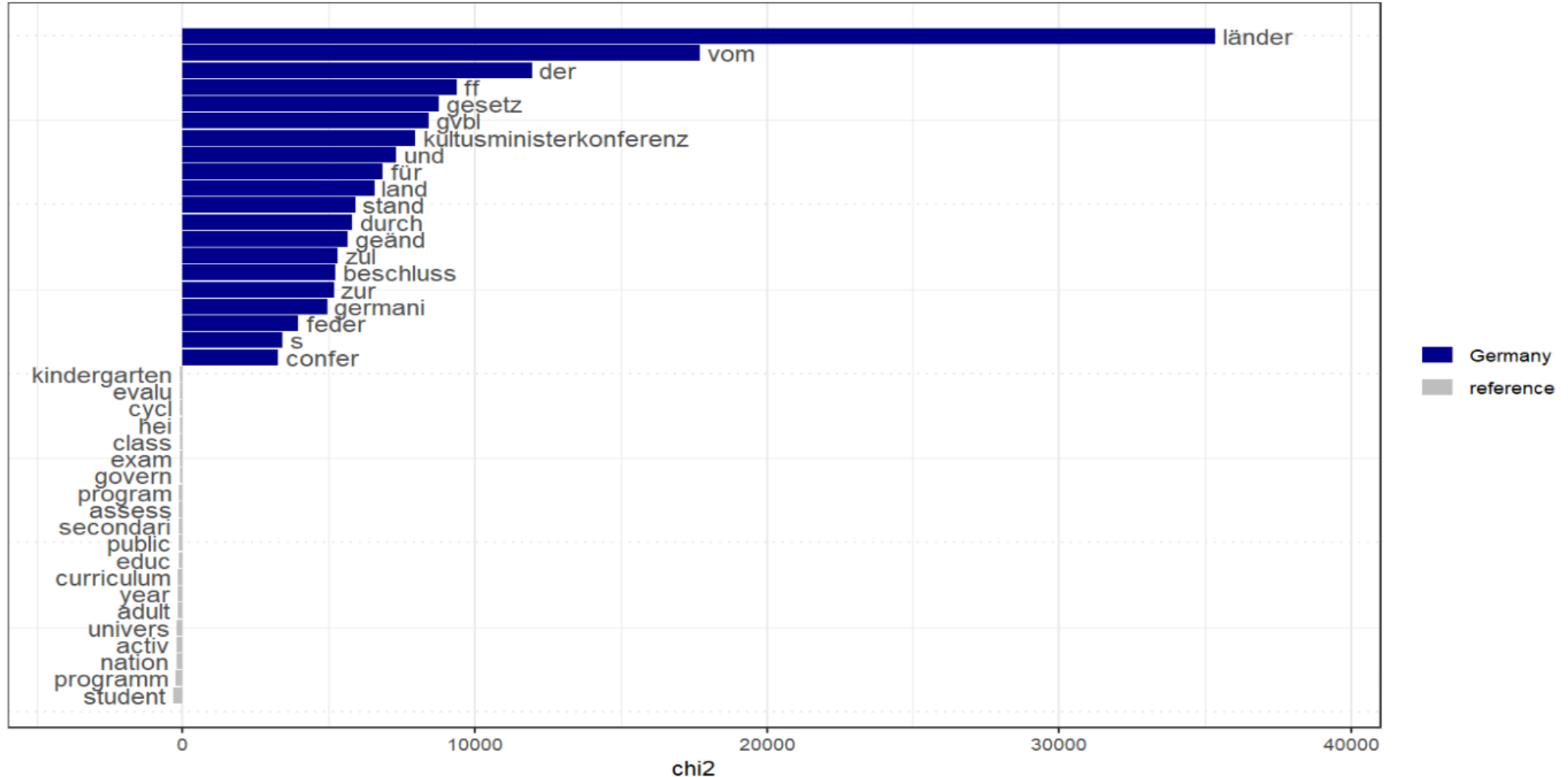
Wordclouds

- Wordclouds are perhaps the most popular visualization of absolute frequencies
- We can create them with a simple line of code:

```
textplot_wordcloud(systems_dfm, max_words = 100)
```

- We can define the number of words the cloud displays, and we can also subset our dfm to create and compare wordclouds for different texts
- What differences do you see between the overall cloud and the one for Germany?

Keyword analysis / keyness

- Relative word frequency: Words that appear more frequent in one text / a group of texts compared to the overall corpus

- Great for understanding differences

- However, easy to confound by particular words (e.g. in another language).

- How to?
  - Define how to group documents
  - Loop function textstat_keyness() in.
  - Define group of interest
  - Head() gives you a list of 10 top words
  - Textplot_keyness() gives you a graph

```
keyness <- dfm_group(systems_dfm,country) %>%
  textstat_keyness("Germany")

head(keyness)

keyness %>%
  textplot_keyness()
```

Co-funded by the
Erasmus+ Programme
of the European Union

How can we interpret this?
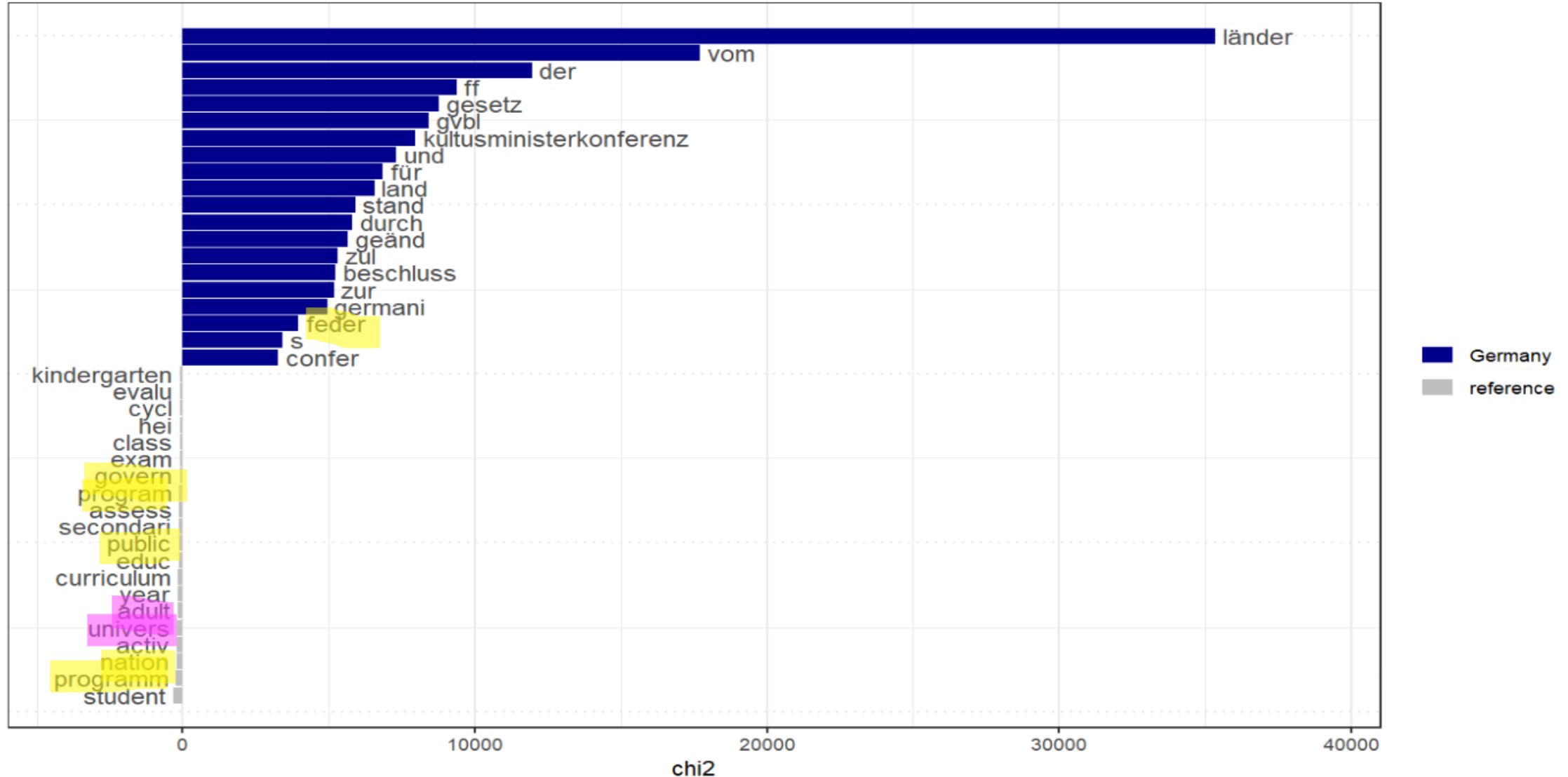
Now you: Script 3 - continued

- Work on Script "3_Pre-Processing & Descriptives"

- Work at your own pace, **ask me or your peers for help**

- In the end you will be able to work with other data, pre-processing it, running analyses and visualizing it

Now you: Script 3 - continued

- Choose a new data set, pre-process it and analyse it.

- I recommend to draw on data from the ParlSpeech database: https://dataverse.harvard.edu/dataverse/ParlSpeech

- Work at your own pace, **ask me or your peers for help**

Enough code – but let's prepare for the next step

- Important paper on how to use and create dictionaries

- Same procedure as yesterday: read it and then we will discuss it together

- How does dictionary analysis work?

- What is the role of human coders?

- Does context matter?

Routledge
Taylor & Francis Group

Check for updates

(Re)Claiming Our Expertise: Parsing Large Text Corpora With Manually Validated and Organic Dictionaries

ASHLEY MUDDIMAN, SHANNON C. MCGREGOR, and NATALIE JOMINI STROUD

Content analysis of large-scale textual data sets poses myriad problems, particularly when researchers seek to analyze content that is both theoretically derived and context dependent. In this piece, we detail the approach we developed to tackle the analysis of the context-dependent content of political incivility. After describing our manually validated organic dictionaries approach, we compare the method to others we could have used and then replicate the method in a different—but still context-dependent—project examining political issue content on social media. We conclude by summarizing the strengths and weaknesses of the approach and offering suggestions for future research that can refine and expand the method.

Project Pitch

- Take 20 minutes to develop a small computational text analysis project based on what you know so far

- Present it to the others

Next week: Methods that get published

Day 3 Text-as-Data Methods

VI.   **Dictionaries and Word Counts (Incl. discussion on Martin 2018! Come prepared)**

VII.   **Supervised- and Unsupervised Machine Learning for Text Analysis. Focus on topic models**

Day 4 Application

VIII.   **In-Class Project: Applied CTA (and the Politics of Education)**

IX.   **Presentation, Reflection and Outlook**

If you would like to work with your own data: Bring text along (in whichever format you have it).