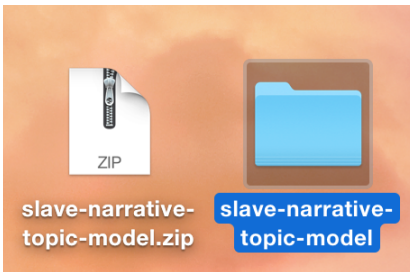
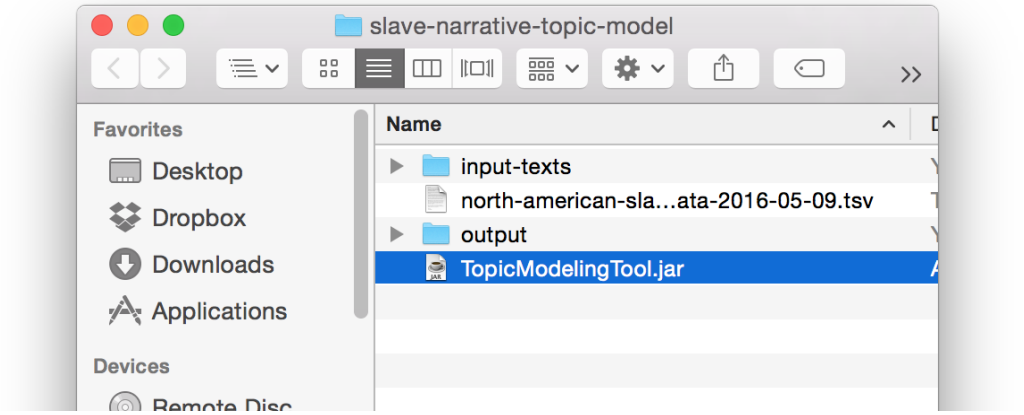
**Using the Topic Modeling Tool**

Download `slave-narrative-topic-model.zip`. Save it to the desktop.

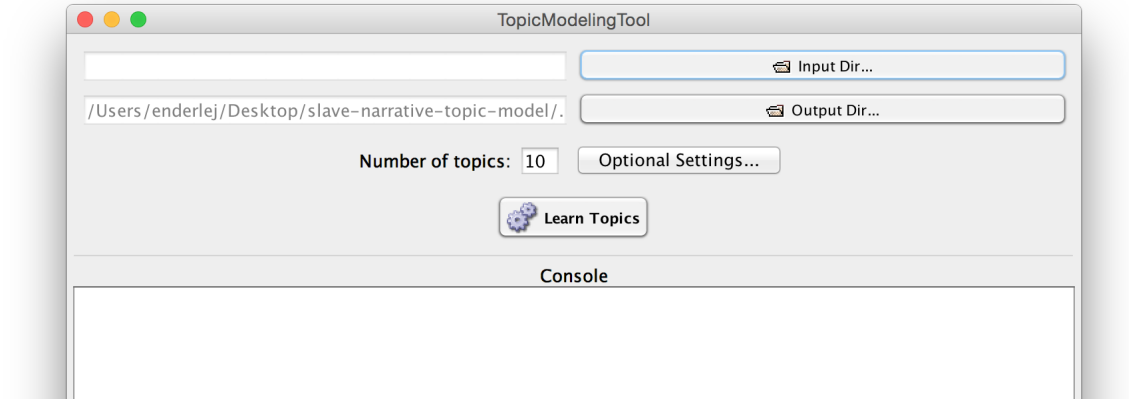

Unzip the file. (On Macs, just double-click the file; on PCs, right-click and select "unzip.")



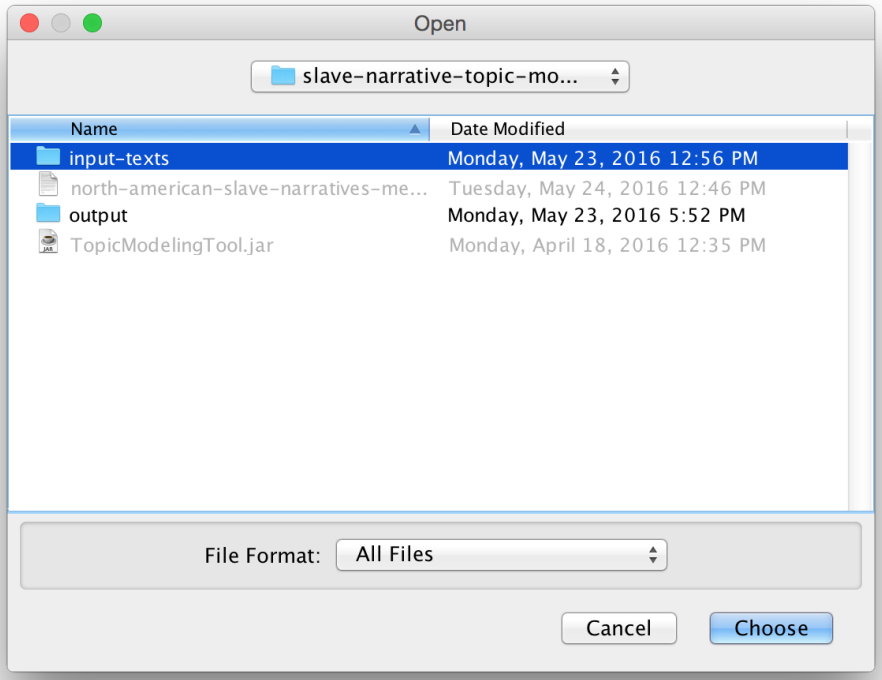

A folder with the same name should appear. Open it.



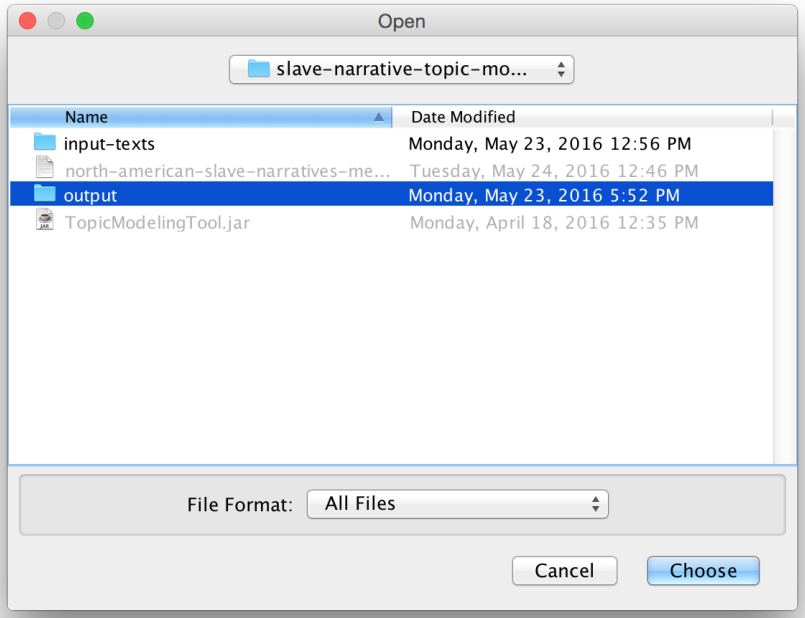

Double-click on `TopicModelingTool.jar`. You might have to right-click and select "open" on Macs.



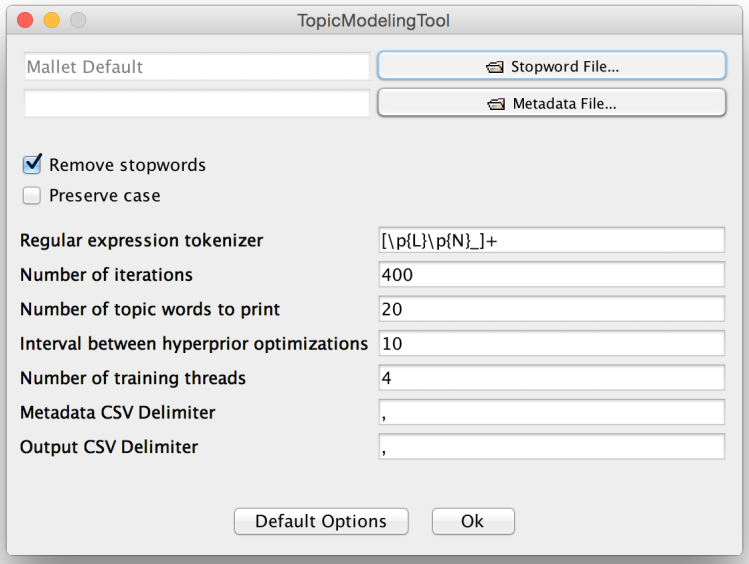

Click on the "Input Dir…" button.



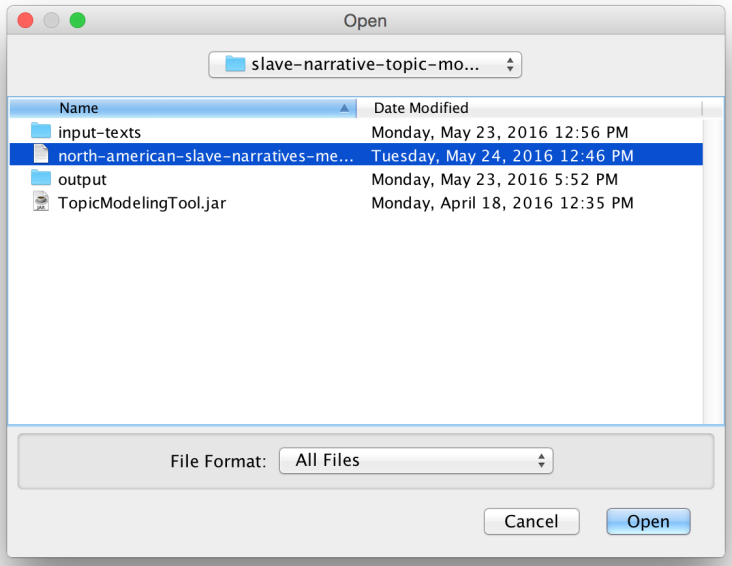

Select the `input-texts` folder. (Do not double-click.) Click the "Choose" button. Click on the "Output Dir…" button.



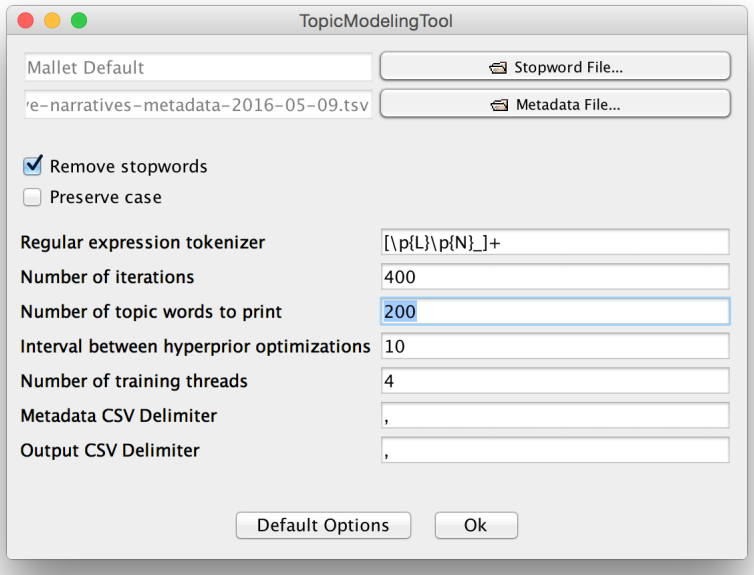

Select the `output` folder. (Do not double-click.) Click the "Choose" button. Click on the "Optional Settings…" button.



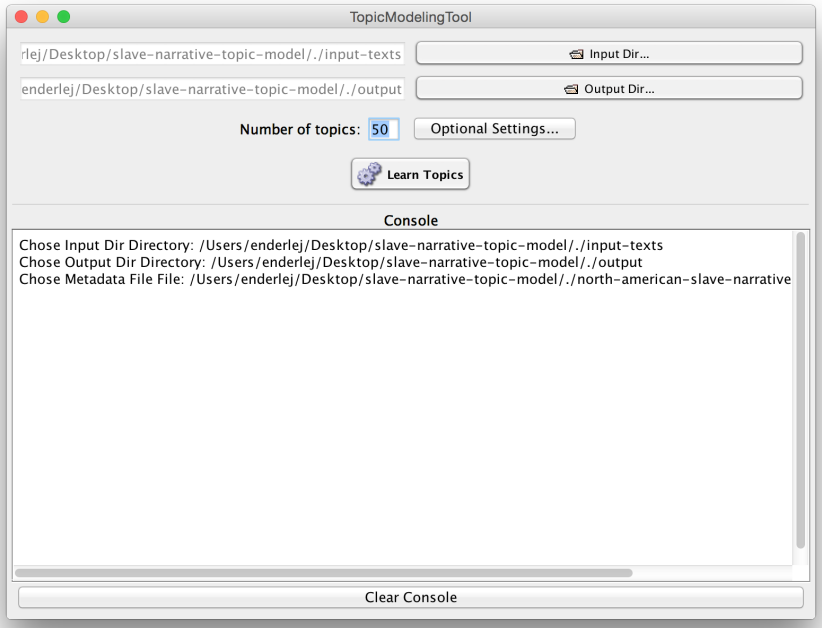

Click on the "Metadata File…" button.



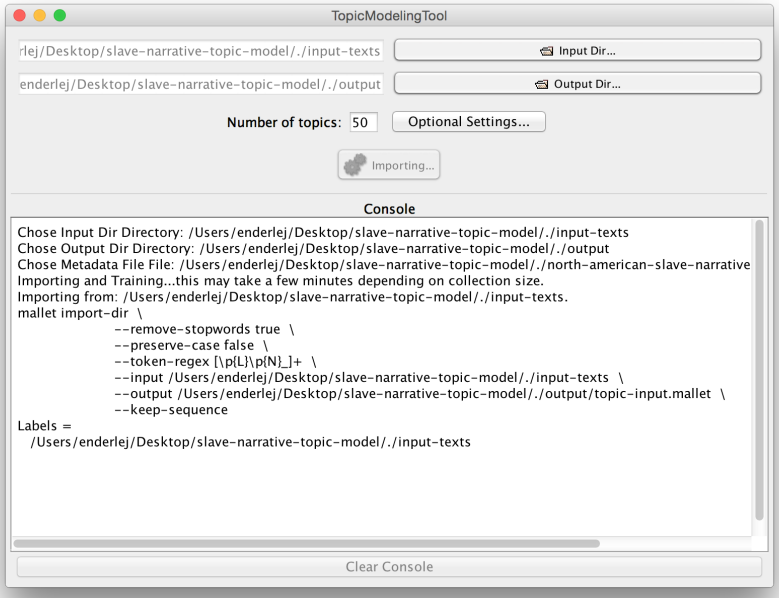

Select the `north-american-slave-narratives-metadata.csv` file. Click the "Open" button.




Change "Number of topic words to print" to 200. Click "OK."

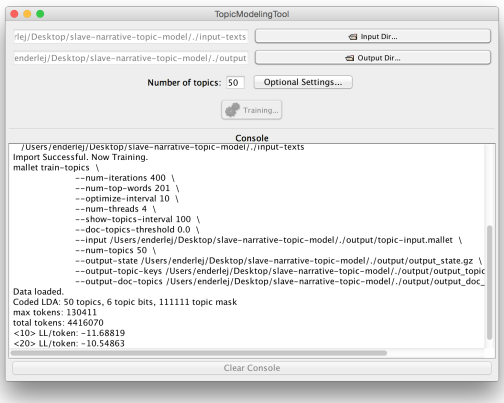


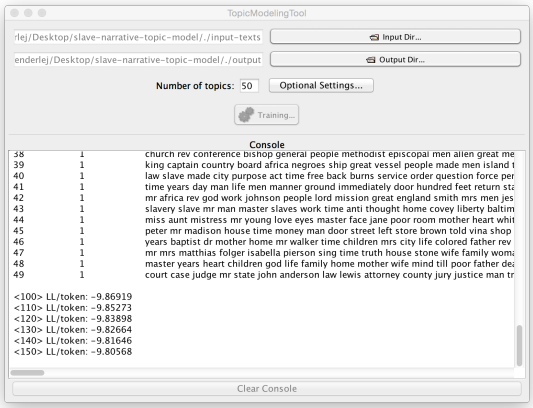

Change "Number of Topics" to 50. Click "Learn Topics"

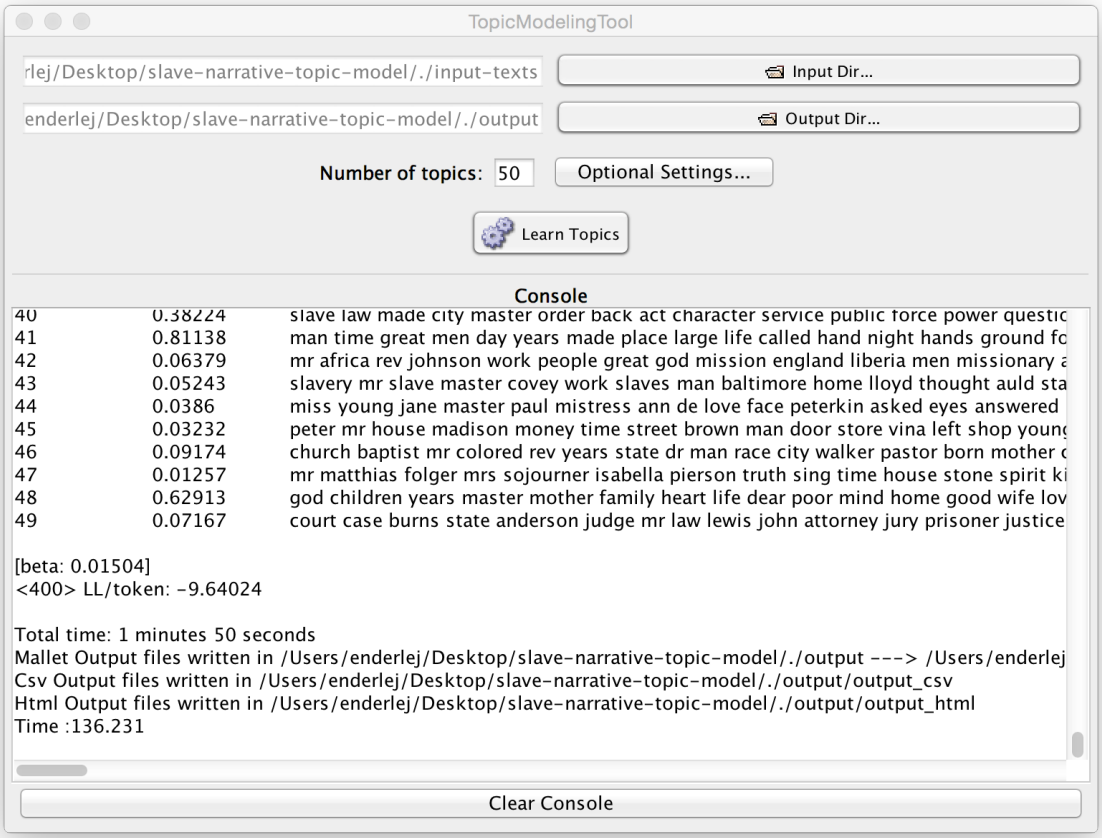



The training process will take about five minutes.

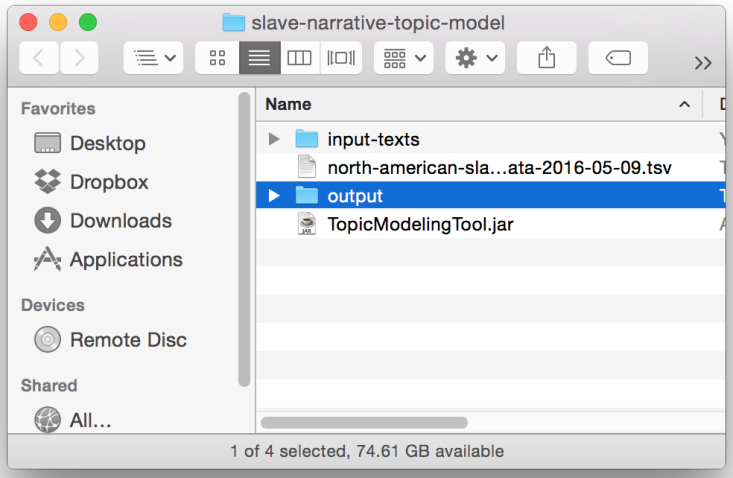
You'll see some progress updates…



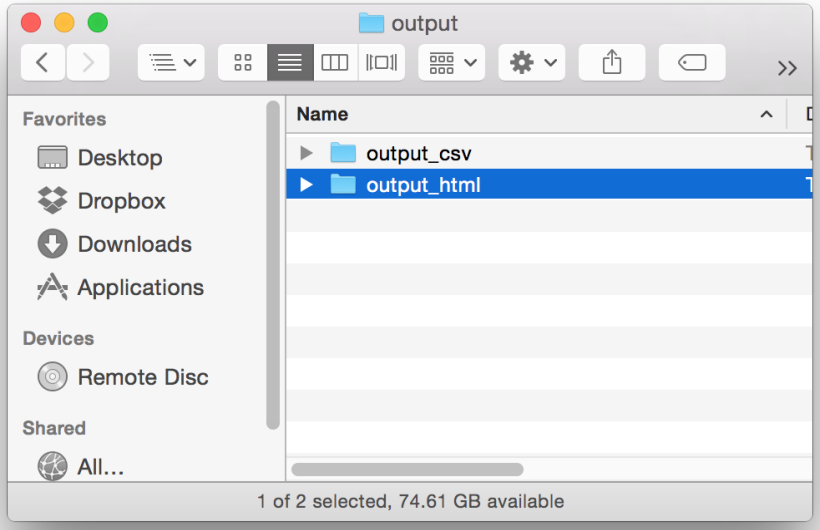



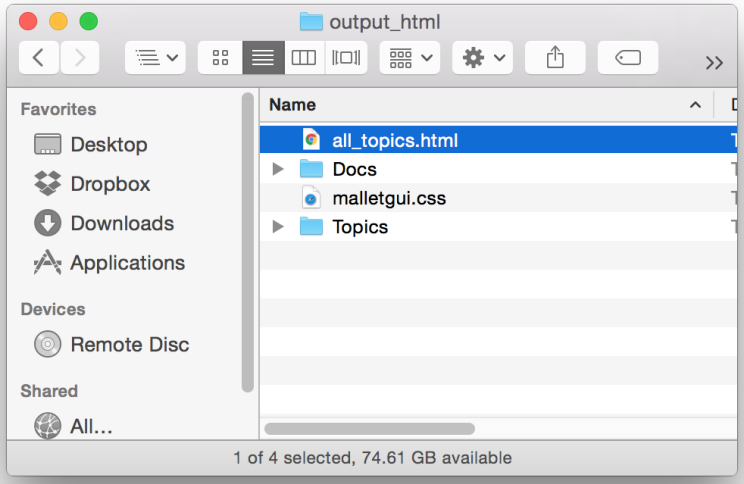

At the end of the process you'll see something like the above window.




Now we can start examining the output. Open the “Output” folder.




The Topic Modeling Tool generates two different kinds of output. The first is raw data stored in CSV files. The second is a set of browsable HTML pages. We’ll take a quick look at the HTML first, just to familiarize ourselves with the output.




Begin by opening the “all_topics.html” file. It should open in your default browser. You’ll see something like this:

**List of Topics**

1. mr good thomas cassy master colonel thing moore overseer carleton till plantation thornton began thought moment montg archy business matter boy presently deal stubbs danger gav husband love feelings money brought whiskey captain chapt indulgence chance favor plantations mind dropped journey a succeeded american find idea carried things understand sun boston vessel dog tavern states tom natural sole rights collec eliza hall mrs fit forward answered servitude gang turned fact happiness bitter meeting table talk gilmore notice committee
2. chaplain cadet officer men army allensworth class point offic soldiers general smith infantry white service war company fro young asked ordered louisville corps sergeant quarters june troops duties captain corporal tent rights honor article time re secretary graduate civil barracks bowling gentlemen adjutant won fight left classes states place position times congress sta assigned thought chaplains uniform rear wanted city commar appointment twenty enemy opinion york month treat due son table american desired true lieut inspection lake serve tents r

If you click on the first topic, it will take you to a page listing the top documents for that topic:

**TOPIC : mr good thomas cassy master colonel thing moore oversee country told carleton till plantation thornton began thought moment neighborhood set girl servant archy business matter boy presently passion sooner hope great woman born husband love feelings mon nature knew labor ready give meadow tone ritty indulgence chance james situation insolence mine struck court fellows succeeded ame distance low beauty crew scarcely served ran intended son boston thy run alarm swamp charleston mere smile caught ll white fields el gained regular misery york proctor torture humanity fever william b**

top-ranked docs in this topic (#words in doc assigned to this topic)

1. (20418) neh-hildreth-hildreth.txt
2. (7074) neh-moore1-moore1.txt
3. (6660) neh-moore2-moore2.txt
4. (1118) neh-jonestom-jones.txt

The documents are sorted by the number of word tokens in them that the algorithm assigned to this topic. By clicking on one of the links, you can look at a preview of the text.

**DOC :neh-hildreth-hildreth.txt**

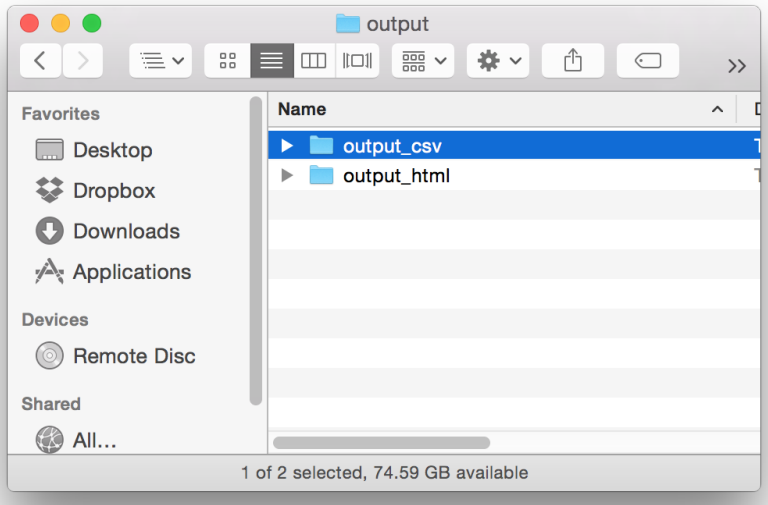
[Cover Image]
[Spine Image]
[Title Page Image]
[Title Page Verso Image]
â€œAll men are by nature equally free and independent, and have certain
INHERENT RIGHTS, of which, when they enter into society, they cannot by any
compact deprive or divest their posterity, namely, the enjoyment of life and
liberty, with the means of acquiring and possessing property, and pursuing
happiness and safety.â€â€”Virginia Bill of Rights, Art. I.
ADVERTISEMENT.
THE earlier chapters of this book were written on a southern plantation, during

Top topics in this doc (% words in doc assigned to this topic)

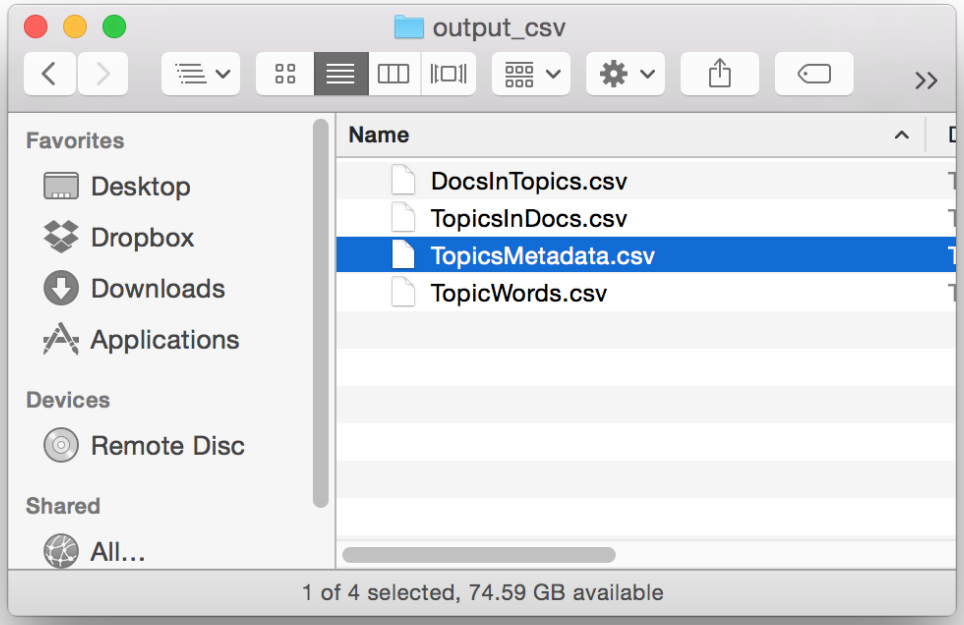
(37%) mr good thomas cassy master colonel thing moore overseer told carleton till plantation thornton began thought moment m set girl servant archy business matter boy presently deal stul hope great woman born husband love feelings money broug labor ready give meadow tone ritty indulgence chance favor insolence mine struck court fellows succeeded american finc beauty crew scarcely served ran intended son boston vessel swamp charleston mere smile caught ll white fields eliza hall misery york proctor torture humanity fever william bed happir

Spend a bit of time browsing through the files to get a feel for the topics that appear in the corpus. Do you see any patterns in the kinds of texts that appear under particular topics? Pick out a couple of topics that you think you might like to explore further.

Now open the “output_csv” file.



Inside, you'll see four CSV files. Open the one named "TopicsMetadata.csv":



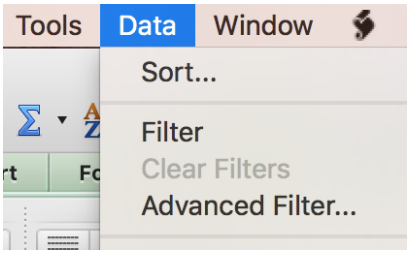

If you have Excel installed, it will load in Excel. Otherwise, you may need to open it using Google Sheets. If you're not sure how to do that, let me know, and I'll help you out.

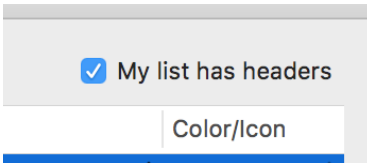
Once you have it open, you should see rows of output data like this:

| | docId | filename | Filename | Author | Title |
|---|---|---|---|---|---|
| 2 | 0 | church-hatch | church-hatch | William E. Ha | John Jasper: |
| 3 | 1 | fpn-ball-ball. | fpn-ball-ball. | Charles Ball | Fifty Years ir |
| 4 | 2 | fpn-brownw- | fpn-brownw- | William Well | Narrative of |
| 5 | 3 | fpn-bruce-br | fpn-bruce-br | Henry Clay B | The New Ma |
| 6 | 4 | fpn-burton-b | fpn-burton-b | Annie L. Burt | Memories o |
| 7 | 5 | fpn-burtont- | fpn-burtont- | Thomas Willi | What Experi |
| 8 | 6 | fpn-ferebee- | fpn-ferebee- | L. R. Ferebee | A Brief Histo |
| 9 | 7 | fpn-grandy-g | fpn-grandy-g | Moses Granc | Narrative of |
| 10 | 8 | fpn-hortonlif | fpn-hortonlif | George Mose | Life of Georg |
| 11 | 9 | fpn-hortonpc | fpn-hortonpc | George Mose | The Poetical |
| 12 | 10 | fpn-hughes-l | fpn-hughes-l | Louis Hughes | Thirty Years |
| 13 | 11 | fpn-jackson-j | fpn-jackson-j | John Andrew | The Experier |

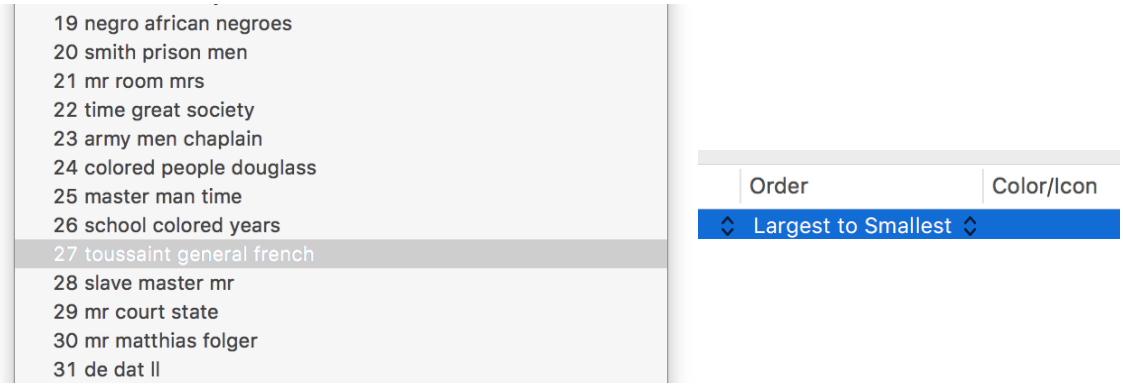
**Exploring data with Excel**

Once you've loaded the topic output, look through the topics and find ones that you think are interesting. Now you can start looking for patterns in the data. First, let's try simply sorting the rows by one of the topic columns. Under the data menu, select "Sort." Check the box in the




upper right corner of the sort window that says "My list has headers."




Select a topic column. Then sort the rows from largest to smallest.

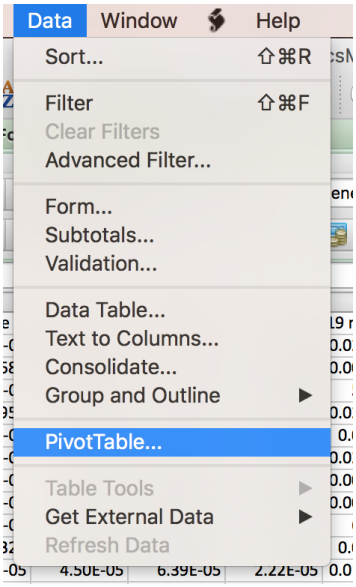
Once the data is sorted, the top rows will contain the data for texts that have the highest proportion of tokens labeled with the topic you picked. Scan through the metadata.

Do you see any patterns? Were most of the texts published in the northern or southern US? Are they clustered around a particular time period? Play with this for a while and see what emerges.

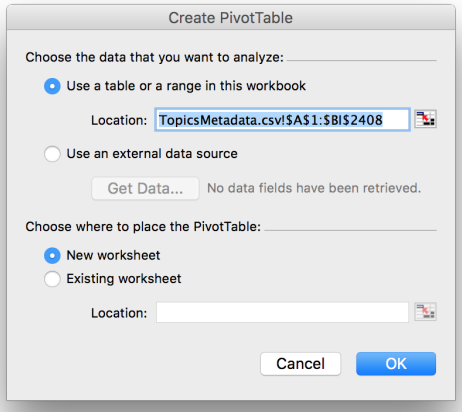
**Creating a Pivot Table**

To do even more complex transformations on a table like this, Excel provides a tool called a pivot table. In addition to allowing you to sort values in more complex ways, pivot tables allow you to group data together and view aggregate values like sums, averages, and counts within groups. For example, suppose you wanted to see how the “court, case, law” topic is distributed across geographic regions. You could use a pivot table to group the rows by publication location and take the average value for that topic across all groups. You could then compare that distribution to the geographic distribution of the “school, colored, years” topic, and you could use the resulting data to build a chart.

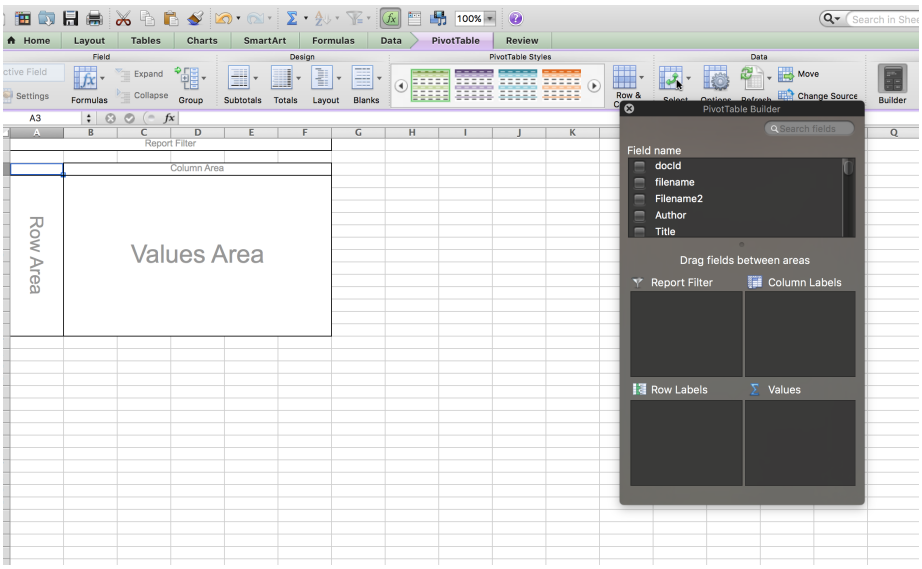
To create a pivot table, first select the PivotTable menu item, which is under the Data menu.




A window will pop up with a few options. You can ignore them all for now, and just click OK.




Excel will then create a pivot table in a new sheet




**Getting Oriented**

Take a moment to look at the table on the left side of the screen. You'll see three areas labeled "Row Area," "Values Area," and "Column Area."

Now look at the dark gray window on the right side of the screen titled "PivotTable Builder." You'll notice that it too has boxes for Rows, Values, and Columns ("Row Labels," "Values," and "Column Labels"). It also has a "Report Filter" box, which you can ignore, and a "Field name" box, which contains a checkbox for every column in the main spreadsheet.

There are three steps to building the table:

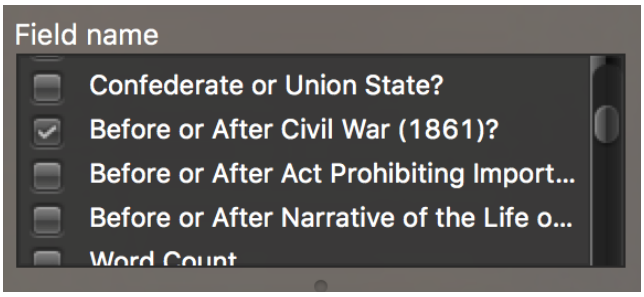
1) Select the fields (i.e. columns from the main spreadsheet) you want to examine.
2) Assign fields to row, column, and value areas in the pivot table.
3) Select the way you want to handle groups of values. (Do you want to sum them? Take their average? Calculate their standard deviation?)

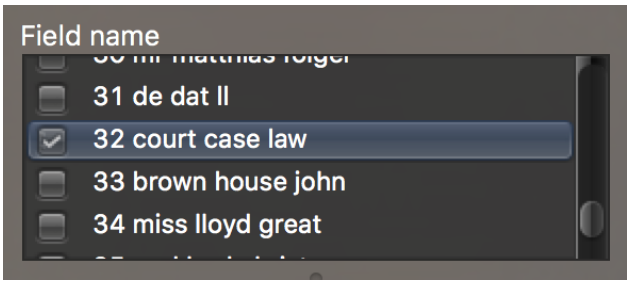
By changing the way you assign fields to those three areas, you can create a huge range of different kinds of tables.

**Creating the Table**

Let's begin by creating something simple -- a table of average values for the topic "court, case, law" in texts published before and after the Civil War.

First, select the "Before or After Civil War?" field and the "?? court case law" field. (I'm using ?? instead of a topic number because the topic numbers will differ between runs!)




Your initial table will look like this. Not very useful.

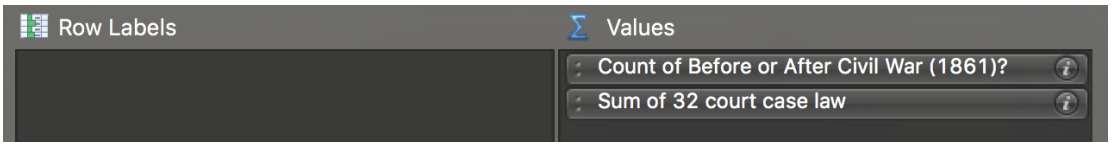
| | Values | |
|---|---|---|
| | Count of Before or After Civil War (1861)? | Sum of 32 court case law |
| **Total** | **2407** | **28.92216784** |

Right now, our pivot table is using default settings. In this case, that means that it's just counting the number of separate values in the first column ("Before or After…"), and summing the values for the second ("court case law").

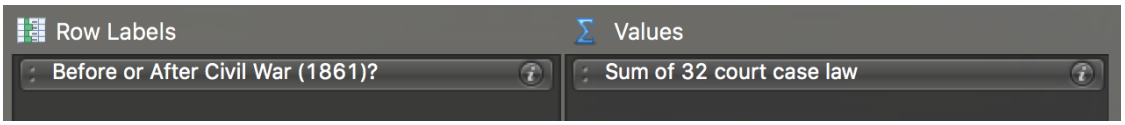
To do something more useful, let's assign the ("Before or After…") field to the Row area. That way, we'll see a row for every unique value for that field. In this case, the values are just "Before," "After," or "Unknown," so there will be just three rows after we're done.

To make the assignment, simply drag the "Before or After…" field in the gray "PivotTable Builder" window from the "Values" box to the "Row Labels" box. (You might want to enlarge the gray window so that you can see the full field names -- you can just resize it in the usual way.)

The window will look something like this at first:




And when you're done it will look like this:




You'll also see that the table has changed!

| Sum of 32 court case law | |
|---|---|
| Row Labels | Total |
| After | 13.56913171 |
| Before | 15.25641901 |
| Unknown | 0.096617112 |
| **Grand Total** | **28.92216784** |

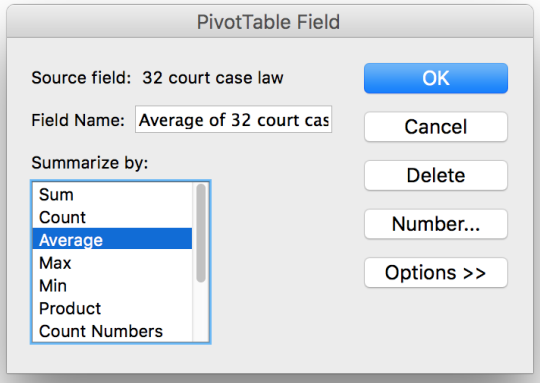
Now the table is telling us something a bit more useful: it's telling us the sum of the values for the "court case law" topic for texts published before and after the Civil War. (If it bugs you that "After" is on top, you can change the sort order by clicking on that little down arrow button on the right side of the "Row Labels" cell.)

We still have a problem though: we don't really want a sum. Each of the topic values being summed over is a proportion, represented as a decimal fraction between 0 and 1. In theory, the sum could tell us something meaningful if an equal number of texts were in each category. But that's not the case. Right now, the "Before" value is larger than the "After" value, but that might just mean that many more of the texts were published before the Civil War.

So we need to find a way to correct for that difference, and a good way to do that is to take the average instead of the sum. That way, we're getting the average proportion of each text devoted to this topic in each category, regardless of how many texts are in the categories.
To make this change, look for a little italicized "*i*" icon on the right side of the "Before or After…"


field in the "PivotTable Builder." It's a little bit hidden.

You'll see a window like this. Select "Average" and click "OK."




Problem solved!

| Average of 32 court case law | |
|---|---|
| Row Labels | Total |
| Unknown | 0.004830856 |
| Before | 0.013831749 |
| After | 0.01056786 |
| **Grand Total** | **0.012015857** |

Here, I've resorted the values so that "Before" is above "After." It turns out that this topic really is more prominent before the Civil War than after -- in this corpus at least!

**A Few Exercises**

Pivot tables are extremely powerful, and it's worth playing around with them for a while to get a feel for how they work. Here's a set of exercises to try:

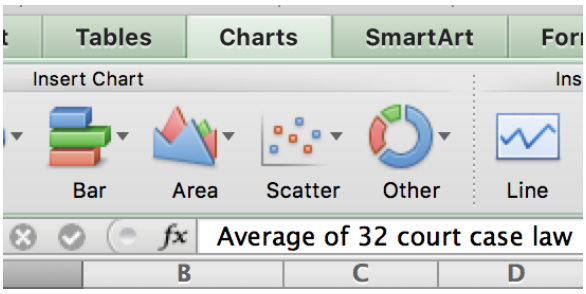
1) Add a new topic to the table by selecting it in the "Field name" box. Leave it in the "Values" box once you've selected it, remembering to change take the average instead of the sum of the values for that topic. What happens? How might this help you develop a hypothesis or a historical argument?
2) Remove the topic by unselecting it in the "Field name" box. In its place, add a second metadata field to the table. (Try "Confederate or Union State?") Is the result useful?
3) Now drag the second metadata field into the "Column Labels" box. What happens? Based on the structure of the table, what do you think the values in the cells correspond to now? What pattern do you notice in the values?

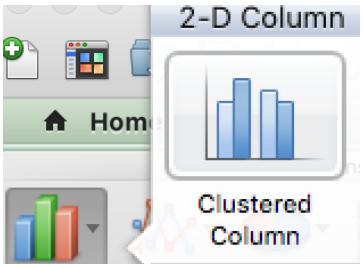
These are just a few of the different configurations you can create with a pivot table -- so keep experimenting!

**Making a Chart**

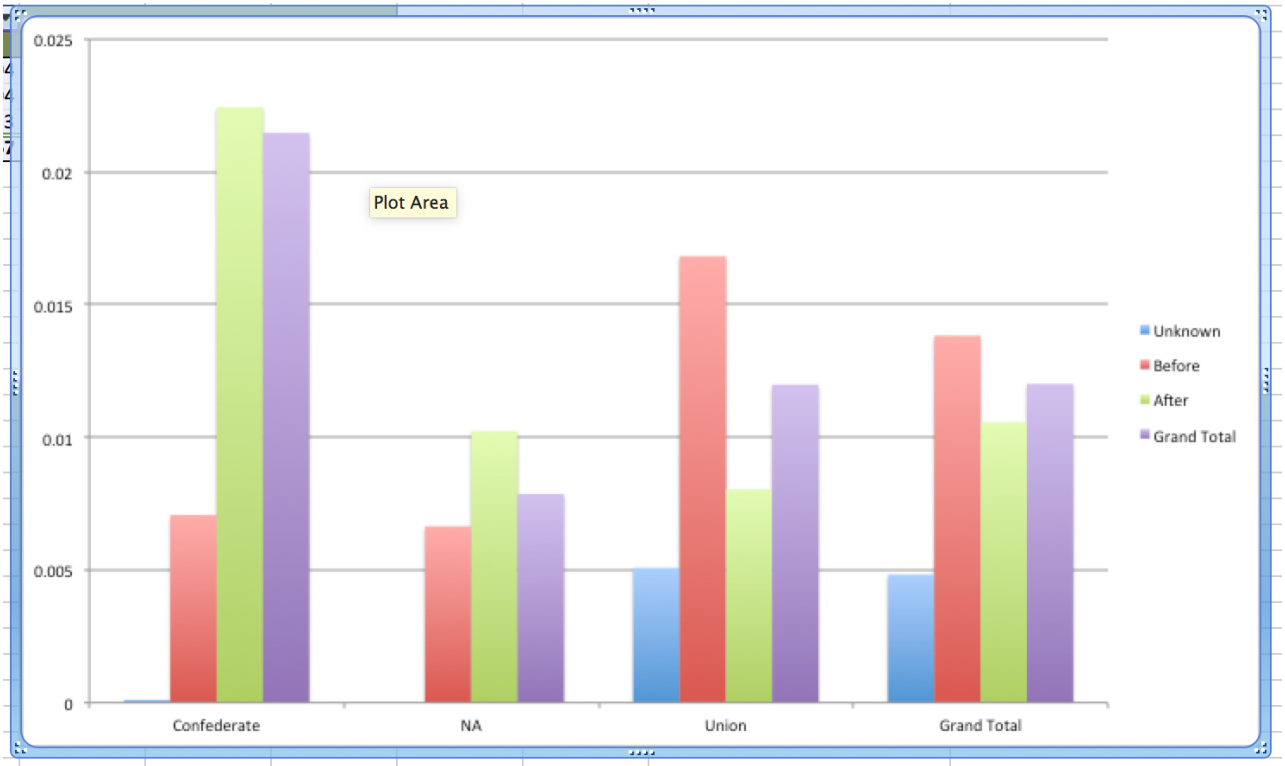
Once you've found an interesting configuration for your pivot table, you might want to visualize the data in a new way. One simple approach would be to create a chart. To start with, close the gray "PivotTable Builder" window. (You can always restore it by clicking the "Builder" button on the right side of the menu bar.) This will give us more screen real estate for the chart.

Select the "Charts" tab. Under the "Column" menu, try the "Clustered Column" option.

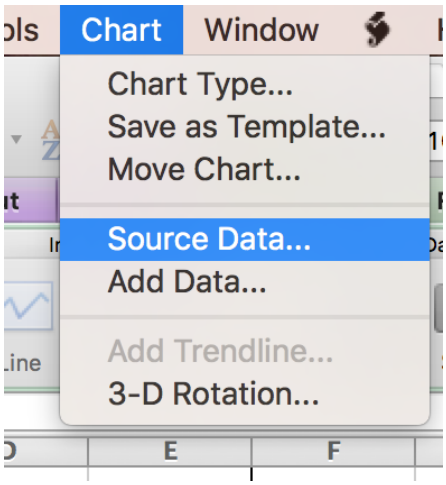


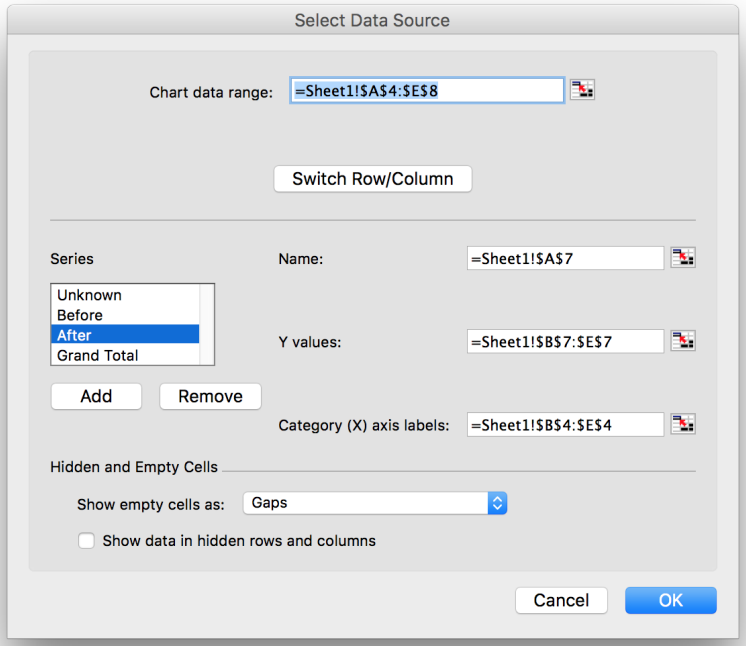



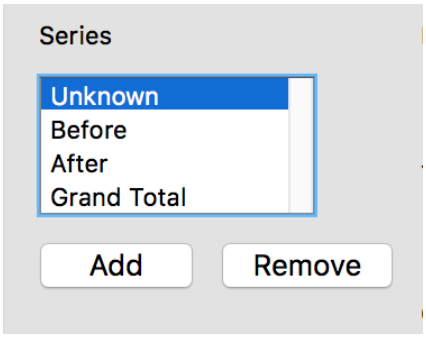

You should get something like this:




Unfortunately, there's a bit of visual noise here -- the "Grand Total" values aren't very helpful, nor are the "Unknown" values. To fix that, we can unselect that data. Under the Chart menu, select "Source Data…"

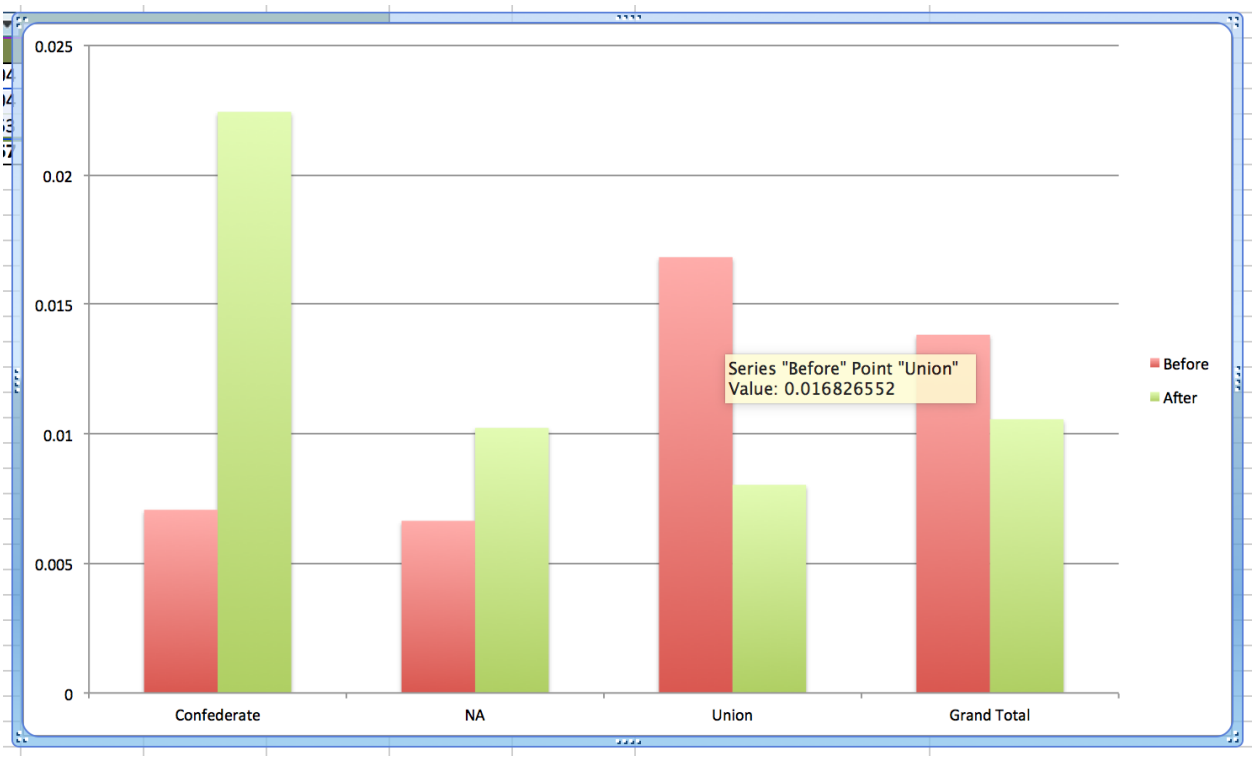



Now look for the “Series” box. Highlight the “Unknown” value and click the “Remove” button. Do the same for “Grand Total.”

Now you should have something that looks like this:




If you've been following along with the same topic ("court case law") or a similar one, you might see the same pattern. What do you make of it?