Instructions for Reviewing and Labeling Topics

(Adapted from Philip Resnik's work on getting labels for topics derived from Reddit posts)

Objective

This project is using a computerized method called a "topic model" to identify topics or themes of discussion in political text. You will be reviewing the output of the computational model to help determine what the topics are (by assigning names or labels to the automatically discovered topics), and the coherence and polarization of those topics.

You will be provided with two excel spreadsheets:
(A) A spreadsheet of 50 unnamed topics and words strongly associated with them ("top words"), and,
(B) A spreadsheet containing documents from the dataset, and their associated topic scores (i.e. which topics the model thinks are present in each document).

You will refer to each topics' top words from spreadsheet A and documents from spreadsheet B to create your own topic label, evaluate the coherence of the topic, and the likelihood of a policy's polarization.

This analysis requires an understanding of American politics and partisan polarization.

Background

A "topic model" is a computational way of automatically analyzing an entire body of text to extract themes. An automatic process analyzes a dataset of texts to produce "topics" or categories present within the collection of text. Examples of text that can be analyzed in this way include a set of floor speeches by Congress members, social media posts, interview transcripts, etc.

However, this approach only differentiates topics and organizes the documents–it does not *name* the topics. A topic looks a lot like a cluster of words and labeling computer-generated topics requires human input. (For example, the "top words" or most characteristic words for some topic might be *dog, cat, gerbil, fish*, and a good label for this topic might be "Pets".)

Thus, the goal here is to use the topical categories that have been proposed automatically for a set of texts authored by politicians, and:

1. Figure out what they are about, by labeling each topic with a meaningful *name* and description.
2. Rate the proposed automated categories for their *coherence:* does that topic represent an easily identifiable category or a meaningful concept?
3. If the proposed category is deemed coherent and represents an identifiable category or meaningful concept, rate the topic on expected ideological *polarization*: do you expect

that category or concept to be discussed differently by liberals and conservatives?

The materials you'll be working with

Along with these instructions, you should have received one zip folder which will have a PDF copy of these instructions and two folders: one marked "speeches" and the other marked "tweets". Both those folders will be structured exactly the same way, and the below instructions apply to both.

Note that the remainder of this section is focused on the materials, with some overview of what you'll be doing, but the precise instructions you should follow are in the next section, titled *Instructions*.

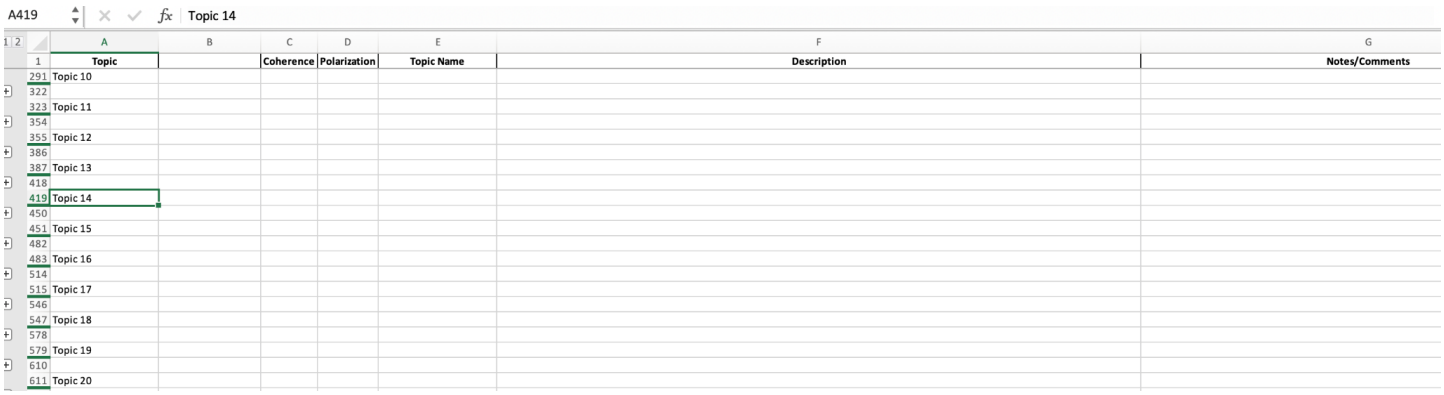
NOTE: The below example is meant to be just an example – your labeling or any other interpretation does not need to mirror this example.

There will be two files (in each folder): 1) **topics_for_annotation.xlsx,** that contains the top words of each topic (hidden by default but the user can expand and see the top words for each topic as explained below), and will be used to document your label and rating, and 2) **document_topics.xlsx**, which contains all of the documents in the dataset. Please refer to both of these files when determining a topic label, coherence, and expected ideological polarization:

1. **topics_for_annotation.xlsx**: The first file will be used to record all annotation work: the topic name (and description), coherence rating, and polarization rating. There is also space for additional comments to supplement the label and ratings. This is the spreadsheet that you will send back (for each folder). It also contains the top words for each topic that you will use for the task (the top documents are in a separate file). The columns (and their descriptions):

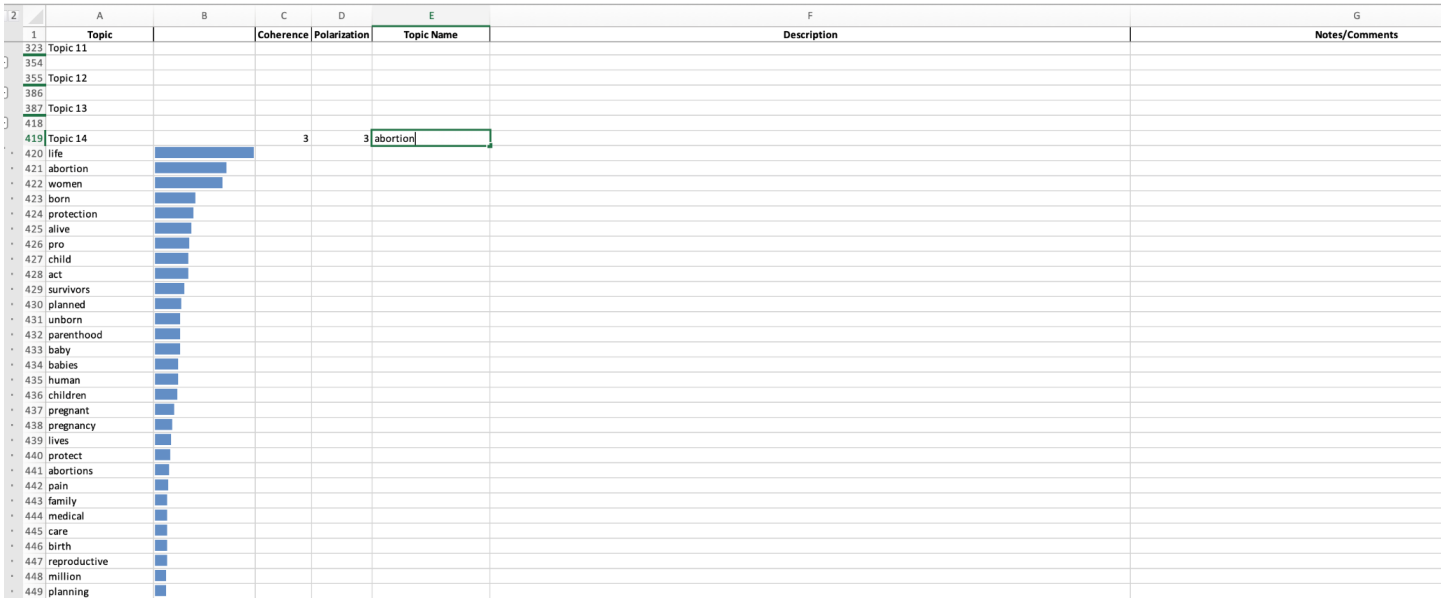
    a. **Topic:** This is the automatically identified category by the topic model without names: Topic 1, Topic 2, etc.

    For each topic, click the "+" just to the left (and slightly below) to see the top words associated with that topic. These should be considered alongside the document text (present in the other file described below) in generating a topic label. After you are done with each topic, you can then hide those words by clicking the "-", Example pre- and post-expansion for a topic (Topic 14) below in Figures 1a and 1b:



	Topic		Coherence	Polarization	Topic Name	Description	Notes/Comments
291	Topic 10						
322							
323	Topic 11						
354							
355	Topic 12						
386							
387	Topic 13						
418							
419	Topic 14						
450							
451	Topic 15						
482							
483	Topic 16						
514							
515	Topic 17						
546							
547	Topic 18						
578							
579	Topic 19						
610							
611	Topic 20						

Fig 1a



	Topic		Coherence	Polarization	Topic Name	Description	Notes/Comments
323	Topic 11						
354							
355	Topic 12						
386							
387	Topic 13						
418							
419	Topic 14		3	3	abortion		
420	life						
421	abortion						
422	women						
423	born						
424	protection						
425	alive						
426	pro						
427	child						
428	act						
429	survivors						
430	planned						
431	unborn						
432	parenthood						
433	baby						
434	babies						
435	human						
436	children						
437	pregnant						
438	pregnancy						
439	lives						
440	protect						
441	abortions						
442	pain						
443	family						
444	medical						
445	care						
446	birth						
447	reproductive						
448	million						
449	planning						

Fig 1b (description is also required)

b. **Coherence:** Rate the topic, on a 1-3 scale, on its coherence (1 being not coherent, 3 being very coherent). Does that topic represent an easily identifiable category or a meaningful concept? The top words for the topic, along with the top documents associated with that topic help make this judgment. Broadly, a set of items can be said to be *coherent* if they enable human recognition of an identifiable category when viewed together.

c. **Polarization**: *If the topic is rated above 1 for coherence*, rate the topic, on a 1-3 scale (1 being not polarized, 3 being polarized), on the expected polarization of this topic: do you expect meaningful ideological differences in the way liberals and conservatives would talk about the category or concept or issue or the stance they would hold on that issue? The top words for the topic, along with the top documents associated with that topic, as well as your personal knowledge of American politics will help make this judgment.

d. **Topic Name**: Name or label the concept/category/issue you think the topic (as per the top documents in the other csv file and the top words in this particular spreadsheet) represents.

e. **Description**: Describe the topic and expand upon the name you chose, i.e., a meaningful short description of the category you identified.

f. **Notes/Comments**: Anything you may want to note for that topic. Might be left blank.

The end output should be a filled spreadsheet. For example, for Topic 14 (top words shown above and top documents shown below), you might label it as "Abortion". If this label or category was easily identifiable or recognizable to you, you might rate the coherence as 3.

Going off of your knowledge about politics in the US, you might rate the topic as 3 for polarization if it represents an ideologically polarized issue to you.

Use the fallback label "DISCARD" for the Topic Name if the topic seems incoherent and collects largely miscellaneous words rather than representing an identifiable category or a particular concept.

<u>You will make these determinations by also consulting the top documents for each Topic:</u>

2. **document_topics.xlsx:** Each row corresponds to one document in the dataset – either a House floor speech or a tweet by an elected representative. The text for each document is in the far right column. Each prior column is labeled Topic 1, Topic 2, etc. corresponding to the first file. 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.

   You will use these numbers to sort the spreadsheet from the documents that are *most strongly* associated with each topic to least associated. This will allow you to quickly read the top documents associated with each topic. For example, if you sort the spreadsheet by the "Topic 14" column, ordering from larger to smaller values in that column, the top rows look like this:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	docID	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16	Topic 17
2	13793	0.001	0.001	0	0.001	0	0	0.001	0.001	0	0.001	0.001	0	0	0.964	0.001	0	0.003
3	14284	0.001	0.001	0	0.001	0	0	0.001	0.001	0	0.001	0.001	0	0	0.964	0.001	0	0.003
4	19982	0.002	0.001	0.001	0.001	0	0.001	0.001	0.001	0	0.001	0.001	0	0	0.959	0.002	0	0.003
5	15216	0.002	0.001	0.001	0.001	0	0.001	0.001	0.001	0	0.001	0.001	0	0	0.957	0.002	0	0.004
6	17940	0.002	0.001	0.001	0.001	0	0.001	0.001	0.001	0	0.001	0.001	0	0	0.956	0.002	0	0.004
7	14414	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0.955	0.001	0	0.021
8	15134	0.002	0.001	0.001	0.001	0	0.001	0.001	0.001	0	0.001	0.001	0	0	0.954	0.002	0.001	0.004
9	20360	0.002	0.001	0.001	0.001	0	0.001	0.001	0.001	0	0.001	0.001	0	0	0.954	0.002	0.001	0.004
10	16410	0.002	0.001	0.001	0.001	0	0.001	0.001	0.001	0.001	0.001	0.001	0	0	0.949	0.002	0.001	0.004
11	14581	0.002	0.001	0.001	0.001	0	0.001	0.001	0.001	0.001	0.001	0.001	0	0.001	0.948	0.002	0.001	0.004
12	13420	0.002	0.001	0.001	0.001	0	0.001	0.001	0.001	0.001	0.001	0.001	0	0.001	0.946	0.002	0.001	0.004
13	13518	0.002	0.001	0.001	0.001	0	0.001	0.001	0.001	0.001	0.001	0.001	0	0.001	0.946	0.002	0.001	0.004
14	13703	0.002	0.001	0.001	0.001	0	0.001	0.001	0.001	0.001	0.001	0.001	0	0.001	0.946	0.002	0.001	0.004
15	13858	0.002	0.001	0.001	0.001	0	0.001	0.001	0.001	0.001	0.001	0.001	0	0.001	0.946	0.002	0.001	0.004
16	13950	0.002	0.001	0.001	0.001	0	0.001	0.001	0.001	0.001	0.001	0.001	0	0.001	0.946	0.002	0.001	0.004
17	14056	0.002	0.001	0.001	0.001	0	0.001	0.001	0.001	0.001	0.001	0.001	0	0.001	0.946	0.002	0.001	0.004

. . .

. . .

You can then scroll to the right to see the top document text:

AX	AY	AZ
Topic 49	Topic 50	text
0.001	0.001	Mr. Speaker, I ask unanimous consent that the Committee on the Judiciary be discharged from further consideration of H.R. 962, the Born-Alive Abortion Survivors Protection Act. This is essential legislation that would protect the lives of children who survive the trauma of an abortion, children who deserve to be given the best medical care, a bill that should not be controversial.
0.001	0.001	Madam Speaker, I ask unanimous consent that the Committee on the Judiciary be discharged from further consideration of H.R. 962, the Born-Alive Abortion Survivors Protection Act, to protect the right to life for innocent children who are born alive instead of allowing the State-sponsored murder after birth, and ask for its immediate consideration in the House.
0.001	0.001	Mr. Speaker, I ask unanimous consent that the Committee on the Judiciary be discharged from further consideration of H.R. 962, the Born-Alive Abortion Survivors Protection Act, which is necessary to save the innocent lives of innocent children, and ask for its immediate consideration in the House.
0.001	0.001	Madam Speaker, if this unanimous consent cannot be entertained, I urge the Speaker and the majority leader to immediately schedule the Born-Alive Abortion Survivors Protection Act, so we can stand up and protect the sanctity of human life, and I would ask all others to join in that request.
0.001	0.001	Madam Speaker, I ask unanimous consent that the Committee on the Judiciary be discharged from further consideration of H.R. 962, the Born-Alive Abortion Survivors Protection Act, an act that protects living babies who survive failed abortions, and ask for its immediate consideration in this House.
0	0	Madam Speaker, a few minutes ago, I rose to call for a vote on the Born-Alive Abortion Survivors Protection Act, which requires any baby who survives an abortion to receive the same medical care that any baby born at the same age would receive. It requires the baby to then be transported to a hospital. As a doctor, I strongly believe that every patient, especially these infants born alive, should be given appropriate medical care. This should not even be a question. New York recently celebrated passing a law that removes protections from babies born alive after an abortion attempt. Other States also fail to protect abortion survivors. Therefore, it is our duty, as Members of Congress, to defend the God-given right to life for every baby in this situation. It is our duty, as compassionate human beings, to ensure that these uniquely vulnerable babies receive the care that they deserve. It is past time to vote on H.R. 962.
0.001	0.001	Mr. Speaker, I ask unanimous consent that the Committee on the Judiciary be discharged from further consideration of H.R. 962, a bill that would protect innocent children, the Born-Alive Abortion Survivors Protection Act, and ask for its immediate consideration in the House.
0.001	0.001	Mr. Speaker, I ask unanimous consent that the Committee on the Judiciary be discharged from further consideration of H.R. 962, the Born-Alive Survivors Protection Act, legislation which protects the sanctity of life for the unborn by ensuring that infants who are born alive receive proper medical care, and ask for its immediate consideration in the House.
0.001	0.001	Mr. Speaker, I ask unanimous consent that the Committee on the Judiciary be discharged from further consideration of H.R. 962, the Born-Alive Abortion Survivors Protection Act, and I ask for its immediate consideration in the House so we can defend life.
0.001	0.001	Mr. Speaker, I rise to ask unanimous consent that the Committee on the Judiciary be discharged from further consideration of H.R. 962, the Born-Alive Abortion Survivors Protection Act, and ask for its immediate consideration in the House.
0.001	0.001	Mr. Speaker, I ask unanimous consent that the Committee on the Judiciary be discharged from further consideration of H.R. 962, the Born-Alive Abortion Survivors Protection Act, and ask for its immediate consideration in the House.
0.001	0.001	Madam Speaker, I ask unanimous consent that the Committee on the Judiciary be discharged from further consideration of H.R. 962, the Born-Alive Abortion Survivors Protection Act, and ask for its immediate consideration in the House.
0.001	0.001	Mr. Speaker, I ask unanimous consent that the Committee on the Judiciary be discharged from further consideration of H.R. 962, the Born-Alive Abortion Survivors Protection Act, and ask for its immediate consideration in the House.
0.001	0.001	Madam Speaker, I ask unanimous consent that the Committee on the Judiciary be discharged from further consideration of H.R. 962, the Born-Alive Abortion Survivors Protection Act, and ask for its immediate consideration in the House.
0.001	0.001	Mr. Speaker, I ask unanimous consent that the Committee on the Judiciary be discharged from further consideration of H.R. 962, the Born-Alive Abortion Survivors Protection Act, and ask for its immediate consideration in the House.
0.001	0.001	Madam Speaker, I ask unanimous consent that the Committee on the Judiciary be discharged from further consideration of H.R. 962, the Born-Alive Abortion Survivors Protection Act, and ask for its immediate consideration in the House.

Fig 2

The corresponding texts when the spreadsheet is sorted on values for Topic 14 (largest to smallest) again show abortion as the issue under discussion, confirming the impression we got

from looking at the top words in the previous file. 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 top words and should guide the description and notes you make.

Note: Sorting by a column value and other helpful tips for excel are provided at the end of this document.

How much time this will take

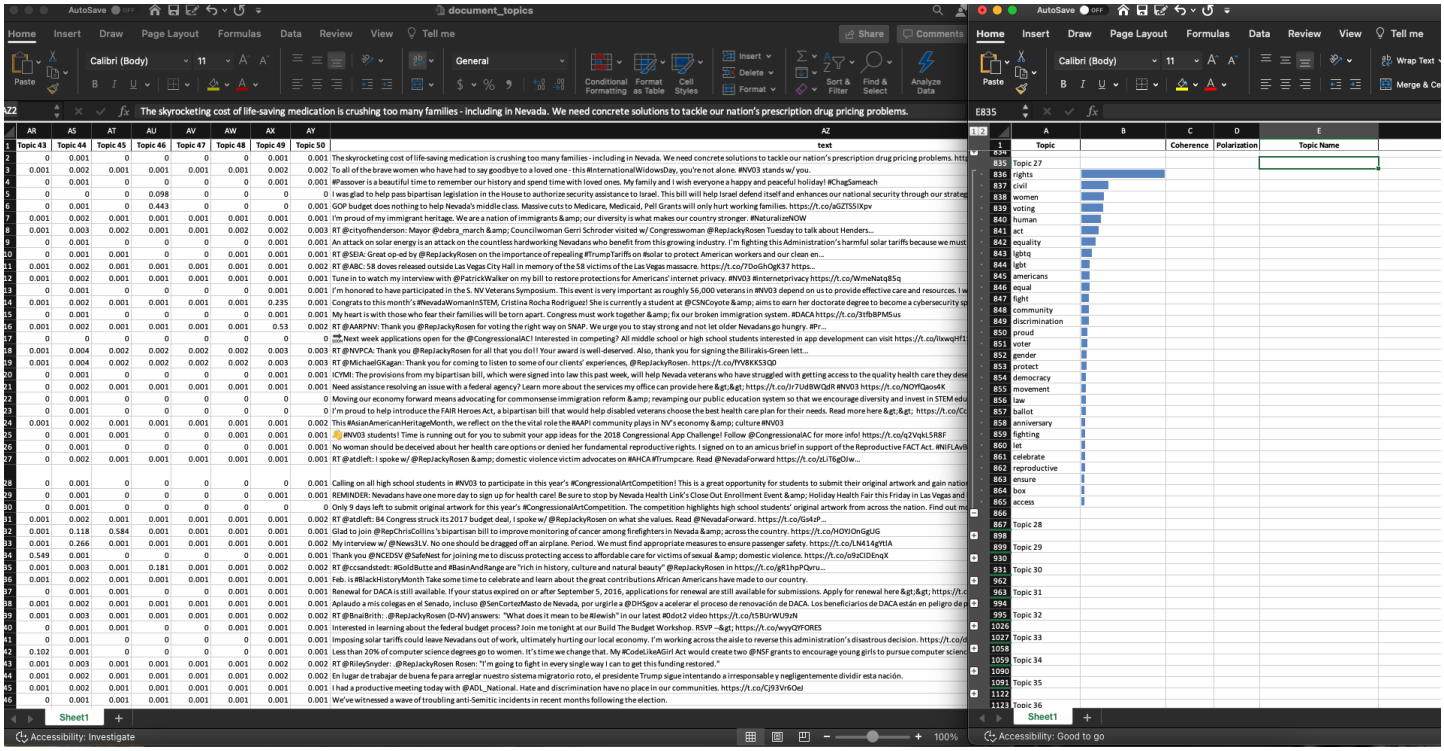
The length of time required for the labeling process should take about 200-250 minutes (per dataset). Each topic should ideally take not more than 2 to 5 minutes for seeing the top words and looking at the top documents and performing the labeling and analysis. The instructions below include setting a timer for the maximum length of time to spend per topic.

Please keep track of the time you've spent working on the whole process and report that to us along with your results.

Instructions

(Note that the steps will be followed twice - once for speeches, and once for tweets.)

1. Open the **document_topics.xlsx** spreadsheet and the **topics_for_annotation.xlsx** to where they are simultaneously viewable on your screen. An example of opening the two windows and having them side-by-side -



2. Treating the column names of the document-topic spreadsheet (e.g. *Topic 1*, *Topic 2*, etc.) as a checklist, go through each one completing the following steps. (Note that you can go

through them in any order you'd like, although the default would just be to do Topic 1 first, then Topic 2, etc.)

a. Use the "+" icon to the left and slightly bottom of the topic in the **topics_for_annotation.xlsx** file to show the top words for that topic. At this point you might already get a sense of what this topic or category is about based on those words, and you might wish to give the theme an initial label or name in the Name column, put information in the Description column to provide any context you may want to for your label or name, and use Notes/Comments for initial additional thoughts you may want to note. Alternatively, you could wait to do these things until after the next step; either way is ok. **Sort** the **document_topics.xlsx** spreadsheet by that topic's column in descending (*largest-to-smallest*) order; as a result, the documents in the *text* column will appear in order with the ones at the top being the ones most strongly associated with this topic. Then *set a timer for 180 seconds for speeches and 120 seconds for tweets*. Look at the *text* column for the top documents (which, in case of long speeches, you may want to not read in full) to just get a sense of what kinds of documents are strong on this particular topic and help refine your sense of what this topic is about. **Note: Before skimming the text for top documents for a particular topic, it's important to set a timer and end your skimming when the timer is up!**
b. Based on reading the top words and top documents for the topic, **assign** the theme its brief label or name in the **Name** column,use the **Description** column to provide any context you may want to for your label or name, and use **Notes/Comments** for any additional thoughts you may want to note. Note: sometimes, the theme might not appear to be very consistent, or perhaps you might identifymultiple or a mix themes and it is a bit hard to settle on a single label or name (even when there does seem to be an identifiable category or meaningful topic) – try to come up with the best label you can, and note any such things in Description and Notes/Comments.

c. **Generate a Coherence score** (1-3).
    i. Are you easily able to identify the category or concept represented by that automatically identified topic? If no, then your rating should be 1; yes, then your rating should be 3.

    Only use the "2" rating if you are unsure in either direction - you seem to identify the possible category but it required some thinking and it was not straightforward. If you are struggling to identify the concept, then the rating will be 1. You can use the Notes/Comments column to provide any context you may want to.

    ii. Note: 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 about what the topic is.

d. **DISCARD topics:** If you just can't make sense of the topic you are reviewing and

you don't think there's a meaningful category at all, make a note of that in the Description column in your notes spreadsheet and use the label/name *DISCARD* for this topic. Coherence of it will likely be 1 and you will not need to rate it for polarization per the process described above.

3. Make another pass through your topics_for_annotation.xlsx file and **Provide a polarization rating (1-3) *only for the topics you rated 2 or 3 on coherence.*** If the topic seems likely to be ideologically polarized in the stances liberals and conservatives would likely systemically hold on that issue, or the way they will approach the issue in their rhetoric, rate the topic as 3. If you do not expect such ideological polarization for that concept or issue in your view, you can rate it as 1. If you think there is a chance of polarization, but are unsure not sure, you can rate 2–however, this should be used sparingly. Use your topic name and description (plus any notes you made earlier) to provide this rating. You can use the Notes/Comments column to provide any further context you may want to.

When all topics have been reviewed and given names, coherence ratings, and polarization ratings when applicable, then:

4. Make a final pass to ensure all topics have been covered, especially for coherence ratings and polarization ratings where the rating is 2, and if the DISCARD labels have been applied appropriately.

Expected OUTPUT -

5. A **completed, filled out topics_for_annotation.xlsx file for each of the two folders** (speeches and tweets). Return the two spreadsheets, labeled **topics_for_annotation.xlsx_lastname_speech/tweet.** In addition, in your email sending the labeled files, please **report the total time you spent working on this task (excluding reading instructions or taking breaks).**

Thank you for your help on this project!

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. https://youtu.be/CiWjGKXvrbI

Note: the document_topics.xlsx files we provide should have wrapped text by default, but if in some case it is not present in the files you download, the link above can help.

- 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

This is what sorting on a particular topic looks like for me in the document_topics.csv (after selecting Topic 14 from the drop down menu for column, and selecting “Largest to Smallest” for the Order):

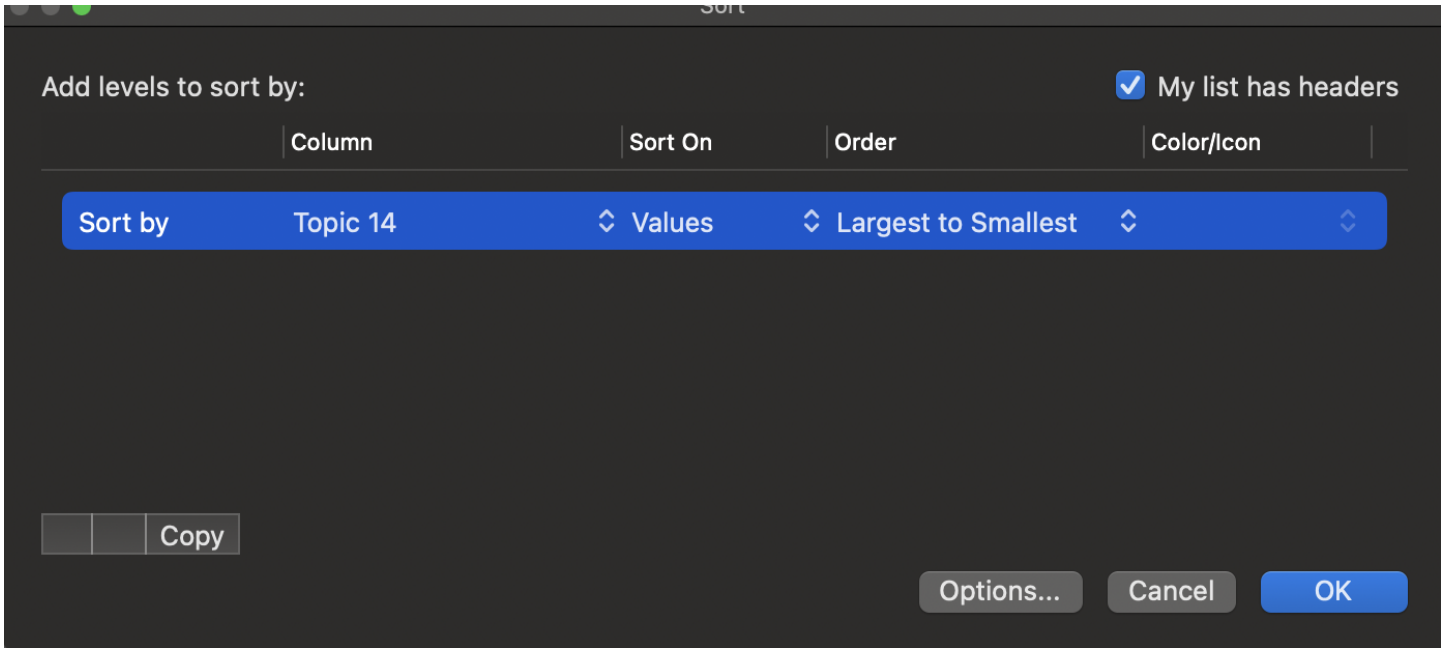
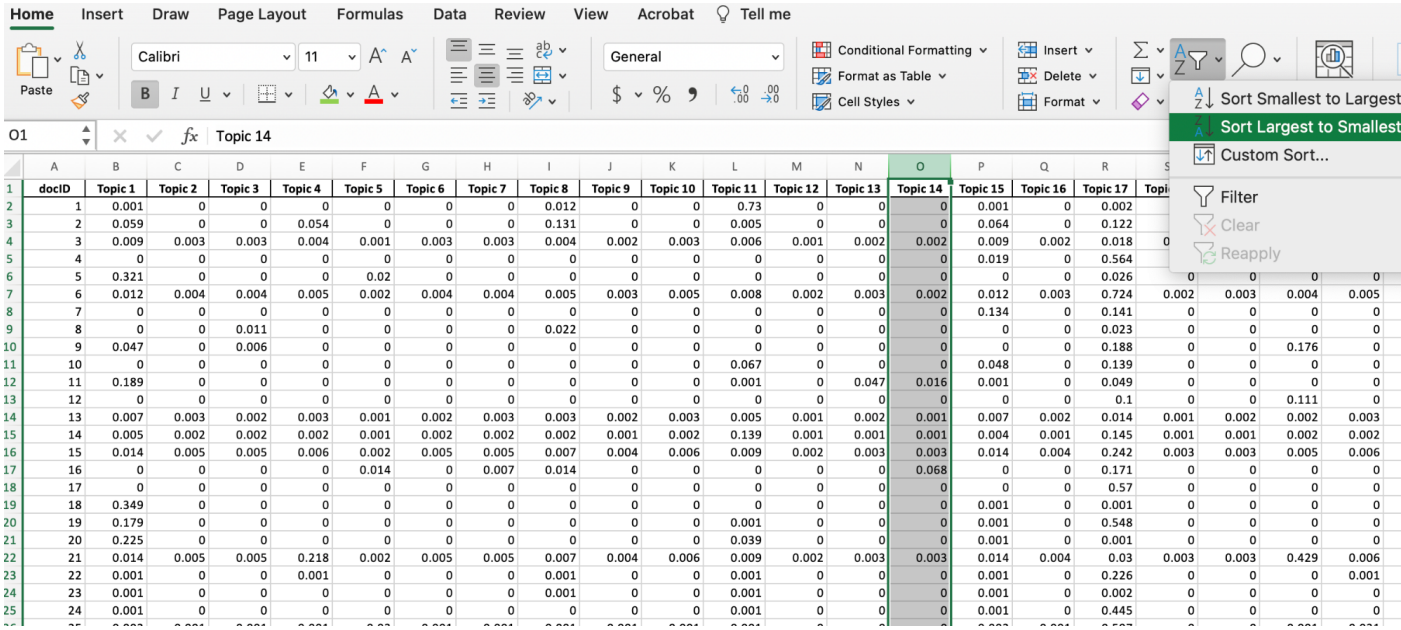


Fig 3

It may also look like this:



docID	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16	Topic 17
1	0.001	0	0	0	0	0	0	0.012	0	0	0.73	0	0	0	0.001	0	0.002
2	0.059	0	0	0.054	0	0	0	0.131	0	0	0.005	0	0	0	0.064	0	0.122
3	0.009	0.003	0.003	0.004	0.001	0.003	0.003	0.004	0.002	0.003	0.006	0.001	0.002	0.002	0.009	0.002	0.018
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.019	0	0.564
5	0.321	0	0	0	0.02	0	0	0	0	0	0	0	0	0	0	0	0.026
6	0.012	0.004	0.004	0.005	0.002	0.004	0.004	0.005	0.003	0.005	0.008	0.002	0.003	0.002	0.012	0.003	0.724
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.134	0	0.141
8	0	0	0.011	0	0	0	0	0.022	0	0	0	0	0	0	0	0	0.023
9	0.047	0	0.006	0	0	0	0	0	0	0	0	0	0	0	0	0	0.188
10	0	0	0	0	0	0	0	0	0	0	0.067	0	0	0	0.048	0	0.139
11	0.189	0	0	0	0	0	0	0	0	0	0.001	0	0.047	0.016	0.001	0	0.049
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1
13	0.007	0.003	0.002	0.003	0.001	0.002	0.003	0.003	0.002	0.003	0.005	0.001	0.002	0.001	0.007	0.002	0.014
14	0.005	0.002	0.002	0.002	0.001	0.002	0.002	0.002	0.001	0.002	0.139	0.001	0.001	0.001	0.004	0.001	0.145
15	0.014	0.005	0.005	0.006	0.002	0.005	0.005	0.007	0.004	0.006	0.009	0.002	0.003	0.003	0.014	0.004	0.242
16	0	0	0	0	0.014	0	0.007	0.014	0	0	0	0	0	0.068	0	0	0.171
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.57
18	0.349	0	0	0	0	0	0	0	0	0	0	0	0	0	0.001	0	0.001
19	0.179	0	0	0	0	0	0	0	0	0	0.001	0	0	0	0.001	0	0.548
20	0.225	0	0	0	0	0	0	0	0	0	0.039	0	0	0	0.001	0	0.001
21	0.014	0.005	0.005	0.218	0.002	0.005	0.005	0.007	0.004	0.006	0.009	0.002	0.003	0.003	0.014	0.004	0.03
22	0.001	0	0	0.001	0	0	0	0.001	0	0	0.001	0	0	0	0.001	0	0.226
23	0.001	0	0	0	0	0	0	0.001	0	0	0.001	0	0	0	0.001	0	0.002
24	0.001	0	0	0	0	0	0	0	0	0	0.001	0	0	0	0.001	0	0.445

Fig 4

We also recommend “freezing” the top row, especially for the **topics_for_annotation.xlsx** spreadsheet (note - this has been done already in all four .xlsx files provided, but if in some case it is not present in the files you download, the top row can be frozen as shown below):

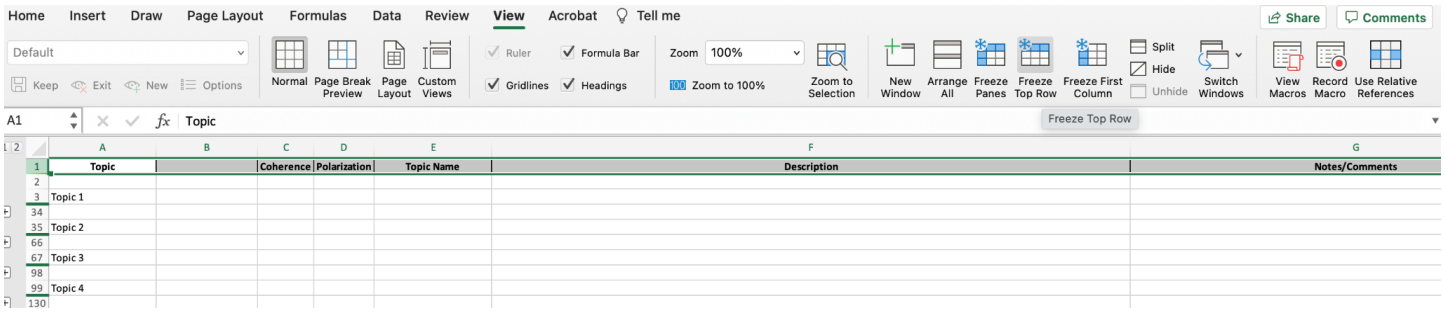

Home Insert Draw Page Layout Formulas Data Review View Acrobat Tell me
Share Comments
Default
Keep Exit New Options
Normal Page Break Preview Page Layout Custom Views
Ruler Formula Bar
Gridlines Headings
Zoom 100%
Zoom to 100%
Zoom to Selection
New Window Arrange All Freeze Panes Freeze Top Row Freeze First Column
Split Hide Unhide
Switch Windows
View Macros Record Macro Use Relative References
A1
Topic
Freeze Top Row
Topic
Coherence
Polarization
Topic Name
Description
Notes/Comments
Topic 1
Topic 2
Topic 3
Topic 4

Fig 5