

Advanced quantitative text analysis (2022W)

,

Day 2

Today

9:45-10:40	Input	
11:00-11:30	Coffee Break	
11:30-12:15	Input	
12:00-13:00	Coding	

Text Processing / Feature Engineering

Day 2 Session 1

Text Preprocessing

- Texts are **highly** dimensional
- When possible, it is nice to reduce this dimensionality
- Ideally, without losing too much information

Danny & Spirling, 2018

- Punctuation
 - Numbers
 - Lowercasing
 - Stemming
 - Stop-words
 - N-grams
 - Removal of words by frequency
-

Punctuation / Numbers / Lowercasing

- Fairly straightforward
- Often we don't care about punctuation and/or numbers – so, might be better to remove them
- We probably do care about the letter case
 - But to what extent?
 - Reduction in dimensions might be worth the reduction in accuracy
 - Can you think of examples when we do /don't care about the case?

Stemming / Lemmatization

- A **stem** is the part of the word responsible for lexical meaning
- A stem is invariable part of the word under inflection
- “wait” is a stem of:
 - “waiting”
 - “waited”
 - “waits”
- A **lemma** is the base / “original” part of the word
- Both are useful for dimension reduction and often produce similar results

Stop Words

- Words that are filtered out before the analysis begins
- Could be any type of words that you do not want in the analysis
- Usually, function words are used as stop words (*FORESHADOWING...*)
 - “The”
 - “Is”
 - “I”
 - “That”
 - etc.
- Domain-specific words are also often excluded from the analysis
- E.g., “Global Warming” in the corpus of texts about Global Warming

N-grams

- So far, we've only looked at "unigrams" – individual words
 - Texts can be broken down into any n-gram sequences
 - "I love ice-cream and bananas"
 - "I" "love" "ice-cream" "and" "bananas"
 - "I love" "love ice-cream" "ice-cream and" "and bananas"
 - 3-grams?
-

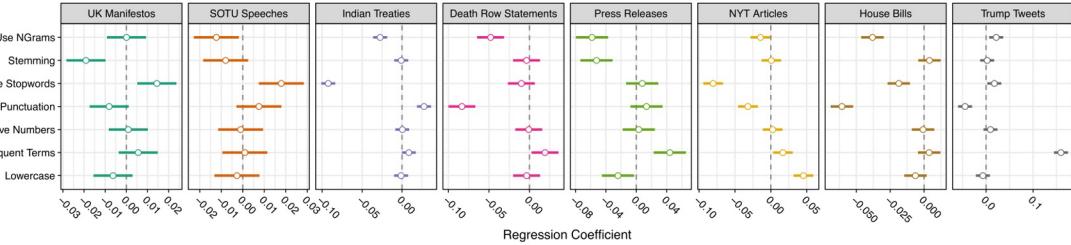
Removal of terms by frequency

- Further removal of dimensionality can be achieved by removing either very frequent or very infrequent terms
- If they are very frequent, they probably don't carry much discriminating information for our analysis (think stopwords)
- If they are very infrequent, they probably carry a lot of discriminating information, but very low statistical power

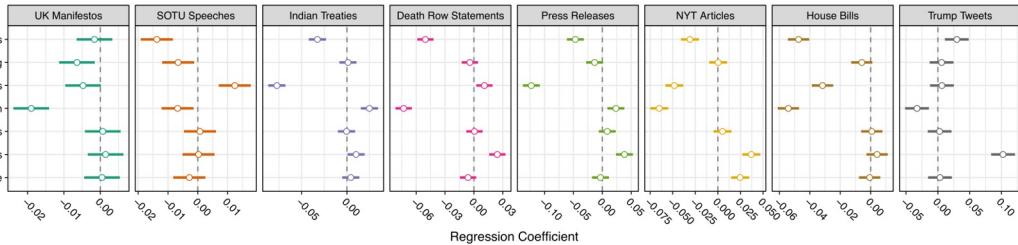
EXAMPLE SENTENCE HERE!!!

Danny & Spirling, 2018

Top 10 Pairs



Top 50 Pairs



Top 100 Pairs

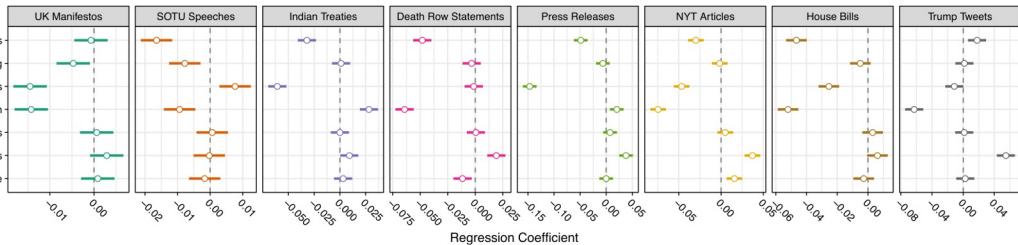


Figure 5. Regression results depicting the effects of each of the seven preprocessing steps on the preText score for that preprocessing combination.

Tf-idf

We can do more than just **count** words

We can transform these counts

Use some sort of a weight in order to transform

Term frequency inverse document frequency in one form of weighting

Term Frequency

$$tf(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

Term Frequency Frequency

$$tf(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

Inverse Document

$$idf = \log \frac{N}{n_j}$$

Term Frequency Frequency

$$tf(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

Inverse Document

$$idf = \log \frac{N}{n_j}$$

Number of Documents

Number of Documents where the term j appears

tfidf

$$W_{ij} \times \log \frac{N}{n_j}$$

What?

What?

- ***Exactly!***
- Introduced in 1972 by a computer scientist (Spärck Jones, 1972)
- No theoretical justification
- Apart from “it seems to work...”

What?

- ***Exactly!***
- Introduced in 1972 by a computer scientist (Spärck Jones, 1972)
- No theoretical justification
- Apart from “it seems to work...”

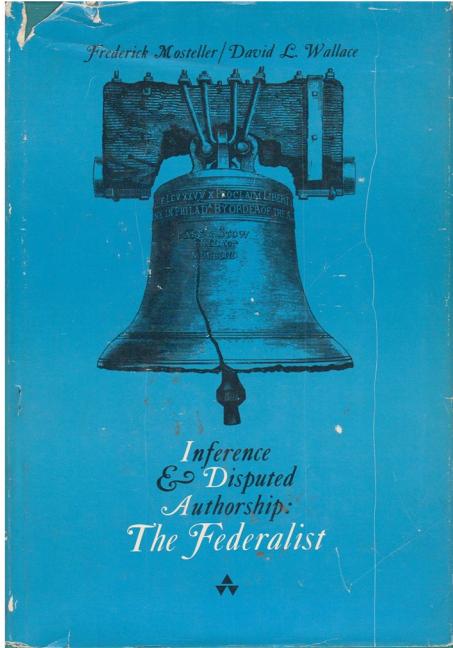
- And *sometimes* it does

Log Odds / Log Odds Ratio

$$\log O_w^i = \log \frac{f_w^i}{1 - f_w^i}$$

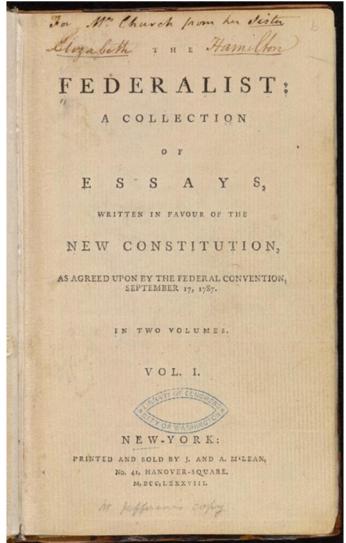
$$\log \frac{O_w^i}{O_w^j} = \log \frac{f_w^i}{1 - f_w^i} / \frac{f_w^j}{1 - f_w^j} = \log \frac{f_w^i}{1 - f_w^i} - \log \frac{f_w^j}{1 - f_w^j}$$

Inference and Disputed Authorship: The Federalist

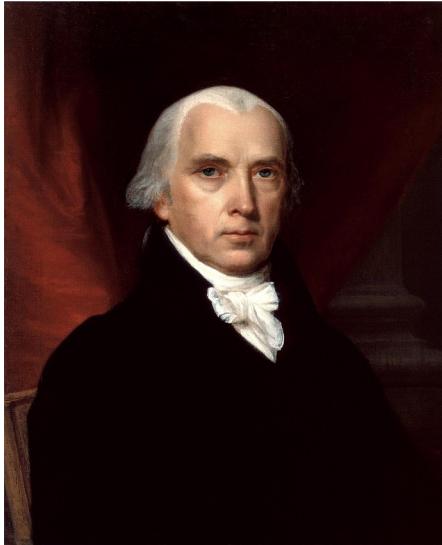
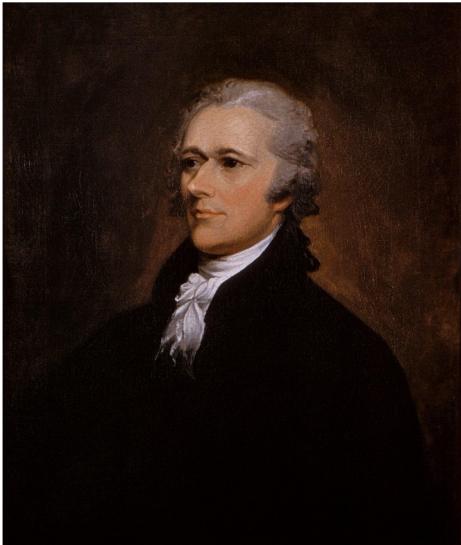
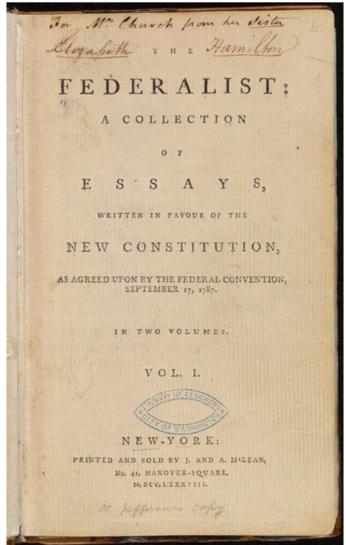


Mosteller & Wallace, 1963

One of the first (if not the first) text-as-data study

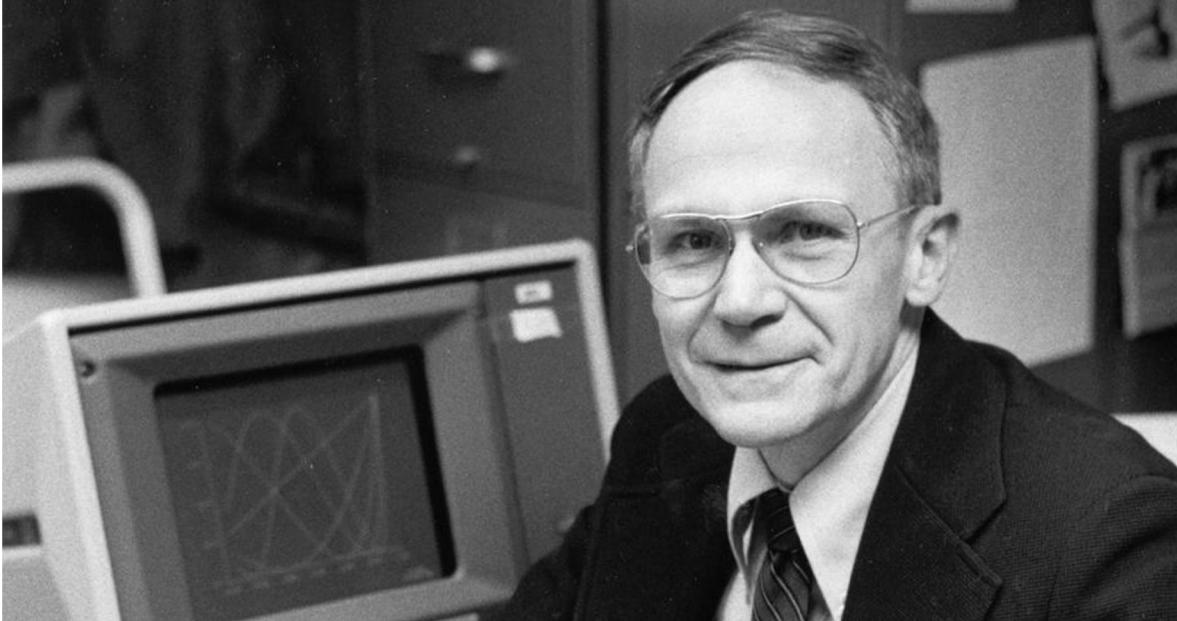


One of the first (if not the first) text analysis study



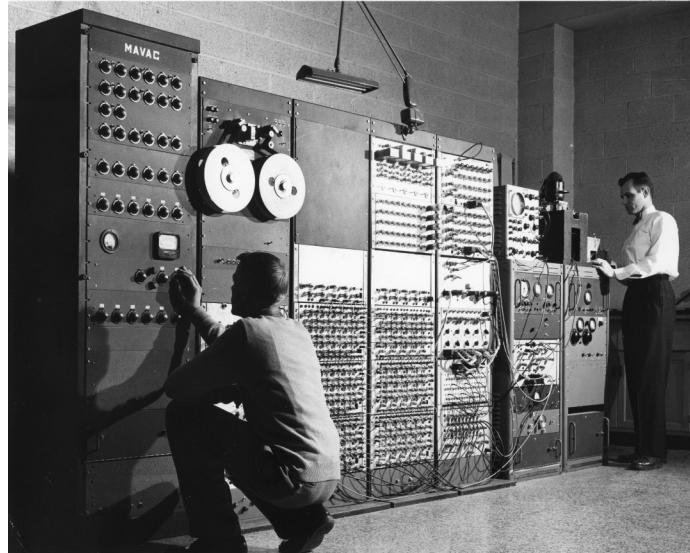
Who wrote them?

- 71 of the essays have a fairly certain authorship
- 12 are disputed
- Big historical debate as to how to ascribe authorship



Computer-assisted text analysis!

Computer-assisted text analysis...?



Dimension Reduction

Remove all the stop-words!

Dimension Reduction

Remove all the stop-words!

Still, too many words...

Dimension Reduction

Remove all the stop-words!

Still, too many words...

Remove **all** words, **but** the stop-words

Dimension Reduction

Remove all the stop-words!

Still, too many words...

Remove **all** words, **but** the stop-words

Maybe there is information in them?

Simplified example from Grimmer et al., 2022

- Focus on:
 - “Man”
 - “By”
 - “Upon”
- The rates with which the authors use these words may indicate authorship

Word Rates

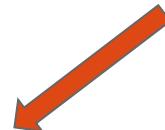
	man	by	upon
Hamilton	102	859	374
Madison	17	474	7
Jay	0	82	1

Word Proportions

	man	by	upon
Hamilton	.076	.643	.28
Madison	.034	.952	.014
Jay	0	.988	.012

Multinomial Language Models

Word Proportions



	man	by	upon
Hamilton	.076	.643	.28
Madison	.034	.952	.014
Jay	0	.988	.012

Disputed Paper

	man	by	upon
Disputed	2	15	0

Disputed Paper

$$p(D|H) = \frac{17!}{2!15!0!} (.076)^2 \times (.643)^{15} \times (.28)^0$$

Total Words

Disputed Paper

Raw Rates from text

$$p(D|H) = \frac{17!}{2!15!0!} (.076)^2 \times (.643)^{15} \times (.28)^0$$

Hamilton Rates

Calculate Jay and Madison yourself

$$p(D|H) = \frac{17!}{2!15!0!} (.076)^2 \times (.643)^{15} \times (.28)^0 = .001$$

$$p(D|J) = \frac{17!}{2!15!0!} (0)^2 \times (.988)^{15} \times (.012)^0 = 0$$

$$p(D|M) = \frac{17!}{2!15!0!} (.034)^2 \times (.952)^{15} \times (.014)^0 = .076$$

Federalist Vector Space Model

In the markdown file...

Data collection

Day 2 Session 2

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

Newspaper Repositories

- LexisNexis
- APA
- Factiva

<https://usearch.univie.ac.at/primo-explore/dbsearch?vid=UWI>

Geben Sie Begriffe oder Quellen ein

Letzte 2 Jahre

Alle Inhaltstypen



Kürzlich ausgeführt und Favoriten

Erweiterte Suche

Suchtipps

"Get a Doc" (US Recht)

Suche mit Filtern

Was suchen Sie?

Nachrichten

eine Publikation

US Entscheidungen

Rechtszeitschriften (Englisch)

Firmeninformationen

Ländersuche

In allen Nachrichten suchen nach

Suchbegriffe eingeben

Einen Datumsbereich auswählen

Alle verfügbaren Daten

Suchen

Search

Search term (empty = all articles in selected search period)

e.g. Salzburg Culture, "United Nations" ([Help](#))

Activate word stem search

SEARCH

[Clear search terms](#)

Search period

from Beginning of previous month

01.10.2022



-

to 7 days ago

06.11.2022



Sources

search for sources



[sources from bundle \(312\)](#)



[selected sources \(18\)](#)

Quellenbündel

- Tageszeitungen, Print
- Zeitschriften und Magazine, Print
- Radio und TV
- Online-Medien
- Branchen
- Fachdatenbanken

Euro (D)

1st FIRST

24sata (HR)

A3 BAU

abendblatt.de

Abendzeitung (D)

Academia

Across

aerotelegraph.com

Alpbach News

andreas-unterberger.at

APA-AußenwirtschaftsNews

APA-BauNews

APA-MobilitätsNews

Der Standard

Die Presse

Heute

Kärntner Tageszeitung (hist. Bestand)

Kleine Zeitung

Kronen Zeitung

Kurier

Medianet

Neue Vorarlberger Tageszeitung

Oberösterreichisches Volksblatt

OÖ Nachrichten

Salzburger Nachrichten

Salzburger Volkszeitung (hist. Bestand)

Tiroler Tageszeitung

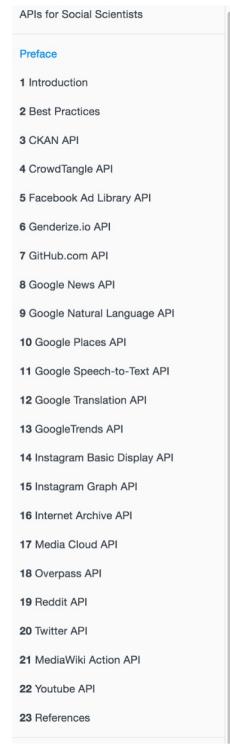
[Show/hide source information buttons](#)

APIs

API = Application Programming Interface

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

API Examples



The screenshot shows a bookdown interface for a collaborative review of APIs for social scientists. At the top, there's a navigation bar with icons for search, refresh, and other document functions. The title "APIs for Social Scientists" is at the top left, followed by a "Preface" section. The main content lists 23 numbered API entries from 1 to 23, each with a brief description. To the right, there's a sidebar with sections for "Current editors" (listing Paul C. Bauer, Camille Landesvatter, and Lion Behrens), "Authors & contributors" (listing many names), and two dates: "First public version: 29 November, 2021" and "This version: 11 Oktober, 2022". Below the main content, there's a "Preface" section with text about the purpose of the book and how to contribute. At the bottom, it says the document was generated with R, RMarkdown, and Bookdown.

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 Twitter API

21 MediaWiki Action API

22 YouTube API

23 References

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. Kaplan, Jana Klein, Markus Konrad, Barbara K. Kreis, Dean Lajic, Camille Landesvatter, Madleen Meier-Barthold, Grace Olzinski, Nina Osenbrügge, Ondrej Pekacek, Pirmin Stöckle, Malte Söhren, Domantas Undžėnas

First public version: 29 November, 2021
This version: 11 Oktober, 2022

Preface

The present online book provide a review of APIs that may be useful for social scientists. Please start by reading the [Introduction](#). The material was/is being developed by various contributors that you can find above and in the contributor section of the corresponding [github repository](#). If you are interested in contributing please check out the Section [How to contribute](#) in the github README.

The material is licensed under a [Apache License 2.0](#) license. Where we draw on other authors material other licenses may apply. We are extremely grateful for feedback and if you find errors please let us know.

This document was generated with [R](#), [RMarkdown](#) and [Bookdown](#).

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

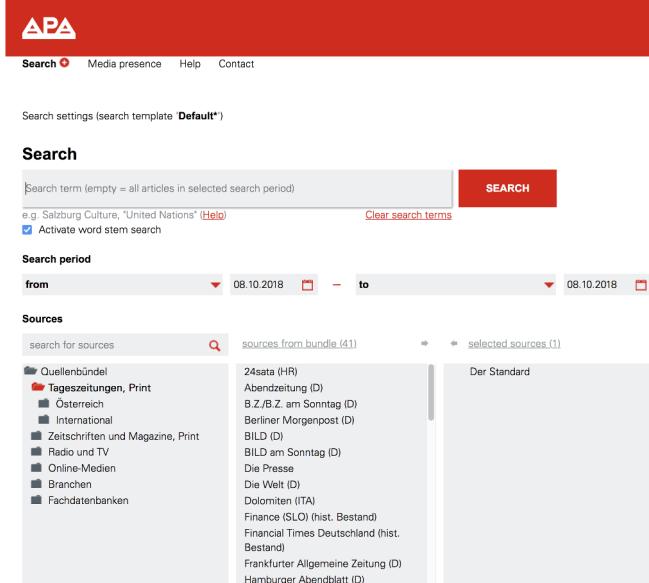
APA-UNIVIE Data project

What: obtain data (Austrian news outlets) for scientific use from the media archive of the Austria Press Agency (APA) via the APA-API

Who: Faculty members of the University of Vienna

How: API Access application

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



The screenshot shows the search interface for the APA OnlineManager Library. At the top, there is a red header bar with the APA logo. Below it, the main search area has a search term input field containing "Salzburg Culture, \"United Nations\"", a "SEARCH" button, and a "Clear search terms" link. Under "Search period", the "from" date is set to "08.10.2018" and the "to" date is also "08.10.2018". The "Sources" section on the left lists categories like Quellenbündel, Tagesschriften, Print, Österreich, International, Zeitschriften und Magazine, Print, Radio und TV, Online-Medien, Branchen, and Fachdatenbanken. On the right, a list of sources is shown, including 24sata (IHR), Abendzeitung (D), B.Z./B.Z. am Sonntag (D), Berliner Morgenpost (D), BILD (D), BILD am Sonntag (D), Die Presse, Die Welt (D), Dolomiten (ITA), Finance (SLO) (hist. Bestand), Financial Times Deutschland (hist. Bestand), Frankfurter Allgemeine Zeitung (D), and Hamburger Abendblatt (D). A sidebar on the right indicates "selected sources (1)".

search for “APA OnlineManager Library” in u:search database

Twitter API

Recommended Tutorial:

<https://github.com/twitterdev/getting-started-with-the-twitter-api-v2-for-academic-research>

Twitter API

<https://developer.twitter.com/en/docs/twitter-api>

V2 Access Levels

Essential

With Essential access, you can now get access to Twitter API v2 quickly and for free!

- Retrieve 500,000 Tweets per month
- 1 Project per account
- 1 App environment per Project
- Limited access to standard v1.1 (**only media endpoints**)
- No access to premium v1.1, or enterprise

Elevated

With Elevated access, you can get free, additional access to endpoints and data, as well as additional App environments.

- Retrieve 2 million Tweets per month
- 1 Project per account
- 3 App environments per Project
- Access to standard v1.1, premium v1.1, and enterprise

Academic Research

If you qualify for our Academic Research access level, you can get access to even more data and advanced search endpoints.

- Retrieve 10 million Tweets per month
- Access to full-archive search and full-archive Tweet counts
- Access to advanced search operators

V2 Access Levels

Essential

With Essential access, you can now get access to Twitter API v2 quickly and for free!

- Retrieve 500,000 Tweets per month
- 1 Project per account
- 1 App environment per Project
- Limited access to standard v1.1 (**only media endpoints**)
- No access to premium v1.1, or enterprise

Elevated

With Elevated access, you can get free, additional access to endpoints and data, as well as additional App environments.

- Retrieve 2 million Tweets per month
- 1 Project per account
- 3 App environments per Project
- Access to standard v1.1, premium v1.1, and enterprise

Academic Research

If you qualify for our Academic Research access level, you can get access to even more data and advanced search endpoints.

- Retrieve 10 million Tweets per month
- Access to full-archive search and full-archive Tweet counts
- Access to advanced search operators

Academic Research

Overview

For academics who have a research project that requires, or would benefit from, studying Twitter's conversational data. Access is free. An application is required.



Your Project has Academic Research access:

General Computational Communication Science Research

Apps

1 environment per project

Tweets

10M Tweets per month / Project

Cost

free

License

For non-commercial use only

~~Facebook~~ Meta

FORT – Facebook Open Research and Transparency
(<https://fort.fb.com/researcher-apis>)

Facebook academic API – Early Access, last update from 2021...

CrowdTangle (<https://www.crowdtangle.com>)

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>

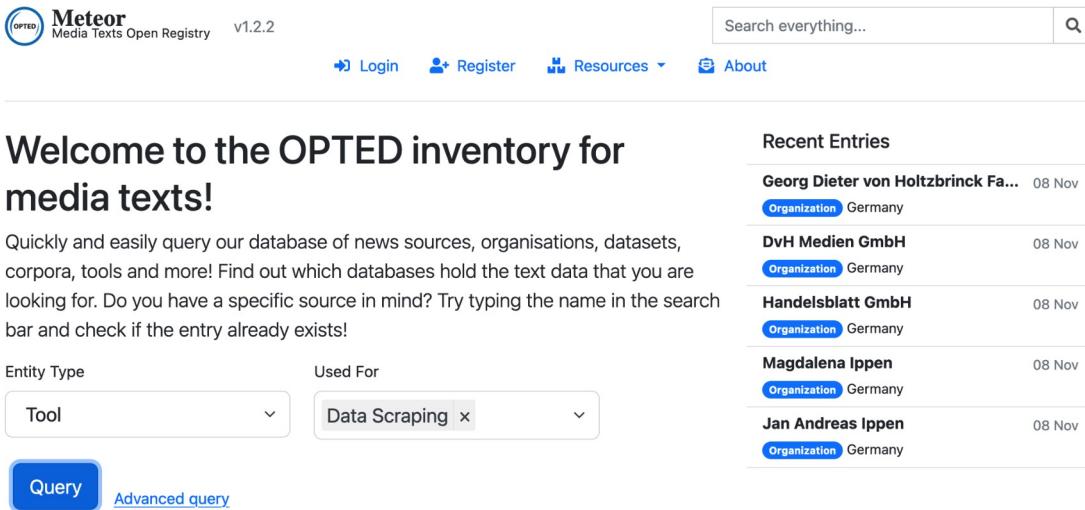


Collection of Data Sources for Communication Science

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

OPTED Tool Collection

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



The screenshot shows the homepage of the Meteor Media Texts Open Registry version 1.2.2. At the top, there is a navigation bar with links for Login, Register, Resources, and About, along with a search bar. Below the header, a large section welcomes users to the OPTED inventory for media texts, explaining its purpose as a database for news sources, organizations, datasets, corpora, tools, and more. It encourages users to search for specific entries. Below this text are two dropdown menus: 'Entity Type' set to 'Tool' and 'Used For' set to 'Data Scraping'. At the bottom of this section are two buttons: a blue 'Query' button and a link to 'Advanced query'. To the right, a 'Recent Entries' sidebar lists several entries with their types (Organization), names (Georg Dieter von Holtzbrinck Fa..., DvH Medien GmbH, Handelsblatt GmbH, Magdalena Ippen, Jan Andreas Ippen), locations (Germany), and dates (08 Nov). Each entry is preceded by a small blue 'Organization' badge.

Meteor
Media Texts Open Registry v1.2.2

Search everything... 

Login Register Resources About

Welcome to the OPTED inventory for media texts!

Quickly and easily query our database of news sources, organisations, datasets, corpora, tools and more! Find out which databases hold the text data that you are looking for. Do you have a specific source in mind? Try typing the name in the search bar and check if the entry already exists!

Entity Type Used For

Tool Data Scraping

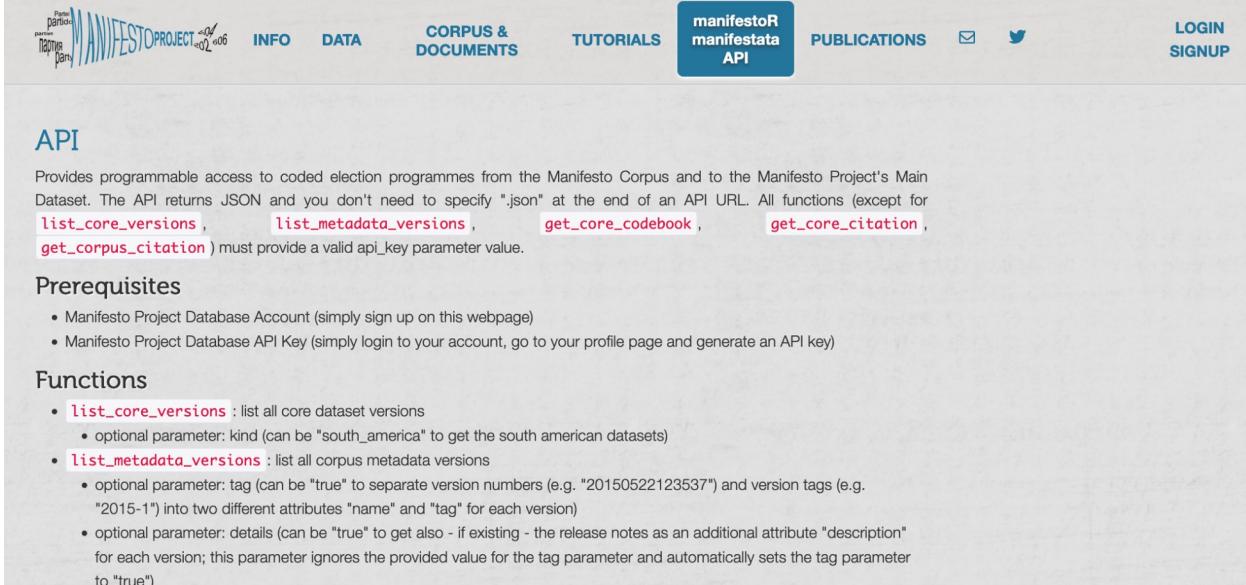
Query [Advanced query](#)

Recent Entries

Georg Dieter von Holtzbrinck Fa...	08 Nov
Organization Germany	
DvH Medien GmbH	08 Nov
Organization Germany	
Handelsblatt GmbH	08 Nov
Organization Germany	
Magdalena Ippen	08 Nov
Organization Germany	
Jan Andreas Ippen	08 Nov
Organization Germany	

Manifesto project

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 bar, the title "API" is centered above a detailed description of the API's purpose and usage. It explains that the API provides programmable access to election programmes and datasets, returning JSON. It specifies that most functions require a valid api_key parameter. Below this, sections for Prerequisites and Functions are listed with their respective bullet points.

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")

Data Scraping

Rvest package

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 is maybe the moment to make a start
- c) You have data already? Check out other sources for the next project or help others to find ways to access data