Computational Text Analysis and the Politics of Education

University of Marburg
Milan Thies
04.07.2025

Day 3

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

Recap Day 1: Politics of Education

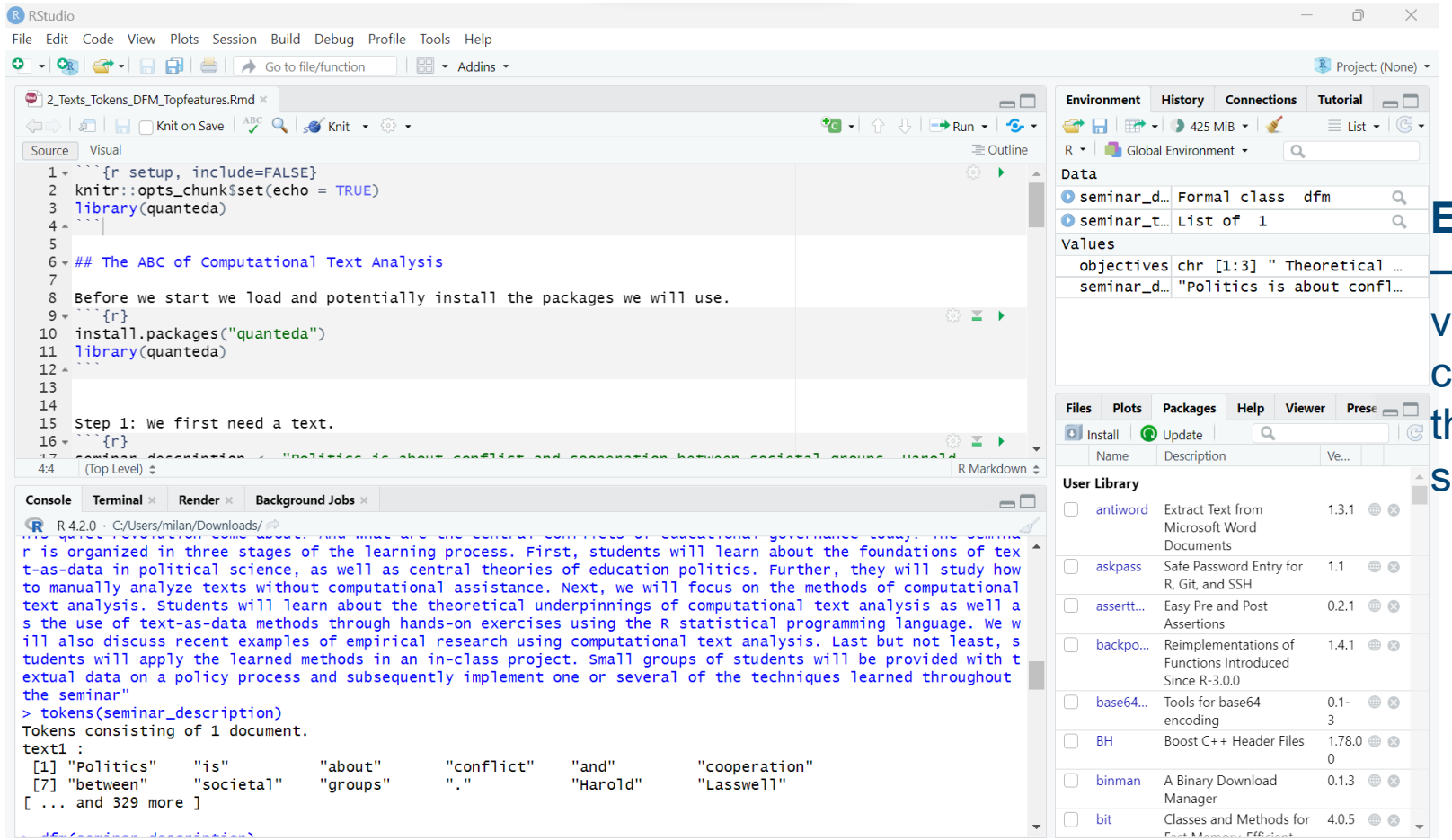
What are questions you could investigate based on the theoretical foundations we discussed on day one?

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

Recap Day 1
Introduction to R, Rmarkdown, Qualitative Text Analysis

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.



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

Recap – Day 2

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

The Bag of Words Representation

Source

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

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

The workflow

- 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

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

Get to know the data: Keywords in context

- Get window of words before and after a keyword
- How?

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

- What do we see?

<chr>	<chr>	<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

3 Steps of Pre-Processing:

1. Remove non-text tokens
2. Remove uninformative text (from dfm)
3. Stemming

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


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

Advanced Transformations: Create a subset / group features

- Sometimes we want to analyse only some part of the data
- Therefore, we need to create a subset of the data

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


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

Topfeatures

- Before pre-processing:

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

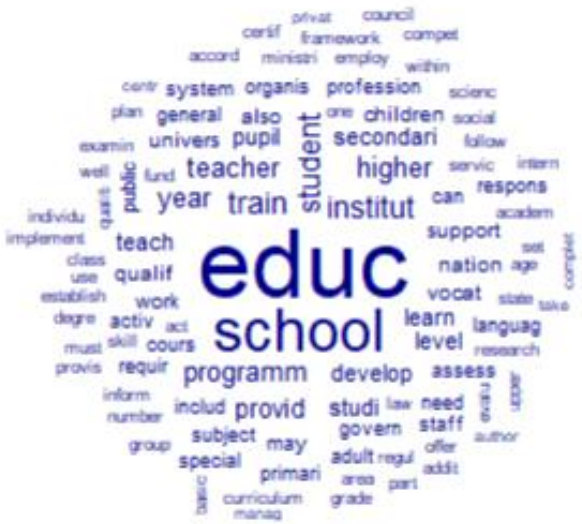
- After pre-processing :

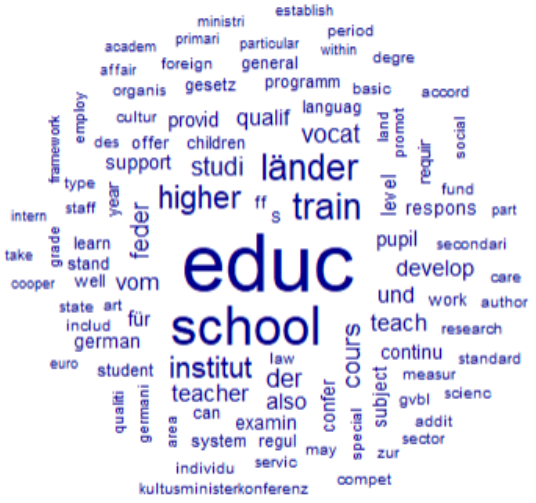
educ	school	student	institut	train	programm
125054	71368	27918	27828	25322	23814
teacher	higher	year	provid		
23326	23166	22670	19144		

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

Wordclouds

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

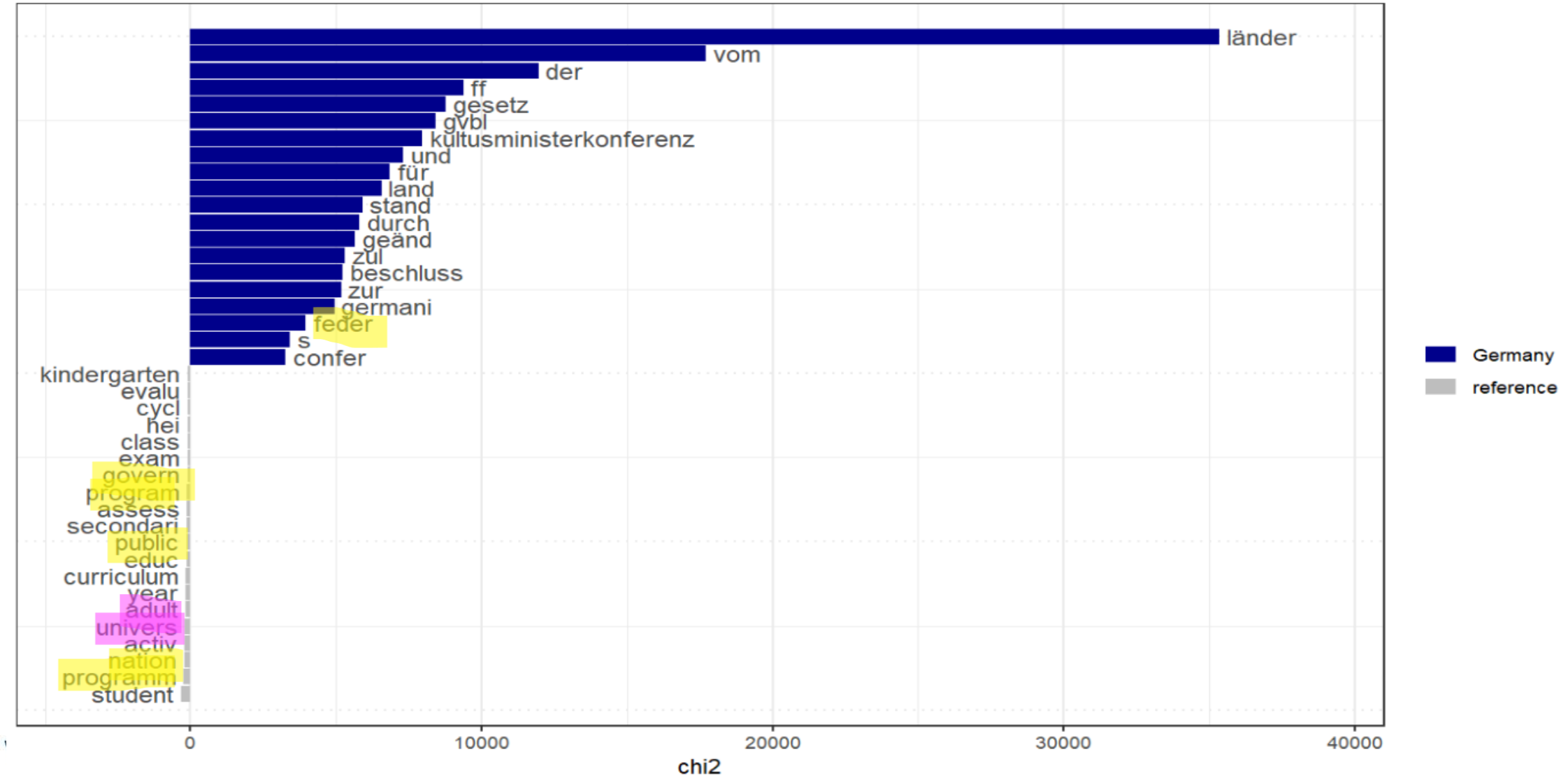




Co-funded by the Erasmus+ Programme of the European Union

EUI DEPARTMENT OF POLITICAL AND SOCIAL SCIENCES

Plotting and interpreting keyness



Seminar Overview

Day 1 Theoretical foundations

I. (Re)visiting the Policy Process & Texts as Traces of Political Conflict and Change
II. The Politics of Education
III. Qualitative Text Analysis

Day 2 Practical Foundations & First Analyses

IV. A Newcomer's Guide to Computational Text Analysis
V. Working with R: Introduction to R & Quanteda
VI. Workflow & Descriptive Analyses

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

Seminar Overview

Day 3 Text-as-Data Methods

VII. Word Counts

VIII. Dictionaries

IX. Supervised- and Unsupervised Machine Learning for Text Analysis

Day 4 Application

X. Optional Input (I propose: Plotting results)

XI. In-Class Project: Applied CTA and the politics of education

XII. Presentation, Reflection and Outlook

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

Questions before we start?

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

What we will do today

Day 3 Text-as-Data Methods

1. Word Frequencies: How simple methods can trump complex ones (Martin 2018)
2. Applying Dictionary analysis
3. Supervised- and Unsupervised Machine Learning for Text Analysis: Focus on topic models

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

Day 3 Session 1

Word Frequencies: How simple methods can trump complex ones (Martin 2018)

IMAGINE ALL THE PEOPLE

Literature, Society, and Cross-National Variation in Education Systems

By CATHIE JO MARTIN*

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

Martin 2018: Literature, Society, and Cross-National Variation in Education Systems

- What is the research question of Martin (2018)?
- What text data does she use?
- Who created the text data? What does this mean for her analysis?
- How does she analyse the text?
- What does she find?
- Do you think this is a good research design? Is the analysis convincing?

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

Lessons from Martin (2018):

1. Find the right method, not the most hyped / novel!
   - The right method depends on your research question and available data
2. No need to choose sides on the qualitative/quantitative divide. We can gain great leverage from using both.
3. Be open about possible data sources.

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

Day 3 Session 2

Dictionaries

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

What are dictionaries?

- **Keys** that stand for a theorized concept
- **Values** that represent the keys empirically

- Example: Food (key); bread, apple, banana, chocolate, potatoes (values)

- How to measure the prevalence of a concept in a corpus?
  - We count the frequency of dictionary features (values)

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

Using dictionaries

- **Advantages:**
  - Easy to learn
  - Time and cost-efficient
  - Perfectly reliable if compared to human coding
- **Disadvantages:**
  - Human involvement might cause bias
  - Creating new dictionaries can be difficult and lead to biased findings
  - Dictionary analyses depend on individual words. In small data sets individual words can have large effects. Also: Problem of negation: “eat animals” or “don’t eat animals”
- **Challenge**: Dictionaries must be exhaustive but also unambiguous. Dictionary creation must be transparent

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

Existing Dictionaries

- Ideologies - Pauwels (2011): Measuring Populism: A Quantitative Text Analysis of Party Literature in Belgium
- Uses frequency of word use to measure if text expresses ideology

Table A2. Dictionary

Dictionary	Dutch words	Translation
Conservatism	christ*; geloof; gezin; kerk; normen; porn*; seks*; waarden	christ*; belief; family; church; norm; porn*; sex*; values
Environment	ecol*; groene*; klimaat*; milieu*; opwarming	ecol*; green*; climate*; environment*; heating
Immigration	marok*; turk; allocht*; asiel*; halal*; hoofddoek*; illega*; immigr*; islam*; koran; moslim*; vreemd*	moroc*; turk; allocht*; asylum*; halal*; scarf*; illega*; immigr*; islam*; koran; muslim*; foreign*

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

Existing Dictionaries

- Are recommendation letters gendered?
- Schmader et al. (2007): A Linguistic Comparison of Letters of Recommendation for Male and Female Chemistry and Biochemistry Job Applicants

Study-Defined Dimension Dictionaries Standout words: excellen*, superb, outstanding, unique, exceptional, unparalleled, *est, most, wonderful, terrific*, fabulous, magnificent, remarkable, estraordinar*, amazing, supreme*, unmatched

Ability words: talent*, intell*, smart*, skill*, ability, genius, brilliant*, bright*, brain*, aptitude, gift*, capacity, propensity, innate, flair, knack, clever*, expert*, proficient*, capable, adept*, able, competent, natural*, inherent*, instinct*, adroit*, creative*, insight*, analytical

Grindstone words: hardworking, conscientious, depend*, meticulous, thorough, diligen*, dedicate, careful, reliab*, effort*, assiduous, trust*, responsib*, methodical, industrious, busy, work*, persist*, organiz*, disciplined

Teaching words: teach, instruct, educat*, train*, mentor, supervis*, adviser, counselor, syllabus, syllabus, course*, class, service, colleague, citizen, communicate*, lectur*, student*, present*, rapport

Research words: research*, data, study, studies, experiment*, scholarship, test*, result*, finding*, publication*, publish*, vita*, method*, scien*, grant*, fund*, manuscript*, project*, journal*, theor*, discover*, contribution*

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

Disclaimer:

Too much code and too quick? Do not worry, you go through the same code with lots of time

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

Workflow: Working with existing dictionaries – Step 1: Load and Pre-Process

1.
2.
3.
4.
5. & 6.

```
debate_data <- read.csv("C:/Users/milan/OneDrive - Istituto Universitario Europeo/Promotion/6. Side Projects and J
obs/2023 Teaching CTA in Marburg/Teaching Material/Data & Scripts/us_election_2020_1st_presidential_debate.csv")

corp_debate <- corpus(debate_data)

toks_debate <- tokens(corp_debate,  remove_punct=T,  remove_numbers=T, remove_symbols=T)
dfm_debate <- dfm(toks_debate)
dfm_debate <- dfm_remove(dfm_debate,stopwords("en"))
dfm_debate <- dfm_wordstem(dfm_debate,"en")
```

What do numbers 1-6 stand for?

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

Data: US Presidential Election Debate 2020

	speaker	minute	text
1	Chris Wallace	01:20	Good evening from the Health Education Campus of Case Western Reserve University and the Cleveland Clinic. I'm Chris Wallace of Fo
2	Chris Wallace	02:10	This debate is being conducted under health and safety protocols designed by the Cleveland Clinic, which is serving as the Health Sec
3	Vice President Joe Biden	02:49	How you doing, man?
4	President Donald J. Trump	02:51	How are you doing?
5	Vice President Joe Biden	02:51	I'm well.
6	Chris Wallace	03:11	Gentlemen, a lot of people been waiting for this night, so let's get going. Our first subject is the Supreme Court. President Trump, you
7	President Donald J. Trump	04:01	Thank you very much, Chris. I will tell you very simply. We won the election. Elections have consequences. We have the Senate, we have
8	President Donald J. Trump	04:53	And we won the election and therefore we have the right to choose her, and very few people knowingly would say otherwise. And by t
9	Chris Wallace	05:22	President Trump, thank you. Same question to you, Vice President Biden. You have two minutes.
10	Vice President Joe Biden	05:29	Well, first of all, thank you for doing this and looking forward to this, Mr. President.
11	President Donald J. Trump	05:34	Thank you, Joe.
12	Vice President Joe Biden	05:36	The American people have a right to have a say in who the Supreme Court nominee is and that say occurs when they vote for United S
13	Vice President Joe Biden	06:12	Now, what's at stake here is the President's made it clear, he wants to get rid of the Affordable Care Act. He's been running on that, he
14	Vice President Joe Biden	07:08	And that ended when we, in fact, passed the Affordable Care Act, and there's a hundred million people who have pre-existing conditio
15	President Donald J. Trump	07:34	There aren't a hundred million people with pre-existing conditions. As far as a say is concerned, the people already had their say. Okay,
16	Vice President Joe Biden	08:01	He's elected to the next election.
17	President Donald J. Trump	08:02	During that period of time, during that period of time, we have an opening. I'm not elected for three years. I'm elected for four years. J

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

Workflow: Dictionaries – Step 2: Load Dictionary

- If working with pre-existing dictionary, safe dictionary in working directory
- Load with dictionary function
- Here: Newsmap geography dictionary

```
newsmap_dict <- dictionary(file = "english.yml",
                           format = "YAML")
```

```
newsmap_dict
```

```
## Dictionary object with 5 primary key entries and 3 nested levels.
## - [AFRICA]:
##   - [EAST]:
##     - [BI]:
##       - burundi, burundian*, bujumbura
##     - [DJ]:
##       - djibouti, djiboutian*
##     - [ER]:
##       - eritrea, eritrean*, asmara
##     - [ET]:
##       - ethiopia, ethiopian*, addis ababa
##     - [KE]:
##       - kenya, kenyan*, nairobi
##     - [KM]:
##       - comoros, comorian*, moroni
##     [ reached max_nkey ... 13 more keys ]
##   - [MIDDLE]:
##     - [AO]:
##       - angola, angolan*, luanda
##     - [CD]:
##       - democratic republic congo, dr congo, drc, democratic republic congolese, dr congolese, kinshasa
```

Workflow: Dictionaries – Step 2: Load Dictionary

- Words in dictionary: countries, nationalities, capitals
- What kind of questions could we answer with this this dictionary?

```
newsmap_dict
```

```
## Dictionary object with 5 primary key entries and 3 nested levels.
## - [AFRICA]:
##   - [EAST]:
##     - [BI]:
##       - burundi, burundian*, bujumbura
##     - [DJ]:
##       - djibouti, djiboutian*
##     - [ER]:
##       - eritrea, eritrean*, asmara
##     - [ET]:
##       - ethiopia, ethiopian*, addis ababa
##     - [KE]:
##       - kenya, kenyan*, nairobi
##     - [KM]:
##       - comoros, comorian*, moroni
##     [ reached max_nkey ... 13 more keys ]
##   - [MIDDLE]:
##     - [AO]:
##       - angola, angolan*, luanda
##     - [CD]:
##       - democratic republic congo, dr congo, drc, democratic republic congolese, dr congolese, kinshasa
```


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

Workflow: Dictionaries – Step 3: Apply Dictionary

- Use dfm_lookup() command to match dictionary words with dfm
- What do we see here? For each document we test if a feature from the dictionary is present or not

```
dict_dfm_results <- dfm_lookup(dfm_debate,newsmap_dict)
dict_dfm_results[610:615,111:119]
```

```
## Document-feature matrix of: 6 documents, 9 features (96.30% sparse) and 2 docvars.
##          features
## docs      AMERICA.NORTH.GL AMERICA.NORTH.PM AMERICA.NORTH.US ASIA.CENTER.KG
##   text610                0                0                2               0
##   text611                0                0                0               0
##   text612                0                0                1               0
##   text613                0                0                0               0
##   text614                0                0                0               0
##   text615                0                0                0               0
##          features
## docs      ASIA.CENTER.KZ ASIA.CENTER.TJ ASIA.CENTER.TM ASIA.CENTER.UZ
##   text610              0              0              0              0
##   text611              0              0              0              0
##   text612              0              0              0              0
##   text613              0              0              0              0
##   text614              0              0              0              0
##   text615              0              0              0              0
##          features
```


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

Workflow: Dictionaries – Step 4: Frequency analysis

- Using the textstat_frequency() command we can get the frequency and a ranking of the prevalence of regions
- So: What might be the geopolitical focus of the 2020 US presidential campaigns?
- We can apply dictionaries to either dfms or tokens

```
dict_dfm_results %>% textstat_frequency()
```

```
##                feature frequency rank docfreq group
## 1   AMERICA.NORTH.US        44    1      35   all
## 2       ASIA.EAST.CN        10    2       9   all
## 3     EUROPE.EAST.RU         6    3       6   all
## 4      ASIA.SOUTH.IN         2    4       2   all
## 5  AMERICA.CENTER.MX         1    5       1   all
## 6   AMERICA.SOUTH.BR         1    5       1   all
## 7       ASIA.EAST.JP         1    5       1   all
## 8       ASIA.WEST.IQ         1    5       1   all
## 9    EUROPE.NORTH.IE         1    5       1   all
## 10    EUROPE.WEST.DE         1    5       1   all
```


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

Beyond frequencies: Dictionaries with levels

- Some dictionaries have levels
- Newsmap has 3 levels: Continents, regions and countries
- In script, you can explore levels further

```
dfm_debate %>% dfm_lookup(newsmap_dict, levels = 2) %>%
  textstat_frequency()
```

```
##   feature frequency rank docfreq group
## 1   NORTH        45    1      35   all
## 2    EAST        17    2      13   all
## 3   SOUTH         3    3       3   all
## 4    WEST         2    4       2   all
## 5  CENTER         1    5       1   all
```


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

Analysing by group: Comparing Trump's and Biden's geopolitical focus

```
dfm_debate %>% dfm_lookup(newsmap_dict, levels = 1) %>%
  textstat_frequency(groups=speaker)
```

```
##    feature frequency rank docfreq                        group
## 1 AMERICA        11    1      10                 Chris Wallace
## 2    ASIA        11    1       8 President Donald J. Trump
## 3 AMERICA         8    2       6 President Donald J. Trump
## 4  EUROPE         7    3       7 President Donald J. Trump
## 5 AMERICA        27    1      20  Vice President Joe Biden
## 6    ASIA         3    2       3  Vice President Joe Biden
## 7  EUROPE         1    3       1  Vice President Joe Biden
```

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

And now you: Download script “4_Dictionaries” and file “us_election_2020_1st_presidential_debate”

We have covered everything up to “dictionary creation”

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

Dictionary creation

- Technically (like below): dictionary(list(x=c(“x”),x=c(“x”)))
- Substantively (not like below): Exhaustive and unambiguous

```
education_dictionary <- dictionary(list(higher_education=c("universit*","higher education","professor*"),school_ed
ucation=c("school*","teache*")))
education <- dfm_lookup(dfm_debate,education_dictionary)
education %>% textstat_frequency()
```

```
##            feature frequency rank docfreq group
## 1 school_education        12    1       8   all
## 2 higher_education         2    2       1   all
```

Dictionary creation: Best practice

1. Creating a top-features list
   - From data
2. Deductively selecting features
   - Which of these explain the theoretical concept well (unambiguously)
3. Manual validation
   - Human coders determine whether feature should be included or not

Political Communication

ISSN: 1058-4609 (Print) 1091-7675 (Online) Journal homepage: https://www.tandfonline.com/loi/upcp20

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

Ashley Muddiman, Shannon C. McGregor & Natalie Jomini Stroud

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

Day 3 Session 3

Supervised and unsupervised machine learning for text analsysis: Focus on topic modelling

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

Computational Text Analysis and Human Involvement

- (Computational) Text Analysis requires different levels of human involvement depending on the selected method
- 100% human involvement: Qualitative Content Analysis (Day 2)
- 1-99% human involvement: Supervised methods (dictionaries, supervised machine learning)
- (Almost) no human involvement (unsupervised machine learning)

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

Machine Learning

- "A computer program is said to learn from **experience E** with respect to some class of **tasks T** and **performance measure P** if its performance at tasks in T, as measured by P, improves with **experience E**“ Tom Mitchell, Machine Learning, 1997:
- In text analysis generally used for classification tasks
- Task T = classification of texts
- Experience E = pre-classified text
- Performance measures P = correctly classified texts

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

Machine Learning: Classifiers

- We learn which text features predict categories of interest through classifiers (algorithms)
- Logic is simple and similar to regressions:
  - Regressions: Does the number of universities predict a highly skilled population?
  - Dictionary: Does the word ‘university’ predict a highly skilled population?
  - Classifier: Does the word ‘university’ predict a highly skilled population? Does the word ‘house’ predict a highly skilled population? Does the word…?

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

Machine learning in text-as-data

- **Generalization**: Classifiers learn how to correctly predict output from known inputs but can also classify unknown data.
- **Overfitting**: By learning from data, it might work very well for the known sample, but not so well for unseen data.
- Always have in mind what data a model was trained with when interpreting results.

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

Machine learning in text-as-data

- **Unsupervised Machine Learning:** Requires no prior information, training set, labelled texts or seed words
  - Pick a model/classifyer, input data, (potentially adjust parameters like number of topics in topic modelling)
- **Supervised Machine Learning:** Some degree of human involvement. Often, we label data to train a model and then use it to classify large amounts of data
- Today: **Topic models** – there are both supervised and unsupervised forms, so you will see what there different forms of machine learning to in practice

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

Topic Models

- Topic models are algorithms for discovering the main "themes" in an unstructured corpus
- What is a topic?
  - Google: “a matter dealt with in a text, discourse, or conversation; a subject”
  - Topic model definition: a probability distribution over a fixed word vocabulary
- Practically this means topics are made up of words. Each word has a probability that estimates how likely it belongs to a certain topic.

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

Topic Models: Example

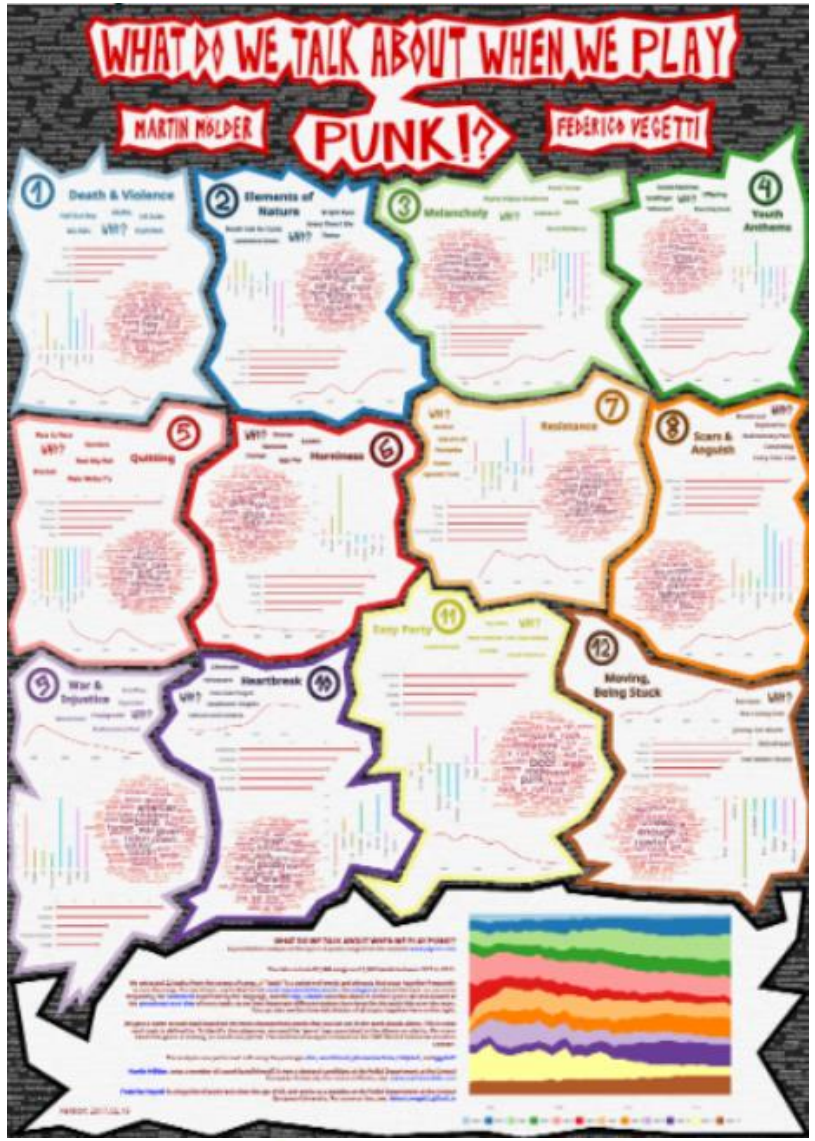
- Imagine a corpus that includes the following words
  - Gene, dna, genetic, data, number, computer
- Texts about genetics will more frequently use the words gene, dna and genetic, but less frequently use data, number and computer.
- Texts about computation might more frequently use the words data, number and computer
- Probabilities of each word to belonging to either the topic genetics or computation:

Topic	gene	dna	genetic	data	number	computer
Genetics	0.4	0.25	0.3	0.02	0.02	0.01
Computation	0.02	0.01	0.02	0.3	0.4	0.25

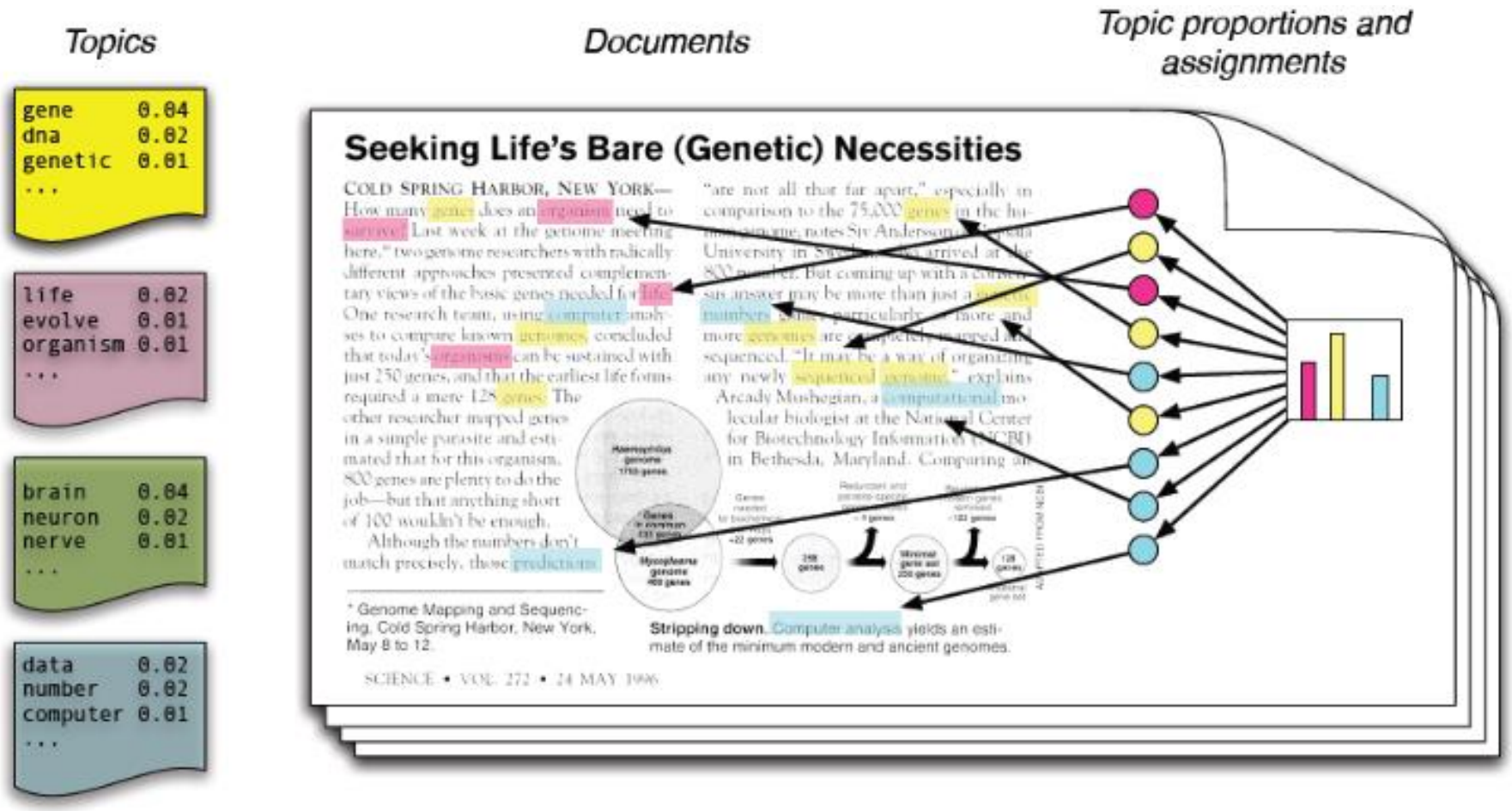
Co-funded by the Erasmus+ Programme of the European Union

Topic Models: Typical use

- To study trajectories of large topics
  - In scientific journals, archives, social media etc
- To study different frames of the same debate
  - Highlighting different topics, using different words
- To get first impression of large amounts of data
- One example: Punk lyrics!

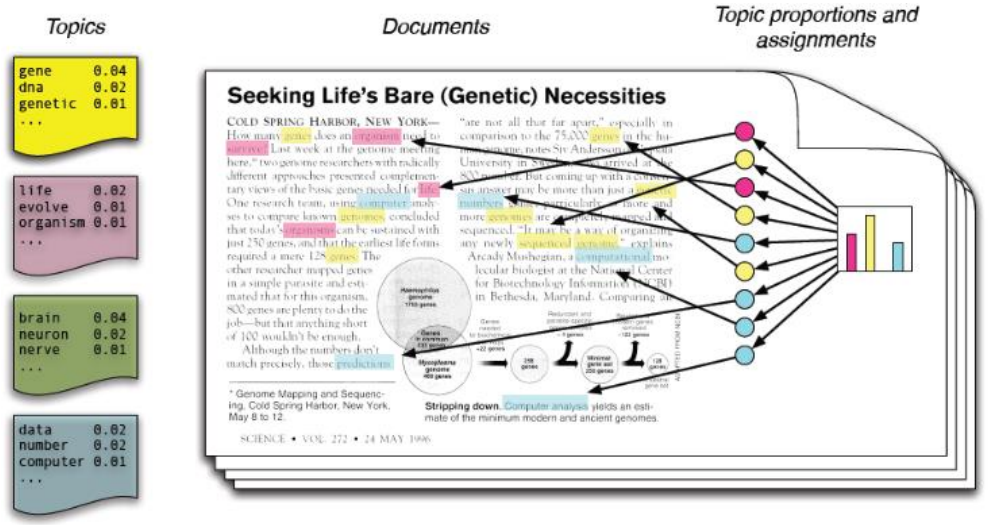


Topic Models: Latent Dirichlet Allocation (LDA)



Topic Models: Latent Dirichlet Allocation (LDA)

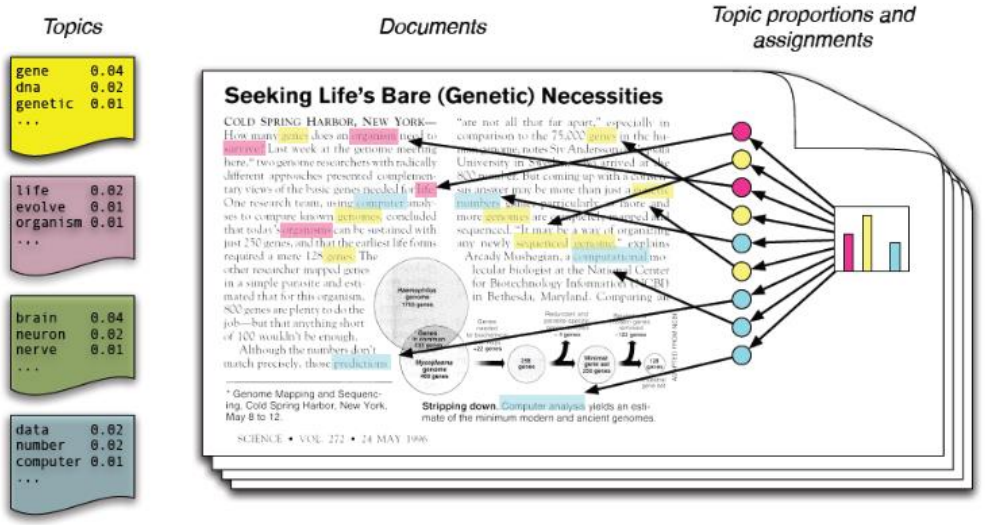
- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics



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

Topic Models: Latent Dirichlet Allocation (LDA)

- LDA topic modelling has two goals:
  1. For each document, allocate its words to as few topics as possible.
  2. For each topic, assign high probability to as few words as possible.
- Modell optimizes this trade-off
- Result will depend in data input, not theoretical assumptions
- Can be unsupervised or supervised (seeded). We will learn about both.



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

LDA Topic Models: Workflow

- First step: Selecting data set, loading data and pre-processing
- We work with the Eurydice data again in the presentation as you already know the data
- What are the steps of pre-processing?

```
eurydice <- readRDS("C:/Users[redacted]
Scripts/eurydice.RDS")

corp_eurydice <- corpus(eurydice)

toks_eurydice <- tokens(corp_eurydice,  remove_punct=T,  remove_numbers=T, remove_symbols=T)

dfm_eurydice <- dfm(toks_eurydice)
dfm_eurydice <- dfm_remove(dfm_eurydice,stopwords("en"))
dfm_eurydice <- dfm_wordstem(dfm_eurydice,"en")
```


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

LDA Topic Models: Workflow

- Creating topic models without further supervision requires just two decisions:

1. How do I need to clean the data?
2. How many topics should my model predict?
   - There are different ways to determine the optimal number of topics. We will table this question for now to keep it simple.

- Next execute the command x <- textmodel_lda(x, k = x)
- What do the different x stand for?

```
# run the model. This will take time
tmod_lda <- textmodel_lda(dfm_eurydice, k = 10)
```


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

LDA Topic Models: Workflow

- The model determines the optimal distribution of words over topics for the given dfm.
- The first output we get is a list of topics and the words that with a high probability predict that a text belongs to a topic.
- We can get list of terms associated with each topic to see if it makes sense.
- Remember: Unsupervised models are guided by an optimization problem and the data we provide, not theoretical conceptions.

```
# what are the terms associated with each topic
terms(tmod_lda, 20)
```


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

Topic Models

- What can we learn about the data base from this unsupervised topic modelling?

	topic1	topic2	topic3	topic4	topic5	topic6
[1,]	"teacher"	"school"	"educ"	"educ"	"educ"	"educ"
[2,]	"staff"	"educ"	"institut"	"train"	"de"	"develop"
[3,]	"school"	"pupil"	"school"	"vocat"	"institut"	"european"
[4,]	"teach"	"secondari"	"qualiti"	"adult"	"access"	"project"
[5,]	"work"	"year"	"evalu"	"qualif"	"school"	"mobil"
[6,]	"profession"	"student"	"higher"	"learn"	"republ"	"languag"
[7,]	"educ"	"subject"	"law"	"provid"	"o"	"programm"
[8,]	"employ"	"class"	"govern"	"cours"	"czech"	"countri"
[9,]	"head"	"special"	"ministri"	"programm"	"slovak"	"intern"
[10,]	"servic"	"primari"	"respons"	"develop"	"avail"	"nation"
[11,]	"year"	"grade"	"nation"	"profession"	"act"	"cooper"
[12,]	"manag"	"teach"	"council"	"level"	"la"	"support"
[13,]	"appoint"	"teacher"	"public"	"skill"	"et"	"student"
[14,]	"may"	"languag"	"system"	"employ"	"du"	"activ"
[15,]	"salari"	"curriculum"	"bodi"	"compet"	"isc"	"new"
[16,]	"train"	"assess"	"establish"	"institut"	"des"	"increas"
[17,]	"requir"	"general"	"regul"	"system"	"v"	"learn"
[18,]	"posit"	"compulsori"	"state"	"guidanc"	"provid"	"particip"
[19,]	"condit"	"upper"	"assur"	"centr"	"facil"	"cultur"
[20,]	"period"	"need"	"develop"	"work"	"last"	"strategi"

	topic7	topic8	topic9	topic10
[1,]	"educ"	"higher"	"feder"	"educ"
[2,]	"school"	"student"	"canton"	"school"
[3,]	"provid"	"studi"	"der"	"children"
[4,]	"learn"	"educ"	"länder"	"fund"
[5,]	"support"	"programm"	"und"	"year"
[6,]	"develop"	"univers"	"für"	"provid"
[7,]	"includ"	"institut"	"tel"	"public"
[8,]	"govern"	"degre"	"austria"	"institut"
[9,]	"qualif"	"cours"	"vom"	"age"
[10,]	"skill"	"academ"	"austrian"	"privat"
[11,]	"level"	"research"	"german"	"support"
[12,]	"framework"	"year"	"univers"	"parent"
[13,]	"fund"	"qualif"	"liechtenstein"	"grant"
[14,]	"need"	"requir"	"websit"	"municip"
[15,]	"act"	"examin"	"confer"	"child"
[16,]	"colleg"	"scienc"	"die"	"care"
[17,]	"nation"	"level"	"s"	"fee"
[18,]	"set"	"teach"	"swiss"	"kindergarten"
[19,]	"provis"	"doctor"	"educ"	"state"
[20,]	"also"	"profession"	"last"	"famili"

- Topic 1 seems to be about the **teaching profession**
- Topic 2 seems to be about **school education** but there is room for ambiguity
- Topic 3 is about **governance**
- Topic 4 about the education as means to create **human capital on the labor market**
- Topic 6 seems to be **European integration**
- Topic 8 is about **higher education**.
- Topic 9 is about federalism
- Topic 10 seems to be about the perspective of **families** and early childhood education and care (**ECEC**)
- Topic 5 and 7 hard to interpret because of language diversity in data set – lots of gibberish and very frequent words. Common denominator: It is about **education**. More cleaning could help.

	topic1	topic2	topic3	topic4	topic5	topic6
[1,]	"teacher"	"school"	"educ"	"educ"	"educ"	"educ"
[2,]	"staff"	"educ"	"institut"	"train"	"de"	"develop"
[3,]	"school"	"pupil"	"school"	"vocat"	"institut"	"european"
[4,]	"teach"	"secondari"	"qualiti"	"adult"	"access"	"project"
[5,]	"work"	"year"	"evalu"	"qualif"	"school"	"mobil"
[6,]	"profession"	"student"	"higher"	"learn"	"republ"	"languag"
[7,]	"educ"	"subject"	"law"	"provid"	"o"	"programm"
[8,]	"employ"	"class"	"govern"	"cours"	"czech"	"countri"
[9,]	"head"	"special"	"ministri"	"programm"	"slovak"	"intern"
[10,]	"servic"	"primari"	"respons"	"develop"	"avail"	"nation"
[11,]	"year"	"grade"	"nation"	"profession"	"act"	"cooper"
[12,]	"manag"	"teach"	"council"	"level"	"la"	"support"
[13,]	"appoint"	"teacher"	"public"	"skill"	"et"	"student"
[14,]	"may"	"languag"	"system"	"employ"	"du"	"activ"
[15,]	"salari"	"curriculum"	"bodi"	"compet"	"isc"	"new"
[16,]	"train"	"assess"	"establish"	"institut"	"des"	"increas"
[17,]	"requir"	"general"	"regul"	"system"	"v"	"learn"
[18,]	"posit"	"compulsori"	"state"	"guidanc"	"provid"	"particip"
[19,]	"condit"	"upper"	"assur"	"centr"	"facil"	"cultur"
[20,]	"period"	"need"	"develop"	"work"	"last"	"strategi"

	topic7	topic8	topic9	topic10
[1,]	"educ"	"higher"	"feder"	"educ"
[2,]	"school"	"student"	"canton"	"school"
[3,]	"provid"	"studi"	"der"	"children"
[4,]	"learn"	"educ"	"länder"	"fund"
[5,]	"support"	"programm"	"und"	"year"
[6,]	"develop"	"univers"	"für"	"provid"
[7,]	"includ"	"institut"	"tel"	"public"
[8,]	"govern"	"degre"	"austria"	"institut"
[9,]	"qualif"	"cours"	"vom"	"age"
[10,]	"skill"	"academ"	"austrian"	"privat"
[11,]	"level"	"research"	"german"	"support"
[12,]	"framework"	"year"	"univers"	"parent"
[13,]	"fund"	"qualif"	"liechtenstein"	"grant"
[14,]	"need"	"requir"	"websit"	"municip"
[15,]	"act"	"examin"	"confer"	"child"
[16,]	"colleg"	"scienc"	"die"	"care"
[17,]	"nation"	"level"	"s"	"fee"
[18,]	"set"	"teach"	"swiss"	"kindergarten"
[19,]	"provis"	"doctor"	"educ"	"state"
[20,]	"also"	"profession"	"last"	"famili"

LDA Topic Models: Workflow

- Let's get a list of the prevalent topics of the first 20 documents using the command 'head(topics(x), 20)'
- Many of the first texts are classified as topic 7 which was a mixture of general education words (education, school, learn*, provid*, support*, develop*). This makes sense as the first chapters are general overview chapters

```
# get prevalent topics of first 20 documents (texts)
head(topics(tmod_lda), 20)
```

```
##  text1  text2  text3  text4  text5  text6  text7  text8  text9 text10 text11
## topic7 topic7 topic6 topic7 topic7 topic6 topic6 topic1 topic6 topic6 topic6
## text12 text13 text14 text15 text16 text17 text18 text19 text20
## topic6 topic6 topic2 topic2 topic6 topic9 topic2 topic2 topic2
## 10 Levels: topic1 topic2 topic3 topic4 topic5 topic6 topic7 topic8 ... topic10
```


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

LDA Topic Models: Workflow

- We can also count how prevalent the topics are in the overall corpus by counting how many times a topic is predicted to be dominant in a document
- By far the most revalent topics are school education (topic 2) and skills / education as human capital (topic 4), followed by higher education (topic 6). Topic 9, federalism is the least frequent.

topic1	topic2	topic3	topic4	topic5	topic6	topic7	topic8	topic9	topic10
402	833	494	728	89	612	349	484	36	429

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

LDA Topic Models: Workflow

- To get these results we need to integrate the dominant topic of each document into the document feature matrix or database
- We assign a new variable ‘topic’ to the existing dfm and define that the content of this new variable is the topics of our lda model.
- With the table function we the frequency of each topic
- Use the $ between dfm and variable name to specify a variable of the dfm

```
# assign topic as a new document-level variable
dfm_eurydice$topic <- topics(tmod_lda)

# cross-table of the topic frequency
table(dfm_eurydice$topic)
```


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

Now you! Work on Script 5 “Topic Models”. Do not start with seeded topic modelling yet.

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

LDA Topic Models in Martin (2018): British writers more individualist and Denmark more collective?

TABLE 2
TOPICS IN BRITAIN AND DENMARK 1720–1770[a]

Britain

1. learn, knowledge, word, nature, man, life, great, mind, part, men, world, education, prince
2. read, history, great, book, lady, chap, Dr., trim, knowledge, Mrs., work
3. word, book, hand, write, uncle, read, count, half, head, eye, turn, world
4. word, history, great, man, good, never, lady, think, thought, miss, give, letter, write, answer, book
5. write, read, dear, sir, good, lady, letter, hut, word, think, give, hope, love, happiness

Denmark

1. people, learn, meaning, truth, give, wise, right, word, wild/lost, hand, hold, rector
2. discuss, part, Greek, writing, translate, Latin, book, philosophy, learning, indicate
3. man, wild/lost, right, same/together, old, number, people, hold, calling, learning, learn
4. king, fault, learn, God, church, hold, right, less, follow, give, love, wise
5. hand, poem, learning, name, Latin, man, philosophy, book, write, rector

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

Seeded LDA Topic Models

- In seeded LDA topic modelling, you can predefine topics using a dictionary of “seed words”.
- But we only nudge the model: Jagarlamudi et. al 2012, 205:

"importantly, we only encourage the model to follow the seed sets and do not force it. So if it has compelling evidence in the data to overcome the seed information then it still has the freedom to do so.”

- More on the theory of seeded LDA: Watanabe et al (2020: Theory-Driven Analysis of Large Corpora: Semisupervised Topic Classification of the UN Speeches

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

Seeded LDA Topic Models: Workflow

- First, we need a dictionary. You can create it just like in the dictionary analysis.
- Example in Eurydice: Public or private governance in education. Are state actors or private companies more important in education?
- General idea: More state means less stratification; more private actor involvement increases labout market match.
- Dictionary is an incomplete mixture of terms from the lexicoder policy agendas dictionary and frequent words from the text that are (relatively) unambiguous.

```
# Define a dictionary
dict_PubvPriv <- dictionary(list(public = c("public","govern*", "state","regulat*", "control",
"nation*", "assist", "benefit", "care"), private = c("privat*","enterpr*", "firm*","industr*",
"work*", "marke*","competit*", "dereg*", "autonom")))

print(dict_PubvPriv)
```

```
## Dictionary object with 2 key entries.
## - [public]:
##   - public, govern*, state, regulat*, control, nation*, assist, benefit, care
## - [private]:
##   - privat*, enterpr*, firm*, industr*, work*, marke*, competit*, dereg*, autonom
```

Seeded LDA Topic Models: Workflow

- Run the seeded lda topic model (slda) using the textmodel_seededlda() command.
  - Define the name of the model
  - Define the dfm the model uses
  - Define the dictionary that provides the seedwords. The number of topics is defined by the number of topics in your dictionary. In the example it is only two. Usually there are more topics in texts.

```
tmod_slda <- textmodel_seededlda(dfm_eurydice, dictionary = dict_PubvPriv)
terms(tmod_slda, 20)
```


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

Seeded LDA Topic Models: Workflow

- With the command “terms(tmod_slda, 20)” we can again see the 20 top terms for each topic.
- Note that two topics can be too few. Example: The topic private is captured by much more prevalent topic of school education.
- Contrastingly, the second concept (public) seems to be captured quite well. What do you think?
- The terms should be understood as a warning: The model does not really capture public versus private topic modeling. Therfore, the method is not useful for the question at hand.

	public	private
[1,]	"educ"	"school"
[2,]	"nation"	"educ"
[3,]	"public"	"work"
[4,]	"higher"	"teacher"
[5,]	"govern"	"year"
[6,]	"institut"	"student"
[7,]	"state"	"pupil"
[8,]	"train"	"secondari"
[9,]	"programm"	"teach"
[10,]	"univers"	"children"
[11,]	"develop"	"studi"
[12,]	"student"	"privat"
[13,]	"adult"	"train"
[14,]	"provid"	"programm"
[15,]	"fund"	"subject"
[16,]	"qualiti"	"provid"
[17,]	"learn"	"special"
[18,]	"qualif"	"level"
[19,]	"research"	"vocat"
[20,]	"support"	"primari"

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

Seeded LDA Topic Models: Workflow

- You can see which topic is prevalent in which document with the topics() command.
- With the head(topics()) command you get a glimpse at the first 20 topics

```
topics_slda <- topics(tmod_slda)

head(topics(tmod_slda), 20)
```

```
##   text1   text2   text3   text4   text5   text6   text7   text8   text9  text10
##  public  public  public  public  public  public  public  public  public  public
##  text11  text12  text13  text14  text15  text16  text17  text18  text19  text20
##  public  public  public private  public  public  public  public private private
## Levels: public private
```

Seeded LDA Topic Models: Workflow

- To compare the prevalence of the topics on the dfm, you can add the topics as variable to your dfm and then apply the table(dfm$topic) command.

```
dfm_eurydice$topic <- topics(tmod_slda)

table(dfm_eurydice$topic)
```

```
##
##  public private
##    2261    2195
```


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

Seeded LDA Topic Models: Best practice

- What can seeded topic modelling deliver?
- Example from Watanabe et al. (2020), “Theory-Driven Analysis of Large Corpora: Semi-supervised Topic Classification of UN General Assembly Speeches.”
- What to we see in this figure?

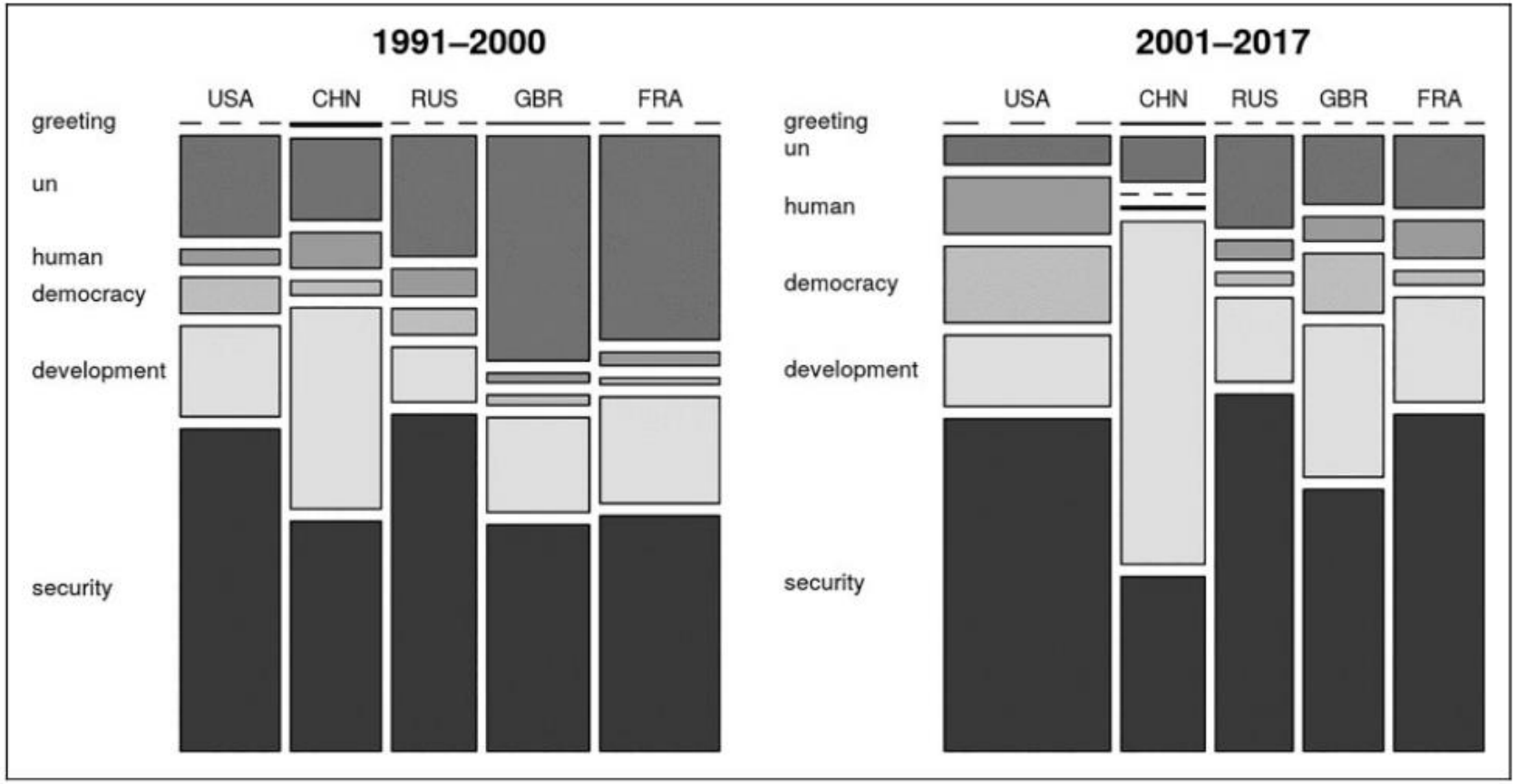


Figure 8. Topic in speeches by the Security Council permanent members before and after the 9/11 attack. *Note*. Width indicates length of speeches in number of sentences.

- Changes in UN Security Council speeches after 9/11:
  - US speeches become longer compared to other members
  - Security topic becomes more central among the speeches of the USA, GBR and FRA, but clearly less salient in the Chinese speeches

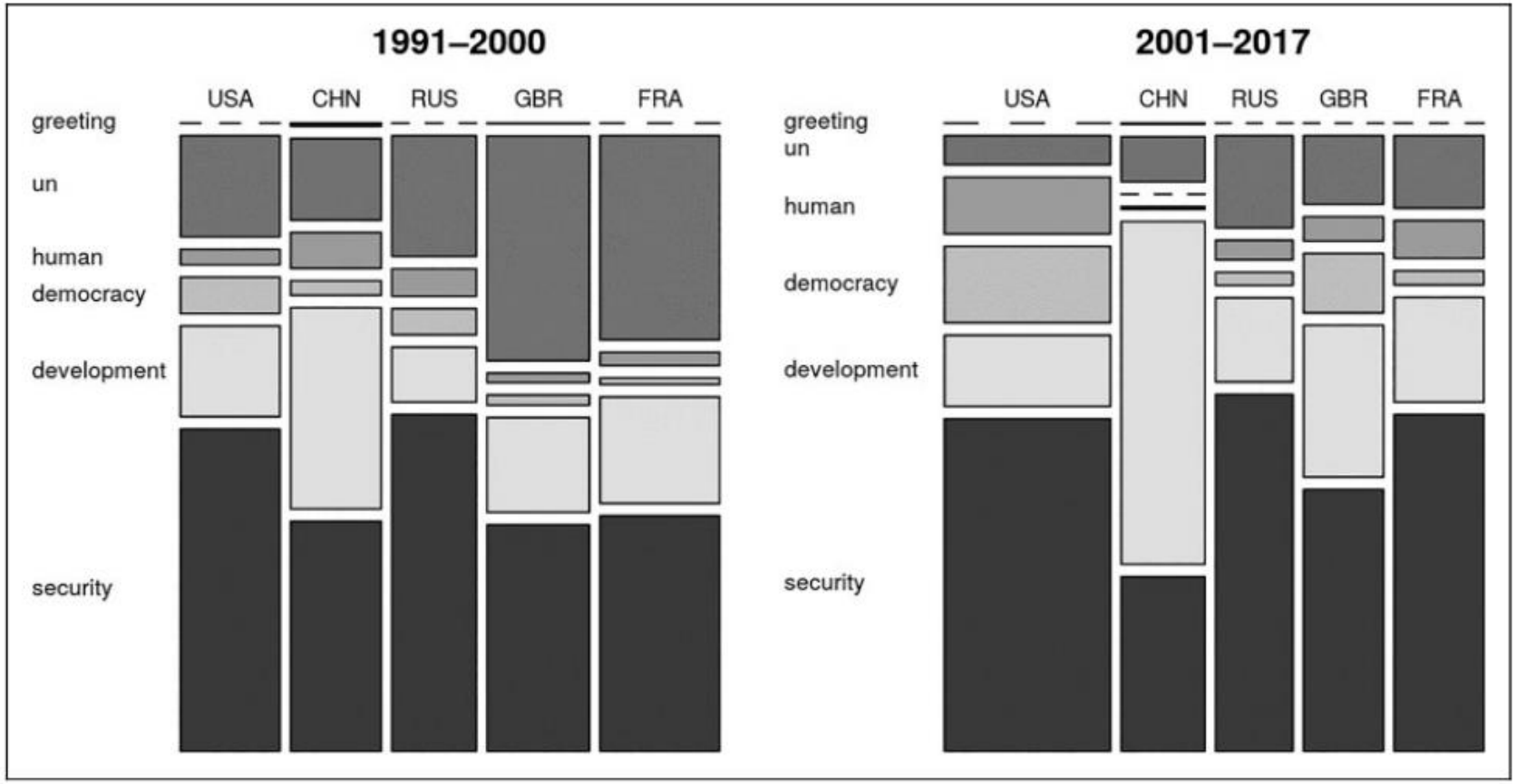


Figure 8. Topic in speeches by the Security Council permanent members before and after the 9/11 attack. *Note*. Width indicates length of speeches in number of sentences.

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

Discuss with your neighbour:

- What is the difference between supervised and unsupervised machine learning? How does this affect topic modelling?
- What are the decisions we have to take when we do seeded LDA topic modelling?

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

Now you! Work on Script 5 Topic Models

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

What are the advantages / disadvantages of the different methods we have learned?

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

Optional Input for Day 4

- Tomorrow there is some time for additional inputs. Would you like to
  - Learn how to visualize your data (I propose a focus on the packages ggplot and gt)?
  - Learn more about alternative topic models?
  - Explore advanced methods of text analysis?
  - Revisit something, we already discussed?
  - Maximize the time for project work?

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

In-Class Project:

- This could be part of your bachelor thesis / master thesis / project studies
- … or it could be just to practice what you have learned.
- The general idea: Noone can really teach you the method through slides. You learn it by applying it.
- You can bring your own data or use the data I provide
- Do not worry of you have not understood everything yet. You have a whole day and my assistance at all times.

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

Your task:

- This is not just about coding. This is about applying new methods in a research context. Therefore

1. Define a research question that has a theoretical foundation (remember the first day or remember why you are interested in your topic).
2. Import, clean and organize your data. This might be the biggest challenge for those with own data.
3. Apply those methods you think are helpful to answer your research question.
4. Interpret the findings: What have you learned?
5. Outline the limitations of your analysis and describe how you could improve the analysis.
6. Visualize your findings in a nicely knit markdown file. Prepare a short presentation of 5-10 minutes guiding us through the html file.

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

Data to work with

- Who will bring own data?
- Important: Download data before seminar. I will upload it tonight and you should have downloaded it before the seminar starts.
- Options for you:
  - UN General Assembly Speeches:
    - Corpus of texts of General Debate statements from 1970 (Session 25) to 2016 (Session 71). Docvars include country, year and session
  - ParlSpeech:
    - Full text corpora of parliamentary speeches in Austria, the Czech Republic, Germany, Denmark, the Netherlands, New Zealand, Spain, Sweden, and the United Kingdom
  - Vocational Education Reform Corpus:
    - Would contain documents of two recent VET reforms so you can compare the topics and language used in the different reforms (4 years difference).

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

Rest of today

- Catch-up time: We all have different speeds, so keep working on the scrips
- Only if you are done with all of them: Start with your project. My advice: Don’t directly start coding. Think about your research question, your data, the theoretical underpinning. Maybe read up on your topic a bit before you formulate the research question.

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