universität wien

Advanced quantitative text analysis (2022W)

,

Day 2

universität wien

Today

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

universität wien

Text Processing / Feature Engineering

Day 2 Session 1

universität wien

Text Preprocessing

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

universität wien

Danny & Spirling, 2018

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

universität wien

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?

universität wien

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

universität wien

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

universität wien

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?

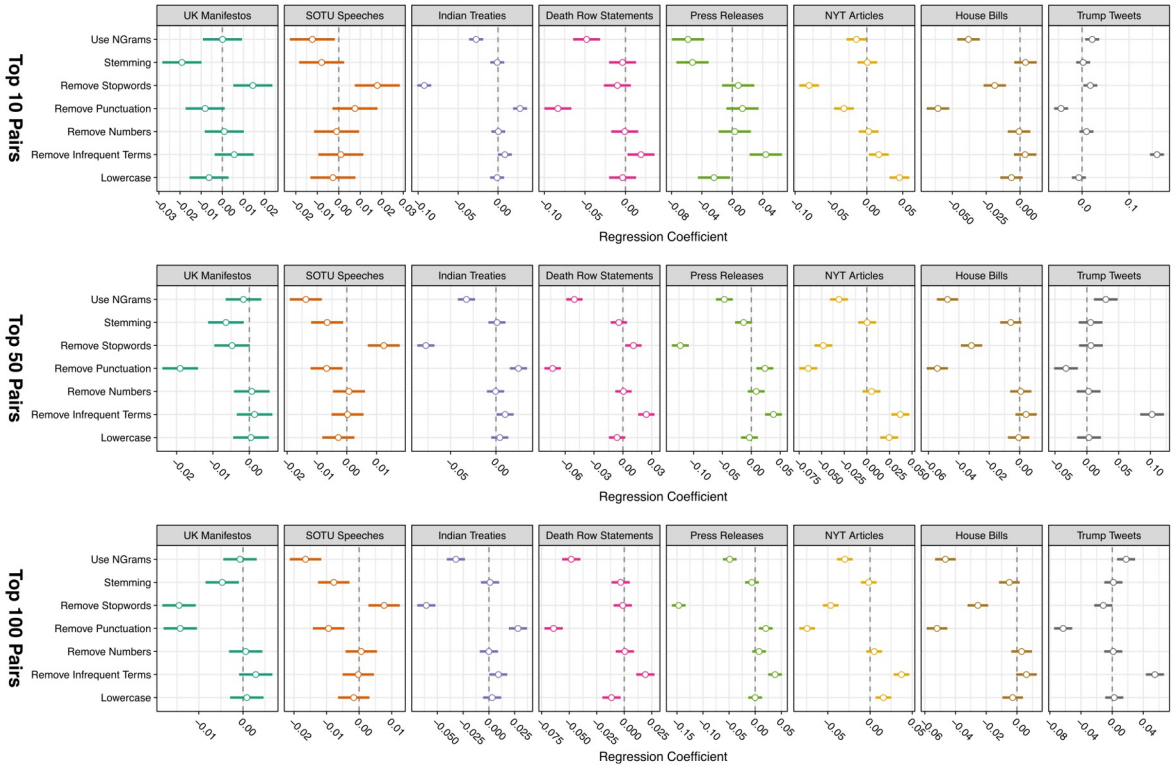
universität wien

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

universität wien

EXAMPLE SENTENCE HERE!!!



Danny & Spirling, 2018

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

universität wien

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

universität wien

Term Frequency

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

universität wien

Term Frequency Frequency

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

Inverse Document

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

universität wien

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

universität wien

tfidf

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

universität wien

What?

universität wien

What?

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

universität wien

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

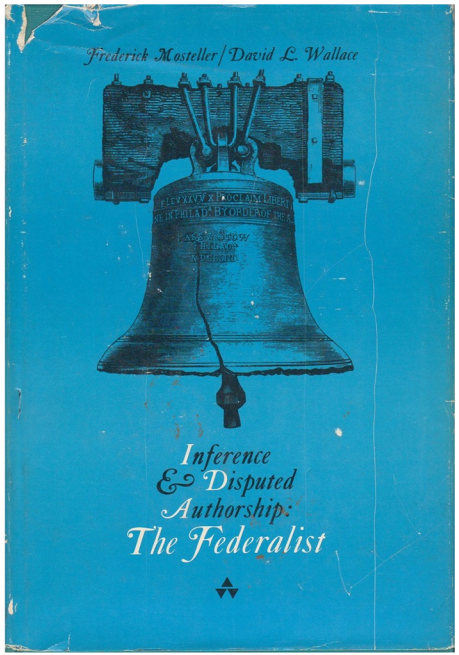
universität wien

Log Odds / Log Odds Ratio

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

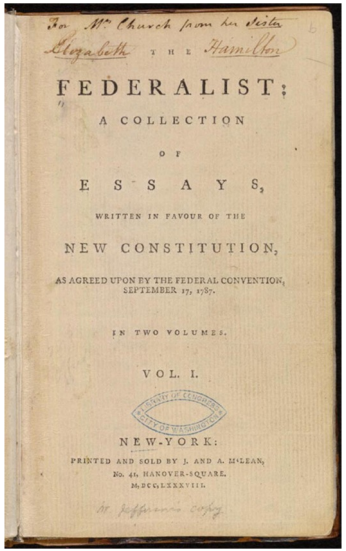
$$\log \frac{O_w^i}{O_w^j} = \log \frac{f_w^i}{1 - f_w^i} / \frac{f_w^j}{1 - f_w^j} = \log \frac{f_w^i}{1 - f_w^i} - log \frac{f_w^j}{1 - f_w^j}$$

Inference and Disputed Authorship: The Federalist



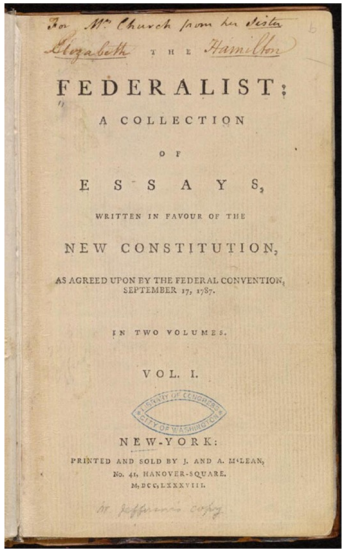
Mosteller & Wallace, 1963

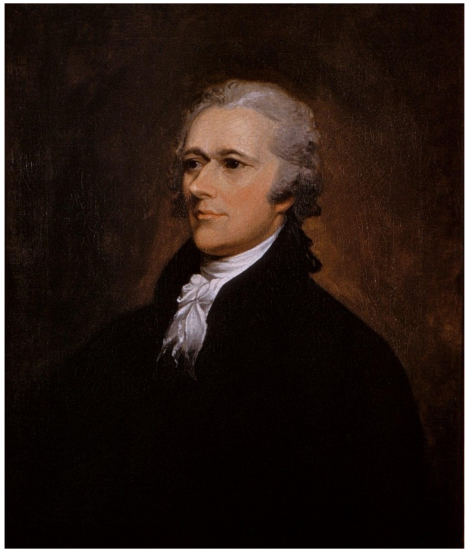
One of the first (if not the first) text-as-data study

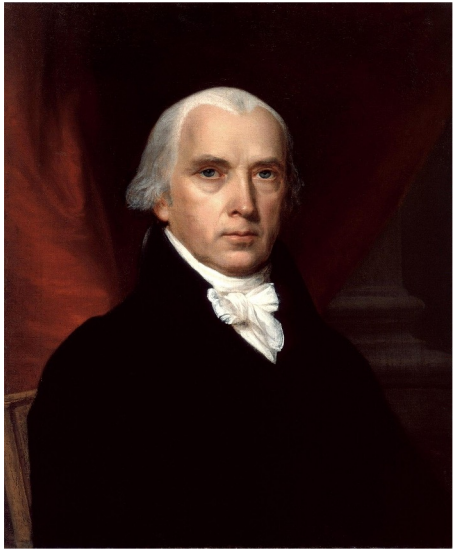


universität wien

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


For Mrs Church from her Sister
Elizabeth
THE FEDERALIST:
A COLLECTION
OF
ESSAYS,
WRITTEN IN FAVOUR OF THE
NEW CONSTITUTION,
AS AGREED UPON BY THE FEDERAL CONVENTION,
SEPTEMBER 17, 1787.
IN TWO VOLUMES.
VOL. I.
NEW-YORK:
PRINTED AND SOLD BY J. AND A. M'LEAN,
No. 41, HANOVER-SQUARE.
M,DCC,LXXXVIII.
Mr Jefferson's Copy



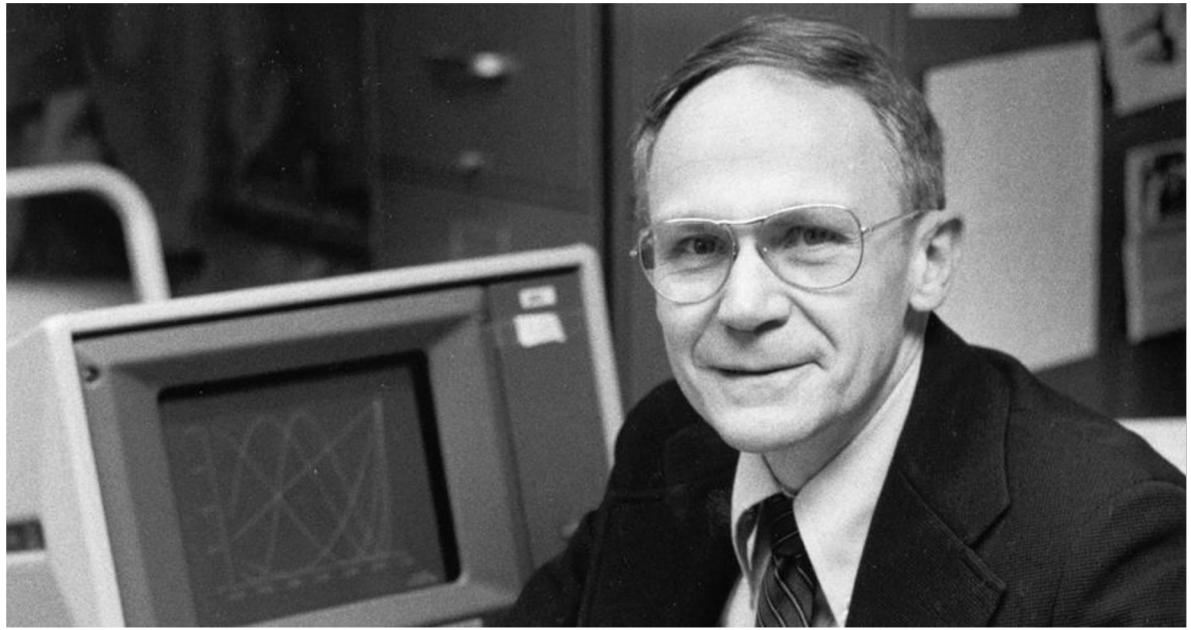


universität wien

Who wrote them?

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

universität wien

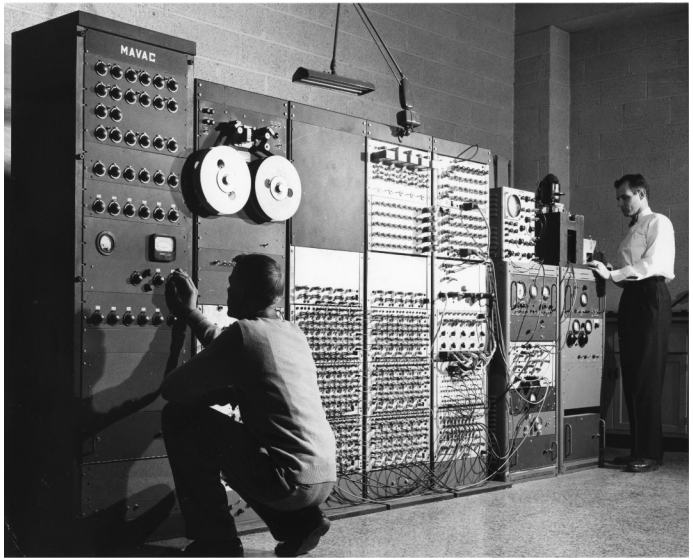


universität wien

Computer-assisted text analysis!

universität wien

Computer-assisted text analysis…?



universität wien

Dimension Reduction

Remove all the stop-words!

universität wien

Dimension Reduction

Remove all the stop-words!

Still, too many words…

universität wien

Dimension Reduction

Remove all the stop-words!

Still, too many words…

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

universität wien

Dimension Reduction

Remove all the stop-words!

Still, too many words…

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

Maybe there is information in them?

universität wien

Simplified example from Grimmer et al., 2022

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

universität wien

Word Rates

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

universität wien

Word Proportions

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

universität wien

Multinomial Language Models

Word Proportions

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

universität wien

Disputed Paper

	man	by	upon
Disputed	2	15	0

universität wien

Disputed Paper

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

universität wien

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

universität wien

Calculate Jay and Madison yourself

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

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

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

universität wien

Federalist Vector Space Model

In the markdown file…

universität wien

Data collection

Day 2 Session 2

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

Newspaper Repositories

- LexisNexis
- APA
- Factiva

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

Nexis Uni Menu

DE - Deutsch Verlauf Hilfe Anmelden | Registrieren

New to Nexis Uni?

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

universität wien

APA

Search settings (search template '**Default**')

Search

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

SEARCH

e.g. Salzburg Culture, "United Nations" (Help) Clear search terms

Activate word stem search

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

€uro (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-eMobilitätsNews

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

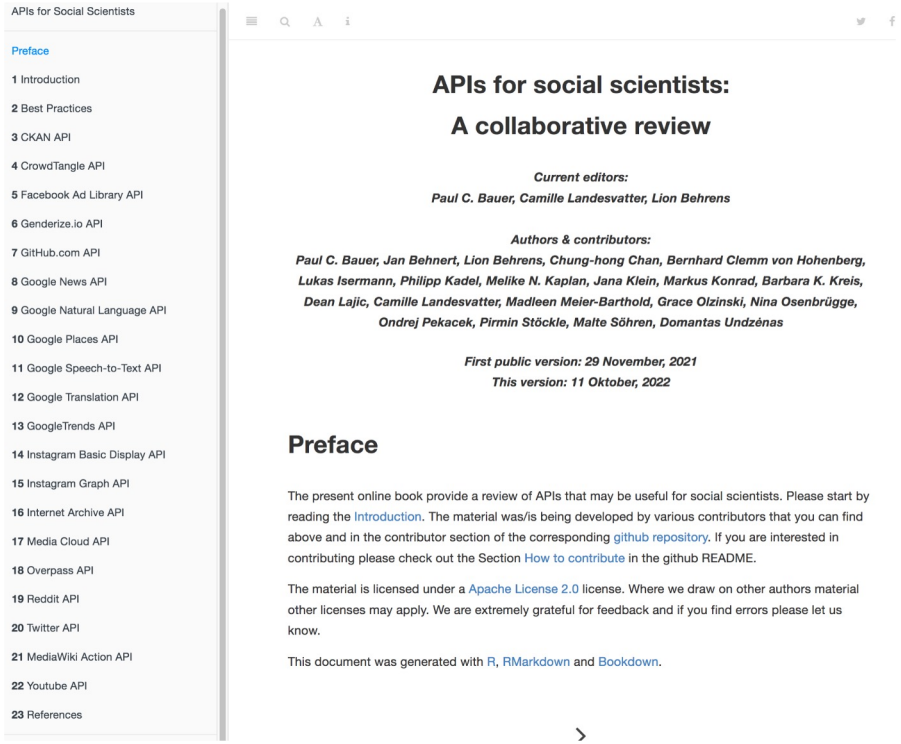
universität wien

APIs

API = Application Programming Interface

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

API Examples



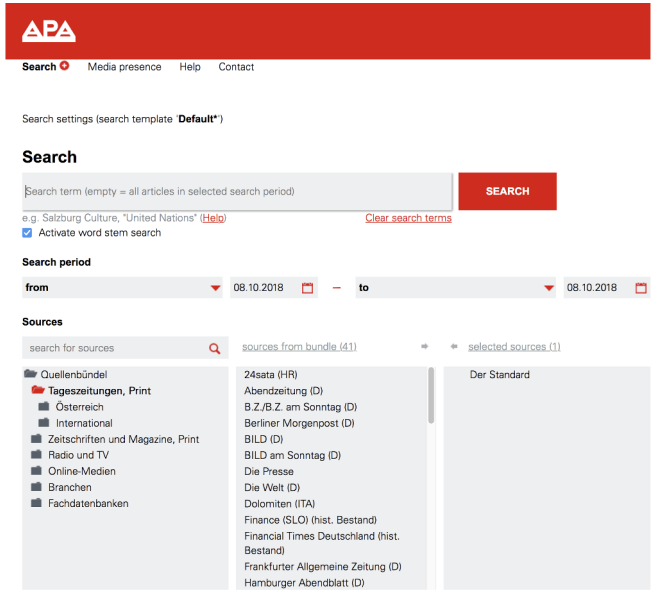
https://bookdown.org/paul/apis_for_social_scientists/

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



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

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

universität wien

Twitter API

Recommended Tutorial:

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

universität wien

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

universität wien

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**

universität wien

~~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)

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

CCL Vienna › 513 Facebook Pages

All Page Lists

Search your lists for any of these words or phrases

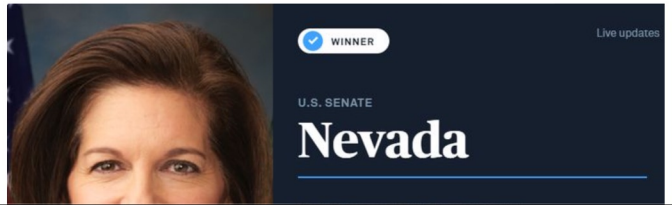
Posts | Leaderboard | Manage

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



universität wien

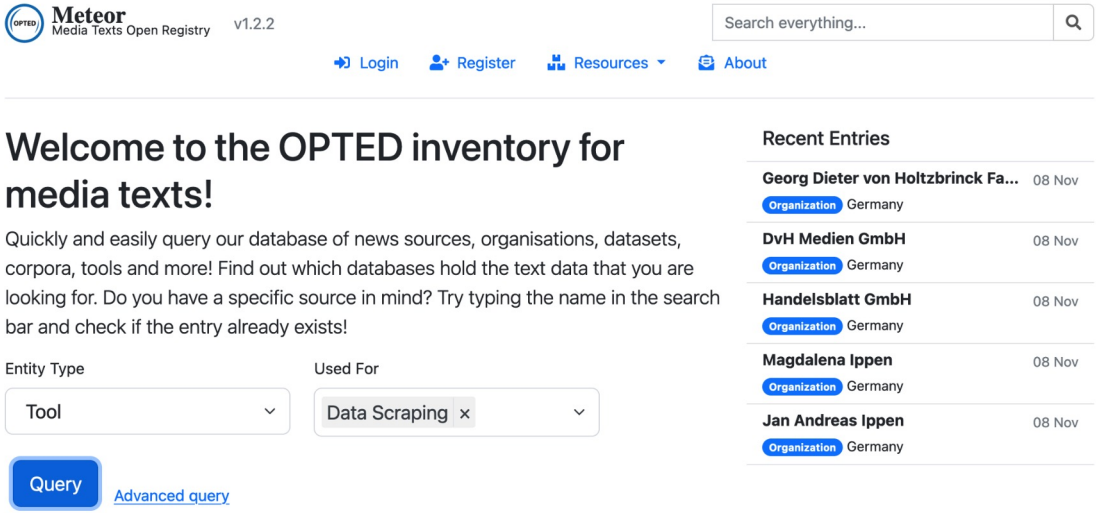
Collection of Data Sources for Communication Science

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

universität wien

OPTED Tool Collection

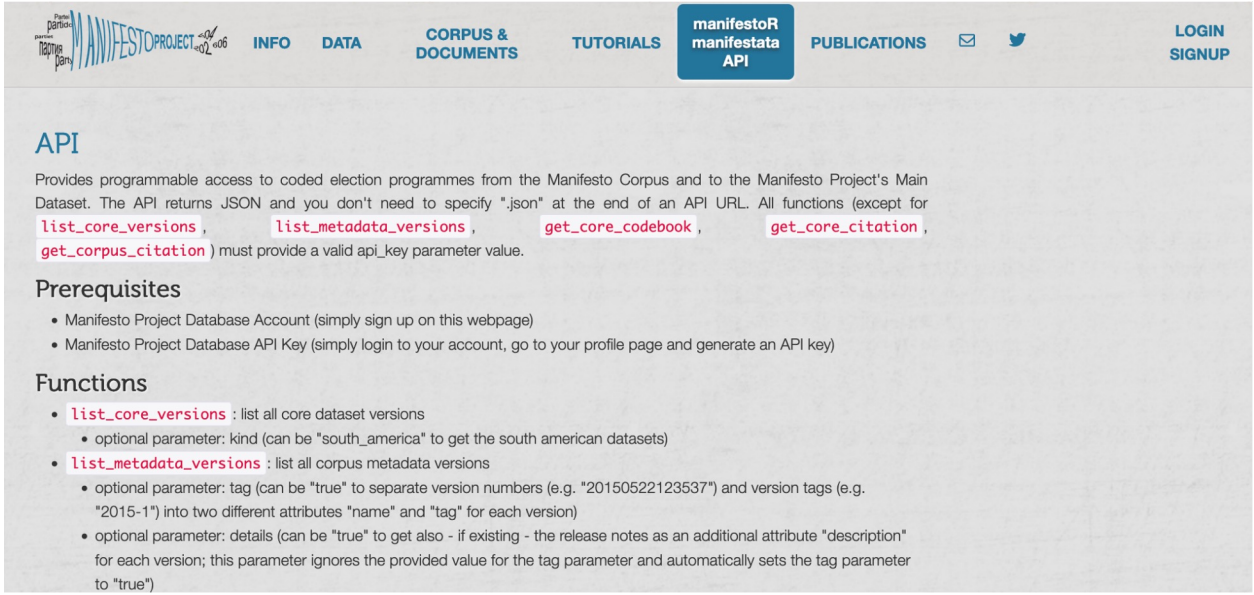
Visit: https://meteor.opted.eu/



universität wien

Manifesto project

manifestoR



universität wien

Data Scraping

Rvest package

universität wien

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