Instructions for Reviewing and Labeling Topics

Version of 23 November 2023

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 collections 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. Following standard convention in topic modeling, we refer to any individual item in such a collection as a *document*, or sometimes as a *text*. In the process of computational topic modeling, an automatic process analyzes the whole dataset to produce "topics" (sometimes also called "themes") that represent categories or concepts found in the collection of text. But it doesn't *name* them or provide descriptions; it just labels them Topic 1, Topic 2, etc. That's where you come in.

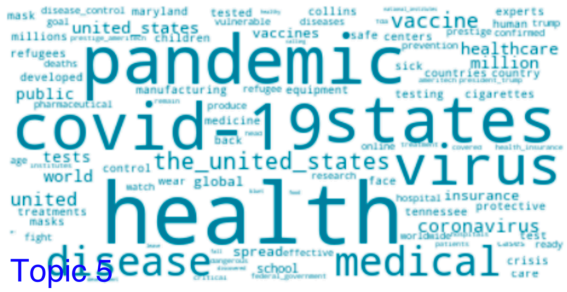
You can think of topic modeling as making an automatic start on the first step of the familiar content analysis process of coding human language, where an analyst reads through the body of material being coded, takes and organizes notes, and identifies meaningful categories that will be used in further analysis. The goal here is to start with the topical categories that have been proposed automatically, and make them more understandable and meaningful by figuring out what they're about. Your task is to give each of Topic 1, Topic 2, etc. a meaningful name and description.

If you have prior experience with coding and coding scheme development, you can think of what we're asking you to do as using the automatic topical categories in a systematic way, in order to efficiently organize your analysis of the dataset to produce a set of meaningful codes.

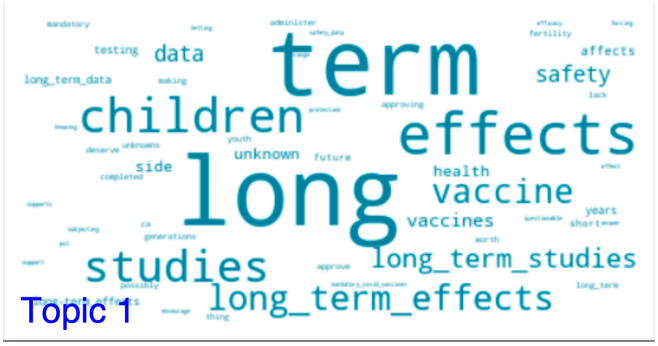
The materials you'll be working with

Along with these instructions, you should have received several files. We recommend that you look at each one as we're describing it in this section. Again, the goal of this process is to replace the names Topic 1, Topic 2, etc. with meaningful labels and create meaningful descriptions.

Clouds PDF file. The first file is a PDF whose name would typically end in *clouds.pdf*. This provides one way of looking at what the topics are. Each page is showing you a cloud of words or phrases associated with a single topic or category that was discovered in the automatic analysis of the whole data collection. Size in this cloud can be interpreted as how *important* that word or phrase is to your understanding of what the topic is about. For example, we did an analysis of speeches in the U.S. Senate from 2020, and the cloud for Topic 5 appears at left. Based on the words that are largest, and therefore most important for this topic – for example *health*, *pandemic*, *covid-19*, *virus*, *medical*, etc. – it is looks like this automatically-discovered topic or category in Senate speeches is connected with the COVID-19 pandemic.



<u>Categories</u> Excel file. The second file is a spreadsheet, and the file name will typically end in *categories.xlsx*. This file also shows you the set of topics/categories from the automatic analysis. For each topic, it shows two things. First, the blue bars in the spreadsheet are showing you the same thing as the clouds in the PDF: 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 use both formats because some people like one better than the other. The example below eis from an analysis of public comments to the US Food and Drug Administration about emergency authorization of a COVID-19 vaccine.)



Topic
Topic 1
long
term
effects
children
studies
long_term_effects
vaccine
long_term_studies
data
safety
vaccines
side
health
unknown
affects
years
long-term_effects
long_term_data
testing
short

In addition, below that bar chart with the top words, you'll see the top *texts*, i.e. the documents in the dataset that are most "about" that topic, according to the model. (For people familiar with qualitative content analysis, you can think of this as a ranking of verbatims by how well this topic applies to them.)

years	
long-term_effects	
long_term_data	
testing	
short	
No vaccines for ANYONE, especially children until long term studies in the effects have been completed! We keep hearing there were	
I am commenting in opposition to mandatory COVID vaccines for children. We have no long term (5-10 years or more) data on affect	
I think this is a mistake without further testing. You can't tell anyone what long term side affects could be caused by this. Do not let th	
NO LONG TERM STUDIES! EVERY OTHER SINGLE DRUG & VACCINE GOES THROUGH YEARS OF SAFETY STUDIES. I cannot trust a produ	
I am very much against the approval of the Covid vaccine for children. The long term effects are not known and will not be for many	
Long term studies cannot be possible because this is new technology. Our kids have been the most resilient in this whole pandemic, a	
There is no reason to push this vaccine on children. The vaccine requires much more long-term testing. We do not know the long-tern	
Please do not implement COVID 19 vaccines for children ages 5-12. There are still too many unknowns on these vaccines to put them	
Long term studies are needed especially in children. Please prolong this decision until longer term, wider scale studies have been com	
I don't not believe in this mandate we don't not know the long term effects this has on our children.	
These vaccines should not be mandated for children at all. There are not enough long term studies done. I am not assured that these	
Please do not approve this. The benefits do not outweigh the risks. There are no long-term studies of this new technology.	
Until we know the long range effects of these vaccines we should NOT expose our children to them! This is simply common sense!!!!	

<u>All documents</u> Excel file. The third file is a spreadsheet with complete category information for each document in the collection that was analyzed. Typically the file name will end with *alldocs.xlsx.* Each row corresponds to one document in the dataset – in the example below, one turn speaking on the floor of the Senate – which is shown in the text column at the far right. You will see columns labeled Topic 1, Topic 2, etc. (Below we're just showing the columns for Topics 5-7 for simplicity.)

	docID	...	Topic 5	Topic 6	Topic 7	...	text
2	1		0.00011	0.10066118	0.00014374		Madam President, I would like to offer my congratulations to Donna Kelly-Williams as she retires as president of the Massachusetts Nurses Associatic
3	2		0.5376	1.81E-05	0.08808783		Mr. President, just 3\1/2\ months ago, a sneaky, dangerous virus turned our country and the world upside down. It is hard to believe that it was just
4	3		0.0002	0.00020803	0.15835818		In August, I held 49 county meetings as part of my annual 99-county set of meetings in Iowa. At every meeting, the format is the same: Iowans set th
5	4		0.00228	0.00240318	0.0030459		Mr. President, I ask unanimous consent that the order for the quorum call be rescinded.
6	5		0.5099	4.19E-05	5.31E-05		Mr. President, I rise today to highlight the heroes of my home State of Maryland who are working on the frontlines to fight COVID-19. On January 21,
7	6		0.00254	0.00267455	0.00338983		Mr. President, I ask unanimous consent that the text of the bill be printed in the Record.
8	7		0.00031	0.00032963	0.00041778		Madam President, section 36(b) of the Arms Export Control Act requires that Congress receive prior notification of certain proposed arms sales as de

The number in each column (a value between 0 and 1) indicates to what extent the item in that row contains discussion of that topic or, to put it another way, the extent to which the topical category applies to that item. For example, in the spreadsheet above, you can see that document IDs 2 and 5 (the floor speeches in lines 3 and 6 in the spreadsheet) both have higher values for Topic 5. For those familiar with coding, you can think of the topic as a proposed code, and the number between 0 and 1 as a score indicating how much of the language in the item "represents" that code.

Thanks to the numbers in the table, using Excel it's possible to sort (that is, re-order) the spreadsheet in order to look at the documents that are *most strongly* associated with each topic. For example, if you sort the spreadsheet by the "Topic 5" column, ordering from larger to smaller values in that column, the top rows look like this:

docID	...	Topic 5	Topic 6	Topic 7	...	text
2		0.5376	1.81E-05	0.08808783		Mr. President, just 3\1/2\ months ago, a sneaky, dangerous virus turned our country and the world upside down. It is hard to believe that it was just 3
380		0.53459	3.88E-05	4.92E-05		Madam President, I want to report on an important hearing the Senate HELP Committee just completed. Senator Murray and I organized it. We heard
5		0.5099	4.19E-05	5.31E-05		Mr. President, I rise today to highlight the heroes of my home State of Maryland who are working on the frontlines to fight COVID-19. On January 21,
498		0.49203	0.00046574	0.0005903		Mr. President, I thank Senator Blunt and Senator Murray for their cooperation this month in this series of six hearings that we have had on COVID-19.
588		0.47756	9.83E-05	0.07485811		Mr. President, I am delighted to be here today with my friend, the Senator from California, with whom I have worked so closely on so many issues. W
599		0.47239	6.29E-05	7.97E-05		Madam President, as chairman of the Committee on Small Business and Entrepreneurship, each week I recognize a small business that exemplifies th
327		0.42049	5.72E-05	7.25E-05		Mr. President, we are now several months into a global pandemic that has caused terrible human and economic suffering. Here in the United States a
617		0.40069	0.01531691	0.00308916		Madam President, as I noted earlier, Samuel Johnson once noted that there is nothing that focuses the mind like the prospect of a hanging. I would sa
664		0.37863	0.0002636	0.00033409		Madam President, yesterday afternoon, South Dakota Governor Kristi Noem notified me that multiple residents of South Dakota have tested positive
163		0.34777	0.00010042	0.00012728		Mr. President, I want to thank my colleagues, Senators Baldwin and Murphy, for their urgent words on this most important issue. I am proud to join th

In this case inspection of the texts near the top suggests they generally are discussing the pandemic, confirming the impression we got just by looking at the cloud of words in the cloud PDF file. That suggests a good name or label for Topic 5 might be, say, *COVID-19 Pandemic*. Sometimes reading the texts that have been sorted to the top might lead to a more nuanced view than you can get just by looking at the cloud of words; for example, if you saw a lot of discussion about infection rates in different states, you might decide that *Pandemic Infection Rates* might be a more accurate label.

What you'll be doing

What follows below is a structured, step-by-step process that you should follow in order to turn the automatically-discovered "Topic 1", "Topic 2", etc. into a set of labeled categories that are understandable and meaningful for this dataset.

The length of time required for this process will vary depending on the number of topics, the number and length of the documents reviewed, the reviewer's reading speed, and the time devoted for sustained attention to the task. For a model with 30 topics, following the steps below would typically take roughly 90 to 120 minutes.

There are two different ways to go through this process.

Way 1: Steps to Follow Using the Excel All-Documents Workbook

The following steps use the categories Excel file. (If you don't know how to do something in Excel, see "Helpful Hints" at the end of this document. Also, remember that, just like for anything you do on a computer, you should save your work frequently rather than waiting until the end!) Note that for very large collections of items, more than 10,000 rows or so, this process might be a little clunky and you might want to consider using Way 2 (below) instead.

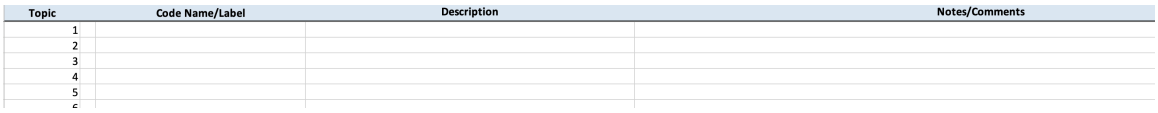
We recommend you make a note of what times you start and end working on this, to keep track of how long the whole process takes. Knowing how much time you've spent is very valuable, since it enables a comparison with traditional manual analysis.

Here are the steps.

1. If you like the cloud format for topics, open the clouds PDF file (or print it out for easy reference). If you prefer the bar-chart format, open the *categories.xlsx* spreadsheet instead. Or open both.

2. Open the all-documents (*alldocs.xlsx)* spreadsheet on the left side of the screen.

3. On the right side of the screen, open a fresh copy of the coding scheme (*coding_scheme.xlsx*) spreadsheet, which will contain the coding scheme you're creating, guided by the topics. There is one row for each topic. In addition to a column with the topic's number, there are additional columns: *Code Name/Label* is the label for the code or category. *Description* is a description of the category. And the *Notes* column is for any other optional notes that might be useful.



Topic	Code Name/Label	Description	Notes/Comments
1			
2			
3			
4			
5			

4. Treating the column names of the all-documents spreadsheet (e.g. *Topic 1*, *Topic 2*, etc.) as a checklist, please go through each one completing the following steps:

   a. Sort the spreadsheet by that topic's column in descending (largest-to-smallest) order.

b. Set a timer for 120 seconds. This is simply to keep you from reading all the text responses in case they're interesting. You can use a cell phone, web browser or other timer. ("Siri [or Alexa], set a timer for two minutes!")

c. Skim down the text column for the top documents to get a sense of what kinds of responses are strong on this particular topic. Also look at this topic's top words (via the cloud or the bar-chart or both). You might want to jot down common words or ideas under *Notes* in the labels-remarks spreadsheet (discussed in Step 3, above). For people familiar with developing qualitative coding schemes, notice that this is very much like the traditional "emergent coding" process of reading through verbatims and developing your notes/ideas for codes in a sructured way:, you're being guided by the automatic process to look efficiently, one at a time, at categories that it is proposing.

d. You may find that documents you're looking at for this topic are actually about more than one thing. That's ok: you're not labeling documents here, you're looking at the documents to get a clearer understanding in your mind about what *this column's topic* is about.

e. Once you've formed your concept for what this topic is about by looking at the top words and top documents, consolidate any notes you took in the *Notes* column to create a description in the *Description* column, and give the category a brief label or code name in the *Name/Label* column.

f. Replace the original column header in the all-documents spreadsheet (e.g. *Topic 1*) with your new name/label. Don't forget to save your files as you continue working!

g. If the topic you are reviewing contains responses that are consistently similar to another topic you've read, it's possible that the model may have discovered two (or more) categories that you really consider to be the same category. If that's the case:

   Copy the information in the row from that other topic in your coding scheme to this row in the coding scheme (except for the topic number).

   Replace the column name header in the document-topic spreadsheet (e.g. *Topic 2*) with the same name you used for the previous, similar topic.

   Use copy/paste rather than re-typing to avoid typos and ensure that all the information for the two equivalent categories is identical.

h. If the topic you are reviewing seems like it has some meaning to it, but you can't settle on a single category description or label (e.g. it does not seem to have a consistent theme, or contains multiple themes that you have identified in other topics), you can use the label "Misc" and provide notes on about it in the Description and Other Notes column in your notes spreadsheet. It's ok to use "Misc" more than once.

i. If you just can't make sense of the topic you are reviewing and you don't think there's a meaningful category there at all, use the label *DISCARD* for this topic.

j. Sometimes a category is coherent and understandable, but it's irrelevant for your analysis. For example, if there are a small number of Spanish items mixed in with a mostly-English dataset, sometimes you might get a category with top words *el, la, un, una,* etc.; that is, a "theme" consisting of frequent Spanish words. Optionally, you might want to consider using a structured label to indicate this, e.g. *IRREL-Spanish*. You can take a similar strategy of structured code labels if there are other distinctions you'd like to make; for example, you might find that there are two different topics that are *about* the same thing, but one expresses positive attitudes and the other one expresses negative attitudes. In such cases it's also often useful to assign a structured multi-part label. For example, in a set of responses about presidential approval, you might see topics that are appropriate to label *Immigration-POSITIVE* and *Immigration-NEGATIVE*.

When all topics have been reviewed and given labels and descriptions, then:

5. Make a final pass over the categories that you identified to see if any are close in their meaning and should be combined. If so, follow the directions in 2g above.

6. Once you're done, you have two "products" of this process:

   Coding scheme. The coding-scheme spreadsheet now lists your codes and their descriptions. In addition, the top documents for each topic give you an easily accessible list of "verbatims" that can be useful in reporting on your analysis.

   First-pass coding. The all-documents spreadsheet can be thought of as a first-pass "soft" coding of items using your new coding scheme, where the number between 0 and 1 in each cell can be interpreted as "how well" the code applies to that item. (Note that multiple codes can apply well to a single item.) We are working on technical tools to support efficient human review and correction of this first-pass coding process, but for the moment you can find initial recommendations for an efficient process using Excel in the TOPCAT documentation (README.md).

Helpful hints for Excel

- One-minute video on how to hide and unhide columns in Excel, so that you can look at a smaller number of columns at a time. https://youtu.be/trk1MIOynm8

- One-minute video on how to "wrap" long text in Excel, so that instead of the text just going outside the edges of the cell, instead the cell will expand to fit all the text into it. It van be helpful to make a column wider first, and then "wrap" for easier readability. https://youtu.be/CiWjGKXvrbI

- Two-minute video on how to sort (i.e. re-order) the rows in your spreadsheet by the values in one of the columns. https://youtu.be/9KjkVDH3_ig