Selecting a Topic Model for Curation (v6)

Background

A "topic model" is a computational way of automatically analyzing an entire body of text to identify categories of things being discussed. Examples of text that can be analyzed in this way might include a set of open-ended text survey responses, social media posts, conversational turns in interview transcripts, or public comments to a federal agency. In the process of computational topic modeling, an automatic process analyzes the whole dataset to produce "topics" that represent categories or concepts found in the collection of text. But it doesn't *name* them; it just labels them Topic 1, Topic 2, etc.

You can think of topic modeling as making an automatic start on the first step of the familiar qualitative content analysis process, where an analyst reads through the body of material being coded, takes and organizes notes, and identifies meaningful categories that will be used in further analysis. Our process of "curating" a topic model starts with the topical categories that have been proposed automatically, and makes them more understandable and meaningful by having human subject matter experts figure out what they're about, i.e. assigning each automatically-obtained topic a meaningful name and description (or possibly deciding it should be discarded).

Selecting an automatic analysis for curation

There are a variety of levers and dials involved in the automatic part of the process – for example, when you run an automatic topic model, you have to tell it what level of granularity you'd like in the analysis, i.e. how many topics it should look for. Different levels of granularity can give you different results. For example, if you ask for too few topics, you can wind up with muddled results that try to combine multiple categories into one when it's better to keep them separate. Conversely, if you ask for too many topics, you can wind up with multiple categories that are just small variations of each other and should have been combined.

Right now there is no reliable and automatic way to automatically pick the correct granularity. Therefore, instead we will provide analyses with multiple granularities, and we're asking you to start by identifying which one seems like the best starting point for human curation. There are more and less formal/rigorous ways of deciding which granularity is the best starting point for the human curation process.

> Least rigorous and fastest. For each granularity K, look at its *clouds.pdf* file. Each cloud shows the words that are most important in characterizing the topic. For example, if you see a cloud where the largest words are *dog*, *cat*, *gerbil*, *vet, goldfish*, it's likely that the model has discovered a reasonably coherent category relate to pets. Skim through the topic clouds for each granularity to form a subjective opinion of how coherent the analysis is. Note that with this method you're relying entirely on the words associated with the topic, and not looking at how the model connects topics/categories with documents; be aware that sometimes the top words don't tell the whole story.
>
> More rigorous. For each granularity, one at a time, open the *categories.xlsx* spreadsheet in Excel Optionally also open the *clouds.pdf* file, if you find the cloud visualization useful. The blue bars in the spreadsheet are showing you the same thing as the clouds: the bigger the blue bar for a word, the more important it is in telling you what the category is about – so this corresponds to the size of the word in the cloud. (We include both formats because some people like one better than the other.) In addition, below the top most important words for each topic, you'll see texts in the dataset that are most "about" that topic, according to the model. (For people who do qualitative content analysis, you can think of this as ranking verbatims by how good they are at representing the topic.) Look at the top words, *skim* some of the top documents from the top down, and then assign a rating from 1 to 3 in the yellow box in column C ("coherence") for that topic. *It is absolutely essential that you do not give into the temptation most qualitative analysts feel to start reading the texts in detail. You also do not need to skim through all the top texts for the topic; typically the first 10-20 top texts, together with the top words, are enough to*

form a sense of how "good" the category is. This process is intended to be a rapid first pass, and the process for full curation of the model comes later. *If you are spending more than 30-60 seconds per topic you are probably taking too long.* Once you have gotten through all the topics in a spreadsheet, click on the column letter C at the very top of the spreadsheet to highlight the whole column . If you look at the bottom of your Excel window, you'll see it's giving you a value for the average of the numbers in this column:

You can view that average as the "score" for this model, and the model with the highest score is the one you should start with for curation.

Most rigorous. Have two (or ideally three) people follow the directions for "More rigorous", above, independently on their own copies of the spreadsheet. For each model, its final score is the average of those people's scores for that model.

In general, you're looking for the analysis in which the categories seem to be the most coherent and well defined, and where the top documents are most "on topic" for that category, although even in the best analysis some categories may not seem like they make much sense and some documents, especially short ones, may not be a good fit for the category (and that's ok).

Please note again that the selection of one analysis is *not* intended to be the same as curating the topics – there are separate instructions for that. You don't need to try to come up with names or descriptions of the categories you're looking at. The intent here is to do an efficient first pass to decide which automated analysis is the best starting point for the next step.

Granularities that are most likely to be best

There are no hard and fast rules governing which granularities are going to be best for any given dataset. This can vary a great deal depending on things like the dataset size, diversity of vocabulary and subject matter being discussed, and length of the texts. (It's worth noting that very short texts are a challenge for standard topic models. There are some specialized topic models in the literature designed specifically to deal with shorter texts.) However, the following are rules of thumb based just on dataset size that we and others have found tend to work reasonably well.

For a document collection with fewer than 500 texts, we would typically try granularities of 5, 10 and 15, though the automatic process may or may not produce anything of use at all for collections that small.

For 500-to-1000 texts: granularities of 10,15,20 or 10,20,30. LDA

1000-to-10000 texts: granularities of 15,20,40 or 20,30,50

10000- to-100000+: granularities of 75,100,150

These are anecdotally consistent with what we have heard from a number of other frequent topic model practitioners. Crucially, the human curation process reduces the burden to identify any particular model size as optimal; in general we tend to err mildly on the side of more rather than fewer topics since our curation process permits less-good topics to be discarded and fine-grained topics can be merged under a single label and description. Automated model selection for content analysis is a topic of ongoing work.