Recipe: Examining words and phrases

Tommy Shane / First Draft / King's College London @tommyshane
Marc Tuters / UvA
Tom Willaerts / VUB

Introduction

Corpus

Selecting words, phrases and concepts

Recipe

References

Introduction

While conspiracy theories have long been understood as irrational narratives produced by extremists in the margins of political and social life, a body of humanities scholarship considers them as critical responses to the complexities and uncertainties of (post)modern life (Knight 2000). While conspiracy theory is often imagined to be a right-wing preoccupation, with the pandemic we have seen the emergence of new "diagonal movements" that cut across traditional left/right distinctions sharing the conviction that all power is conspiracy (Slobodian & Callison 2021).

To explore the parallels between conspiracy theory and more "legitimate" forms of critical thought, researchers can attend to the role of words and phrases as they travel across domains and communities. By combining close and distant reading techniques, this approach bridges gaps between data analytics and discourse analysis, an approach that can be called data hermeneutics:

> data hermeneutics centers on the symbolic analysis of the meaning structures of online conversations … data hermeneutics' chief concern is the synthetic aim of interpreting, reconstructing and explaining the overarching narratives that underpin social media conversations. (Gerbaudo 2016)

Instead of charting the relationships between accounts or posts, this approach focuses on interpreting the relationships between words, the ideas they articulate, and the discourses they construct.

By asking what verbal tools conspiracy communities use and where they got them from, it is possible to answer the question: if conspiracy theorising is a form of social critique, what critical tools do they use, how do they use them, and what kind of social critique do they make possible?

Verbal boundary objects

Like conspiracy theories themselves, words can connect disparate domains and communities. For example, the word ‘gaslight’ might discursively link domains such as social justice campaigning with conspiracy theorising, and the ‘woke’ left with the libertarian right.

In cases like this, words are boundary objects: “objects which are both plastic enough to adapt to local needs and the constraints of the several parties employing them, yet robust enough to maintain a common identity across sites” (Star and Griesemer, 1989, p. 393).”

Verbal boundary objects, such as ‘gaslight’, tend to originate from one domain and community, and get co-opted by another. This can be a powerful disinformation tactic, illustrated by the success of Donald Trump’s co-optation of the phrases “fake news” and “big lie” for his own purposes, and confusing the discursive clarity of the terms and thus its ability to critique him.

By attending to how a single word or phrase moves between communities, it is possible to interrogate the tactics used by conspiracy communities and the kinds of social critique they afford.

Corpus

The corpus is of Twitter posts related to conspiracy theory during the period of the pandemic. The original corpus was collected using a long list of conspiracy theory hashtags and keywords (e.g. #BillGatesisEvil). Together with domain experts a list of words was developed in four categories (see below). Using NLP, other words were then identified that were deemed to be used similarly to those in the expert list. The first and second “snowballed” lists of words were then used to once again scrape Twitter (when those words co-occurred with covid OR coronavirus OR covid19).

The first corpus is of about 15M Tweets and is called “Fabio’s Secret Tweet Stash”. The second is currently only about 4M but is still scraping.[1] Note that working with the second dataset

[1] ['knowthyself', 'prophesied', 'regain', 'tummy', 'authoritarianism', 'binacom', 'non-binary', 'disputing', 'misrepresented', 'open-minded', 'demand-side','facts', 'make-believe', 'sinister','incontrovertible', 'awakenings', 'activism', 'chinsespolice','connecting', 'subjugated', 'patriarchy', 'whining', 'deceive',

will probably return many more instances of the use of a given term, but because of how they were collected, we can't assume them to be *as* "conspiratorial".

Selecting words, phrases and concepts

We have identified epistemic keywords in four high level themes:

1. **Critical theory** (eg 'biopolitics')
2. **Social justice** (eg 'gaslight')
3. **New Age** (eg 'awakening')
4. **Epistemic** (eg 'truth')

All of the words are contained in column in this spreadsheet, which also features other "snowballed" words (see above)

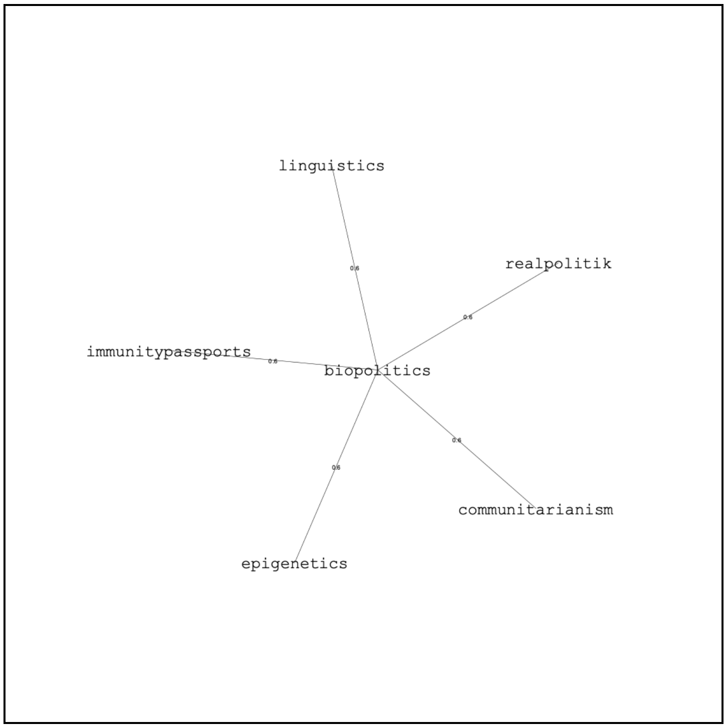


Figure: example of 'snowballed' similar terms for 'biopolitics' (based on gensim w2v word embeddings)

Consider these features when selecting your words, phrases and concepts:

'chauvinism','contrived','uncovering', 'misdirection', 'fabrication', 'wakeupamerica', 'prove','polarization','maybe', 'racial', 'perhaps','immeasurable','ayurveda','trusting', 'fraud-', 'mindset','distorting','reveals','restore','capitalist'] + [coronavirus OR covid OR covid19]

- **Quantity**: ideally keywords should feature above a minimum number of instances to ensure there are enough posts and posters to analyse, but below a maximum to facilitate meaningful close reading.
- **Relevance:** keywords should be relevant to you and your community's interests. For example, a word might have a particularly interesting history, or its co-optation might be particularly surprising (eg 'gaslight' being used by the libertarian right).
- **Centrality:** keywords that link different domains or communities (i.e. have high 'centrality' in the network) may indicate a particularly multi-valent word that has many meanings and uses for different communities (see step 14 & 15).

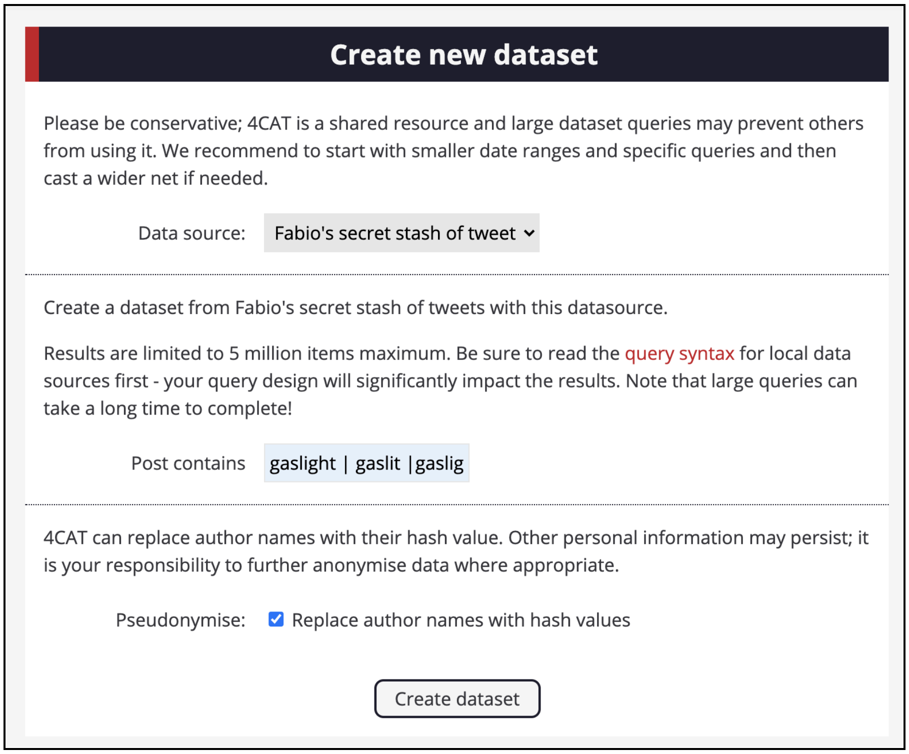
Recipe

1. Choose your keyword(s). For example, 'gaslight', 'gaslit', 'gaslighted', 'gaslighting'

2. Create a worksheet based on this template (feel free to adjust).

3. Create a subfolder in this gdrive directory w/ naming convention: Word_Surname

4. Go to 4CAT, and select 'Fabio's secret stash of tweets' (the second dataset will also be accessible at a given point, see above), and enter your keywords in the correct notation. For example, 'gaslight | gaslit | gaslight | gaslighting'. Use the proper query syntax, and use all the conjugations of the keyword - eg not just 'gaslight', but 'gaslight | gaslit | gaslight | gaslighting'. Un-select the default privacy settings in order to retain "author names" for analyses of authors in step 13.



Create new dataset

Please be conservative; 4CAT is a shared resource and large dataset queries may prevent others from using it. We recommend to start with smaller date ranges and specific queries and then cast a wider net if needed.

Data source: Fabio's secret stash of tweet

Create a dataset from Fabio's secret stash of tweets with this datasource.

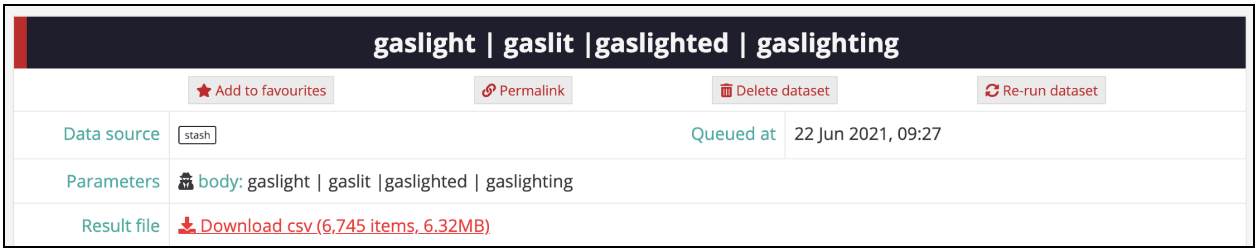
Results are limited to 5 million items maximum. Be sure to read the query syntax for local data sources first - your query design will significantly impact the results. Note that large queries can take a long time to complete!

Post contains: gaslight | gaslit |gaslig

4CAT can replace author names with their hash value. Other personal information may persist; it is your responsibility to further anonymise data where appropriate.

Pseudonymise: ☑ Replace author names with hash values

5. Create dataset, download the csv and upload to the appropriate folder in the [gdrive directory](#) (in settings you can select to automatically "convert uploaded files to Google Docs editor")



gaslight | gaslit |gaslighted | gaslighting

Data source: stash

Queued at: 22 Jun 2021, 09:27

Parameters: body: gaslight | gaslit |gaslighted | gaslighting

Result file: Download csv (6,745 items, 6.32MB)

6. Begin by ranking the posts in the csv based on the number of retweets. Copy across the top 10 tweets into the worksheet. Analyse the key themes.
    a. Order all of your datasets in such a way that it can then be used as the basis for a set of data visualizations. Create a spreadsheet for the entire list of words that your group is exploring. Following the diagram below, enter labels for tweets in column "m". This is an *axial coding scheme*, in which you try to relate data together in order to reveal categorinums. Working in a team, then consult with others as you develop this scheme.
    b. Position the tweets (at least one or two) in the *political compass*. In order to do so please assign a value from -10 to +10 in column "k" and column "l". -10x =

extreme left. +10x = extreme right. -10y = extreme libertarian. +10y = extreme authoritarian. (0x + 0y = ideologically neutral ;-).

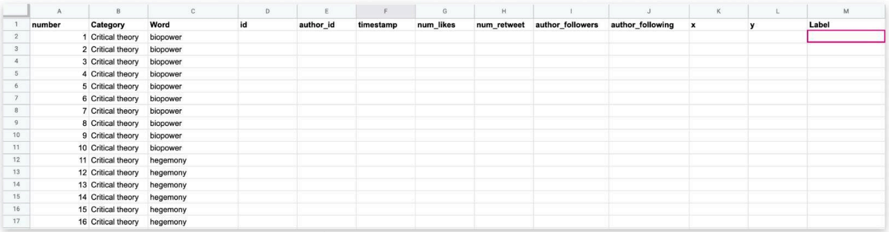
HOW TO FILL THE DATASET

Category

word analyzed

Copy this data from the downloaded dataset

Tematic label assigned to the tweet (eg. yoga, pro-family, ecc..)

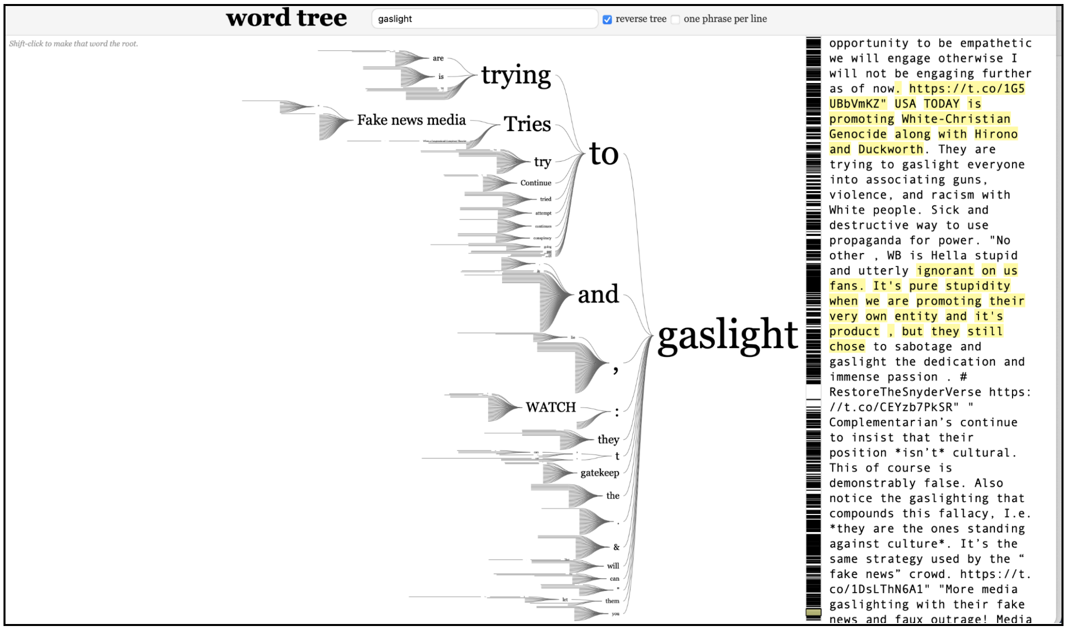


	A	B	C	D	E	F	G	H	I	J	K	L	M
1	number	Category	Word	id	author_id	timestamp	num_likes	num_retweet	author_followers	author_following	x	y	Label
2	1	Critical theory	biopower										
3	2	Critical theory	biopower										
4	3	Critical theory	biopower										
5	4	Critical theory	biopower										
6	5	Critical theory	biopower										
7	6	Critical theory	biopower										
8	7	Critical theory	biopower										
9	8	Critical theory	biopower										
10	9	Critical theory	biopower										
11	10	Critical theory	biopower										
12	11	Critical theory	hegemony										
13	12	Critical theory	hegemony										
14	13	Critical theory	hegemony										
15	14	Critical theory	hegemony										
16	15	Critical theory	hegemony										
17	16	Critical theory	hegemony										

Each row is for one tweet

7. Select all posts in the dataset, copy them, and paste into the interactive word tree tool. You may need to paste into a text editor to strip the formatting from the CSV.

8. Generate the word tree, and begin to explore. Type in your keywords into the box at the top to make that the root of the word tree. Click a word to combine it with the root word to form a phrase. The size of the word or phrase is proportional to the number of times that word (or phrase) precedes or succeeds your root word, so pay particular attention to larger words and phrases.

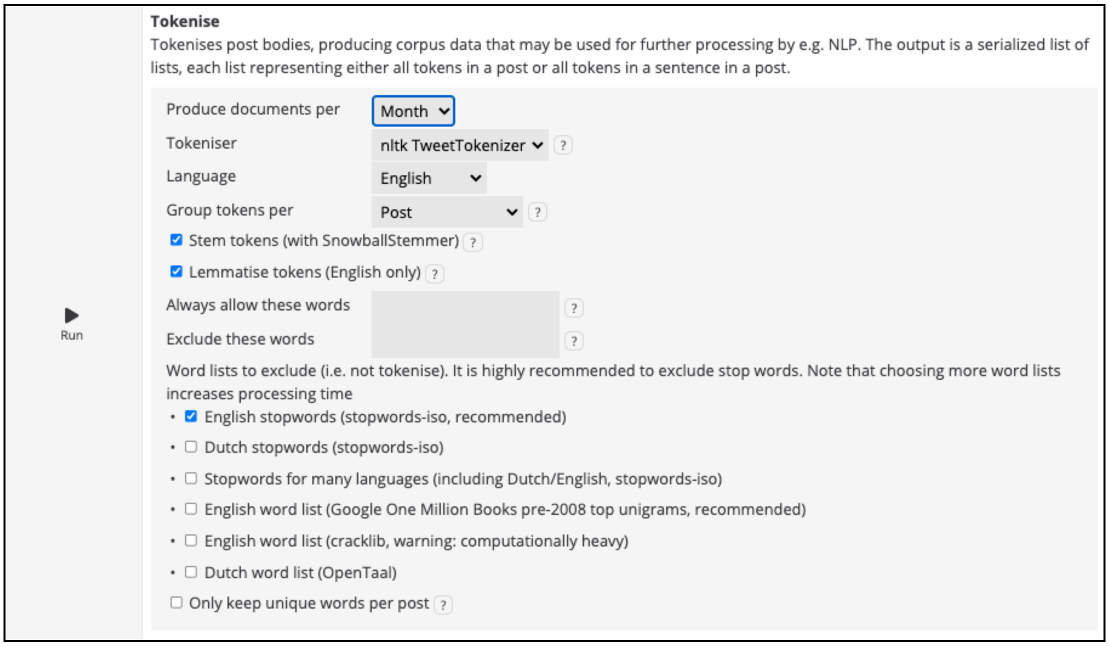


opportunity to be empathetic
we will engage otherwise I
will not be engaging further
as of now. https://t.co/1G5
UBbVmKZ" USA TODAY is
promoting White-Christian
Genocide along with Hirono
and Duckworth. They are
trying to gaslight everyone
into associating guns,
violence, and racism with
White people. Sick and
destructive way to use
propaganda for power. "No
other , WB is Hella stupid
and utterly ignorant on us
fans. It's pure stupidity
when we are promoting their
very own entity and it's
product , but they still
chose to sabotage and
gaslight the dedication and
immense passion . #
RestoreTheSnyderVerse https:
//t.co/CEYzb7PkSR" "
Complementarian's continue
to insist that their
position *isn't* cultural.
This of course is
demonstrably false. Also
notice the gaslighting that
compounds this fallacy, I.e.
*they are the ones standing
against culture*. It's the
same strategy used by the "
fake news" crowd. https://t.
co/1DsLThN6A1" "More media
gaslighting with their fake
news and faux outrage! Media

9. Log the key items - entities, actions, themes and claims - that emerge in your word tree, creating the screen grabs to be included into the worksheet.

10. Conduct 'zoom-ins' into anything that's confusing, unexpected or weird, by looking at occurrences in individual posts in the CSV (via using its text search function). This step may require some additional research in order to interpret results. For example, the phrase 'girlboss genocide' commonly follows the word 'gaslight', which can be interpreted with reference to individual posts and a meme. Collect the key items (see step 8). Log the key 'zoom-ins' in your worksheet. You may want to see the comments below a given tweet for which you can try googling the entire tweet in order to see if you can find your way back to it and its comment thread.

11. Preparing the text for time-series analysis. Before we can do this we need to tokenize the text, producing documents *monthly* and *stemming* and *lemmatizing* the words, and *removing stopwords*. Having done so the text is then prepared for time-series analysis



Tokenise

Tokenises post bodies, producing corpus data that may be used for further processing by e.g. NLP. The output is a serialized list of lists, each list representing either all tokens in a post or all tokens in a sentence in a post.

Produce documents per: Month

Tokeniser: nltk TweetTokenizer

Language: English

Group tokens per: Post

☑ Stem tokens (with SnowballStemmer)

☑ Lemmatise tokens (English only)

Always allow these words:

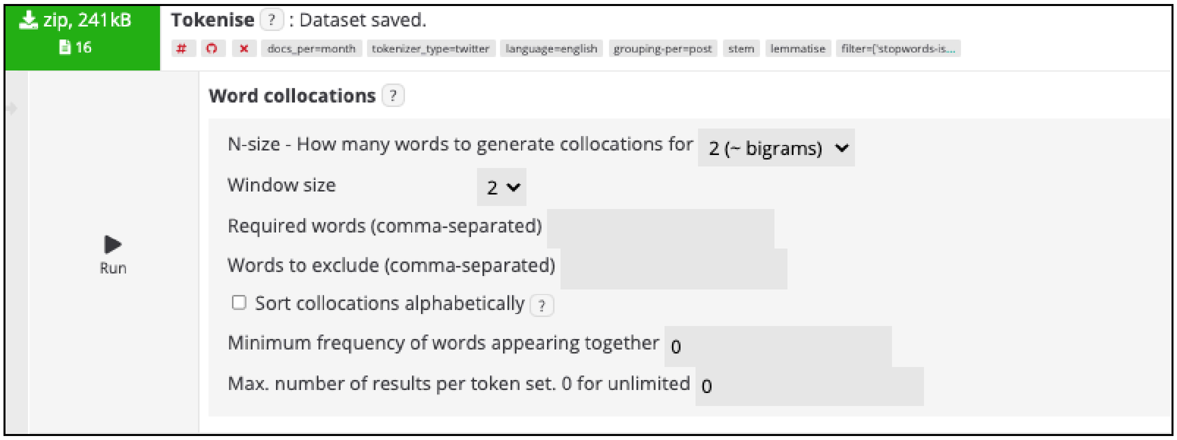
Exclude these words:

Word lists to exclude (i.e. not tokenise). It is highly recommended to exclude stop words. Note that choosing more word lists increases processing time

- ☑ English stopwords (stopwords-iso, recommended)
- □ Dutch stopwords (stopwords-iso)
- □ Stopwords for many languages (including Dutch/English, stopwords-iso)
- □ English word list (Google One Million Books pre-2008 top unigrams, recommended)
- □ English word list (cracklib, warning: computationally heavy)
- □ Dutch word list (OpenTaal)

□ Only keep unique words per post

12. Time series analysis: bi-grams over time. This step is a bit of a variation on wood trees. It lets you see how words co-occur over time, which may give an idea of some other popular concepts (e.g. “fake news”) or valences of the key term under analysis (i.e. “gaslight girlboss”)



zip, 241 kB 16

Tokenise : Dataset saved.

docs_per=month tokenizer_type=twitter language=english grouping-per=post stem lemmatise filter=['stopwords-is...

Word collocations

N-size - How many words to generate collocations for: 2 (~ bigrams)

Window size: 2

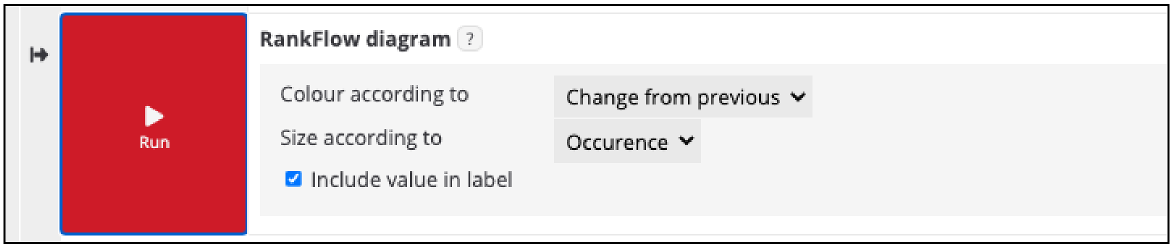
Required words (comma-separated):

Words to exclude (comma-separated):

□ Sort collocations alphabetically

Minimum frequency of words appearing together: 0

Max. number of results per token set. 0 for unlimited: 0

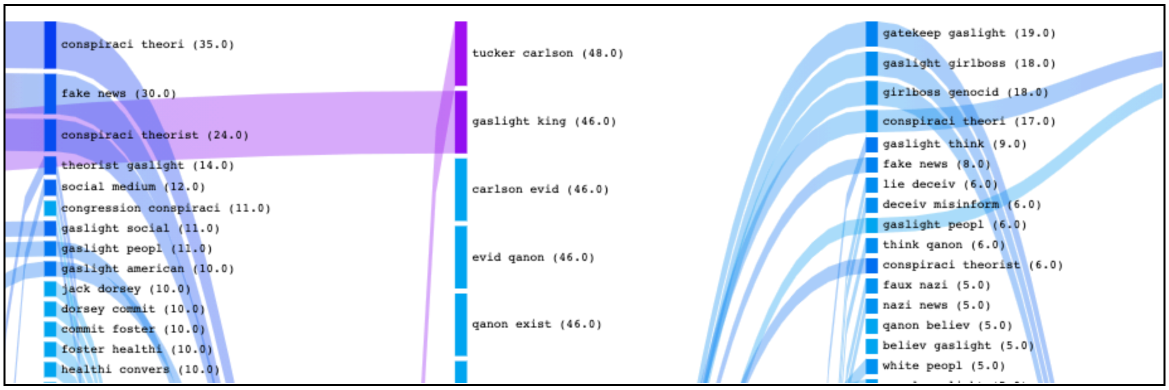


RankFlow diagram

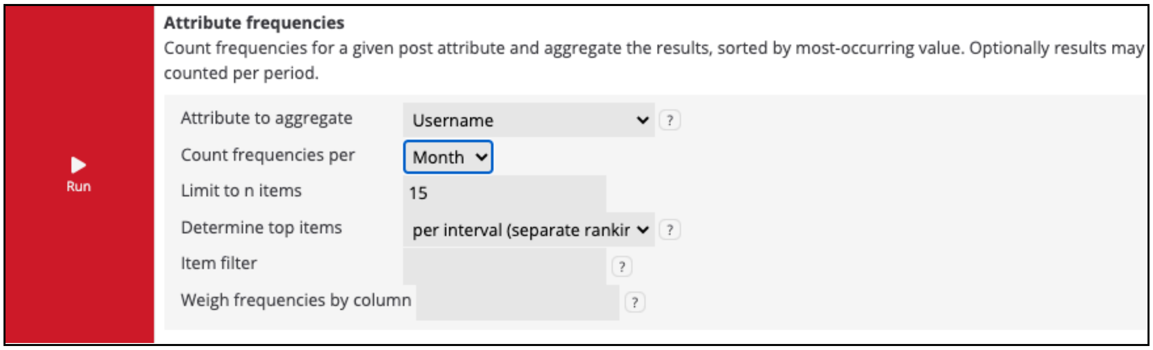
Colour according to: Change from previous

Size according to: Occurence

☑ Include value in label



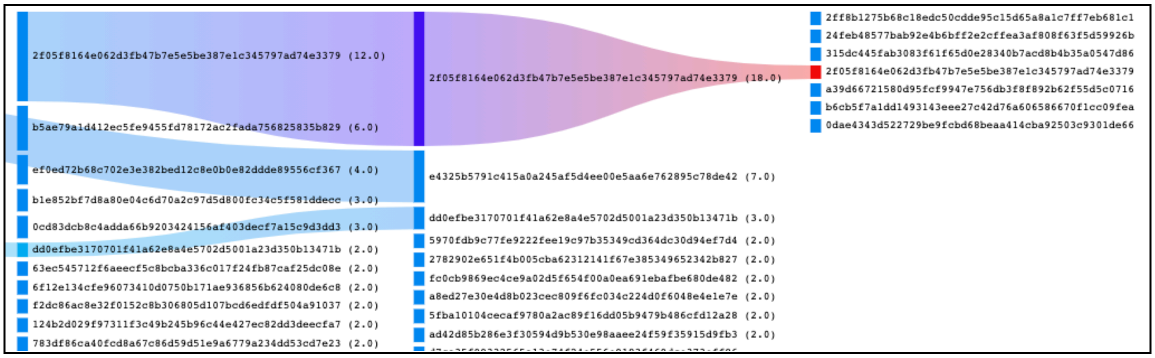
13. Time series analysis: authors over time. This gives you a sense of how many users are using this term, and how widely adopted it is or if it's just some 'rando' who is always using it.


▶ Run

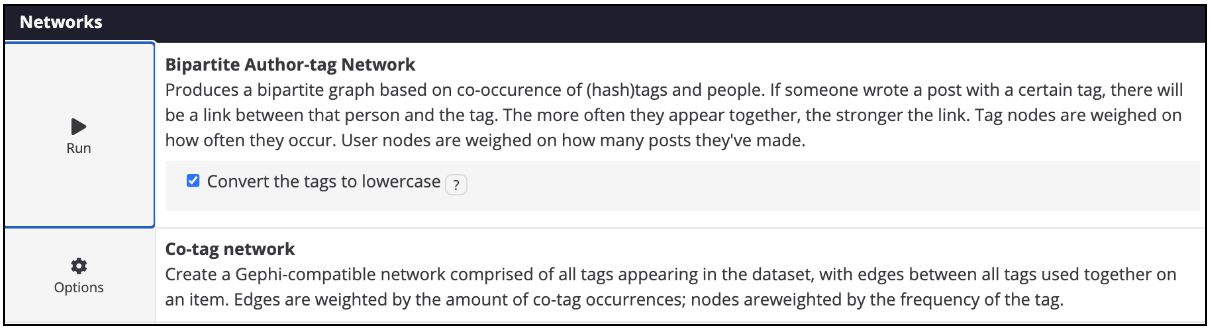
Attribute frequencies

Count frequencies for a given post attribute and aggregate the results, sorted by most-occurring value. Optionally results may counted per period.

Attribute to aggregate	Username
Count frequencies per	Month
Limit to n items	15
Determine top items	per interval (separate rankir
Item filter	
Weigh frequencies by column	



14. Network analysis: **Depending on the size of the network**, conduct an analysis of a *Co-tag network*. For this you will need to have installed Gephi and follow a basic tutorial. *Co-tag network* may show you clusters of related hashtags. This will draw edges between hashtags to give you a *panoramic view of what people think they are talking* about when they use this term.



Bipartite Author-tag Network

Produces a bipartite graph based on co-occurence of (hash)tags and people. If someone wrote a post with a certain tag, there will be a link between that person and the tag. The more often they appear together, the stronger the link. Tag nodes are weighed on how often they occur. User nodes are weighed on how many posts they've made.

☑ Convert the tags to lowercase ?

Co-tag network

Create a Gephi-compatible network comprised of all tags appearing in the dataset, with edges between all tags used together on an item. Edges are weighted by the amount of co-tag occurrences; nodes areweighted by the frequency of the tag.

15. Network analysis (idea entrepreneurs). You can also run a *Bipartite Author-tag network* which may show you communities of users using the term. If you find clusters here you can think of the *Bipartite Author-tag* as networks of "idea entrepreneurs", whose use of the term clusters into different communities. In doing so perhaps you can identify clear different interpretations of the concept. Here we would refer to the linguistic pragmatics ideas that meaning is use and uses are local. In order to characterize these uses properly you may have to move back into the CSV to see how the authors at the core of different clusters use the term differently

16. Write a 250 analysis of the keyword, considering the top posts, keyphrases and items, with screengrabs and examples where helpful.

17. Present your analyses at a given point (at the end of each day) with the aim of positioning these terms as diagonal connections across the political compass

References

Knight P (2000) Conspiracy Culture: From Kennedy to the X-Files. London: Routledge.

Slobodian, Q. & W. Callison. 2021. Coronapolitics from the Reichstag to the Capitol. Boston Review

Star, S., & Griesemer, J. (1989). "Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39". Social Studies of Science, 19 (3): 387–420. doi: 10.1177/030631289019003001