Instructions for Reviewing and Labeling Topics

Version of 23 February 2022

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” (sometimes also called “themes”) 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. That’s where you come in.

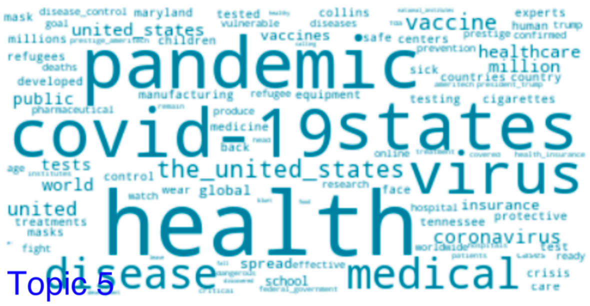
You can think of topic modeling as making an automatic start on the first step of the familiar content analysis process of coding human language, where an analyst reads through the body of material being coded, takes and organizes notes, and identifies meaningful categories that will be used in further 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 codeset 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.

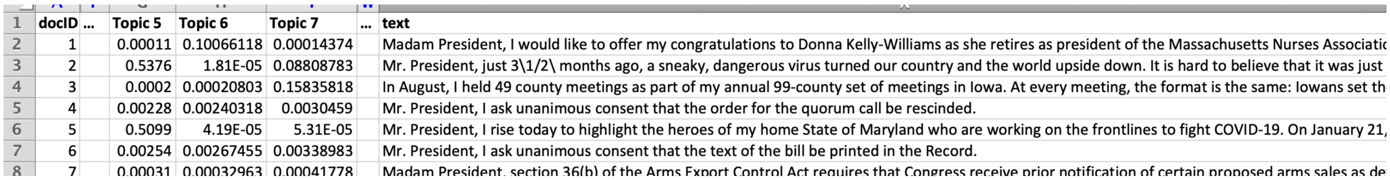
The materials you’ll be working with

Along with these instructions, you should have received two 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.

PDF file. The first file is a PDF, with 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.



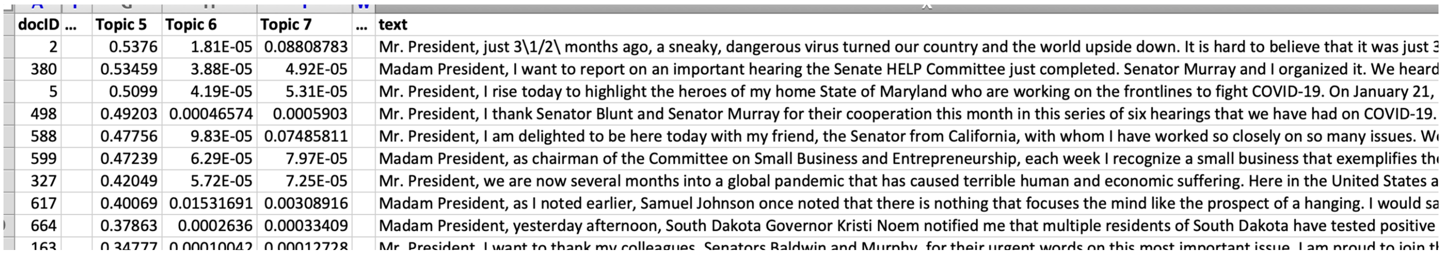
Spreadsheet. The second file you're receiving is a spreadsheet. Each row corresponds to one document in the dataset – in this case, one turn speaking on the floor of the Senate – which is shown in the text column at the far right. You will see columns labeled Topic 1, Topic 2, etc. Below is an example. (We're just showing the columns for Topics 5-7 for simplicity.)



1	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 text in that row contains discussion of that topic. 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.

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. 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.

How much time this will take

The length of time required for 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.

Steps to Follow Using an Excel Workbook

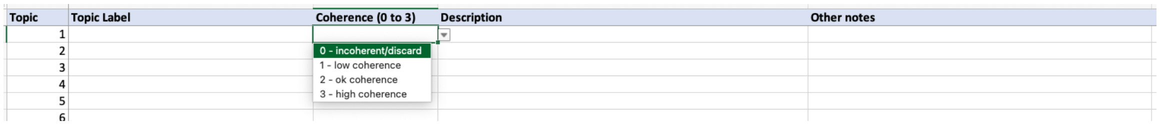
If you are using an Excel workbook for this process, please proceed according to the following steps. If you don't know how to do something in Excel, see "Helpful Hints" at the end of this document, or just ask us and we'll show you! (Also, note that, just like for anything you do on a computer, you should save your work frequently rather than waiting until the end!)

Please 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 information for this study.

1. Either open the PDF file with the topic clouds, or print it out for easy reference.

2. Open the document-topics spreadsheet on the left side of the screen.

3. On the right side of the screen, open the labels-remarks spreadsheet, which is the codebook you're creating, as guided by the topics. There is one row for each topic (e.g. *Topic 1, Topic 2, etc.*). In addition to a column with the topic's number, there are additional columns: *Topic Label* is the label for the code/category. *Description* is a codebook-style description of the category. *Coherence* is a rating, from 0 to 3, of how coherent or interpretable the topic is, which you can also think of as how confident you are that the automatically-generated topic corresponds to an understandable category; a value of 0 means that you think the automatically-generated topic is not useful and should be discarded. The *Other Notes* column is for any other optional notes that might be useful.



Topic	Topic Label	Coherence (0 to 3)	Description	Other notes
1				
2				
3				
4				
5				
6				

4. Treating the column names of the document-topic 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 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 cloud in the PDF. You might want to jot down common words or ideas under Other Notes in the labels-remarks spreadsheet (discussed in Step 3, above).

   d. You may find that documents you're looking at are about more than one topic. 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. Consolidate the notes you took in the Other Notes column to create a description in the Description column, and give the topic category a brief label or name in the Topic Label column. Replace the original column header in the document-topic spreadsheet (e.g. *Topic 1*) with your new label.

f. 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 notes table (discussed in Step 3, above). 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.

g. 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.

h. 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, assign a Coherence score of 0 in your notes spreadsheet and use the label *DISCARD* for this topic.

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

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

6. Send us back the two spreadsheets (document-topics and labels-remarks). Please also copy/paste the following questions into your cover email and provide answers:

   a. How long did you spend on the entire process, including the instructions (but leaving out any breaks you took)?
   b. Did you do it all in one sitting or did you take breaks?
   c. On a scale from 1 (terrible) to 5 (great), how was the quality of the topics that you were given to work with?
   d. If you've engaged in more traditional codeset development before, please comment on how this compares with that previous experience.
   e. Please identify any aspects of the process that you particularly liked.
   f. Please identify any aspects of the process that you think could be improved, and we welcome any specific suggestions for improving it.

Thanks very much!

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