universität wien

Advanced quantitative text analysis (2023W)

Petro Tolochko, Fabienne Lind

Day 2

universität wien

Contents (smaller changes possible)

Day	Session 1	Session 2
1	Text as data, Text representation	Feature Engineering
2	Concepts & Data	Dictionaries
3	Supervised machine learning	Unsupervised machine learning
4	Neural network models, transformers	Using LLMs
5	Multilingual text analysis, Wrap-up	Project talks

universität wien

Concepts & Data

Day 2 Session 1

universität wien

Concepts? Data?

- What do we actually want to measure?
- And where do we find it?

Zooming out: A scientific research process

- Problem: phenomenon in need of explanation
- Research question: Specification of the problem
- Connecting with existing knowledge: theoretical foundation
- Research questions/(Hypothesis formation)
- Choice of design and method
- Choice of indicators and operationalization
- Selection of the feature carriers
- Data collection
- Data evaluation and interpretation
- Publication

universität wien

From the research questions/hypotheses to key concepts

Example research question: Comparing the Twitter accounts of the altright influencers Paula Winterfeldt, Lauren Southern and Lana Lokteff, which images of women are conveyed most frequently?

Example hypothesis: On Twitch, male streamers are rated more positively by viewers than female streamers?

- Which concepts must be recorded?

universität wien

Discovery

- Idea from qualitative research
- A “method” to discover new concepts through descriptive analysis
- Grimmer et at., 2022:
    - I. Context Relevance
    - II. No Ground Truth
    - III. Concept vs. Method
    - IV. Data Separation

Principles of Discovery

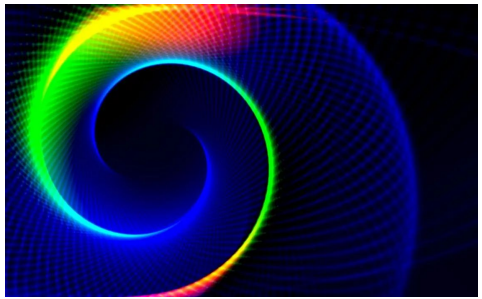
Principle 1: Context relevance. Text as data models complement theory and substantive knowledge. Contextual knowledge amplifies our ability to make computational discoveries

Who could discover relationships between symptoms, diseases, treatments, and outcomes by analyzing large volumes of medical literature, clinical notes, and patient records?

Grimmer & Stewart, 2022, Chapter 10

Principles of Discovery

Principle 2: No ground truth. There is no ground truth conceptualization; only after a concept is fixed can we talk meaningfully about it being right or wrong



Grimmer & Stewart, 2022, Chapter 10

Principles of Discovery

Principle 3: Judge the concept, not the method. The method you used to arrive at a conceptualization does not matter for assessing the concept's value – its utility does

Do the following four groups help to group customers in useful categories? And for whom and what is this of use?

Grimmer & Stewart, 2022, Chapter 10

Principles of Discovery

Principle 4: Separate data is best.
Ideally after data is used for discovery it should be discarded in favor of new data for confirming/testing discoveries.

Grimmer & Stewart, 2022, Chapter 10

universität wien

Identify concepts/interests of your own work

Based on your currently defined research question(s)/hypothesis(s)/interests:

- Which concepts/variables will you record empirically with text analysis?
  - How do you define these concept(s)?
  - What options/categories do your concept(s) have (if applicable)
- What are your discovery interests?
- Take notes
- If necessary, specify your research question(s)/hypothesis(s) a little further in order to be able to do the exercise or work with the following research question

universität wien

Peer feedback on the variables

- Form teams of 3
- Present your currently defined research question(s)/hypothesis(s) and which variables (including classes if applicable) you would like to record to answer them
- Give each other feedback and help each other to specify your research question(s)/hypothesis(s)/concepts further

universität wien

Codebook

The codebook is the tool you use to code your content

It is a kind of questionnaire that you use to inquiry the examined texts/photos/videos

The codebook should be detailed enough so that

- you can apply it again in the same way after some time (intracoder reliability)
- other persons (with a little training) can also use it in the same way as you (intercoder reliability)

universität wien

Codebooks and automated text analysis

Wait, codebooks are certainly relevant in manual content analysis

But what role do they have in automated text analysis?

universität wien

Codebooks and automated text analysis

Wait, codebooks are certainly relevant in manual content analysis

But what role do they have in automated text analysis?

- Validation
- Documentation

universität wien

Codebook development

Iterative process:

- First draft based on the variables identified in the research question(s)/hypothesis(s).
- Tip: take other studies as a model (cite!)
- Code material examples using the draft codebook
- Then refine/edit the codebook

Example:

Key variables

In the following, we will define and describe the key variables of this media analysis as well as the tools that allowed us to measure them and annotate each article accordingly. All our tools make use of a so-called dictionary approach (Lind et al., 2019). Such an approach applies a top-down procedure, where texts are searched based on a predefined list of words and phrases that reflect the concept of interest. The rate at which specific keywords then appear (together) in a text are used to classify documents into substantive categories.

Intra-EU mobility

One key variable identifies whether an article refers to "Intra-EU mobility" (`european mobil`). The annotated data set thus contains a variable that shows if an article mentions migration within the EU and/or Schengen area (coded with 1) or whether this is not the case (coded with 0).

Concept Definition

Here, "Intra-EU Mobility" is an umbrella term designed to capture all news stories referring to migration or people migrating (past/present/future) from a territory in the EU or Schengen area to another territory in the EU or Schengen area. Manual coders were provided with a list of EU or Schengen Area States[3], a definition of what was defined as migration (see page 7), and additional instructions such as:

If people from territories that are not included in the list migrate, this is considered "Non-European Migration" and is thus not to be coded as "Intra-EU Mobility". It is important to keep in mind that migrants are not always referred to by mentioning their nationalities. Any references to a population group that is not country-specific but can certainly be attributed to intra-EU or Schengen migration is also to be coded as "Intra-EU mobility". For example, a) Northern/Central/Western Europeans/etc., b) EU-citizens/-people/-families/-nationals/etc., or c) people/families/nationals from the Eurozone or the Schengen area.

Text examples for "Intra-EU Mobility" include:

"A Swedish person moves to Germany"
"Luisa from Poland moves to the UK for work"
"Northern Europeans often migrate to neighbouring countries",
"EU- citizens have the right to move within Europe"
"Ali recently received a German passport and can now study in France"

[3] 1. Austria, 2. Belgium, 3. Bulgaria, 4. Croatia, 5. Cyprus, 6. Czech Republic, 7. Denmark (also Greenland, Faroe Islands), 8. Estonia, 9. Finland, 10. France, 11. Germany, 12. Greece, 13. Hungary, 14. Iceland, 15. Ireland, 16. Italia, 17. Latvia, 18. Lichtenstein, 19. Lithuania, 20Luxembourg, 21. Malta, 22. Netherlands, 23. Norway, 24. Poland, 25. Portugal, 26. Romania, 27. Slovakia, 28. Slovenia, 29. Spain (also Canaries), 30. Sweden, 31. Switzerland, 32. United Kingdom/UK (Northern Ireland, Scotland, Wales, England, Gibraltar)

https://doi.org/10.11587/IEGQ1B

universität wien

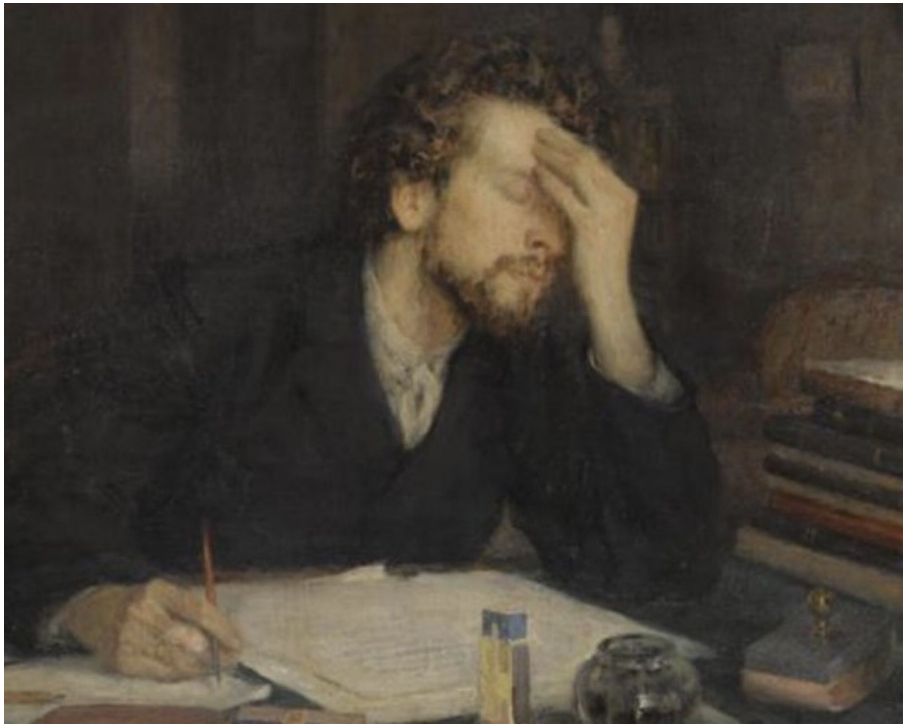
Causal inference in text

universität wien

Causal inference in text

More complex than might seem…

Treatment



Treatment

Outcome

universität wien

Multidimensionality of Text

- Dimension reduction

universität wien

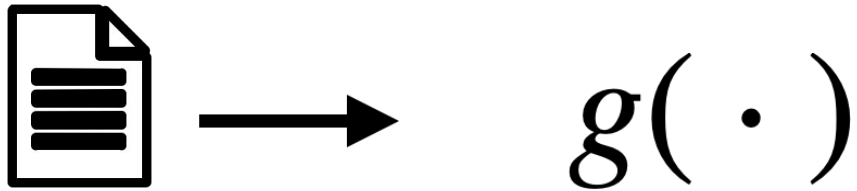
Multidimensionality of Text

- Dimension reduction
- Codebook function (e.g., Egami et al., 2022)

universität wien

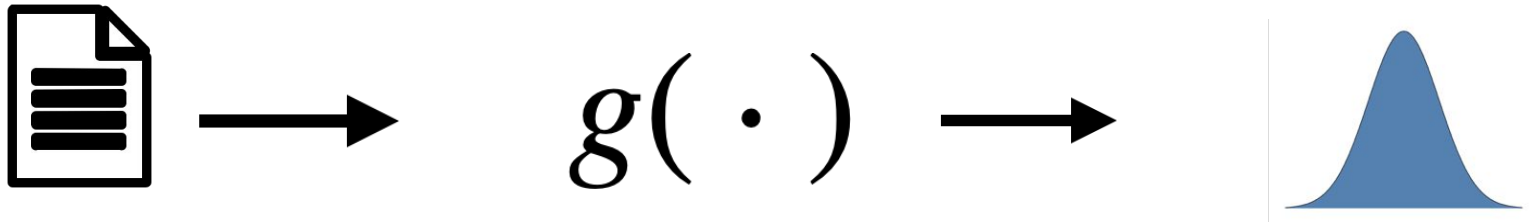
Multidimensionality of Text

- Dimension reduction
- Codebook function (e.g., Egami et al., 2022)



Multidimensionality of Text

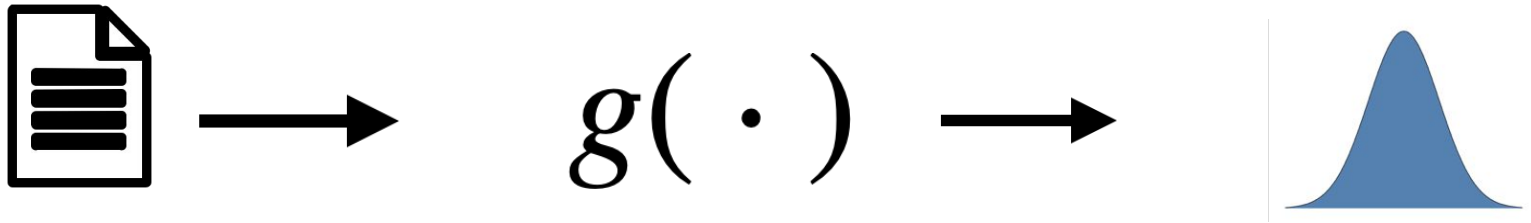
- Dimension reduction
- Codebook function (e.g., Egami et al., 2022)



universität wien

Multidimensionality of Text

- Dimension reduction
- Codebook function (e.g., Egami et al., 2022)



OBJ g(.) is ***emergent***

universität wien

Multidimensionality of Text

- (Non)Linearity
- Traditional causal inference methods assume linear relations
  - “Holding everything else constant…”
- Not always possible for text data

universität wien

Multidimensionality of Text

- Latent Representations
- (Almost) infinite amount of ways to represent a concept
- A lot of concepts represented by texts are ***latent***
- Causal inference with latent concepts is much more complicated

universität wien

Multidimensionality of Text

- Treatment ***OR*** outcome ***OR*** both
- Emergent analysis strategy
- Non-linearity
- Latent representation of concepts

universität wien

Multidimensionality of Text

- Treatment ***OR*** outcome ***OR*** both
- Emergent analysis strategy
- Non-linearity
- Latent representation of concepts

At the same time!

Zooming out: A scientific research process

- Problem: phenomenon in need of explanation
- Research question: Specification of the problem
- Connecting with existing knowledge: theoretical foundation
- Research questions/(Hypothesis formation)
- Choice of design and method
- Choice of indicators and operationalization
- Selection of the feature carriers
- Data collection
- Data evaluation and interpretation
- Publication

universität wien

Data collection plan

A recommended component of any text analysis project

Includes decisions + brief justifications on the following aspects

- Geographic and Time Restrictions
- Data sources
- Units of analysis
- Data access
- Data management
- Data validation

universität wien

Data sources

Different levels

- Channel/platform (e.g. print newspaper, Instagram)
- Outlet (e.g. BBC UK, account of organization X)

universität wien

Units of analysis

The variables are recorded at this level

Examples:

- All posts/retweets/quotes published on specific accounts
- Posts published on specific accounts that contain specific keywords
- All texts that contain certain keywords (anywhere or in the headline?)

universität wien

Data and where to get it

- Many different places
- Many different methods
  - Some are perfectly fine
  - Some are illegal

universität wien

Data and where to get it

- Many different places
- Many different methods
  - Some are perfectly fine
  - Some are ~~illegal~~ unsavoury

universität wien

For newspapers

Repositories, for example

- LexisNexis
- APA
- Factiva
- EMIS

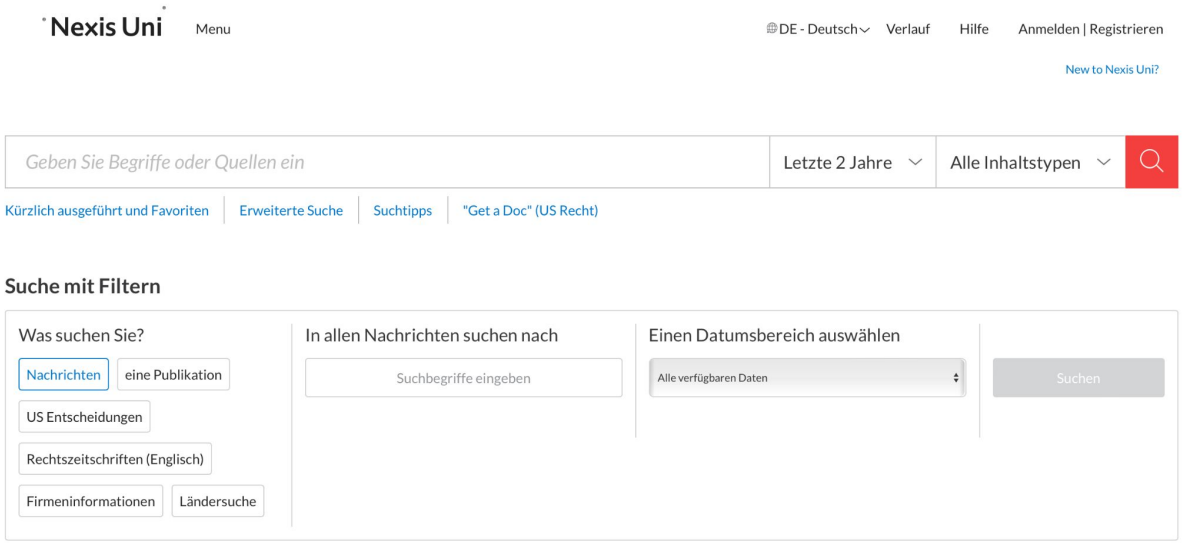
You might have access via your institutions

universität wien

LexisNexis

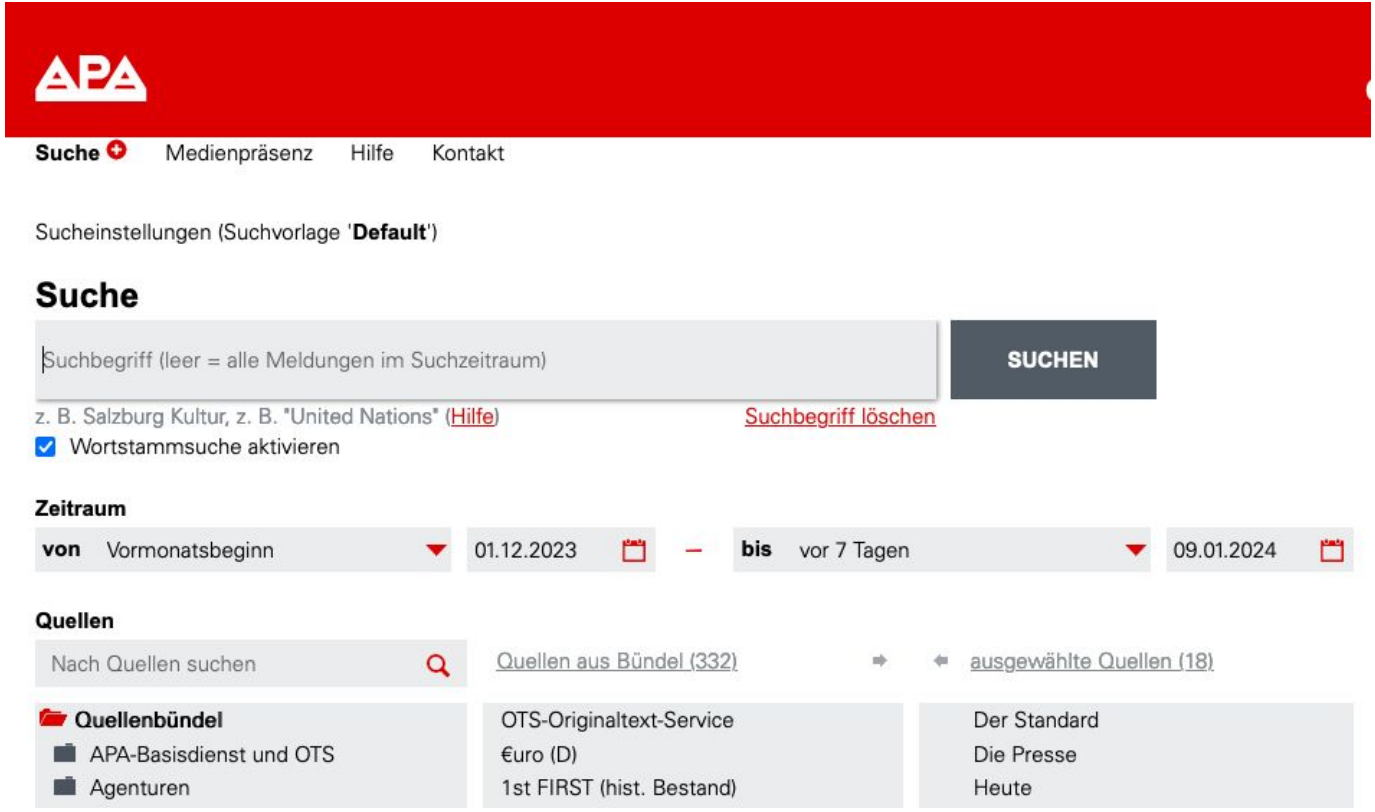
One method:

- You can download in batches of 100 documents
- You can parse the downloaded files with the R package LexisNexisTools (Gruber, 2023)



universität wien

APA



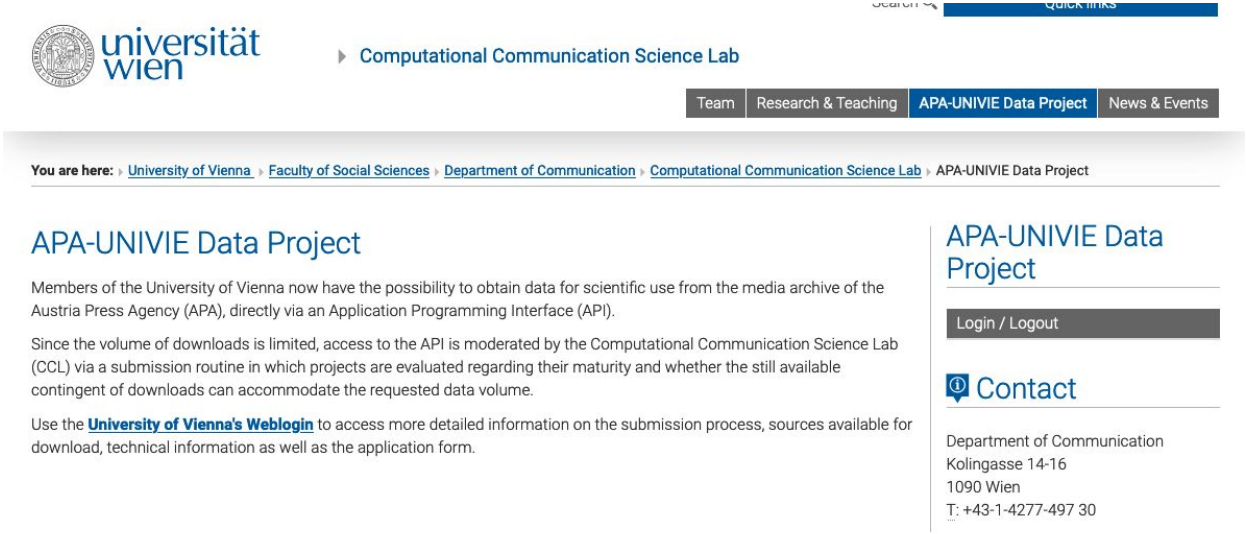
universität wien

APIs

API = Application Programming Interface

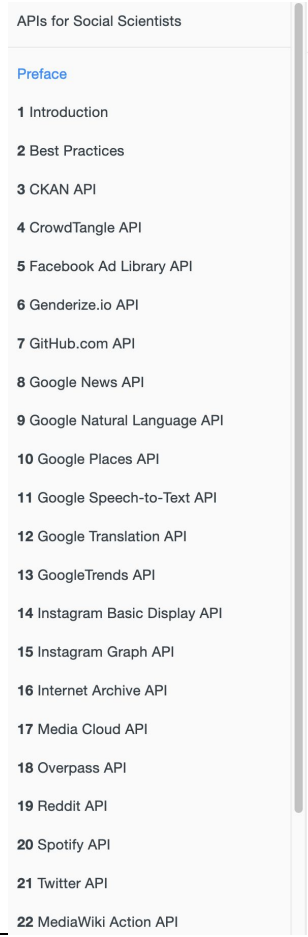
- gather data without scraping
- query a database and receive data

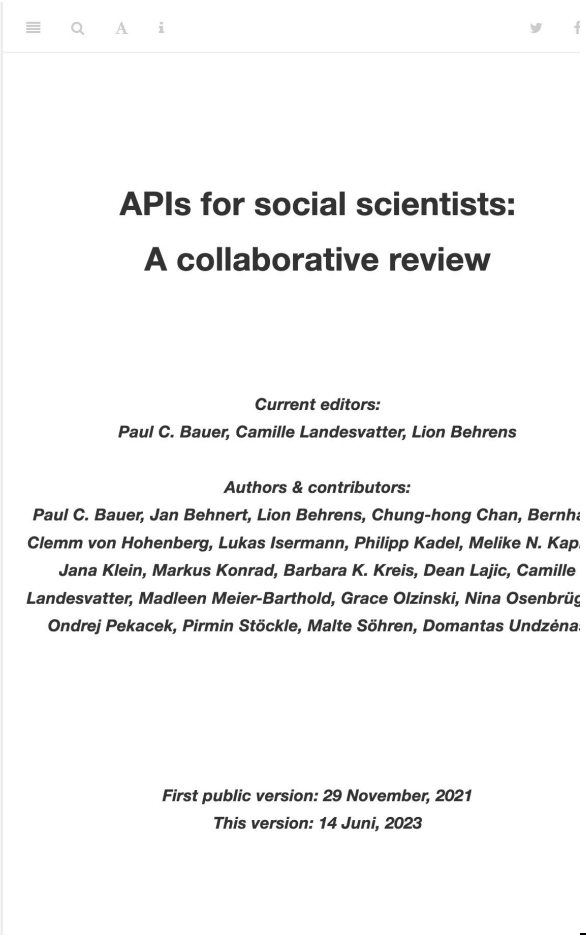
APA API



https://compcommlab.univie.ac.at/apa-univie-data-project/

API Examples





https://bookdown.org/paul/apis_for_social_scientists/

universität wien

Facebook and Instagram

- CrowdTangle: https://help.crowdtangle.com/en/articles/4296372-academics-researchers-user-hub

- Type of information Facebook: user information and contents from public Facebook pages with more than 25K Page Likes or Followers (automated via API), and to public Facebook groups with 95k+ members. Personal Facebook profiles are generally not available in CrowdTangle.

- Type of information Instagram: user information and contents from public Instagram accounts with more than 50K followers, as well as from verified accounts.

universität wien

DASHBOARDS | LIVE DISPLAYS | INTELLIGENCE | SEARCH

Notifications

Explore

Lists

+ Create List

MY FAVORITES

You don't have any favorites!

PAGES

- All Page Lists
- Business Media
- Instagram
- Tech Media
- US College Newspapers
- US General Media

Saved Searches

Saved Posts

Weights

CCC Vienna › 513 Facebook Pages

All Page Lists

Search your lists for any of these words or phrases

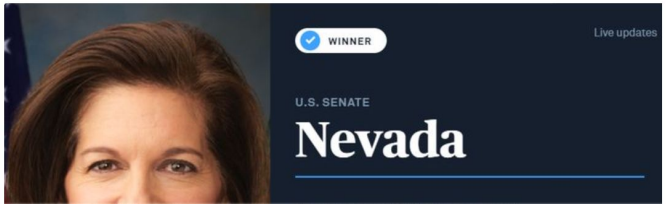
Posts | Leaderboard | Manage

Overperforming | Last 2 Hours | All Posts | More

Posts with the most interactions do not equal posts with the most content views or reach. Check out the Widely Viewed Content Report at Facebook's Transparency Center.

PBS NewsHour
36 minutes ago · 2,248,849 Followers

Democratic Sen. Catherine Cortez Masto, the first Latina ever elected to the U.S. Senate, wins reelection in Nevada, The Associated Press reports.
https://to.pbs.org/3hCjCvA



universität wien

TikTok

- API: https://developers.tiktok.com/products/research-api/

- Type of information: user information and contents of TikTok users that have set their account to public and are aged 18 and over.

universität wien

Telegram

API: https://my.telegram.org/auth?to=apps

Guidelines: https://docs.telethon.dev/en/stable/quick-references/client-reference.html

universität wien

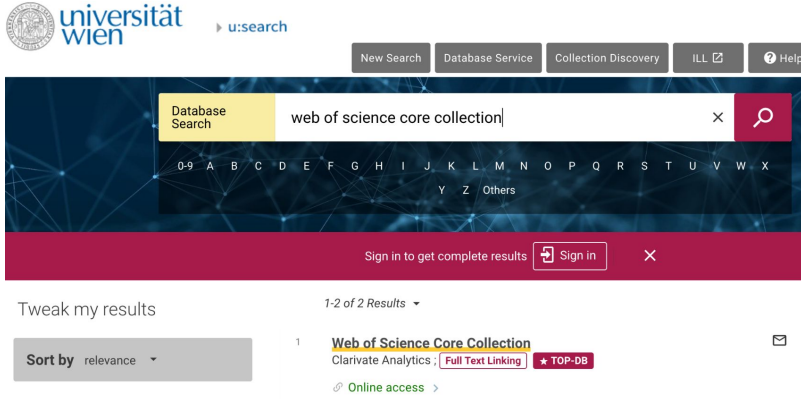
YouTube

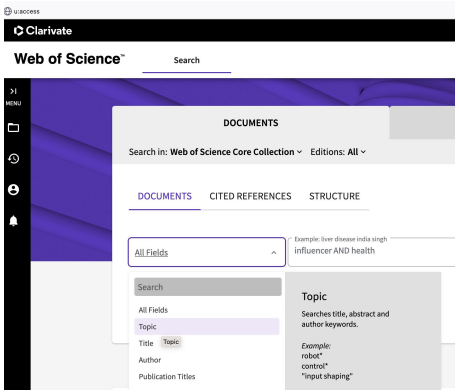
https://developers.google.com/youtube/v3

Guidelines:
https://jingwen-z.github.io/how-to-get-a-youtube-video-information-with-youtube-data-api-by-python/

Scientific Publications

- Download abstracts and meta-data from Web of Science
- Download abstracts and meta-data from Scopus



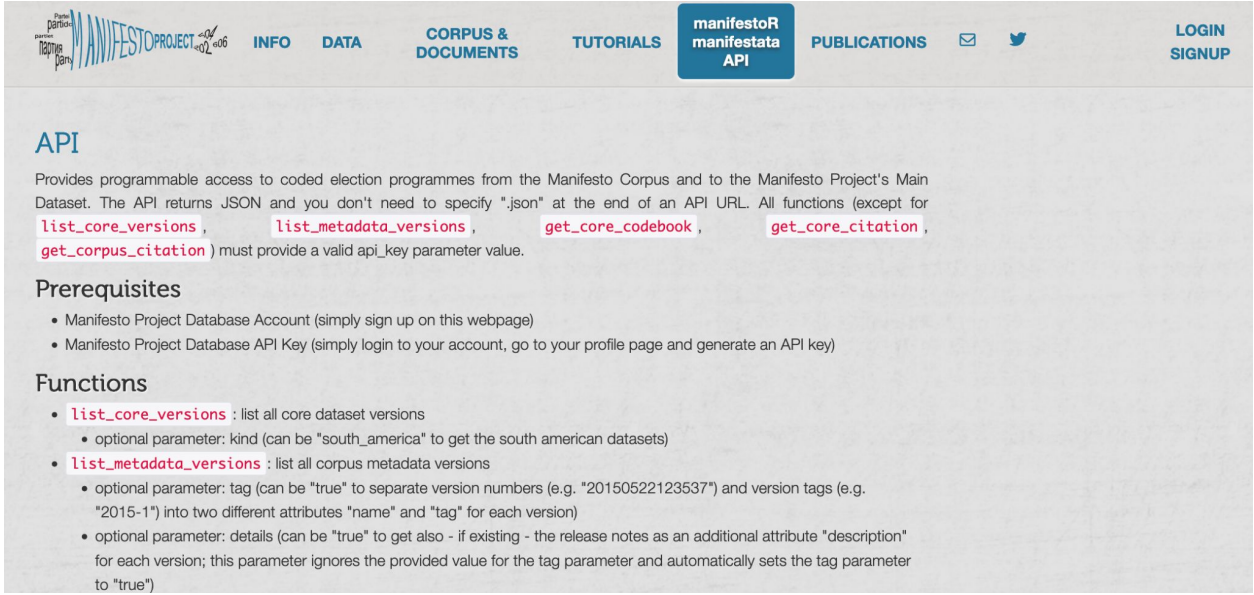


Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2016). Empirical analysis and classification of database errors in Scopus and Web of Science. *Journal of informetrics*, *10*(4), 933-953.

universität wien

Party manifestos

manifestoR



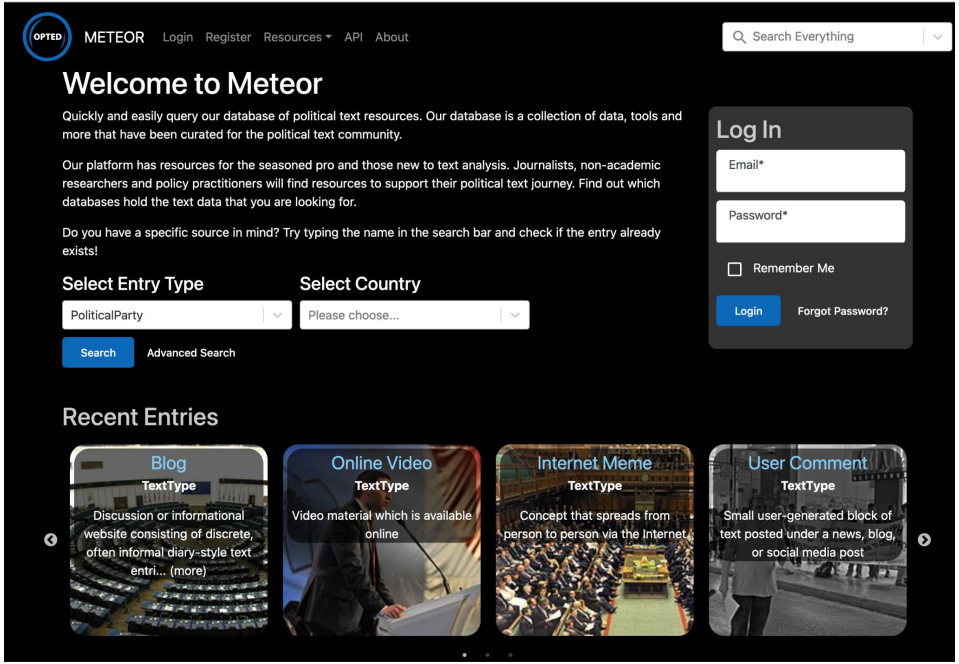
universität wien

Collection of Data Sources

https://docs.google.com/document/d/1pfEDiIU6iDbrbMSnfkTgDsZBAY5h02zmVYGPZIB25gE/edit?usp=sharing

OPTED Data & Tool Collection

Visit: https://meteor.opted.eu/)



universität wien

Data Scraping

Rvest package

universität wien

Data management

- Data security (access to a save cloud?)
- Sensitive data
- Data sharing?
- How to save:
  - Important!! Assign ID per analysis unit (text)
  - Record meta-info: source, date, author for each analysis unit
- Feel free to share your experiences and best practises

universität wien

Data validation

What possibilities do we have to evaluate the quality of our data?

universität wien

Data validation

Motivation: data selection determines results and conclusions

Are the selected data sources and selected data points representative for my target concept or discourse?

- Are they relevant?
- Are they representative?

Stryker et al. 2006

universität wien

Data validation

Validation techniques:

- Define your data universe of interest
- Rely on expert opinion
- Rely on data source selection of similar research
- Search string validation (see next session)

Document these steps

universität wien

Exercise: Gathering data for your project

a) You need data and you need inspiration of where to get it? Browse through the resources and exchange ideas with a partner to develop a data collection strategy
b) You need data, you know where to get it but need help to set up the API? This can be the moment to make a start
c) You have data already? Think about relevant units of analysis. Check out other sources for the next project or help others to find ways to access data

universität wien

Dictionaries

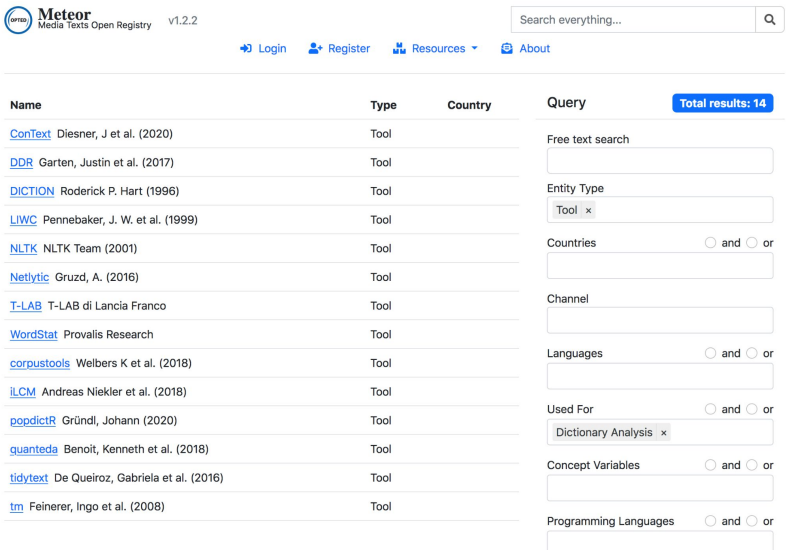
Day 2 Session 2

Dictionary method

Rule-based classification approach when categories are known

- List of words (or phrases) that indicate a category
- Create your own or use/edit existing dictionaries
- Always: validate!

Some existing dictionaries and dictionary application tools



universität wien

Dictionary use cases

- Selecting text data according to some defined category (‘what is relevant data’)

- Classifying documents into known categories

universität wien

Example: Search string for migration news

Table 3 Boolean search strings used for retrieval of migration-related news articles

Country	Language	Search string
Spain	Spanish	asilo* OR inmigra* OR refugiad* OR migrante* OR migratori* OR "sin papeles" OR "campo de desplazados" OR patera* OR emigra* OR "libre circulación" OR "fuga de cerebros"
UK	English	asyl* OR immigrant* OR immigrat* OR migrant* OR migrat* OR refugee* OR foreigner* OR "undocumented worker*" OR "guest worker*" OR "foreign worker*" OR emigrat* OR "freedom of movement" OR "free movement"
Germany	German	asyl* OR immigrant* OR immigriert* OR immigrat* OR migrant* OR migrat* OR flüchtling* OR ausländer* OR zuwander* OR zugewander* OR einwander* OR eingewander* OR gastarbeiter* OR "ausländische arbeitnehmer*" OR emigr* OR auswander* OR ausgewander* OR personenfreizügigkeit* OR arbeitnehmerfreizügigkeit* OR "freier personenverkehr*"

Heidenreich et al., 2020; Lind et al., 2020

universität wien

Example: Classify actor salience

Objective: salience of women and men migrant's in the news across time and outlet

- Dictionary approach to measure mentions of women and men migrants in German news articles

Lind & Meltzer, 2020

Example: Classify actor salience

Keyword list more complex than initially thought

Recall: .67. Precision: .91

```
person_f_migrant_endIn_regex  = c("[Aa]sylantin","[Aa]sylbewerberin", "[Aa]
                                 "[Zz]uwanderin", "[Ee]inwanderin", "[Gg]a
                                 "[Aa]usländische\\w{0,2}\\s[Bb]ürgerin",

f_relation_regex = c("[Mm]eine\\w{0,2}\\s[Ff]rau(en|)(\\s|\\W)", "[Mm]ein\\
                 "[Dd]eine\\w{0,2}\\s[Ff]rau(en|)(\\s|\\W)", "[Dd]ein\\w{
                 "[Ss]eine\\w{0,2}\\s[Ff]rau(en|)(\\s|\\W)", "[Ss]ein\\w{
                 "[Ii]hr\\w{0,2}\\s[Ff]rau(en|)(\\s|\\W)", "[Ii]hr\\w{0,2
                 "[Uu]ns\\w{0,2}\\s[Ff]rau(en|)(\\s|\\W)", "[Uu]ns(er|re)
                 "[Ee]u(er|re)\\w{0,2}\\s[Ff]rau(en|)(\\s|\\W)", "[Ee]u(e

f_relation_regex <- paste(apos_closed, f_relation_regex, "", sep = "") #add
df_f_relation_regex <- data.frame(f_relation_regex) #as. dataframe
```

Lind, F., & Meltzer, C. E. (2020).

Table 2. Dictionary subcategories for the measurement of the concepts: migrant women's and migrant men's salience.

	Subcategory name	Description	Measured concepts (examples)	
			Migrant women's salience	Migrant men's salience
1.	Impersonal designation	General person nominations for migrants	(immigrant, asylum seeker, etc.) with female or gender-neutral word endings (-ln,/-in, etc.)	(immigrant, asylum seeker, etc.) with generic masculine form or gender-neutral word endings (-ln,/-in, etc.)
2.	Origin: Single words	Nominations that refer to the territorial origin of a person [a]	(French, African, Syrian, etc.) with female or gender-neutral word endings (-ln, -/in, etc.)	(French, African, Syrian, etc.) with generic masculine form or gender-neutral word endings (-ln, -/in, etc.)
3.	Origin: Combinations of different word groups [b]			
a		General gender-related person nominations + from + general territorial denominations [c]	(women, girl, mother, sister) + from + (France, Africa, Syria, etc.)	(man, boy, father, brother) + from + (France, Africa, Syria, etc.)
b		General territorial denominations (adjectives) [c] + General gender-related person nominations	(French, African, Syrian, etc.) + (women, girl, mother, sister)	(French, African, Syrian) + (man, boy, father, brother)
4.	In relation	Relational expressions: possessive pronouns + General gender-related person nominations	(my, your, her, his, our, yours, their) + (women, girl, mother, sister)	(my, your, her, his, our, yours, their) + (man, boy, father, brother)
5.	As phrase	Phrases	e.g., "women and children"	e.g., "men and children"
6.	Other expressions		e.g., "women from abroad"	e.g., "men from abroad"

As general notes, the dictionary includes the singular and plural version of all words, word endings (e.g., for prepositions) consider the different cases used in the German language.

[a]Downloaded from the CLDR (Unicode Common Locale Data Repository) http://cldr.unicode.org/, which holds standard name translations of countries and regions (version v33.1).

[b]Measured at the sentence level.

[c]Manually compiled by a native speaker for all CLDR territorial denominations, which the German language allows (e.g., no separate word for many smaller islands, e.g., Isle of Man, Curaçao). Assisted by the preeminent German language dictionary duden.de.

universität wien

Validate, validate, validate

universität wien

Data validation

Motivation: source and data selection determines results and conclusions

Are the selected data sources and selected data points representative for my target concept or discourse?

- Are they relevant?
- Are the representative?

Stryker et al. 2006

universität wien

Side note: Data source validation

Validation techniques:

- Rely on expert opinion
- Rely on data source selection of similar research

universität wien

Relevance of search string validation

- Sampling based on search strings popular (Stryker et al. 2016) and recommended (Barberá et al., 2021)
- Reviews of search string validation procedures
  - out of 83 content analyses, 39% stated the search terms they used, and only 6% discussed their validity (Stryker et al. 2016)
  - out of 105 content analysis studies, 73.3% stated the search terms they used, only 12.4% reported validity metrics (Mahl et al., 2022)
- Careless application of non-validated search terms may lead to noisy inferences (Mahl et al., 2022)

Mahl, D., von Nordheim, G., & Guenther, L. (2023). Noise pollution: A multi-step approach to assessing the consequences of (not) validating search terms on automated content analyses. *Digital Journalism*, *11*(2), 298-320.

Stryker, Jo Ellen, Ricardo J. Wray, Robert C. Hornik, and Itzik Yanovitzky. 2006. "Validation of Database Search Terms for Content Analysis: The Case of Cancer News Coverage." *Journalism & Mass Communication Quarterly* 83 (2): 413–430

universität wien

Key validation approach

How close is an automated measurement to a more trusted measurement:

Human understanding of text

universität wien

Dictionary validation with manually created baseline

Steps

- Code a subset manually (consider intercoder reliability)
- Compare manual decisions with automated classification decisions (via recall, precision, F1)
- Iterative dictionary improvement
- Ideally: manual coding and dictionary development is performed by different persons

universität wien

Creation of a manual baseline

Steps:

- Codebook creation
- Who codes manually?
  - Expert coders: Coder recruitment and training sessions
  - Crowdcoders: test questions, majority choice
- Quality assessment: e.g., Inter-coder reliability of involved coders, majority vote
  - How reliable? Consider *valid disagreement* (Baden et al., 2023)
- Documents selected for baseline should be representative for target discourse (e.g., random selection or artificial week)

Baden, C., Boxman-Shabtai, L., Tenenboim-Weinblatt, K., Overbeck, M., & Aharoni, T. (2023). Meaning multiplicity and valid disagreement in textual measurement: A plea for a revised notion of reliability. *SCM Studies in Communication and Media*, *12*(4), 305-326.

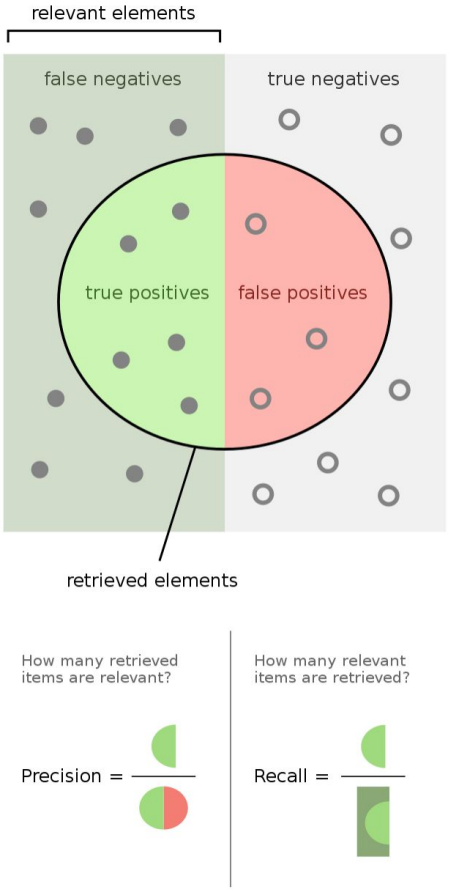
Lind, F., Gruber, M., & Boomgaarden, H. G. (2017). Content analysis by the crowd: Assessing the usability of crowdsourcing for coding latent constructs. *Communication methods and measures*, *11*(3), 191-209.

Recall, precision, F1?

Metrics frequently used to express the validity of a search string & more generally also of automated classification methods

- Precision (P)
- Recall ®(R)
- F1 = 2*(P * R)/R+R)

Scharkow, 2012, 133–36



universität wien

Tools for manual coding

- Google sheets
- AmCAT: https://vu.amcat.nl/accounts/login/?next=/
- AnnoTinder: https://github.com/ccs-amsterdam/CCS_annotator
- Label Studio: https://labelstud.io/

Example: Baseline creation templates For search strings

Create sampling plan (goal: representative for universe of texts)

Database	Date	Outlet	Number of all articles published that day
LexisNexis	Mon 8.10.2018	Standard	136
LexisNexis	Tue 9.07.2019	Standard	87
LexisNexis	Wed 12.02.2020	Standard	89
LexisNexis	Thr 15.04.2021	Standard	98
LexisNexis	Fri 27.05.2022	Standard	94
Note: Ideally repeat this procedure for each outlet included; cover the full range of time period investigated			

Collect articles

Code manually

Search with search string

Article id	Date	Text	Manually perceived as relevant (1=yes, 0 = No)	Perceived as relevant by search string (1=yes, 0 = No)
1	Mon 8.10.2018	Kern ist an sich selbst gescheitert. Die SPÖ braucht jetzt mehr Geradlinigkeit und weniger Gockelhaftigkeit …	1	1
2	Mon 8.10.2018	Impressum und Offenlegung: Herausgeber: Oscar Bronner…	0	0
3	Mon 8.10.2018	Einseitiger Vorschlag. Zu viele Waffen in der Hand der Bürger sind gefährlich. Ein Blick in die USA zeigt, warum. Im Kern geht es…	0	1
…	…	…	…	

Calculate recall and precision

universität wien

Dictionary

Pros

- Often needed to select data (search strings)
- High reliability and control
- High transparency and reproducibility

Cons

- Difficulty increases with the latency of the construct
- Language nuances

universität wien

Regular Expressions (regex)

universität wien

Regular Expressions (regex)

- How to pronounce?

universität wien

Regular Expressions (regex)

- How to pronounce?
- /gɪf/
- /dʒɪf/

universität wien

Regular Expressions (regex)

- How to pronounce?
- /gɪf/
- /dʒɪf/
- /ɹɛ.gɛks/
- /ɹɛ.dʒɛks/

universität wien

Regular Expressions (regex)

- How to pronounce?
- ***/gɪf/***
- /dʒɪf/
- ***/ɹɛ.gɛks/***
- /ɹɛ.dʒɛks/

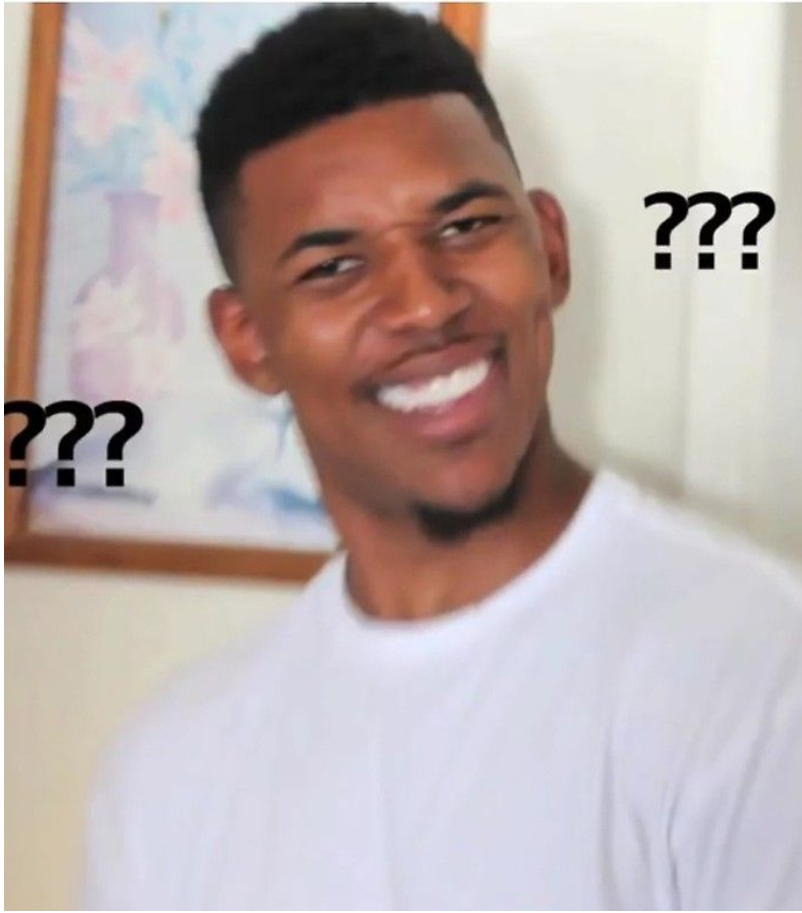
universität wien

Regular Expressions (regex)

- How to pronounce?
- /gɪf/
- ***/dʒɪf/***
- /ɹɛ.gɛks/
- ***/ɹɛ.dʒɛks/***

Regular Expressions (regex)

- How to pronounce?
- /gɪf/
- ***/dʒɪf/***
- /ɹɛ.gɛks/
- ***/ɹɛ.dʒɛks/***



universität wien

regex

- Formal language to specify search strings

universität wien

regex

- Formal language to specify search strings
- Insanely difficult

universität wien

regex

- Formal language to specify search strings
- INSANELY difficult

universität wien

regex

- Formal language to specify search strings
- ***INSANELY*** difficult

universität wien

regex

- Formal language to specify search strings
- ***INSANELY*** difficult
- Nobody can remember anything

universität wien

regex

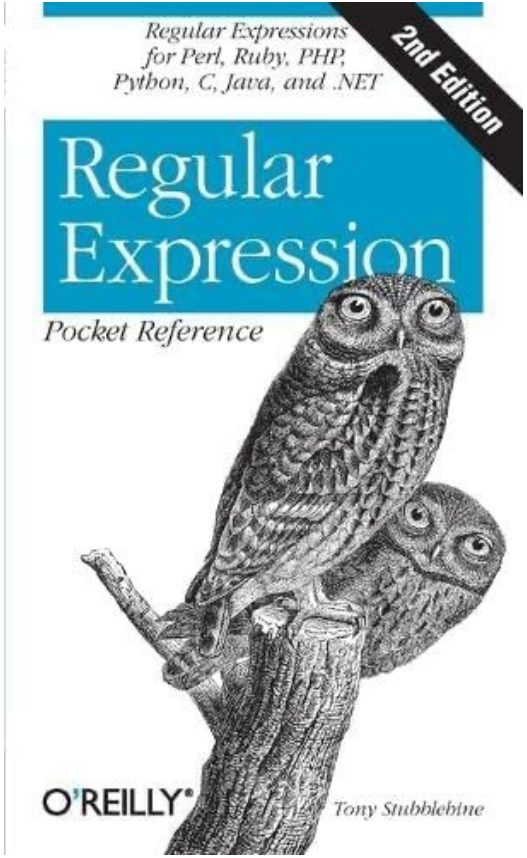
- Formal language to specify search strings
- ***INSANELY*** difficult
- Nobody can remember anything
- Different ***flavours***

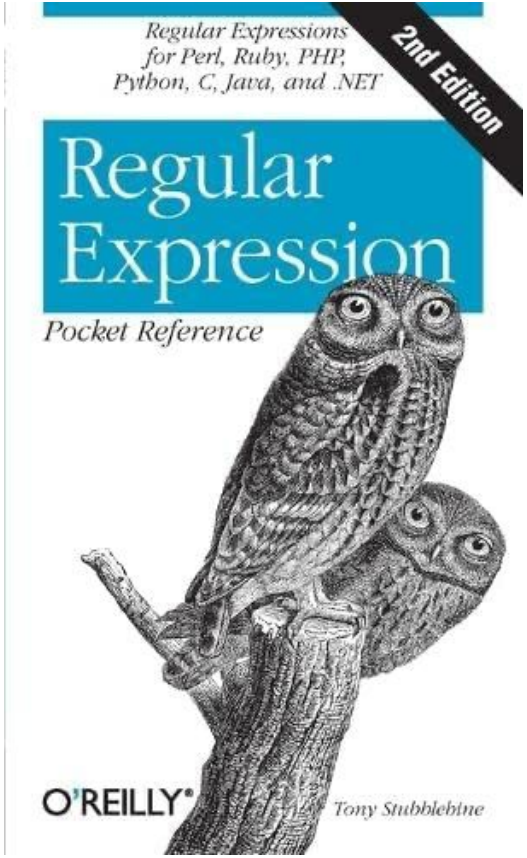
universität wien

regex

- Formal language to specify search strings
- ***INSANELY*** difficult
- Nobody can remember anything
- Different ***flavours***

- “Some people, when confronted with a problem, think “I know, I’ll use regular expressions.” Now they have two problems.” *Jamie Zawinski*


Regular Expressions
for Perl, Ruby, PHP,
Python, C, Java, and .NET
2nd Edition
Regular
Expression
Pocket Reference
O'REILLY
Tony Stubblebine


Regular Expressions
for Perl, Ruby, PHP,
Python, C, Java, and .NET
2nd Edition
Regular
Expression
Pocket Reference
O'REILLY
Tony Stubblebine

128 pages!!!

universität wien

Grimmer / Jurafsky
Cheat-sheet

- Disjunctions

RE	Match	Example Patterns Matched
`[mM]oney`	Money or money	"Money"
`[abc]`	'a', 'b', *or* 'c'	"Investing in Iran"
		"is dangerous business"
`[1234567890]`	any digit	"sitting on $7.5 billion dollars"
		"2005 and 2006, more than "
		"$150 million dollars"
`[\.]`	A period	" 'Run!', he screamed."

- Ranges

RE	Match	Example Patterns Matched
`[A-Z]`	an upper case letter	"Rep. Anthony Weiner
		(D-Brooklyn & Queens)"
`[a-z]`	a lower case letter	"ACORN's"
`[0-9]`	a single digit	"(9th CD) "

universität wien

Grimmer / Jurafsky
Cheat-sheet

- Negations

RE	Match	Example Patterns Matched
`[^A-Z]`	not an upper case letter	"ACORN**'s**"
`[^Ss]`	neither 'S' nor 's'	"**ACORN'**s"
`[^\.]`	not a period	" 'Run!', he screamed."

- Optional Characters: `?`, `*`, `+`

RE	Match	Example Patterns Matched
`colou?r`	Words with `u` 0 or 1 times	"color" or "colour "
`oo*h!`	Words with `o` 0 or more times	"oh!" or "ooh!" or "oooh!"
`o+h!`	Words with `o` 1 or more times	"oh!" or "ooh!" or "oooooh!" or

universität wien

Grimmer / Jurafsky
Cheat-sheet

- Start of the line anchor ^, end of the line anchor **$**

RE	Match	Example Patterns Matched
`^[A-Z]`	Upper case start of line	"Palo Alto" "the town of **P**alo Alto"
`^[^A-Z]`	Not upper case start of line	"the town of Palo Alto" "**P**alo Alto"
`^.`	Start of line	"Palo Alto" "the town of Palo Alto"
`.$`	Identify character that ends a line	"Wait**!**" "This is the end**.**"

- "Or"| statements, Useful short hand

RE	Match	Example Patterns Matched
yours|mine	Matches"yours" or "mine"	"it's either yours or mine"
\ d	Any digit	"1-Mississippi"
\ D	Any non-digit	"1-Mississippi"
\ s	Any whitespace character	"1, 2"
\ S	Any non-whitespace character	"1, 2"
\ w	Any alpha-numeric	"1-Mississippi "
\ W	Any non-alpha numeric	"1-Mississippi"

universität wien

How difficult to regex an email

universität wien

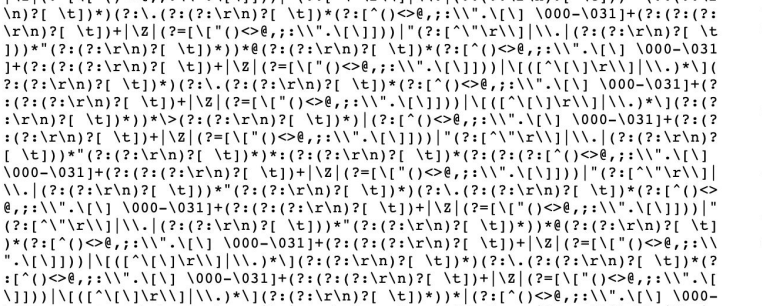
How difficult to regex an email

Rather

universität wien

How difficult to regex an email

```
(?:[a-z0-9!#$%&'*+/=?^_`{|}~-]+(?:\.[a-z0-9!#$%&'*+/=?^_`{|}~-]+)*|"(?:[\x01-
\x08\x0b\x0c\x0e-\x1f\x21\x23-\x5b\x5d-\x7f]|\\[\x01-\x09\x0b\x0c\x0e-\x7f])*")@(?:
(?:[a-z0-9](?:[a-z0-9-]*[a-z0-9])?\.)+[a-z0-9](?:[a-z0-9-]*[a-z0-9])?|\[(?:(?:(2(5[0-
5]|[0-4][0-9])|1[0-9][0-9]|[1-9]?[0-9]))\.){3}(?:(2(5[0-5]|[0-4][0-9])|1[0-9][0-9]|[1-
9]?[0-9])|[a-z0-9-]*[a-z0-9]:(?:[\x01-\x08\x0b\x0c\x0e-\x1f\x21-\x5a\x53-\x7f]|\\
[\x01-\x09\x0b\x0c\x0e-\x7f])+)\])
```




universität wien

Helpers

https://regex101.com

Recommended R tutorial (Wickham & Grolemund, 2017):
https://r4ds.had.co.nz/strings.html

To test regular expressions quickly:
https://spannbaueradam.shinyapps.io/r_regex_tester/

universität wien

Exercise in R

RStudio

- Dictionary creation: Visibility of political actors in news headlines
- Calculate recall, precision, F1, Krippendorff's alpha