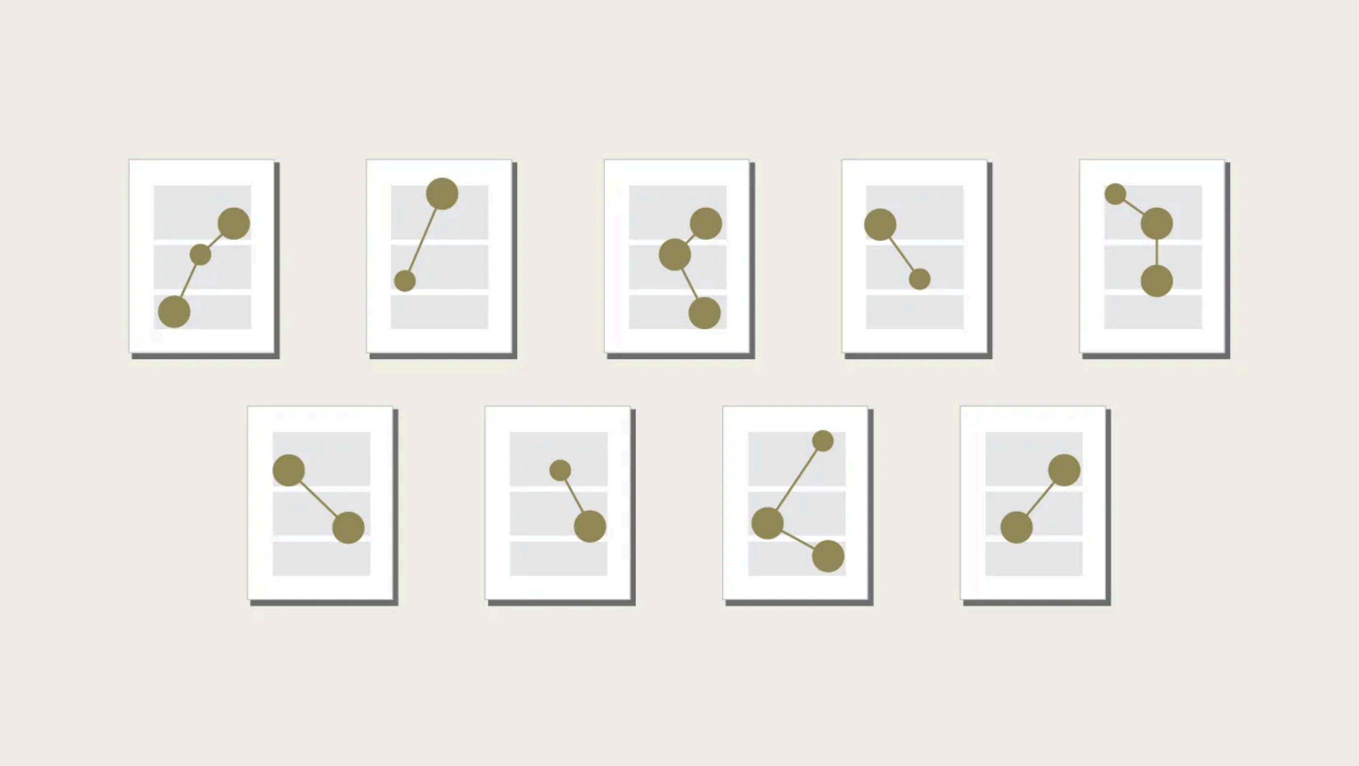
Making sense of topic models

Patrick van Kessel · Follow

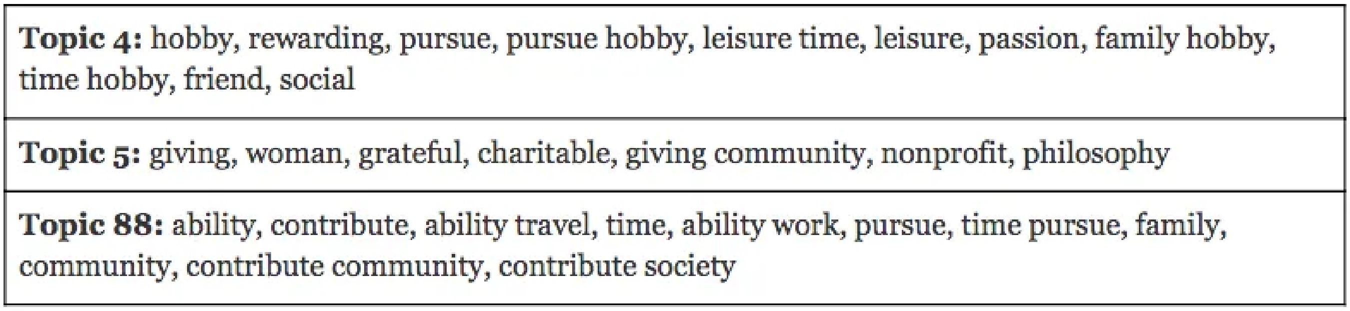
Published in Pew Research Center: Decoded · 7 min read · Aug 13, 2018

253 3



(Related posts: An intro to topic models for text analysis, Overcoming the limitations of topic models with a semi-supervised approach, Interpreting and validating topic models, How keyword oversampling can help with text analysis and Are topic models reliable or useful?)

In my first post about topic models, I discussed what topic models are, how they work and what their output looks like. The example I used trained a topic model on open-ended responses to a survey question about what makes life feel fulfilling and examined three topics in particular:



Topic 4: hobby, rewarding, pursue, pursue hobby, leisure time, leisure, passion, family hobby, time hobby, friend, social
Topic 5: giving, woman, grateful, charitable, giving community, nonprofit, philosophy
Topic 88: ability, contribute, ability travel, time, ability work, pursue, time pursue, family, community, contribute community, contribute society

By looking at the top words for each topic, we can clearly see that Topic 4 is related in some way to hobbies; Topic 5 seems to pertain to charity and giving back to society; and Topic 88 has something to do with spending time pursuing and contributing to more serious things, like work. However, we also notice some quirks, which raise an important question: How do we figure out what these topics *are*, exactly? As it turns out, that can be a very difficult question to answer.

Translating topics to concepts

Let’s say we find Topic 5 present in enough documents that it’s clearly capturing something unique about a subset of our corpus — something we’d like to analyze. Again, the most common words in Topic 5 are “giving,” “woman,” “grateful,” “charitable,” “giving community,” “nonprofit,” and “philosophy.” What exactly could such an analysis tell us? What *is* Topic 5? If we’re interpreting this topic as being about charity and giving, does the word “woman” really belong in it? What about “philosophy”? When working with topic models, it’s important to remember that these algorithms cannot guarantee that the words in each topic will be related to one another *conceptually* — only that they frequently occur together in your data *for some reason.*

Text data naturally involve the variety, complexity, and context of language. So, if you think about it, it's completely reasonable that topic modeling algorithms may pick up on noise or quirky, idiosyncratic correlations of the words in our documents. However, if we wish to use our topic model to make classifications and measure different concepts in our documents, then it's also reasonable to be rather concerned about false positives arising from words in our topics that *don't align with how we want to interpret them*. Since they don't align with the general concept that we think the model is picking up — the concept that we want to use the model to measure — we might consider words like these to be *conceptually spurious*. For one reason or another, they are associated in our data with the other words in the topic, but these associations are driven by something other than semantics. That can be a result of where the data came from, perhaps, or how the data were collected.

When analyzing survey responses, for example, conceptually spurious words can appear not because they're related to other words in a topic semantically, but because they're demographically correlated. That is, certain types of respondents may mention some themes more than others, and these themes can sometimes coincide together frequently in their responses even though they're not conceptually related.

Another topic in our model highlights this problem particularly well, consisting of the following words: "grandchild", "grand", "child grandchild", "grand child", "child", **"florida"**, "child grand", "grandchild great", and "great grandchild". Clearly, one of these words is not like the others. Having trained our topic model on open-ended responses from a nationally representative survey, in this analysis we're interested in using our topics to measure *what* respondents talk about so that we can then characterize the population in terms of *who* talks about what. Of course, that means that we need to make sure that our measures don't conflate the *who* and the *what.* So while it's insightful to find that certain respondents commonly mention both grandchildren and Florida, we probably don't want both of these concepts present together in a single topic.

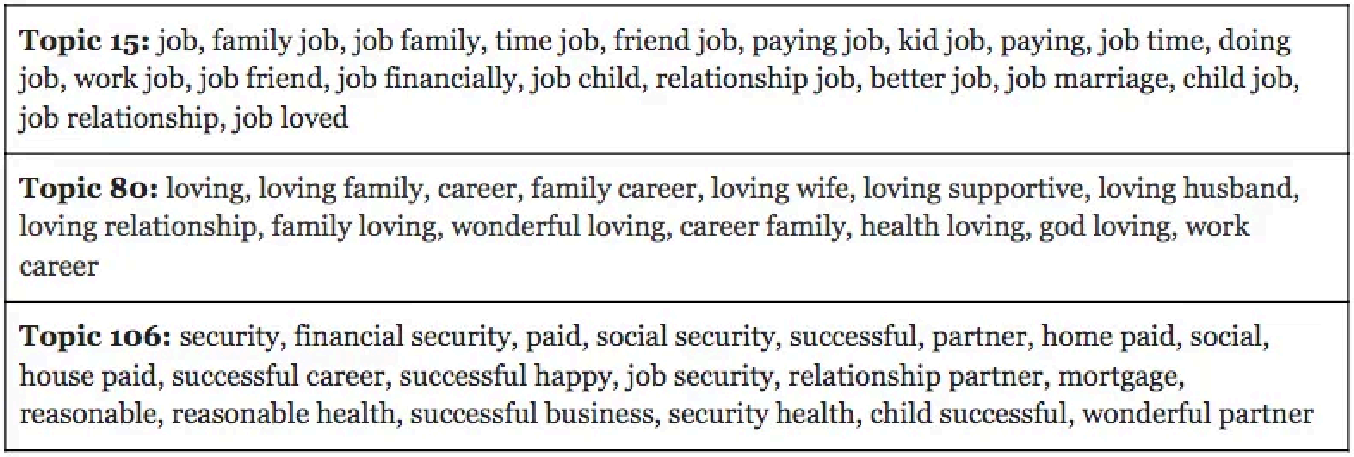
Conceptually spurious words may seem like a minor problem, but they can profoundly impact the results of your analysis. Even just a few extra words can have surprisingly dramatic consequences. Assuming that we want to use Topic 5 to measure whether or not a respondent mentions charity or giving, we probably want to remove several out-of-place terms: "woman," "grateful," and "philosophy." Unfortunately, with most common topic modeling algorithms, this is not an easy task. The words and their distributions across each topic are difficult to modify without a fair amount of programming and statistical knowledge. For the sake of simplicity, though, let's assume that we can roughly approximate whether a document has a topic or not simply by checking to see if it contains any of the top words in the topic.

If we do this for the top words in Topic 5 (giving, woman, grateful, charitable, giving community, nonprofit, philosophy), we find that there are 163 responses in our corpus that match to at least one of these keywords. But if we remove the three words that seem unrelated to the concept of giving or charity (woman, grateful, philosophy), our findings change dramatically. Just 60 documents now match — *less than half* the number that we got from the words in the raw topic model output.

This highlights a major limitation of topic modeling: the "topics" they produce are often not necessarily topics in the traditional sense of the term, and even one out-of-place word can make the difference between a model that accurately measures whether documents mention a coherent concept and one that vastly overestimates how often that concept actually appears in the corpus.

There is no magic number of topics

Conceptually spurious words are not the only potential problem with topic models. Another issue arises from the simple fact that researchers must specify a predetermined number of topics without knowing what that number should be. Regardless of the number you pick, many of your topics may come out looking like this:



Topic 15: job, family job, job family, time job, friend job, paying job, kid job, paying, job time, doing job, work job, job friend, job financially, job child, relationship job, better job, job marriage, child job, job relationship, job loved

Topic 80: loving, loving family, career, family career, loving wife, loving supportive, loving husband, loving relationship, family loving, wonderful loving, career family, health loving, god loving, work career

Topic 106: security, financial security, paid, social security, successful, partner, home paid, social, house paid, successful career, successful happy, job security, relationship partner, mortgage, reasonable, reasonable health, successful business, security health, child successful, wonderful partner

Topic 106 above is an example of what some of us at Pew Research Center have come to colloquially refer to as an “undercooked” topic. Rather than capturing one coherent theme, we can see a variety of themes embedded in this topic, all mixed together. How might we interpret this group of words as a single theme? Is this topic about security, finance, success, health, or relationships? Here, our model appears to be grouping multiple distinct topics into a single topic. Accordingly, we may look at this topic and decide to re-run the model with a greater number of topics so it has the space to break these topics apart.

However, in the *very same model*, we also have Topic 15, an example of an "overcooked" topic. Rather than being too broad or confounded to be interpretable, this topic is too granular. We might be inclined to give it a label like "mentioning your job or career," but in that case, we would probably want to include other words like "career," "work," and "profession." In fact, the word "career" actually exists elsewhere in our model in Topic 80 — another potentially "undercooked" topic — and if we want to bring those words together into a single topic, we might actually need to *reduce* the total number of topics to encourage the model to merge them.

Even then, the topics may not condense the way we want them to. In fact, "job" and "career" may simply not co-occur frequently enough in our documents for the very reason we want them to be in the same topic together — they're more or less synonyms, and likely to be substituted for one another. Depending on our data, the use of both words in a single document could be redundant and therefore uncommon, making it impossible for our model to connect the dots between the two words. Indeed, based on the output of this model, it seems that both words occur frequently alongside a wide variety of other terms ("family" and "work" in particular), but not alongside each other enough for the model to pick up on it.

No matter which particular algorithm and parameters we've used, or how many topics we've asked a model to produce, we've found that undercooked and overcooked topics are largely unavoidable. Metrics like perplexity and coherence can help guide you towards a number of topics that minimizes this problem, but no "perfect" number of topics exists that can eliminate it entirely. This is because while topic modeling algorithms are great at identifying clusters of words that frequently co-occur, they do not actually understand the context in which those words occur.

As researchers, we're interested in measuring the distribution of certain themes in our documents, but some of those themes may be general, while others might be nuanced, detailed and specific. Topic models can find useful exploratory patterns, but they're unable to reliably capture context or nuance. They cannot assess how topics conceptually relate to one another; there is no magic number of topics; and they can't say how the topics should be interpreted. Of course, supervised classification algorithms can make use of more sophisticated linguistic models to overcome some of these limitations, but this requires training data that can be time-consuming and costly to collect. And the very promise of unsupervised topic modeling is that it's fast, easy, and avoids the need for manual coding.

In future posts, I'll explore potential ways to overcome the limitations of topic models and assess the extent to which we can actually use them to reliably measure concepts in a corpus.

Patrick van Kessel is a senior data scientist at Pew Research Center.

Text Analytics Topic Modeling Machine Learning Data Analysis Research

Published in Pew Research Center: Decoded

2K Followers · Last published Dec 7, 2022

Follow