

# Advanced quantitative text analysis (2023W)

Petro Tolochko, Fabienne Lind

Day 2



# 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

# Concepts & Data

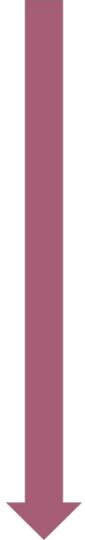
Day 2 Session 1

---

# 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

# 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?

# Discovery

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

# Principles of Discovery

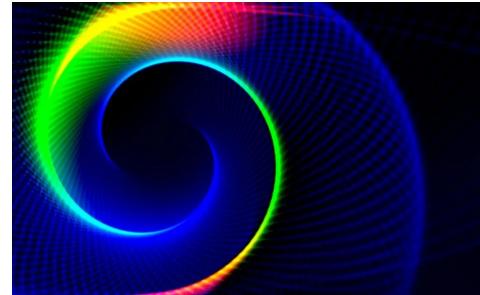
**Principle 1: Context relevance.** Text as data models complement theory and substantive knowledge. Contextual knowledge amplifies our ability to make computational discoveries



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

# 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



# 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?



# 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.



## 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

## 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
-

# 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)
-

## Codebooks and automated text analysis

Wait, codebooks are certainly relevant in manual content analysis

But what role do they have in automated text analysis?

## 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

# 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<sup>3</sup>, a definition of what was defined as migration (see page 7), and additional instructions such as:

<https://doi.org/10.11587/IEGQ1B>

---

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”

---

<sup>3</sup> 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, 20. Luxembourg, 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)

# Causal inference in text

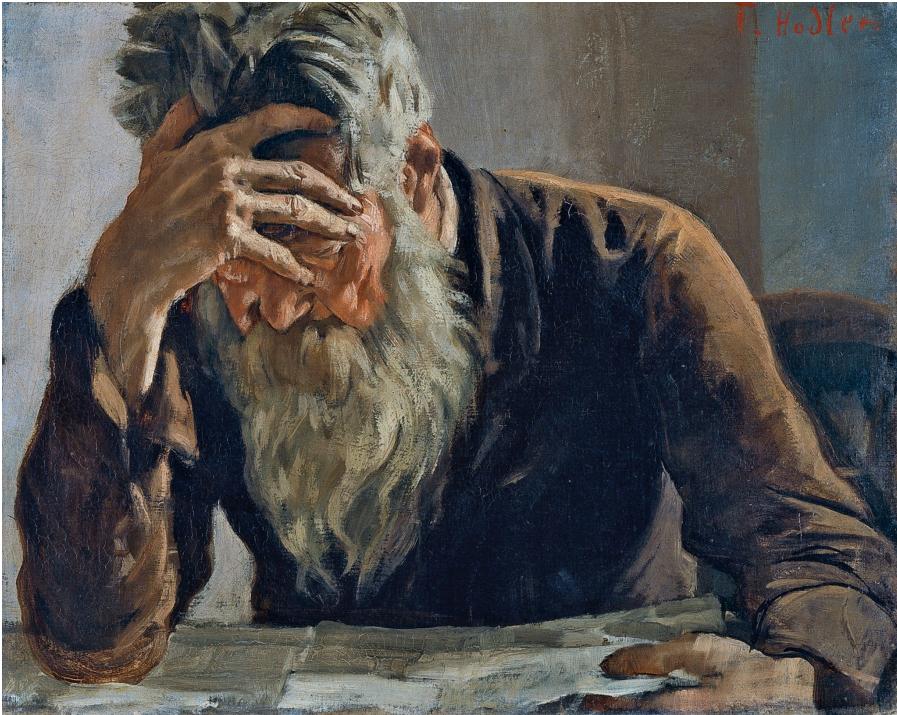
# Causal inference in text

More complex than might seem...

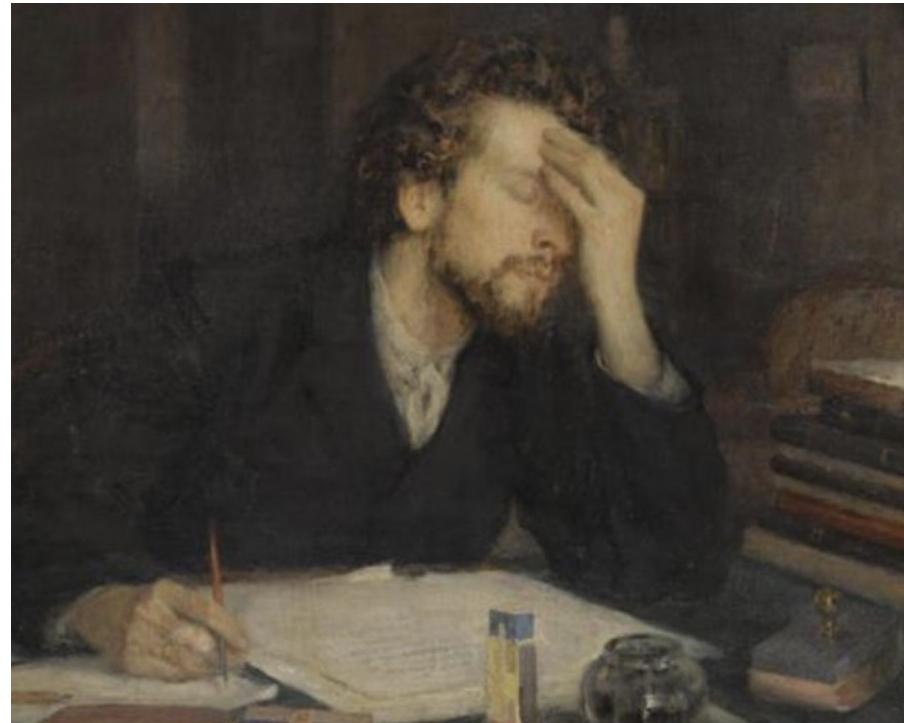


## Treatment

---



Treatment



Outcome

# Multidimensionality of Text

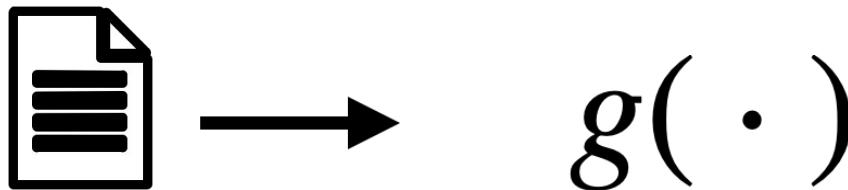
- Dimension reduction

# Multidimensionality of Text

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

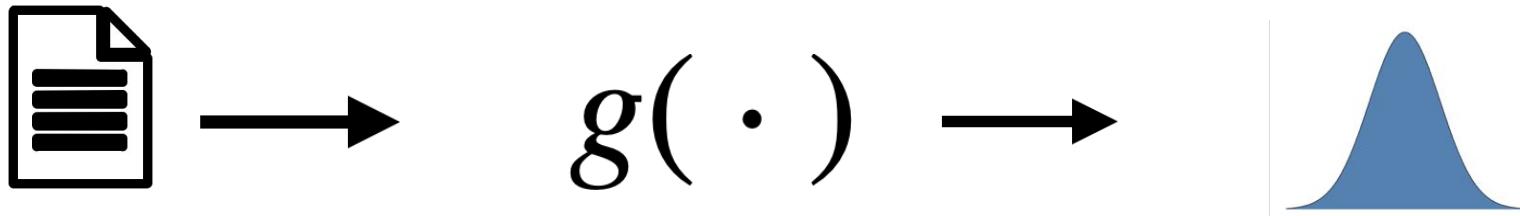
# Multidimensionality of Text

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



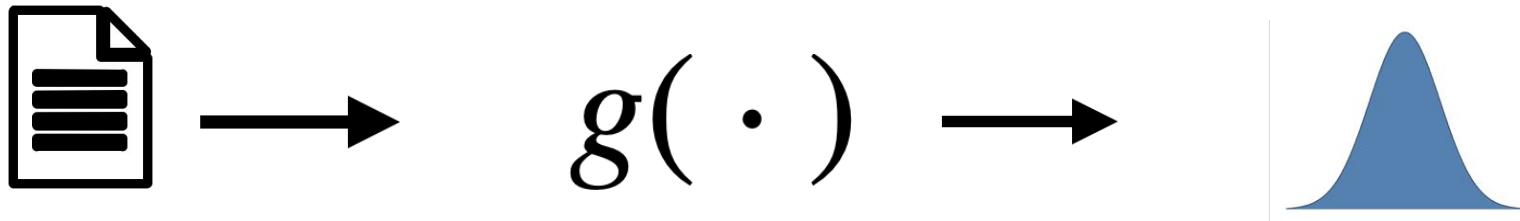
# Multidimensionality of Text

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



# Multidimensionality of Text

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



[OBJ]  $g(\cdot)$  is **emergent**

# Multidimensionality of Text

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

# 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

# Multidimensionality of Text

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

# 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

# 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

# Data sources

Different levels

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

## 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?)

## Data and where to get it

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

## Data and where to get it

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

## For newspapers

Repositories, for example

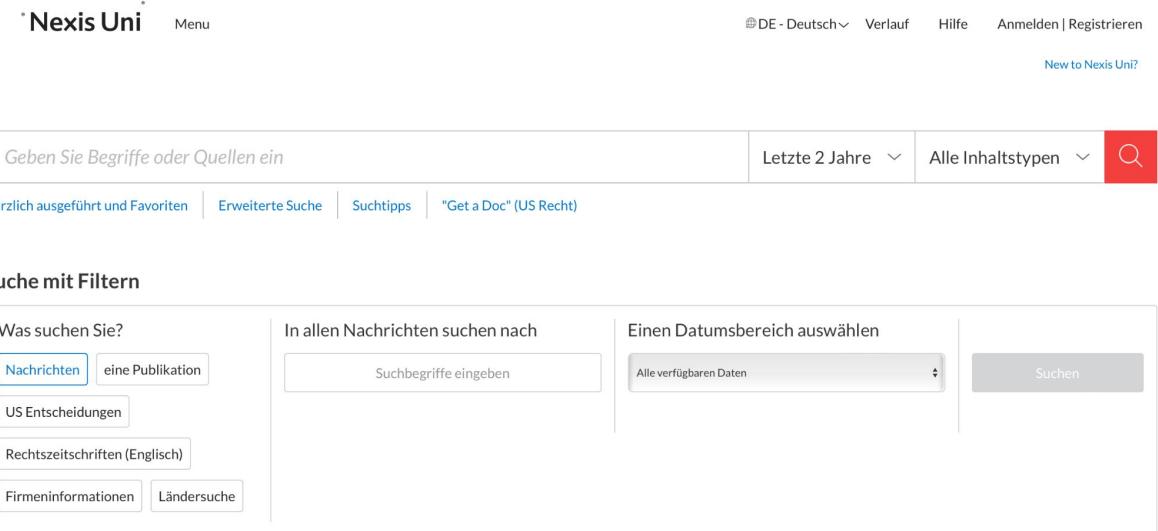
- LexisNexis
- APA
- Factiva
- EMIS

You might have access via your institutions

# LexisNexis

One method:

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



The screenshot shows the LexisNexis Uni search interface. At the top, there is a navigation bar with the text "Nexis Uni" and "Menu". On the right side of the top bar are links for "DE - Deutsch", "Verlauf", "Hilfe", "Anmelden | Registrieren", and "New to Nexis Uni?". Below the top bar is a search bar containing the placeholder text "Geben Sie Begriffe oder Quellen ein". To the right of the search bar are dropdown menus for "Letzte 2 Jahre" and "Alle Inhaltstypen", followed by a red search button with a white magnifying glass icon. Below the search bar are four blue links: "Kürzlich ausgeführt und Favoriten", "Erweiterte Suche", "Suchtipps", and "\"Get a Doc\" (US Recht)". The main search area is titled "Suche mit Filtern". It contains three sections: "Was suchen Sie?", "In allen Nachrichten suchen nach", and "Einen Datumsbereich auswählen". The "Was suchen Sie?" section has several buttons: "Nachrichten" (which is highlighted in blue), "eine Publikation", "US Entscheidungen", "Rechtszeitschriften (Englisch)", "Firmeninformationen", and "Ländersuche". The "In allen Nachrichten suchen nach" section has a text input field labeled "Suchbegriffe eingeben". The "Einen Datumsbereich auswählen" section has a dropdown menu set to "Alle verfügbaren Daten" and a "Suchen" button.

# APA

[Suche](#)  [Medienpräsenz](#) [Hilfe](#) [Kontakt](#)

Sucheinstellungen (Suchvorlage '**Default**')

## Suche

SUCHEN

z. B. Salzburg Kultur, z. B. "United Nations" ([Hilfe](#)) [Suchbegriff löschen](#)

Wortstammsuche aktivieren

### Zeitraum

von Vormonatsbeginn ▼ 01.12.2023 📅 – bis vor 7 Tagen ▼ 09.01.2024 📅

### Quellen

Nach Quellen suchen 🔍 Quellen aus Bündel (332) ➡ ⬅ [ausgewählte Quellen \(18\)](#)

 <b>Quellenbündel</b>  APA-Basisdienst und OTS  Agenturen	OTS-Originaltext-Service Euro (D) 1st FIRST (hist. Bestand)	Der Standard Die Presse Heute
--	---	-------------------------------------

# APIs

API = Application Programming Interface

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

# APA API



The screenshot shows the top navigation bar of the website. It includes the University of Vienna logo and name, a search bar, and a dropdown menu labeled "QUELLEN". Below the search bar, there are four main navigation links: "Team", "Research & Teaching", "APA-UNIVIE Data Project" (which is highlighted in blue), and "News & Events".

You are here: › [University of Vienna](#) › [Faculty of Social Sciences](#) › [Department of Communication](#) › [Computational Communication Science Lab](#) › APA-UNIVIE Data Project

## APA-UNIVIE Data Project

Members of the University of Vienna now have the possibility to obtain data for scientific use from the media archive of the Austria Press Agency (APA), directly via an Application Programming Interface (API).

Since the volume of downloads is limited, access to the API is moderated by the Computational Communication Science Lab (CCL) via a submission routine in which projects are evaluated regarding their maturity and whether the still available contingent of downloads can accommodate the requested data volume.

Use the [University of Vienna's Weblogin](#) to access more detailed information on the submission process, sources available for download, technical information as well as the application form.

## APA-UNIVIE Data Project

[Login / Logout](#)

### Contact

Department of Communication  
Kolingasse 14-16  
1090 Wien  
T: +43-1-4277-497 30

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

# API Examples

[https://bookdown.org/paul/apis\\_for\\_social\\_scientists/](https://bookdown.org/paul/apis_for_social_scientists/)

## Preface

- 1 Introduction
- 2 Best Practices
- 3 CKAN API
- 4 CrowdTangle API
- 5 Facebook Ad Library API
- 6 Genderize.io API
- 7 GitHub.com API
- 8 Google News API
- 9 Google Natural Language API
- 10 Google Places API
- 11 Google Speech-to-Text API
- 12 Google Translation API
- 13 GoogleTrends API
- 14 Instagram Basic Display API
- 15 Instagram Graph API
- 16 Internet Archive API
- 17 Media Cloud API
- 18 Overpass API
- 19 Reddit API
- 20 Spotify API
- 21 Twitter API
- 22 MediaWiki Action API

## APIs for social scientists: A collaborative review

### Current editors:

*Paul C. Bauer, Camille Landesvatter, Lion Behrens*

### Authors & contributors:

*Paul C. Bauer, Jan Behnert, Lion Behrens, Chung-hong Chan, Bernhard Clemm von Hohenberg, Lukas Isermann, Philipp Kadel, Melike N. Kap, Jana Klein, Markus Konrad, Barbara K. Kreis, Dean Lajic, Camille Landesvatter, Madleen Meier-Barthold, Grace Olzinski, Nina Osenbrüggen, Ondrej Pekacek, Pirmin Stöckle, Malte Söhren, Domantas Undzéna.*

*First public version: 29 November, 2021*

*This version: 14 Juni, 2023*

# 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.

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  CCL 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>

 WINNER

U.S. SENATE

**Nevada**

Live updates

# 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.

# Telegram

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

Guidelines:

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

# 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/>

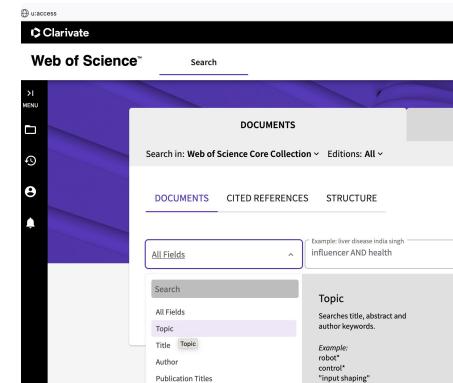
---

# Scientific Publications

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



The screenshot shows the u:search interface for the University of Vienna. At the top, there is a navigation bar with links for "New Search", "Database Service", "Collection Discovery", "ILL", and "Help". Below this is a search bar containing the text "web of science core collection". To the right of the search bar is a magnifying glass icon. The interface features a dark background with a network of nodes and connections. A red banner at the bottom contains the text "Sign in to get complete results" and a "Sign in" button. Below the banner, there is a "Tweak my results" section and a "Sort by relevance" dropdown menu. At the very bottom, there is a link to "Web of Science Core Collection" with the text "Clarivate Analytics | Full Text Linking | TOP-DB" and an "Online access" link.

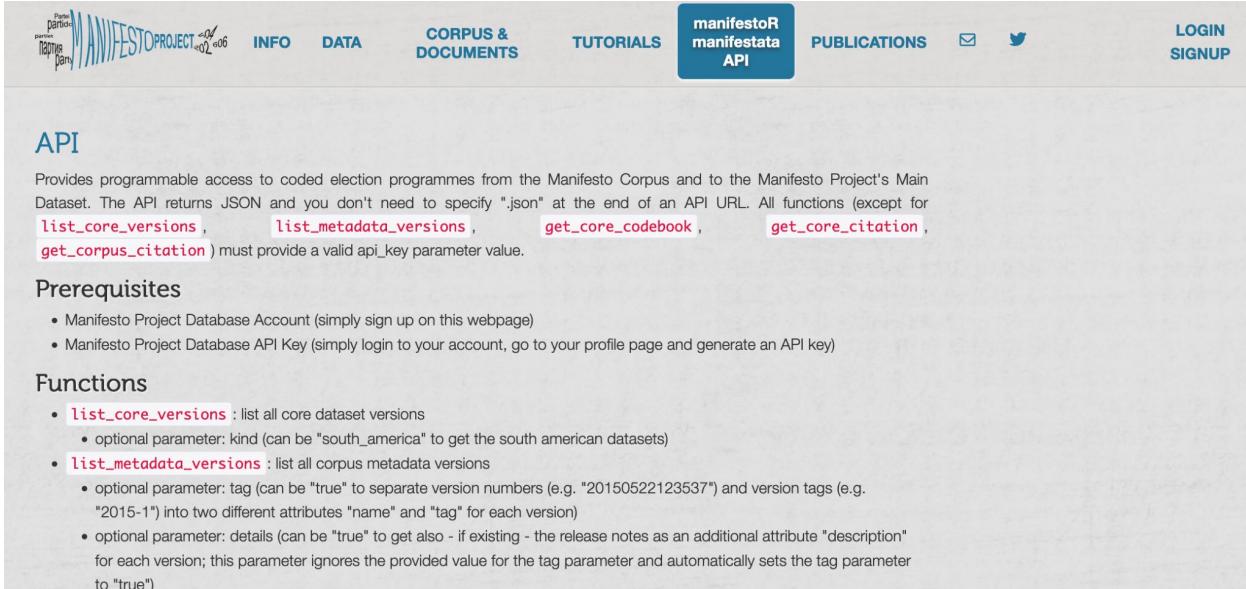


The screenshot shows the Web of Science search interface. At the top, it says "v:access Clarivate Web of Science". Below this is a search bar with the placeholder "Search in: Web of Science Core Collection Editions: All". There are three tabs: "DOCUMENTS", "CITED REFERENCES", and "STRUCTURE", with "DOCUMENTS" being the active tab. On the left, there is a sidebar with a "MENU" icon, a "Search" input field, and sections for "All Fields", "Topic", "Title", "Topic", "Author", and "Publication Titles". A help text box on the right provides examples for "Topic" searches. The main search results area is currently empty.

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

# Party manifestos

manifestoR



The screenshot shows the manifestoR API documentation page. At the top, there's a navigation bar with links for INFO, DATA, CORPUS & DOCUMENTS, TUTORIALS, manifestoR manifestata API (which is highlighted in blue), PUBLICATIONS, and social media icons for email and Twitter. On the far right are LOGIN and SIGNUP buttons. Below the navigation, there's a section titled "API" with a sub-section "Prerequisites" containing two bullet points about account creation and API keys. The main "Functions" section lists four functions: `list_core_versions`, `list_metadata_versions`, `get_core_codebook`, and `get_core_citation`. Each function entry includes a brief description and a bulleted list of parameters.

## API

Provides programmable access to coded election programmes from the Manifesto Corpus and to the Manifesto Project's Main Dataset. The API returns JSON and you don't need to specify ".json" at the end of an API URL. All functions (except for `list_core_versions`, `list_metadata_versions`, `get_core_codebook`, `get_core_citation`, `get_corpus_citation`) must provide a valid api\_key parameter value.

### Prerequisites

- Manifesto Project Database Account (simply sign up on this webpage)
- Manifesto Project Database API Key (simply login to your account, go to your profile page and generate an API key)

### Functions

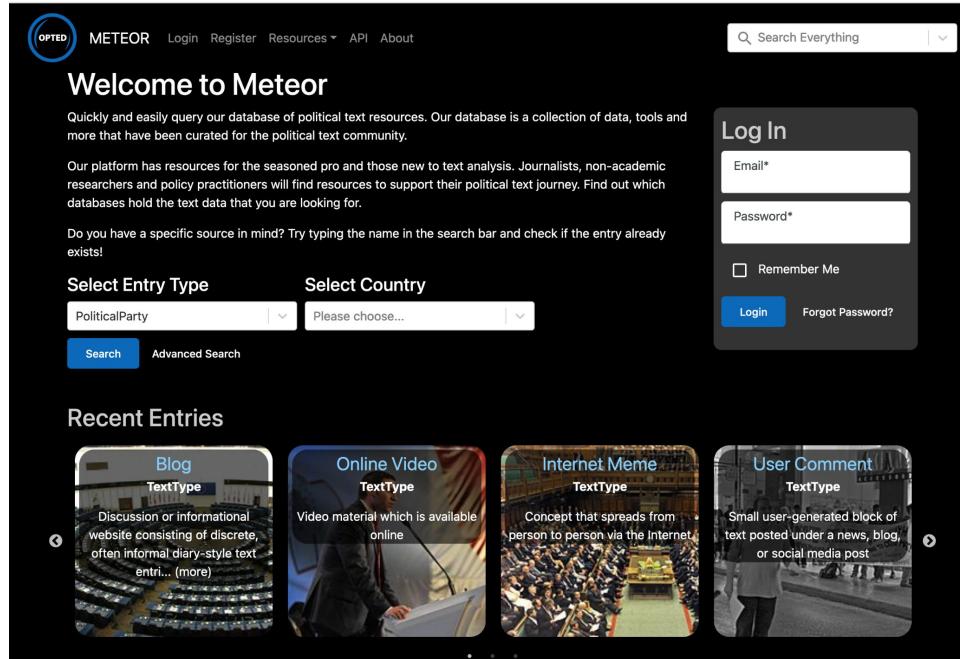
- `list_core_versions` : list all core dataset versions
  - optional parameter: kind (can be "south\_america" to get the south american datasets)
- `list_metadata_versions` : list all corpus metadata versions
  - optional parameter: tag (can be "true" to separate version numbers (e.g. "20150522123537") and version tags (e.g. "2015-1") into two different attributes "name" and "tag" for each version)
  - optional parameter: details (can be "true" to get also - if existing - the release notes as an additional attribute "description" for each version; this parameter ignores the provided value for the tag parameter and automatically sets the tag parameter to "true")

# Collection of Data Sources

<https://docs.google.com/document/d/1pfEDiU6iDbrbMSnfkTgDsZBAY5h02zmVYGPZIB25gE/edit?usp=sharing>

# OPTED Data & Tool Collection

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



The screenshot shows the METEOR platform homepage. At the top, there is a navigation bar with the OPTED logo, the word METEOR, and links for Login, Register, Resources, API, and About. A search bar is located at the top right. Below the navigation, a large heading says "Welcome to Meteor". A sub-headline explains: "Quickly and easily query our database of political text resources. Our database is a collection of data, tools and more that have been curated for the political text community." Another sub-headline states: "Our platform has resources for the seasoned pro and those new to text analysis. Journalists, non-academic researchers and policy practitioners will find resources to support their political text journey. Find out which databases hold the text data that you are looking for." A note below says: "Do you have a specific source in mind? Try typing the name in the search bar and check if the entry already exists!" On the left, there are two dropdown menus: "Select Entry Type" (set to PoliticalParty) and "Select Country" (set to Please choose...). Below these are "Search" and "Advanced Search" buttons. On the right, there is a "Log In" form with fields for Email\*, Password\*, and Remember Me, along with "Login" and "Forgot Password?" buttons. At the bottom, there is a section titled "Recent Entries" featuring four cards: "Blog" (TextType), "Online Video" (TextType), "Internet Meme" (TextType), and "User Comment" (TextType). Each card includes a small thumbnail image and a brief description.

# Data Scraping

**Rvest** package

# 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
-

# Data validation

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

# 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?

# 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

---

## 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
-

# Dictionaries

Day 2 Session 2

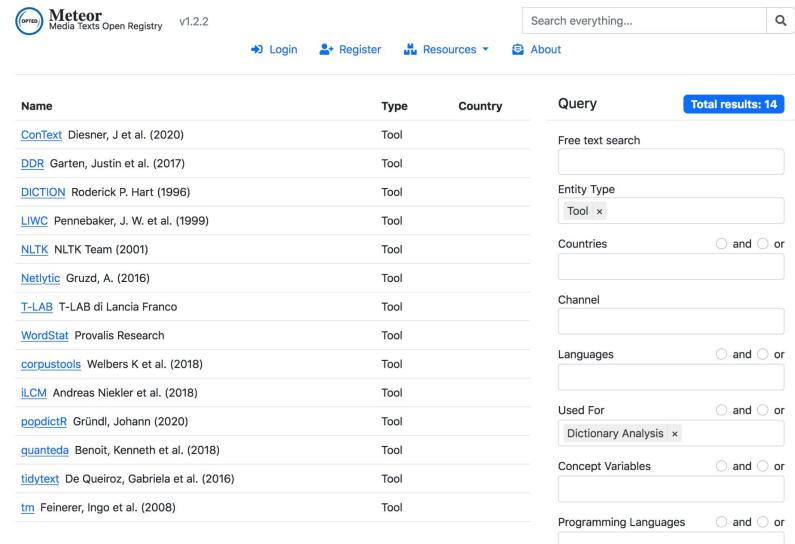
---

# 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



Name	Type	Country	Query
<a href="#">ConText</a> Diesner, J et al. (2020)	Tool		Free text search
<a href="#">DDR</a> Garten, Justin et al. (2017)	Tool		Entity Type
<a href="#">DICTION</a> Roderick P. Hart (1996)	Tool		Tool <input checked="" type="checkbox"/>
<a href="#">LWIC</a> Pennebaker, J. W. et al. (1999)	Tool		Countries <input type="radio"/> and <input type="radio"/> or
<a href="#">NLTK</a> NLTK Team (2001)	Tool		Channel
<a href="#">Netlytic</a> Gruzd, A. (2016)	Tool		Languages <input type="radio"/> and <input type="radio"/> or
<a href="#">T-LAB</a> T-LAB di Lancia Franco	Tool		Used For <input type="radio"/> and <input type="radio"/> or
<a href="#">WordStat</a> Provalis Research	Tool		Dictionary Analysis <input checked="" type="checkbox"/>
<a href="#">corpuTools</a> Welbers K et al. (2018)	Tool		Concept Variables <input type="radio"/> and <input type="radio"/> or
<a href="#">ilCM</a> Andreas Niekler et al. (2018)	Tool		Programming Languages <input type="radio"/> and <input type="radio"/> or
<a href="#">popdictR</a> Gründl, Johann (2020)	Tool		
<a href="#">quanteda</a> Benoit, Kenneth et al. (2018)	Tool		
<a href="#">tidytext</a> De Queiroz, Gabriela et al. (2016)	Tool		
<a href="#">tm</a> Feinerer, Ingo et al. (2008)	Tool		

## Dictionary use cases

- Selecting text data according to some defined category ('what is relevant data')
- Classifying documents into known categories

# 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 inmigrَا* 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

[Heidenreich et al., 2020](#); [Lind et al., 2020](#)

## 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



# Example: Classify actor salience

Keyword list more complex than initially thought

Recall: .67. Precision: .91

```

349
350
351
352 person_f_migrant_endIn_regex = c("[Aa]sylantin", "[Aa]sylbewerberin", "[Aa]"
353           "[Zz]uwanderin", "[Ee]inwanderin", "[Gg]a"
354           "[Aa]usländische\\w{0,2}\\s[Bb]ürgerin",
355
356
357 f_relation_regex = c("[Mm]eine\\w{0,2}\\s[Ff]rau(en)(\\s|\\W)", "[Mm]ein\\"
358           "[Dd]eine\\w{0,2}\\s[Ff]rau(en)(\\s|\\W)", "[Dd]ein\\w{"
359           "[Ss]eine\\w{0,2}\\s[Ff]rau(en)(\\s|\\W)", "[Ss]ein\\w{"
360           "[Ii]hr\\w{0,2}\\s[Ff]rau(en)(\\s|\\W)", "[Ii]hr\\w{0,2"
361           "[Uu]ns\\w{0,2}\\s[Ff]rau(en)(\\s|\\W)", "[Uu]ns(er)re"
362           "[Ee]u(er)re)\\w{0,2}\\s[Ff]rau(en)(\\s|\\W)", "[Ee]u(e"
363
364
365
366 f_relation_regex <- paste(apos_closed, f_relation_regex, "", sep = "") #add
367 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.

Measured concepts (examples)			
Subcategory name	Description	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 (-In, -in, etc.)	(immigrant, asylum seeker, etc.) with generic masculine form or gender-neutral word endings (-In, -in, etc.)
2. Origin: Single words	Nominations that refer to the territorial origin of a person <sup>a</sup>	(French, African, Syrian, etc.) with female or gender-neutral word endings (-In, -in, etc.)	(French, African, Syrian, etc.) with generic masculine form or gender-neutral word endings (-In, -in, etc.)
3. Origin: Combinations of different word groups <sup>b</sup>	General gender-related person nominations + from + general territorial denominations <sup>c</sup>	(women, girl, mother, sister) + from + (France, Africa, Syria, etc.)	(man, boy, father, brother) + from + (France, Africa, Syria, etc.)
a	General territorial denominations (adjectives) <sup>c</sup> + General gender-related person nominations	(French, African, Syrian, etc.) + (women, girl, mother, sister)	(French, African, Syrian) + (man, boy, father, brother)
b	General territorial denominations (adjectives) <sup>c</sup> + 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.

<sup>a</sup>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).

<sup>b</sup>Measured at the sentence level.

<sup>c</sup>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*.

Validate, validate, validate

# 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 they representative?

## Side note: Data source validation

Validation techniques:

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

# 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.

## Key validation approach

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

Human understanding of text

# 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

# 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.

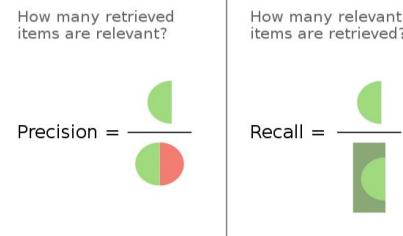
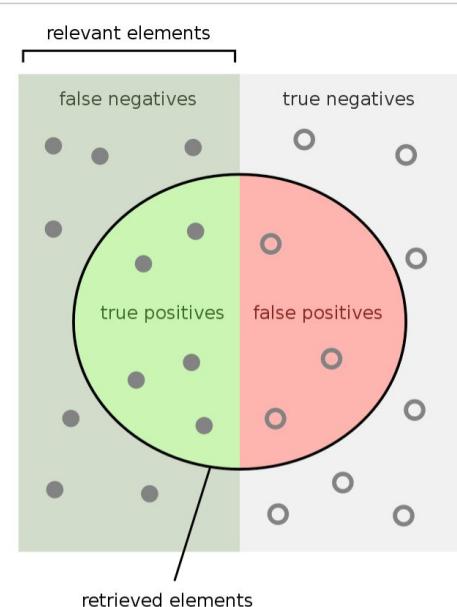
---

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  $\circledast$ (R)
- $F1 = 2^*(P * R) / (R + R)$



Scharkow, 2012, 133–36

# Tools for manual coding

- Google sheets
  - AmCAT: <https://vu.amcat.nl/accounts/login/?next=/>
  - AnnoTinder: [https://github.com/ccs-amsterdam/CCS\\_annotator](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

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 Gerechtigkeit 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
...	...	...	...	...

Code manually  
↓

Search with search string  
↓

Calculate recall and precision  
↓

# 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

# Regular Expressions (regex)

# Regular Expressions (regex)

- How to pronounce?

# Regular Expressions (regex)

- How to pronounce?
- /gɪf/
- /dʒɪf/

# Regular Expressions (regex)

- How to pronounce?
- /gɪf/
- /dʒɪf/
- /ʊɛ.ɡɛks/
- /ʊɛ.dʒɛks/

# Regular Expressions (regex)

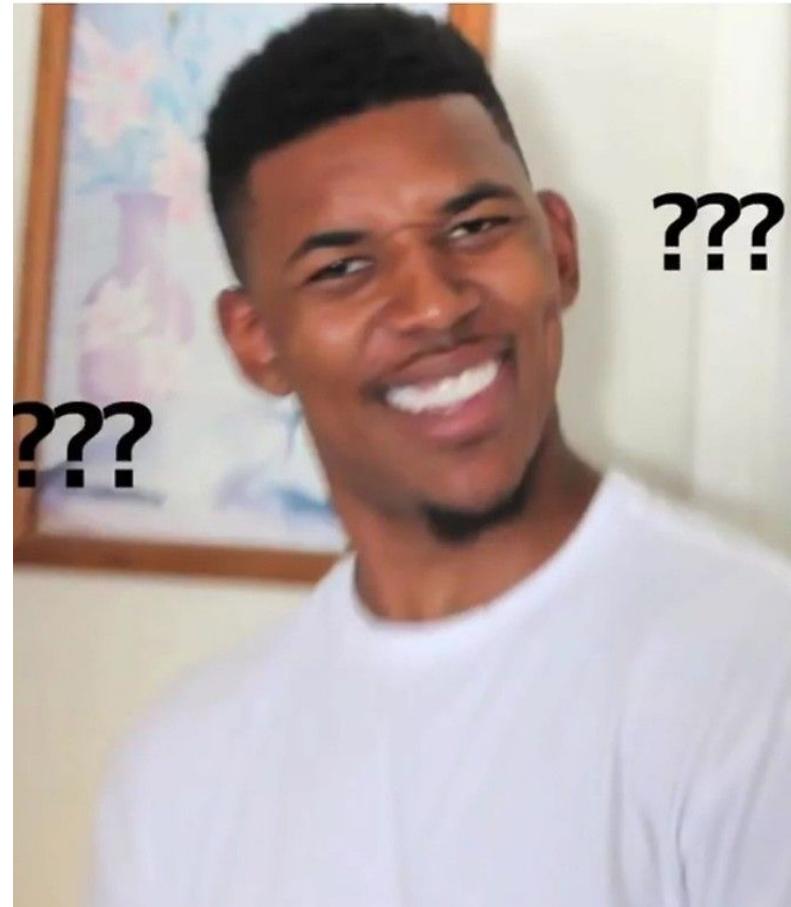
- How to pronounce?
- **/gɪf/**
- /dʒɪf/
- **/rɛ.ɡɛks/**
- /rɛ.dʒɛks/

# Regular Expressions (regex)

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

# Regular Expressions (regex)

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



# regex

- Formal language to specify search strings

# regex

- Formal language to specify search strings
- Insanely difficult

# regex

- Formal language to specify search strings
  - INSANELY difficult
-

# regex

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

# regex

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

# regex

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

# 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!!!

- Disjunctions

Grimmer / Jurafsky  
Cheat-sheet

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

- Ranges

RE	Match	Example Patterns Matched
[A-Z]	an upper case letter	“ <u>Rep.</u> <u>Anthony</u> <u>Weiner</u> ( <u>D</u> - <u>Brooklyn</u> & <u>Queens</u> )”
[a-z]	a lower case letter	“ACORN’s”
[0-9]	a single digit	“(9th CD) ”

- Negations

RE	Match	Example Patterns Matched
[^A-Z]	not an upper case letter	“ACORN’s”
[^Ss]	neither ‘S’ nor ‘s’	“ <u>ACORN</u> ’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

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

RE	Match	Example Patterns Matched
^ [A-Z]	Upper case start of line	“ <u>Palo Alto</u> ” “the town of <u>Palo Alto</u> ” “ <u>the town of Palo Alto</u> ” “ <u>Palo Alto</u> ” “ <u>Palo Alto</u> ” “ <u>the town of Palo Alto</u> ”
^ [^A-Z]	Not upper case start of line	
^. .	Start of line	
. \$	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 <u>yours</u> or <u>mine</u> ”
\ 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 <u>-Mississippi</u> ”
\ W	Any non-alpha numeric	“1-Mississippi”

# How difficult to regex an email

# How difficult to regex an email

Rather

# How difficult to regex an email

```
(?:[a-zA-Z0-9!#$%&'*+/=?^_-`{|}~-]+(?:\.[a-zA-Z0-9!#$%&'*+/=?^_-`{|}~-]+)*|"(?:[\x01-\x08\x0b\x0c\x0e-\x1f\x21\x23-\x5b\x5d-\x7f]|\\"[\x01-\x09\x0b\x0c\x0e-\x7f])*")@(?:(?:[a-zA-9](?:[a-zA-9-]*[a-zA-9])?\.)+[a-zA-9](?:[a-zA-9-]*[a-zA-9])?|\[(?:((?:2(5[0-5]|[0-4][0-9])|1[0-9][0-9]|1[1-9]?[0-9]))\.)\{3\}(?:((?:2(5[0-5]|[0-4][0-9])|1[0-9][0-9]|1[1-9]?[0-9]))|[a-zA-9-]*[a-zA-9]:(?:[\x01-\x08\x0b\x0c\x0e-\x1f\x21-\x5a\x53-\x7f]|\\"[\x01-\x09\x0b\x0c\x0e-\x7f])+)\])
```





## 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/](https://spannbaueradam.shinyapps.io/r_regex_tester/)

---

## Exercise in R

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

