4_1to3

Patrick Schulze, Simon Wiegrebe

June 2020

Contents

1 Results

1.1 Hyperparameter Search and Model Fitting

Throughout this topic analysis we use the *stm* package, which is implemented in the R programming language (Roberts, Stewart, and Tingley (2019)). The most important hyperparameter choice when fitting an STM is the number of topics, K. While there is no *true* or *optimal* number of topics, we explore the hyperparameter space using the *searchK* function 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*.

The *held-out likelihood* approach is based on document completion. The *searchK* function randomly holds out a proportion of some of the documents; both the number of documents from which a portion is held out and the respective held-out proportions can be specified by the user. This gives rise to a set of held-out words, for which the likelihood is calculated, given the trained model. Thus, the higher this held-out likelihood, the more predictive power the model has on average. For more detailed information on held-out likelihood based on document completion and other types of held-out likelihoods, see Wallach et al. (2009).

Regarding the second metric, first introduced by Mimno et al. (2011), a model with K topics is *semantically coherent* whenever those words that characterize a specific topic k (i.e., the most frequent words within topic k) also do appear in the same documents. In order to formally define semantic coherence, let first $D(v)$ be the *document frequency* of word v (that is, the number of documents where v occurs at least once) and let $D(v, v')$ be the *co-document frequency* of words v and v' (that is, the number of documents where both v *and* v' occur at least once). Furthermore, consider the M most probable words in a given topic k. Then, semantic coherence for topic k, C_k, is defined as follows:

$$C_k = \sum_{i=2}^{M} \sum_{j=2}^{i-1} \log(\frac{D(v_i, v_j) + 1}{D(v_j)}).$$

That is, semantic coherence is the sum of (logarithmized) proportions of word co-occurrences to total word occurrences, the additive factor 1 in the numerator just being a smoothness adjustment. It becomes apparent that by having some words that are very frequent across a couple of documents, we could achieve high semantic coherence without our topics being semantically coherent at all once we look beyond those common words (Roberts, Stewart, and Tingley (2019), Mimno et al. (2011)). As a partial remedy, we previously excluded some of these overly frequent words (see section 3.2).

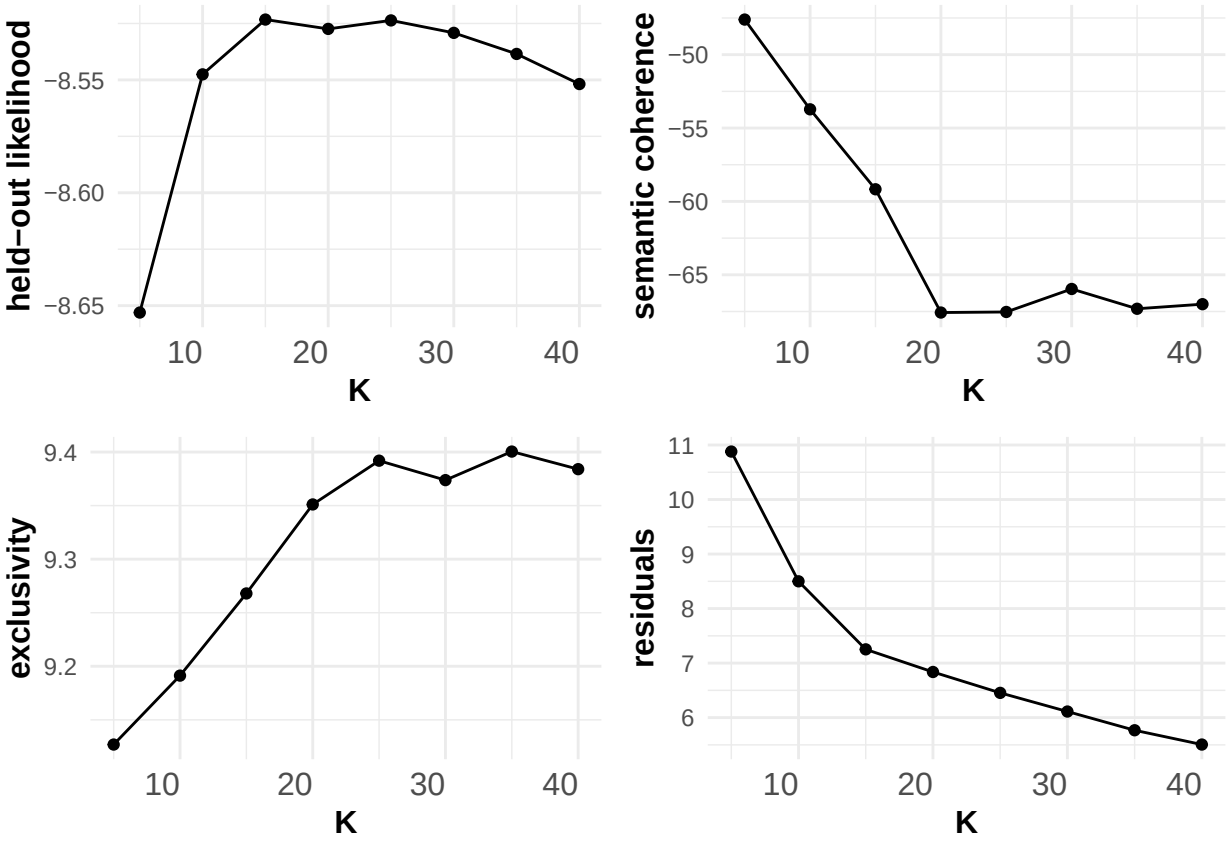
A natural "counter-metric" of semantic coherence is *exclusivity*, which basically tells us to which degree a topic's words *only* occur in that topic. To be specific, for a given word $v \in V$, the empirical frequencies of v in topic k, $\beta_{k,v}$, are normalized across all topics $k \in \{1, ..., K\}$. This way, these normalized frequencies now represent the probability of observing topic k, conditional upon the word being v - that is, the exclusivity of word v regarding topic k. Formally, exclusivity of word v to topic k, $E_{k,v}$, is thus defined as:

$$E_{k,v} = \beta_{k,v} / \sum_{j=1}^{K} \beta_{j,v}.$$

Combining a word's frequency and exclusivity finally yields its Frequency-Exclusivity (*FREX*) score, explained in more detail in section 4.2 below (Bischof and Airoldi (2012)).

Finally, *residuals* is a metric based on residual dispersion. Recall that $z_{d,n}$ is drawn from a K-category multinomial distribution, which is a member of the exponential family. Therefore, its dispersion parameter is equal to one, according to theory. This way, an observed residual dispersion larger than one roughly indicates that the number of topics K was most likely chosen insufficiently small. See Taddy (2012) for a detailed derivation.

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 difficult to label. Furthermore, for large K we would obtain many topics which could be considered sub-topics of the topics we would obtain when using a smaller value for K. The graph 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. As mentioned above, no true or optimal K exists. Taking into account the interpretability aspect, we opt for $K = 15$. For comparison, we

also conducted the subsequent analysis for a small (and easily interpretable) number of topics, $K = 6$, as well as for $K = 20$. In general, topics generated are similar, but for $K = 6$ only around three are clear-cut, while for $K = 20$ some topics could easily be grouped together. This further corroborates our choice that $K = 15$ indeed seems to be a good trade-off.

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).

1.2 Labelling

As a first step after fitting the model, we would like to visually inspect the resulting topics, in particular their most representative words. However, representativeness of words for a given topic depends on the weighting metric used. The STM comes with four topic-word metrics - *highest probability*, *FREX*, *Lift*, and *Score* - which are discussed in the following.

Given a topic k, *highest probability* simply outputs those words in the topic-specific word vector β_k with the highest corpus frequency, i.e, those with the highest absolute frequency across all documents. Using the same notation as in section 4.1 above, let $\beta_{k,v}$ again be the (empirical) frequency of word v within topic k. Then the highest probability word within topic k is simply $\underset{v \in V}{\text{argmax}} \beta_{k,v}$. This relatively simple measure only takes into account how often words occur in absolute terms, but not how specific those words are to the given topic. This is why we observe words like *wichtig*, *berlin*, or *frag* within the highest probability words for several topics. And since such words are very common, unspecific words, they are not particularly useful for distinguishing or labelling topics.

To also account for the degree to which a word *exclusively* belongs to a certain topic, we also consider the top words according to the *FREX* metric. It takes into account not only how frequent but also how exclusive words are. Formally, the FREX score of word v with respect to topic k is calculated as follows:

$$FREX_{k,v} = \left(\frac{\omega}{ECDF(\beta_{k,v} / \sum_{j=1}^{K} \beta_{j,v})} + \frac{1-\omega}{ECDF(\beta_{k,v})}\right)^{-1} = \left(\frac{\omega}{ECDF(E_{k,v})} + \frac{1-\omega}{ECDF(\beta_{k,v})}\right)^{-1},$$

where ω is the weight assigned to exclusivity (set to 0.7 by default in the STM), $E_{k,v}$ is the word's exclusivity as defined in section 4.1, and *ECDF* is the empirical CDF. Thus, for a given topic, $FREX_{k,v}$ is simply the harmonic mean of i) the rank of word v by probability within topic k (frequency rank) and ii) the rank of topic k by the frequency of word v, across all topics $j \in \{1, ..., K\}$ (exclusivity rank). Further information on the estimation of *FREX* can be found in Roberts, Stewart, and Tingley (2019) and in Bischof and Airoldi (2012).

Lift is another topic-word metric, where the frequency of word v within topic k, $\beta_{k,v}$, is weighted by the inverse of v's relative frequency across the entire corpus, i.e., v's empirical corpus probability. Formally:

$$Lift_{k,v} = \beta_{k,v} / (\omega_v / \sum_v \omega_v).$$

where ω_v denotes the word count of word v in the entire corpus. This way, Lift gives higher weight to those words that rarely appear in other topics. Further information on Lift can be found in Taddy (2012).

Finally, the *Score* metric for word v and topic k is formally defined as:

$$Score_{k,v} = \beta_{k,v}(\log \beta_{k,v} - 1/K \sum_{j}^{K} \log \beta_{j,v})$$

Thus, Score weights word v's frequency within topic k, $\beta_{k,v}$, by the difference between v's log frequency within topic k and the average of v's log frequencies across all K topics. This can roughly (but not exactly) be seen as: $\beta_{k,v}$ is weighted by the proportion of v's log frequency within topic k to v's average logarithmic frequency across all topics. For further information on the Score metric, see the R package *lda* (Chang and Chang (2010)).

To get a broad overview of which words characterize each one of the topics, the output below shows, for each topic k, the 5 top words according to 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.

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 geboten. Gegenwärtige Politik widerspricht dem Verstand und auch der Verfassung. Wir müssen zurück zu Recht & Ordnung, wie die #AfD seit fast 3 Jahren fordert!"

The second most representative document, still for topic 1, has a proportion of 98.71%. Its author is the same as for the first document, Hess, Martin, but the date now is 2018-03. The document 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.

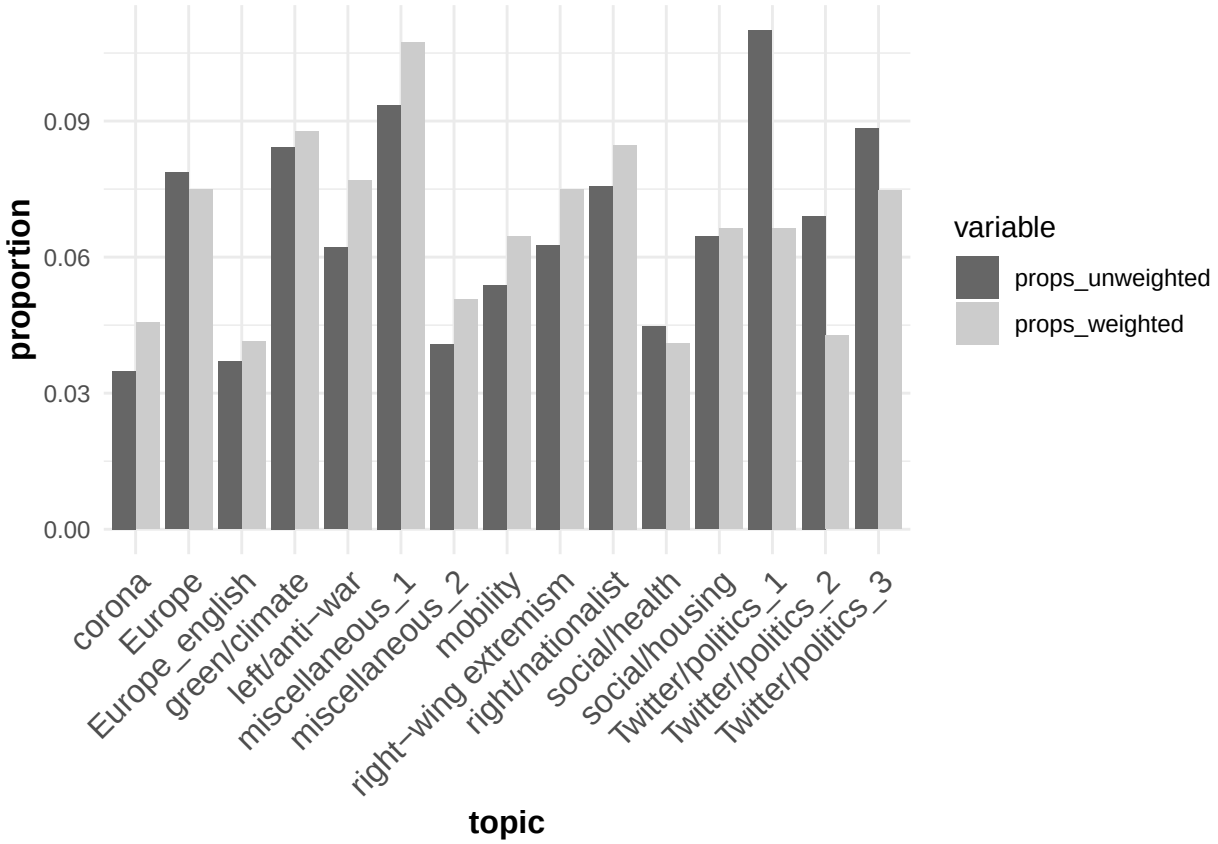
We repeat this 3-step procedure (inspecting top words and word cloud, reading through top documents, assigning a 1- or 2-word label) for all remaining topics, arriving at the following manual labels:.

Topic1	right/nationalist
Topic2	miscellaneous_1
Topic3	green/climate
Topic4	social/housing
Topic5	Europe_english

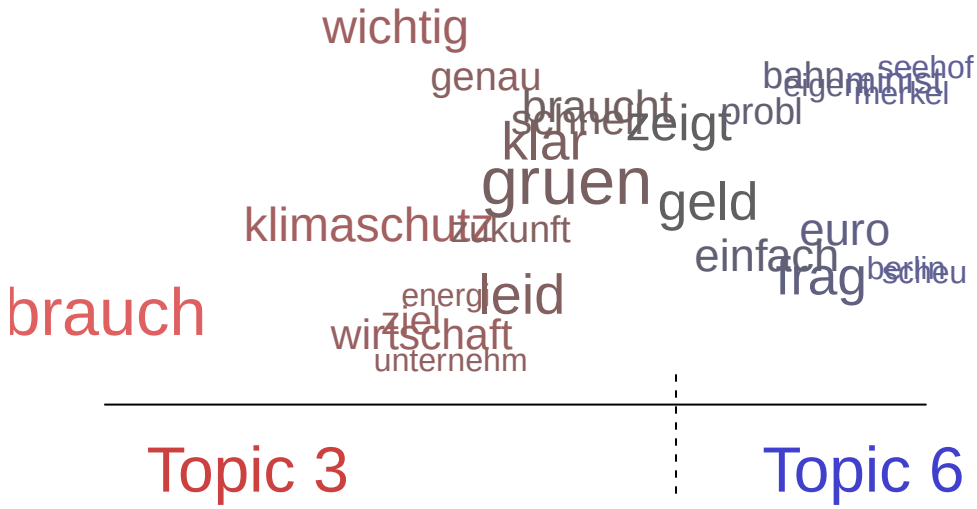
Topic6	mobility
Topic7	Europe
Topic8	corona
Topic9	left/anti-war
Topic10	Twitter/politics_1
Topic11	Twitter/politics_2
Topic12	miscellaneous_2
Topic13	Twitter/politics_3
Topic14	right-wing extremism
Topic15	social/health

1.3 Global-level Topic Analysis

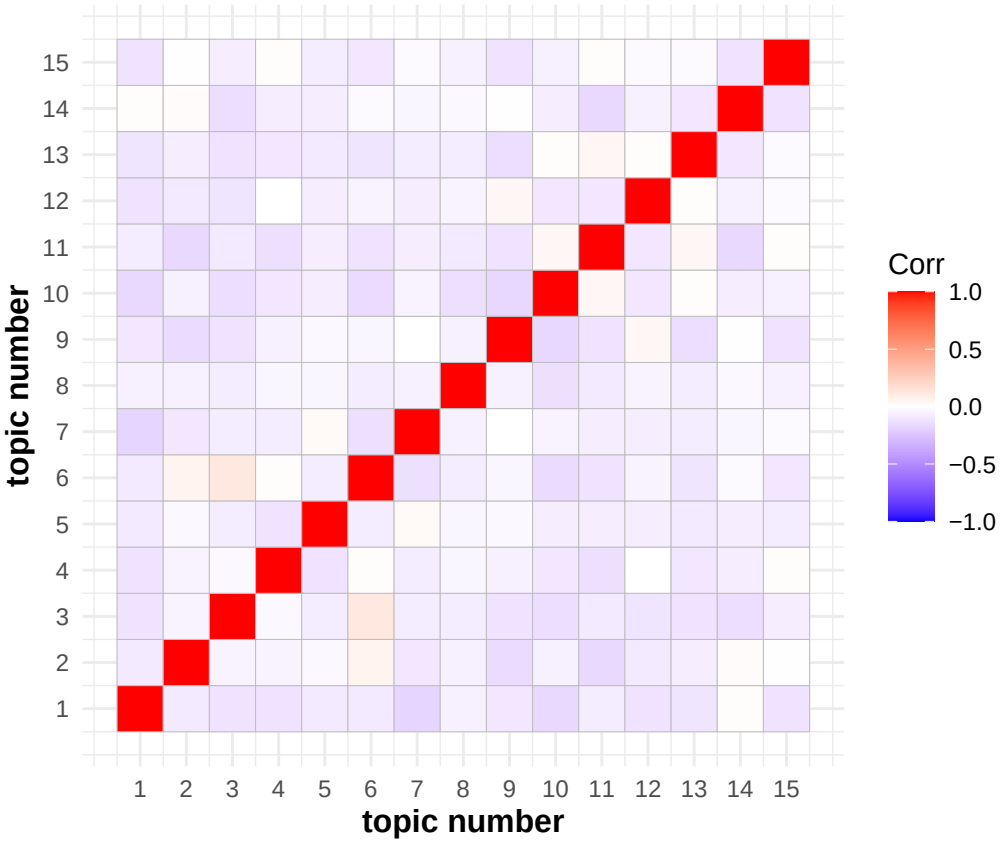
Next, we identify two ways to calculate global topic proportions: either as the simple (unweighted) average of θ_d across all documents (i.e., as the average of MP-level proportions across all MPs); or by weighting each θ_d by the number of words in the respective documents, N_d. The table below shows 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.

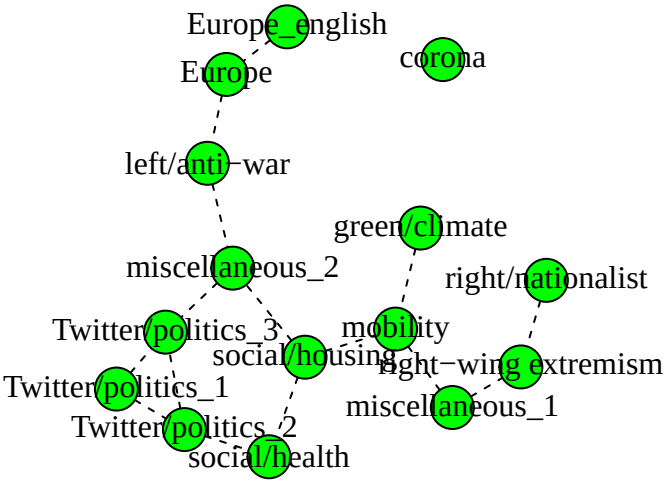


More generally, we can evaluate the connectedness between different topics with the topic correlation matrix of the correlations between document-level topic proportions θ_d. This is visualized in the graph below.



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.



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.

Chang, Jonathan, and Maintainer Jonathan Chang. 2010. "Package 'Lda'." Citeseer.

Mimno, David, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. "Optimizing Semantic Coherence in Topic Models." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 262–72. Association for Computational Linguistics.

Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. 2019. "stm: An R Package for Structural Topic Models." *Journal of Statistical Software* 91 (2): 1–40. https://doi.org/10.18637/jss.v091.i02.

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

Wallach, Hanna M, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. "Evaluation Methods for Topic Models." In *Proceedings of the 26th Annual International Conference on Machine Learning*, 1105–12.