Data Collection, Preparation, and Preprocessing

Simon

May 2020

Introduction

- TBD
- define / translate / introduce German political concepts (Abgeordnete(r), Bundestag, Legislaturperiode, Wahlkreis, Ausschüsse, Parteienlandschaft)

Theoretical Framework

- Introduce topic analysis vocabulary (documents, tokens, . . .)

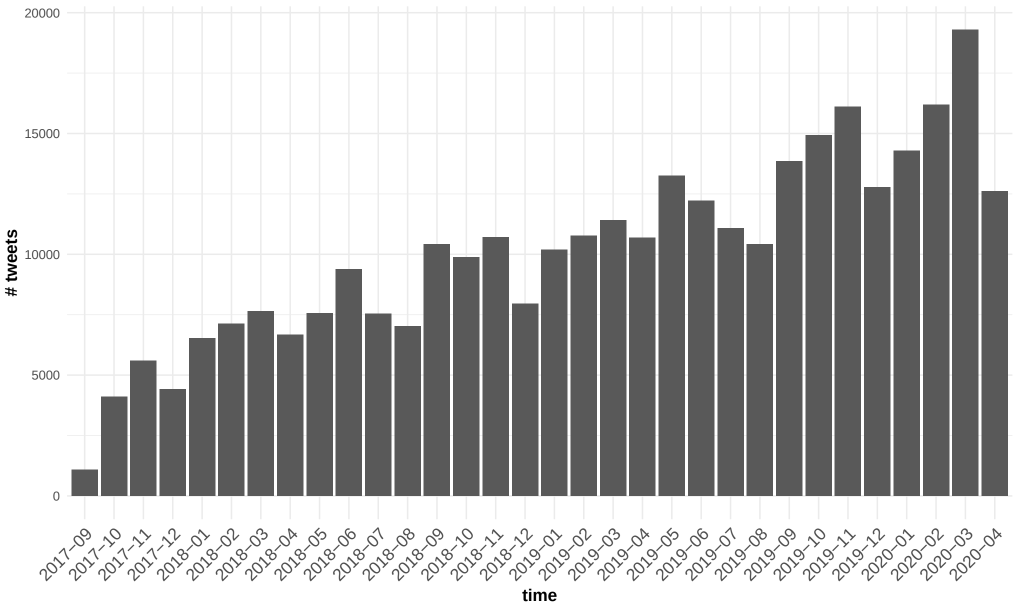
Data

Data (2-3 Seiten): * sentiment analysis: training corpus

As a first step towards applying the STM to German political entities, we constructed a database with personal information about all German MPs. Using Python's BeautifulSoup web scraping tool as well as a selenium webdriver, we gathered data such as name, party, and electoral district from the official parliament website for all of the 709 members of the German parliament during its 19th election period, elected on September 24, 2017. (Footnote: MPs who resigned or passed away since this date were also listed on the website and thus included initially; they were manually excluded from further analysis.)

Since information on social media profiles was scarce and incomplete on the official parliament website, we scraped official party homepages for each of the six political parties represented in the current parliament. MPs who did not provide a Twitter account either on the official parliament website or on their party's official homepage were excluded. Using Python's tweepy library to access the official Twitter API, we scraped all tweets by German MPs from September 24, 2017 through April 24, 2020, i.e., during a total of 31 months. (Footnote: tweepy restricts the total number of retrievable tweets to 3,200. For those MPs with a larger number of tweets, the most recent 3,200 tweets are taken into account. However, this only affects two MPs.) This initially yielded 342542 tweets from a total of 470 members of parliament.

To complement personal data, we also gathered socioeconomic data such as GDP per capita and unemployment rate as well as 2017 election results on an electoral-district level for all of the 299 electoral districts, from the official electoral website. After removing independent MPs as well as MPs without a specific electoral district assigned to them (for matchability with socioeconomic data), the final dataset counted 450 MPs. The corresponding total number of tweets amounted to 323740. The table below shows total monthly tweet frequencies for our period of analysis, September 24, 2017 through April 24, 2020. As can be seen, tweet frequencies - though fluctuating - increase over time, peaking at almost 20,000 in March 2020.



Next, data were grouped and tweets concatenated on a per-user level (thus aggregating tweets across the entire 31 months) as well as on a per-user per-month level, yielding a user-level and a (user-level) monthly dataset. This means that a document represents the concatenation of *all* of a single MP's tweets for the user-level dataset and a single MP's *monthly* tweets for the monthly dataset. This also means that MP-level metadata such as personal information and socioeconomic data (through the electoral district matching) can be used as document-level covariates. For the monthly dataset, the temporal component (year and month) constitutes an additional covariate. At this point, the data preparation was completed, thus marking the starting point of the preprocessing required for topic analysis, which is identical for both datasets.

We used the quanteda package within the R programming language for preprocessing. As a first step, we built a quanteda corpus from all documents, already transcribing German umlauts *ä/Ä*, *ö/Ö*, *ü/Ü* as well as German ligature *ß* as *ae/Ae*, *oe/Oe*, *ue/Ue*, and *ss* and removed hyphens. Next, we transformed the text data into a quanteda document-feature matrix (DFM), which essentially tokenizes texts, thereby convering all characters to lowercase. From the DFM, we removed an extensive list of German stopwords, using the stopwords-iso GitHub repository, as well as English stopwords included in the quanteda package. Moreover, hashtags, usernames, quantities and units (e.g., *10kg* or *14.15uhr*), interjections (e.g., *aaahhh* or *ufff*), terms containing non-alphanumerical characters, meaningless word stumps (e.g., *innen* from the German female plural declension or *amp*, the remainder left after removing the ampersand sign, *&*) were removed. Terms with less than four characters and terms with a term frequency (overall number of occurrences) below five or with a document frequency (number of documents containing the word) below three were excluded. Finally, we manually removed over-frequent terms that would diminish the distinguishability of topics, such as *bundestag* or *polit*.

We also performed word-stemming, which means cutting off word endings to remove discrepancies arising purely from declensions or conjugations - of particular importance for the German language. Due to the nature of the German language, the additional gains of lemmatization (which aims at identifying the base form of each word) would only be small as compared to the large increase in complexity, which is why we decided to use stemming only. Another issue when dealing with German language documents are compound words, which are sometimes hyphenated, basically leading to a distinction where semantically there is none. We addressed this issue by removing hyphens in the very beginning of the preprocessing and converting all terms to lowercase, thus "gluing together" compound words; this way, terms like *Bundesregierung* and *Bundes-Regierung* are both transformed into *bundesregierung* (and, after stemming, into *bundesregier*). Finally, automatic segmentation techniques were not necesssary for the German language (Lucas et al. (2015)).

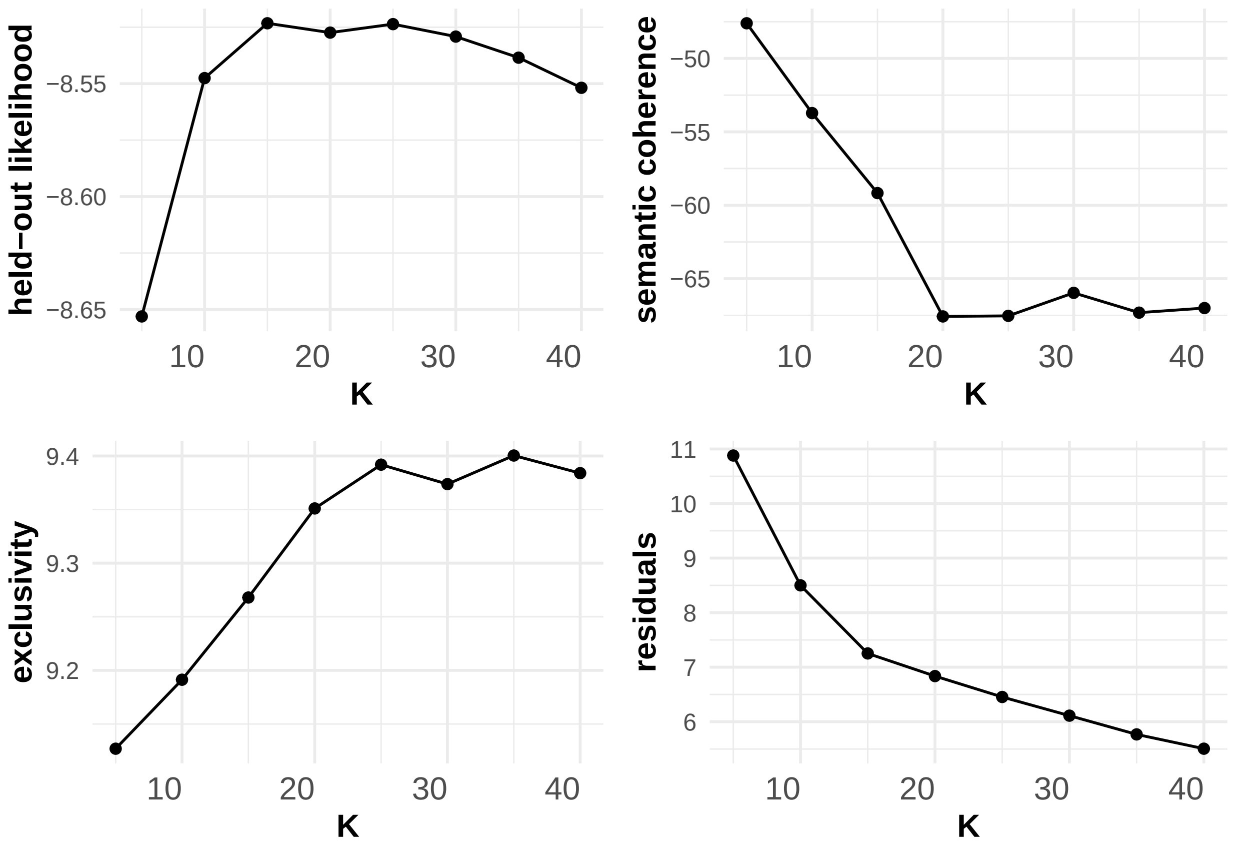
As the result of preprocessing, one empty MP-level document was dropped, so that a total of 10998 MP-level documents were eventually analyzed, each one associated with 90 covariates.

Results (15-20 Seiten insgesamt):

- topic analysis:
  - main analyses w/ graphs, starting with main dataset
  - use party-specific and time-wise dataframes punctually to investigate specific relationships further
  - use input from Thurner & Heumann
  - go into detail regarding estimateEffect, topic prevalence, and topical content estimation (partially relate back to theoretical framework)
- sentiment analysis:
  - TBD

Throughout the topic analysis, we use the stm package, which is implemented in the R programming language. The STM has a single hyperparameter K, the number of topics. While there is no *true* or *optimal* number of topics, we explore the hyperparameter space using the searchK function to get to get an understanding of the impact of K on model fit. We use four of the metrics that come with this function, *held-out likelihood*, *semantic coherence*, *exclusivity*, and *residuals*. As for the first one, the searchK function randomly holds out a proportion of some of the documents. This set of held-out words is then used to evaluate their probability given the trained model, giving rise to the *held-out likelihood.* Regarding the second metric, a model with K topics is *semantically coherent* whenever those words that characterize a specific topic (i.e., the most frequent words within a topic) also do appear in the same documents. *Exclusivity* basically tells us to which degree a topic's word *only* occur in that topic (for more detail, see the *FREX* methodology further below). Finally, *residuals* is a metric based on residual dispersion, which theoretically should be equal to one; so if the observed residuals exceed this value, the number of topics was most likely chosen *insufficiently small.*

Another aspect to be taken into account when choosing K (or, to be precise, when choosing a search grid for searchK) is interpretability. While a large K certainly allows for a more fine-grained determination of topics, the resulting topics might be rather hard to label. Furthermore, for large K we might get many topics which themselves could easily be considered sub-topics of those topics that we would get when using a smaller value for K. The graphs below shows the four metrics, as introduced above, for values of K between 5 and 40 (in steps of 5).



Both 15 and 20 topics seem to be good trade-offs between the metrics used. Taking into account the interpretability aspect, we opt for K = 15.

Before fitting the model, we need to choose the document-level covariates we want to include. Since a topic model is explorative by definition, we simply include those covariates that seem to be most influential *a priori*: party and state (both categorical), date (as smooth effect), as well as percentage of immigrants, GDP per capita, unemployment rate, and the 2017 election results of the MP's respective party (the last four as smooth effects and on an electoral-district level).

The first thing to do after fitting the model is to (visually) inspect the topics, in particular their best words, as demonstrated in the output below. There are different metrics for evaluating which words are most representative of a topic. The STM comes with four such metrics: *highest probability*, *FREX*, *Lift*, and *Score*. The first one, *highest probability*, simply outputs for each topic those words in the topic-specific word vector, $\beta_{d,n}$, with the highest corpus frequency, i.e, those with the highest absolute frequency across all documents. *FREX* takes into account not only how frequent but also how exclusive words are: for a given topic, it is calculated by taking the harmonic mean of i) the word's rank by probability within the topic (frequency) and ii) the topic's rank by word frequency across all topics (exclusivity). Further information on the estimation of *FREX*, *Lift*, and *Score* can be found in Bischof and Airoldi (2012), in the lda package by Jonathan Chang (2015), and in Taddy (2012), respectively.

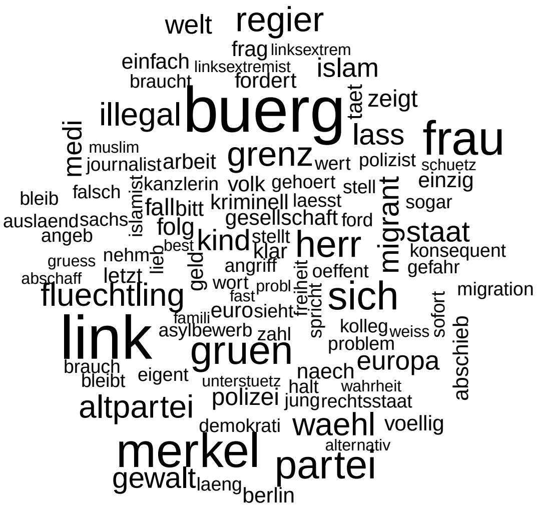
The output below shows, for each topic, the 5 top words for each of the four topic-word evaluation metrics.

```
## Topic 1 Top Words:
##       Highest Prob: buerg, link, frau, merkel, gruen
##       FREX: altpartei, islam, linksextremist, asylbewerb, linksgru
##       Lift: eitan, 22jaehrig, abdelsamad, abgehalftert, afd'l
##       Score: altpartei, islam, linksextremist, frauenkongress, boehring
## Topic 2 Top Words:
```

```
##       Highest Prob: frag, genau, einfach, find, gern
##       FREX: geles, quatsch, sorry, versteh, satz
##       Lift: duitsland, freiraumkonto, garn, kombo, lieblingss
##       Score: tweet, fuerstenberg, sorry, haushaelt, geles
## Topic 3 Top Words:
##       Highest Prob: brauch, gruen, klimaschutz, wichtig, leid
##       FREX: emissionshandel, erneuerbar, klimaziel, fossil, emission
##       Lift: bahnunternehm, betriebskonzept, bewaesser, biogas, biokraftstoff
##       Score: emissionshandel, co2limit, emission, kraftstoff, erneuerbar
## Topic 4 Top Words:
##       Highest Prob: sozial, miet, berlin, brauch, arbeit
##       FREX: miet, mieterinn, wohnung, wohnungsbau, vermiet
##       Lift: baugrundstueck, baustaatssekreta, behrendt, billigflieg, bodenwertzuwachssteu
##       Score: miet, mietendeckel, mieterinn, wohnung, bezahlbar
## Topic 5 Top Words:
##       Highest Prob: europaeisch, thank, good, great, wichtig
##       FREX: important, foreign, policy, discussion, clos
##       Lift: abroad, acknowledg, across, activity, addressed
##       Score: important, need, great, thank, right
## Topic 6 Top Words:
##       Highest Prob: gruen, frag, euro, geld, minist
##       FREX: scheu, verkehrsminist, autoindustri, nachruest, verkehrsministerium
##       Lift: agrarministerin, angstunternehm, aufklaerungsinteress, autoboss, baulueckenkatast
##       Score: schmunzel, scheu, verkehrsminist, perli, pkwmaut
## Topic 7 Top Words:
##       Highest Prob: wichtig, europa, gemeinsam, brauch, europaeisch
##       FREX: integration, partnerschaft, fried, partn, karamba
##       Lift: bahrenfeld, bamako, entrepreneur, erbfeind, friedensmacht
##       Score: integrationsbeauftragt, europa, antisemitismus, transatlant, conduct
## Topic 8 Top Words:
##       Highest Prob: kris, wichtig, brauch, unternehm, massnahm
##       FREX: coronakris, corona, virus, pandemi, coronavirus
##       Lift: 600milliardenfond, abstandhalt, alltagsmask, antikoerp, antikoerpert
##       Score: corona, coronakris, pandemi, coronavirus, virus
## Topic 9 Top Words:
##       Highest Prob: krieg, link, frag, regier, europaeisch
##       FREX: milita, voelkerrechtswidr, aufruest, geheimdien, libysch
##       Lift: abho, airbas, antimilitarist, aufklaerungsdat, aufruestet
##       Score: voelkerrechtswidr, libysch, milita, voelkerrecht, zdebel
## Topic 10 Top Words:
##       Highest Prob: herzlich, glueckwunsch, dank, freu, stark
##       FREX: achim, parteitag, delegiert, gmuend, glueckwunsch
##       Lift: abschlussfoto, borby, dt.israel, ernstwilhelm, hessennord
##       Score: backnang, gmuend, herzlich, glueckwunsch, achim
## Topic 11 Top Words:
##       Highest Prob: berlin, besuch, gespraech, jung, thema
##       FREX: buongiorno, fdpbundestagsabgeordnet, duesseldorf, weiterles, freihold
##       Lift: aero, aign, alois.karl, andreas.scheu, andreas_mattfeldt
##       Score: buongiorno, fdpbundestagsabgeordnet, storjohann, rimkus, freihold
## Topic 12 Top Words:
##       Highest Prob: frau, gruen, sozial, kind, dank
##       FREX: mention, reach, bielefeld, automatically, retweet
##       Lift: barrientos, trainingsplaetz, automatically, unfollowed, aktivenkonferenz
##       Score: mention, unfollowed, automatically, reach, checked
## Topic 13 Top Words:
##       Highest Prob: dank, schoen, freu, berlin, abend
##       FREX: leipzig, nachh, heut, hall, wunderscho
##       Lift: bergenenkheim, mainzbing, sommergrill, altlandsberg, anwohnerinn
##       Score: dank, magdeburg, schoen, freu, abend
## Topic 14 Top Words:
##       Highest Prob: partei, demokrat, klar, link, dank
##       FREX: thuering, hoeck, faschist, kemmerich, ramelow
##       Lift: atrium, epost, kernbereich, kommissionschef, maduroregim
##       Score: kemmerich, faschist, hoeck, ramelow, thuering
## Topic 15 Top Words:
##       Highest Prob: kind, pfleg, wichtig, brauch, versorg
##       FREX: neuwied, organsp, pflegebeduerft, patient, widerspruchsloes
##       Lift: altenkirch, gesundheitsberuf, ahrweil, alltagsheldinn, anglizism
##       Score: neuwied, windhag, patient, altenkirch, nnen
```

A key task of topic analysis is to actually ascribe a meaning to the topics identified, i.e., labelling them. While this is clearly where human judgment should and does come into play, we attempt to conduct the labelling in a more stratetic (and thus less subjective) manner, following a 3-step procedure. This procedure is exemplified using topic 1.

First, we consider the *words* contained in the topic, for instance by simply inspecting the top words (see output above). For a better visualization, we use a word cloud. As shown below, for a given topic (i.e., conditional upon a specific topic being chosen), it shows words weighted by their frequency. For instance, by judging at first sight topic 1 appears to be about right-wing nationalist issues, particularly immigration.



Second, to get a more thorough insight into the topic, we take a look into actual *documents*, specifically into those showing the highest proportion for topic 1.

x	x	x	x	x	x
right/nationalist	miscellaneous_1	green/climate	social/housing	Europe_english	mobility

x	x	x	x	x	x
Europe	Corona	left/anti-war	Twitter/politics_1	Twitter/politics_2	miscellaneous_2

x	x	x
Twitter/politics_3	right-wing extremism	social/health

For instance, the most representative document for topic 1, with a proportion of 99.02% is the one by MP Hess, Martin, a member of the AfD party from Baden-Württemberg, from 2018-06 which starts with:

```
## [1] "Ehem. Verfassungsrichter bestätigt AfD-Forderung: Zurückweisung illegaler Migranten dringend geb
```

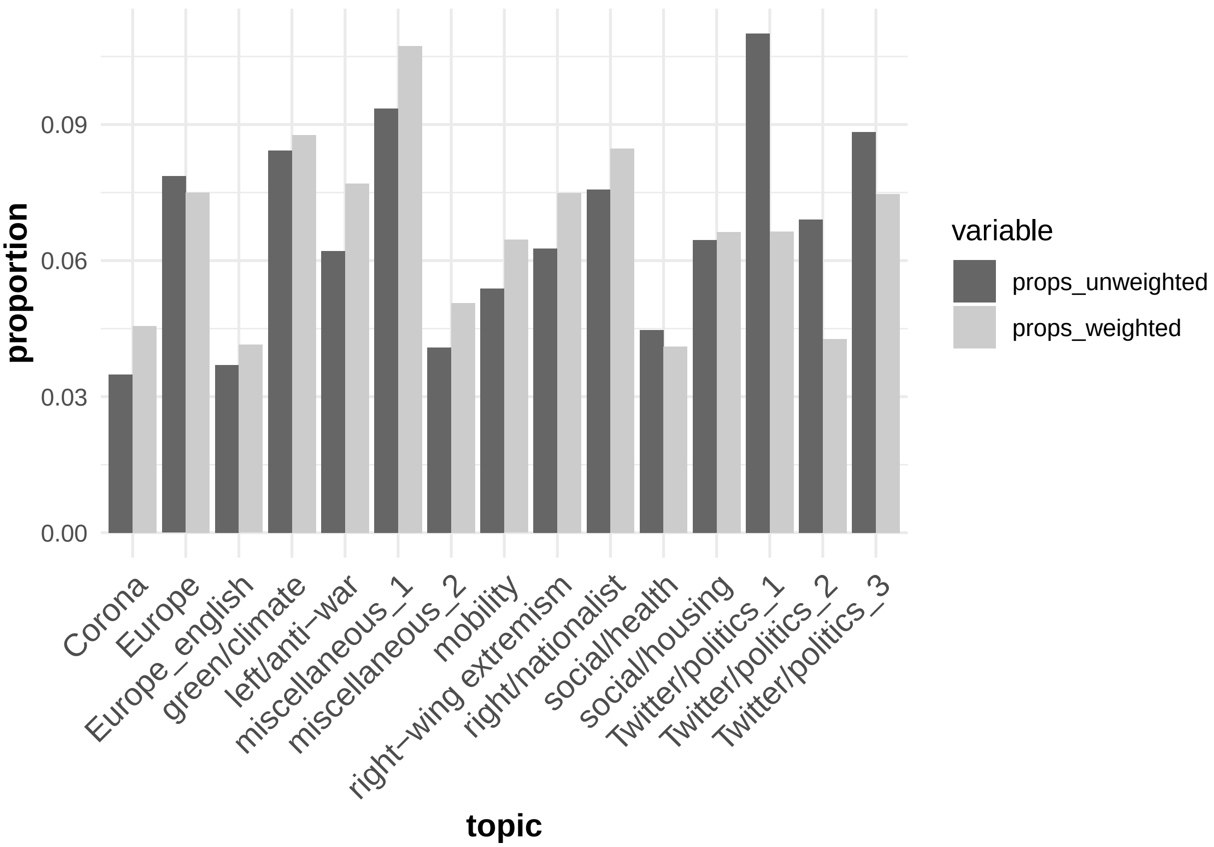
The second most representative document, still for topic 1, with a proportion of 98.71% is the one by MP Hess, Martin, a member of the AfD party from Baden-Württemberg, from 2018-03 which starts with:

[1] "Offenbar handelt das #BAMF nicht im Interesse der Inneren Sicherheit. Die skandalöse Vorgehensweise dieser Behörde muss lückenlos aufgearbeitet werden. Es darf nicht sein, dass die Asyllobby über Unterbehörden Einfluss auf staatliche Entscheidungen nimmt!"

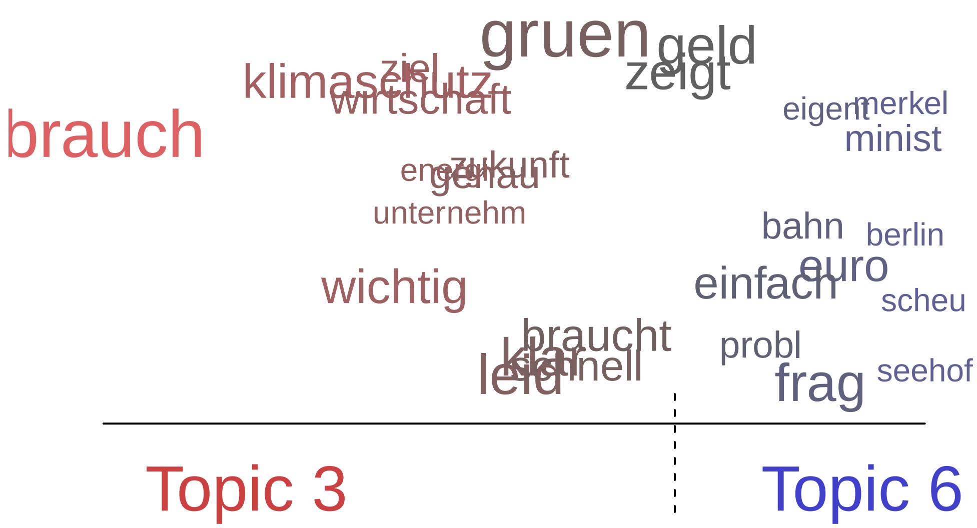
The documents confirm the first impression gained through top words and the word cloud: 1 concerns right-wing nationalist issues, in particular immigration. Thus, as a third step, we finally label the topic: in this case, as right/nationalist.

We repeat this 3-step procedure (inspecting top words and word cloud, reading through top documents, assigning a 1- or 2-word label) for all remaining topics, arriving at the following manual labels:.

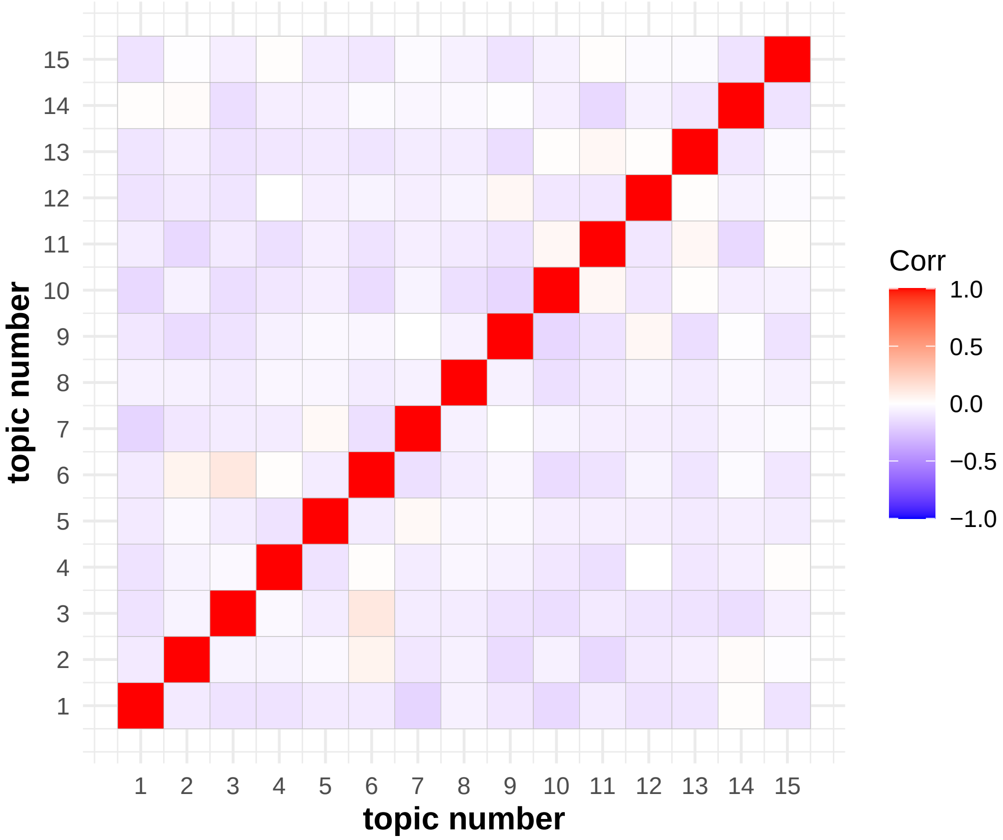
Next, we identify two ways to calculate global topic proportions: either as the simple (unweighted) average of θ_d across all documents (i.e., as the average of MP-level proportions across all MPs); or by weighting each θ_d by the number of words in the respective documents, N_d. The table below shows all topics with their respective global proportions, for both weighting methodologies. We observe that for most topics, weighted and unweighted proportions are rather similar, but there are exceptions. In particular, the topics concerned with everyday political tweets have much higher unweighted than weighted frequencies; this makes sense, however, since such "diplomatic" tweets tend to be shorter than those which actually discuss a specific content.



While labelling tells us which words best represent each topic - and thus, what each topic truly represents - it does not yet tell us to which extent individual topics are related to each other. In the graph below, we visualize the similarity of two topics, Topic 3 (green/climate) and Topic 6 (mobility), in terms of their vocabulary usage. As suggested by the topic labels already, there is a significant overlap in vocabulary usage.

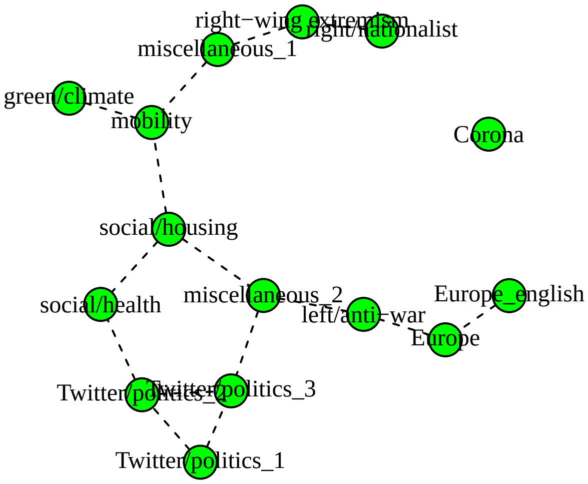


More generally, we can evaluate the connectedness between different topics through the topic correlation matrix, which is simply based on the correlations between document-level topic proportions θ_d. This is visualized in the graph below.



Most topics are negatively correlated with each other, which does not come as a surprise, given the relatively low total number of topics, 15, and that topic proportions are "supplements": the higher one topic proportion, the lower the total of the others. Moreover, most topic correlations are rather weak in absolute size: the strongest negative correlation (-17.57%)is the one between topic 1 (right/nationalist) and topic 7 (right/nationalist), while the strongest positive correlation (11.53%) is the one shown before, between (green/climate) and (mobility).

We can also visualize these correlations using a network graph, where topics are connected whenever they are positively correlated. Most topics are only related to two other topics, while none are related to more than three. The only "isolated" topic is topic 8, Corona, which makes sense since it only entered the public sphere in early 2020, i.e., during the last months of our data collection period. In general, the relationships between the topics, as depicted below, are very much in line with their labelling.



Literature

Bischof, Jonathan, and Edoardo M Airoldi. 2012. “Summarizing Topical Content with Word Frequency and Exclusivity.” In *Proceedings of the 29th International Conference on Machine Learning (Icml-12)*, 201–8.

Lucas, Christopher, Richard A Nielsen, Margaret E Roberts, Brandon M Stewart, Alex Storer, and Dustin Tingley. 2015. “Computer-Assisted Text Analysis for Comparative Politics.” *Political Analysis* 23 (2): 254–77.

Taddy, Matt. 2012. “On Estimation and Selection for Topic Models.” In *Artificial Intelligence and Statistics*, 1184–93.