HOST
Hochschule Stralsund

Data Science

Text Mining
Geodata

Gero Szepannek

Automatisierte Inhaltsanalyse mit R

Einleitung | Grundlagen | Wort- und Textmetriken | Sentimentanalyse | Spezialisierte Lexika | Überwachtes maschinelles Lernen | Themenmodelle

Tagging, Parsing und Entitätenerkennung | Texte und Wörter als Netzwerke

Überblick

Diese Einführung gliedert sich in neun inhaltliche Kapitel, in denen wesentliche Ansätze der automatisierten Inhaltsanalyse mit R anhand von zahlreichen Beispielen vorgestellt werden. Dabei werden sog. R-Notebooks verwendet, die eine Kombination aus Erläuterungen und R-Code enthalten, welcher gemeinsam mit den hier abrufbaren Korpora und weiteren Ressourcen ausgeführt und beliebig angepasst werden kann. Die aktuellste (Entwicklungs-)Fassung der R-Notebooks findet sich auf GitHub.

Inhalt

0. Einleitung
1. Grundlagen von quanteda
2. Wort- und Textmetriken
3. Sentimentanalyse
4. themenspezifische Lexika
5. überwachtes maschinelles Lernen
6. Themenmodelle
7. Tagging, Parsing und Entitätenerkennung
8. Texte und Wörter als Netzwerke
9. Datenimport

http://inhaltsanalyse-mit-r.de/index.html
https://github.com/cbpuschmann/inhaltsanalyse-mit-r.de

Downloads

Sämtliche in dieser Einführung verwendeten R-Notebooks, Korpora und Lexika und können hier heruntergeladen werden.

Text Mining Process

1. Text data
2. Corpus
3. Text Preprocessing
4. Document-Term-Matrix
5. Analysis, e.g.
   - Predictive Modelling
   - Sentiment Analysis
   - Topic extraction

Example: Topic analysis of the program of the AfD (2017 taken from the `praddata` package)

Further reading:
ch. 24-26

Text Preprocessing

original text

1. lower case
2. punctuation
3. tokenization
4. stopwords
5. stemming

Figure taken from: Rabea Aschenbruck, Gero Szepannek (2020): Einsatz von KI zur Qualitätssicherung. Wissenschafft trifft Praxis 13: Digitale Daten, 47-51.

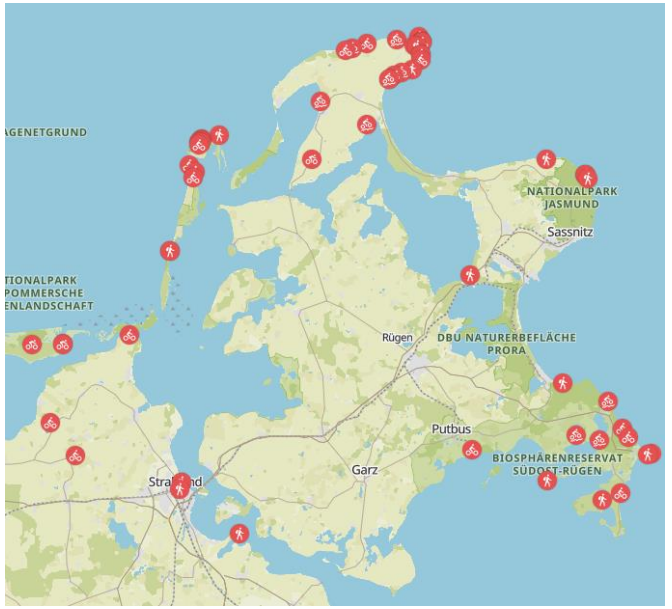
(Some) useful functions for Text Mining

Package	Function	Usage
pdftools	pdf_text()	extract Text from PDF files
stringr	str_sub()	extract substrings
base	gsub()	regular expressions
base	strsplit()	split strings w.r.t. character
quanteda		tex mining utilities
	corpus()	create a corpus
	tokens()	tokenization + preprocessing
	tokens_tolower()	
	tokens_remove()	remove tokens
	stopwords()	stop word list
	dfm()	document term matrix
	topfeatures()	print most frequent tokens
	dfm_select()	filter terms
	dfm_wordstem()	stemming
	dfm_trim()	remove sparse tokens
quanteda.textplots	textplot_wordcloud()	word cloud from dfm
ldatuning	FindTopicsNumber()	find approporiate number of topics
	FindTopicsNumber_plot()	find approporiate number of topics
stm	searchK()	find approporiate number of topics
	stm()	create topic mode (latent Dirichlet allocation)
	stm()$theta	topic memberships of the docs
	labelTopics()	extract representative terms | topic

(Some) useful functions for the analysis of GPS data

komoot

Package	Function	Usage
`sf`	`read_sf()`	read gpx data
`leaflet`	`leaflet()`	create interactive map
	`addPolylines()`	add track data to map
`ggplot2`	`ggplot()`	create plot from data
`gganimate`	`transition_reveal()`	animated plot according to a variable



Menü

KATAPULT

Magazin für ~~Eis~~, Kartografik und Sozialwissenschaft

Abo Shop Über uns

Artikel Themen Hefte Karten

WAHLPROGRAMM

CDU-Wahlgeschenke

Zur Bundestagswahl hat die CDU/CSU ein neues Wahlprogramm vorgelegt. Steuererhöhungen werden ausgeschlossen, Steuergeschenke soll es trotzdem geben.

NACHHALTIGKEIT

Neues Vorbild: Autostädte!

Wieso gab es dort weniger, wo Volkswagen, Audi und BMW sitzen? Weil einige Dienstwagen aus dem Verkehr gezogen wurden und Leute im Homeoffice gearbeitet haben. In den 22 anderen beobachteten Kommunen nahm die Zahl der registrierten Autos zu.

OLYMPIA 2021

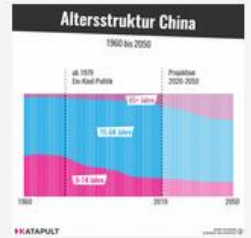
Ungeschützte Sex-Olympiade

Die Teilnehmer der Olympischen Sommerspiele in Tokio bekommen aufgrund der Pandemie keine Kondome geschenkt. Traditionell steht ihnen dieser Service seit 1988 zu. Dieses Jahr müssen alle bis zu ihrer Abreise warten, um die 00 Kondome zu

ERNEUERBARE ENERGIE

Fast alle machen's besser als Deutschland

Der Anteil von Energie aus erneuerbaren Quellen liegt in Deutschland bei 17,4 Prozent. Im Vergleich: Island ist hier bei fast 80 Prozent. An letzter Stelle ist Luxemburg mit nur knapp 7 Prozent.



GEBURTENRATE

Mehr Rentenzahler, bitte!

37 Jahre lang durften chinesische Familien nur ein Kind haben. Jetzt dürfen sie mehr - wollen aber anscheinend nicht.

Göttingen

Feinstaubbelastung in päische A hat 323 Städte

NEUE BANKNOTE

Sklavenhalter und ehemalige Sklavin teilen sich 20-Dollar-Schein

Die Anti-Sklaverei-Aktivistin Harriet Tubman soll in Zukunft den 20-Dollar-Schein der USA zieren.

REGIONALWAHLEN IN FRANKREICH

Bündnis gegen rechts bröckelt

Das rechtspopulistische Rassemblement National könnte erstmals Mehrheitspartei in französischen Regionalparlamenten werden. Ein Vorzeichen für die Präsidentschaftswahlen im kommenden Jahr?

WELTGESUNDHEITSORGANISATION

Na klar, Frauen werden ja eh irgendwann schwanger!

STUDIE

Falscher Name, kein Fußball

Im Rahmen einer Studie wurde erhoben,

SEX

Keine Einzelfälle

56 Mitglieder soll die Chatgruppe gehabt

Further Reading

- **Cornelius Puschmann** (2020): http://inhaltsanalyse-mit-r.de/index.html
- **Sebastian Sauer** (2019): Moderne Datenanayse mit R, Springer.
- **Rabea Aschenbruck, Gero Szepannek** (2020): Einsatz von KI zur Qualitätssicherung. Wissenschafft trifft Praxis 13: Digitale Daten, 47-51.
- **Bettina Grün, Kurt Hornik** (2011): topicmodels: An R Package for Fitting Topic Models, Journal od Statistical Software, https://www.jstatsoft.org/article/view/v040i13
- **Margaret Roberts, Brandon Stewart, Dustin Tingley** (2019): stm: An R Package for Structural Topic Models, Journal od Statistical Software, https://www.jstatsoft.org/article/view/v091i02
- **Robin Lovelace, Jakub Nowosad, Jannes Muenchow** (2021): Geocomputation with R, https://geocompr.robinlovelace.net/intro.html