Tools Guide

By Devin Cornell (dcornell@ucsb.edu)

This guide is to help you navigate the two different kinds of tools available to you for your project: topic modeling and sentiment analysis. The analysis results are similar in format, but they require very different interpretations. Be sure to carefully consider the questions you'd like to answer before you start diving into the tools. Let the questions guide the tools as much as possible.

This document assumes that you've already collected your corpus, sent to me, and received the resulting analysis files. For example, I'll use the Betsy DeVos corpus that I created in the Corpus Preparation document.

If you have any questions about the doc, feel free to email me at dcornell@ucsb.edu.

Introduction

In this guide, I'm going to show you how to do specific analyses using the spreadsheet results from topic modeling and sentiment analysis. Using the analysis, we'll answer the following questions:

- What general topics appear in the documents?
- What kinds of discourse do the generated topics capture?
- How much of each topic appears in the corpus?
- How do topic quantities change over time?
- What kinds of sentiments appear in the documents?
- What kinds of sentiments are related to each topic?

To demonstrate the use of these tools, I'll use ~70 documents collected from Lexis Nexus from the New York Times and Daily News. We can compare the methods of analysis across sources and time.

Topic Modeling

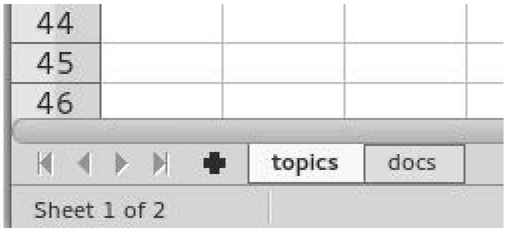
What is topic modeling?

Topic modeling is a way to take a corpus of text data and automatically extract a selected number of topics from them. This can help researchers sort through the data or perform quantitative analysis on to answer specific questions.

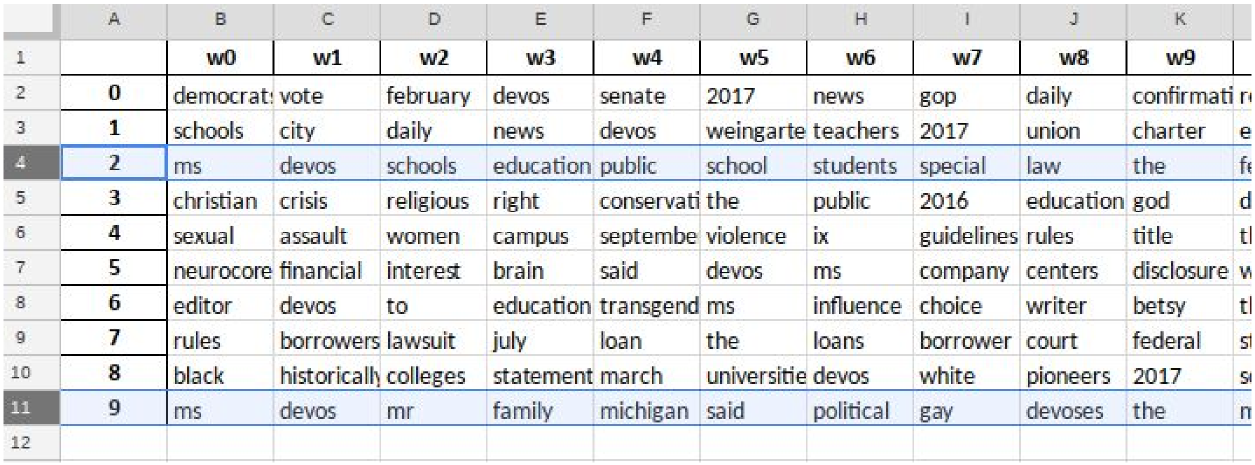
Read more about it to see how it works and what it can tell you:
https://en.wikipedia.org/wiki/Topic_model
https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/

Results

The result of our topic model is a spreadsheet (xlsx file). The spreadsheet contains two 'sheets': one for topic content (topics), and one for document content in terms of topics (docs). Switch sheets using the tabs at the bottom-left of your spreadsheet program (excel, LibreOffice, etc).



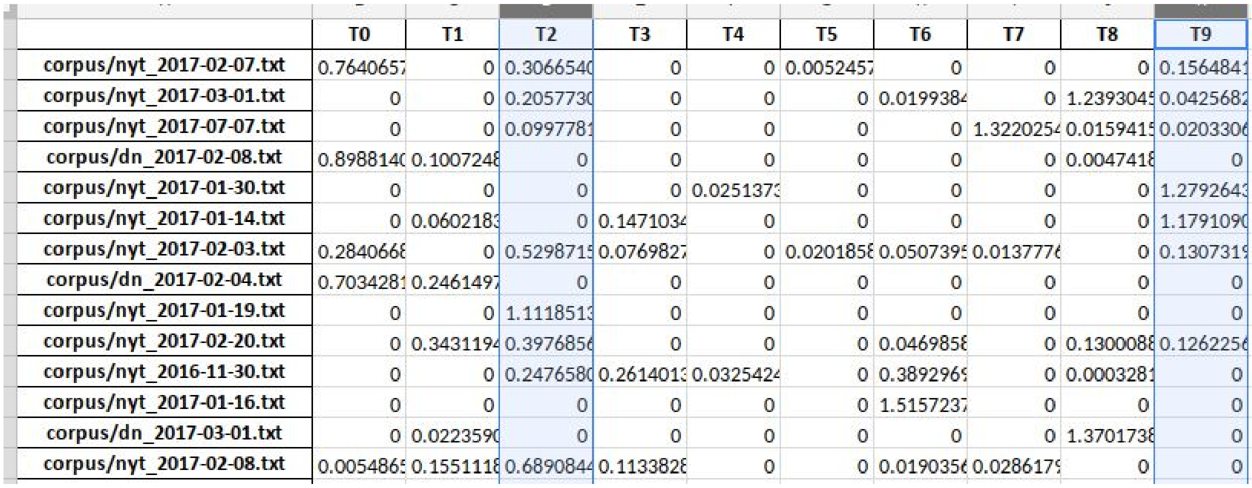
This is an example of the topic ‘sheet’ from an NMF run with T = 10. Each row is a topic, and the words are listed in order of how much they ‘belong to the topic’. The ordering of the topics (rows) is only based on how often they appear in the documents (more frequent topics at top of list); that is, it says nothing of any quality of the topics themselves.



	A	B	C	D	E	F	G	H	I	J	K	
1		w0	w1	w2	w3	w4	w5	w6	w7	w8	w9	
2	0	democrat:	vote	february	devos	senate	2017	news	gop	daily	confirmati	r
3	1	schools	city	daily	news	devos	weingarte	teachers	2017	union	charter	e
4	2	ms	devos	schools	education	public	school	students	special	law	the	fe
5	3	christian	crisis	religious	right	conservati	the	public	2016	education	god	d
6	4	sexual	assault	women	campus	septembe	violence	ix	guidelines	rules	title	tl
7	5	neurocore	financial	interest	brain	said	devos	ms	company	centers	disclosure	w
8	6	editor	devos	to	education	transgend	ms	influence	choice	writer	betsy	tl
9	7	rules	borrowers	lawsuit	july	loan	the	loans	borrower	court	federal	st
10	8	black	historically	colleges	statement	march	universitie	devos	white	pioneers	2017	se
11	9	ms	devos	mr	family	michigan	said	political	gay	devoses	the	m
12												

Making sense of the meanings embedded in topics requires a qualitative interpretation. In the example shown above, I’ve highlighted two topics for comparison. Based on word content alone, it appears that topic 2 is indicative of DeVos’ policy work in 2017 and topic 9 is about DeVos’ history working in Michigan. Because topic models are based purely on empirical data, they don’t necessarily capture the exact streams of discourse and dialog that social scientists are interested in. As such, it is important to carefully analyze the content of the topics in *relation* to the documents that they appear in. The ‘docs’ sheet may help with some of that analysis.

The ‘docs’ sheet contains a table that shows how much of each topic is present in each document. Each row is a document, and each column is a topic corresponding to the numbers in the ‘topics’ sheet. We can see that the first document (row) in the table contains 0.3 of topic 2 (ms, devos, schools, education, public) and 0.16 of topic 9 (ms, devos, mr, family, michigan).



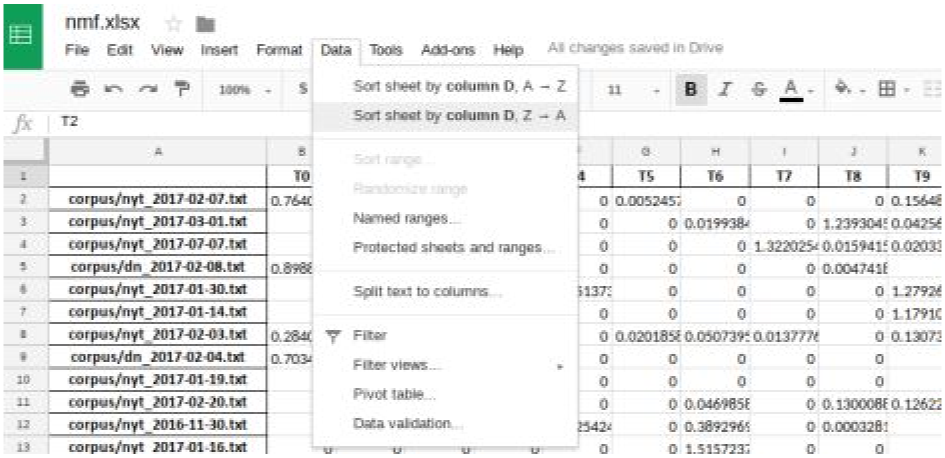
	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9
corpus/nyt_2017-02-07.txt	0.7640657	0	0.3066540	0	0	0.0052457	0	0	0	0.1564841
corpus/nyt_2017-03-01.txt	0	0	0.2057730	0	0	0	0.0199384	0	1.2393045	0.0425682
corpus/nyt_2017-07-07.txt	0	0	0.0997781	0	0	0	0	1.3220254	0.0159415	0.0203306
corpus/dn_2017-02-08.txt	0.8988140	0.1007248	0	0	0	0	0	0	0.0047418	0
corpus/nyt_2017-01-30.txt	0	0	0	0	0.0251373	0	0	0	0	1.2792643
corpus/nyt_2017-01-14.txt	0	0.0602183	0	0.1471034	0	0	0	0	0	1.1791090
corpus/nyt_2017-02-03.txt	0.2840668	0	0.5298715	0.0769827	0	0.0201858	0.0507395	0.0137776	0	0.1307319
corpus/dn_2017-02-04.txt	0.7034281	0.2461497	0	0	0	0	0	0	0	0
corpus/nyt_2017-01-19.txt	0	0	1.1118513	0	0	0	0	0	0	0
corpus/nyt_2017-02-20.txt	0	0.3431194	0.3976856	0	0	0	0.0469858	0	0.1300088	0.1262256
corpus/nyt_2016-11-30.txt	0	0	0.2476580	0.2614013	0.0325424	0	0.3892969	0	0.0003281	0
corpus/nyt_2017-01-16.txt	0	0	0	0	0	0	1.5157237	0	0	0
corpus/dn_2017-03-01.txt	0	0.0223590	0	0	0	0	0	0	1.3701738	0
corpus/nyt_2017-02-08.txt	0.0054865	0.1551118	0.6890844	0.1133828	0	0	0.0190356	0.0286179	0	0

This document-topic content can then be measured qualitatively or quantitatively.

Analysis

Topic Interpretation

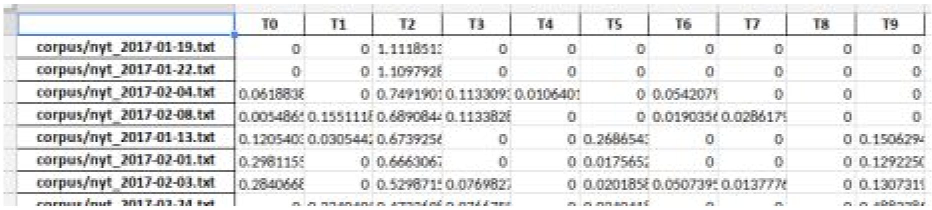
First, every topic modeling analysis should start with a ‘thick description’ or close reading of the topics. It is always important to go ‘back to the text’ with this work - analyzing the words associated with the topics is not enough. That said, the topic content table can help you sort through the data more quickly. In google sheets, you can sort by a column by going to Data->Sort (choose z->a to place the largest first).



fx T2

	A	B	F	G	H	I	J	K
1		T0	4	T5	T6	T7	T8	T9
2	corpus/nyt_2017-02-07.txt	0.764(	0	0.0052457	0	0	0	0.1564€
3	corpus/nyt_2017-03-01.txt		0	0	0.0199384	0	1.2393045	0.04256
4	corpus/nyt_2017-07-07.txt		0	0	0	1.3220254	0.0159415	0.02033
5	corpus/dn_2017-02-08.txt	0.8988	0	0	0	0	0.0047418	
6	corpus/nyt_2017-01-30.txt		51373	0	0	0	0	1.27926
7	corpus/nyt_2017-01-14.txt		0	0	0	0	0	1.17910
8	corpus/nyt_2017-02-03.txt	0.2840	0	0.0201858	0.0507395	0.0137776	0	0.13073
9	corpus/dn_2017-02-04.txt	0.7034	0	0	0	0	0	
10	corpus/nyt_2017-01-19.txt		0	0	0	0	0	
11	corpus/nyt_2017-02-20.txt		0	0	0.0469858	0	0.1300088	0.12622
12	corpus/nyt_2016-11-30.txt		25424	0	0.3892969	0	0.0003281	
13	corpus/nyt_2017-01-16.txt		0	0	1.5157233	0	0	

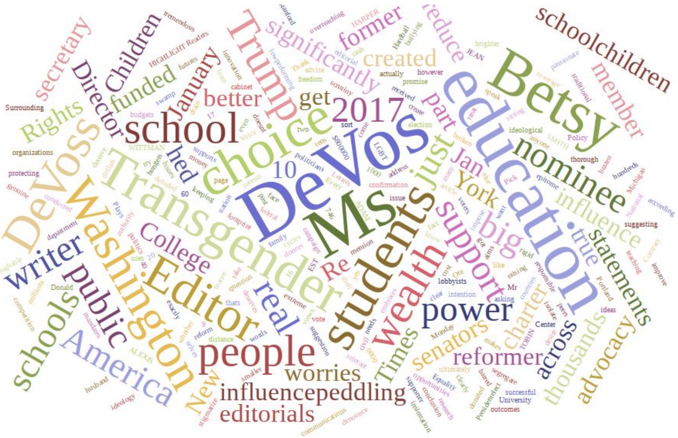
And after sorting we observe that the NYT articles seem to contain most of the content related to this topic -> interesting in itself! The document with the filename ‘nyt_2017-01-19.txt’ appears to be almost entirely composed of Topic 2. Open the filename and read the document to get an idea about what the topic is about.



	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9
corpus/nyt_2017-01-19.txt	0	0	1.1118513	0	0	0	0	0	0	0
corpus/nyt_2017-01-22.txt	0	0	1.1097928	0	0	0	0	0	0	0
corpus/nyt_2017-02-04.txt	0.0618838	0	0.7491901	0.1133093	0.0106401	0	0.0542079	0	0	0
corpus/nyt_2017-02-08.txt	0.0054865	0.1551118	0.6890844	0.1133828	0	0	0.0190356	0.0286179	0	0
corpus/nyt_2017-01-13.txt	0.1205403	0.0305443	0.6739256	0	0	0.2686543	0	0	0	0.1506294
corpus/nyt_2017-02-01.txt	0.2981155	0	0.6663067	0	0	0.0175652	0	0	0	0.1292250
corpus/nyt_2017-02-03.txt	0.2840668	0	0.5298715	0.0769827	0	0.0201858	0.0507395	0.0137776	0	0.1307311

I recommend examining the top 5 or 6 documents related to each topic. Be sure to carefully document the trends that you find in the documents related to each topic of interested. This detailed description will be the basis on which you interpret all of the quantitative analyses.

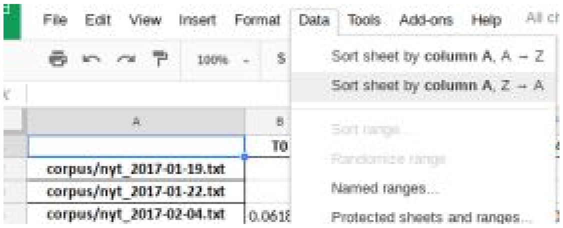
To even further strengthen the argument you make in your paper, you can include word clouds for a couple documents that capture topics appropriately. Because Topic Modeling is built only on the word frequencies that appear in the documents, this might be an appropriate way to understand how the computer reads the documents. To make a word cloud, you can copy-paste the raw text from your article into the text box on https://www.jasondavies.com/wordcloud/. The word cloud generated from the document most relevant to topic 2 might also just be a good visualization for the topic itself.



Quantitative

After you have established exactly the content of each topic and the discourse the topic is capturing, you can proceed to perform quantitative analysis in the spreadsheet. The quantitative approach of text analysis tools can be used for descriptive or hypothesis testing purposes. It is encouraged that you develop hypotheses or theories about what kinds of discourse are present and when before you begin the analysis.

The quantitative analysis can be done with excel's built-in features. It is easy (and there are many guides on the web) to sort by text filename (and thus date) or add/average the presence of a particular topic. Start by sorting by document name. This will organize the Betsy DeVos corpus first by source, then by date.



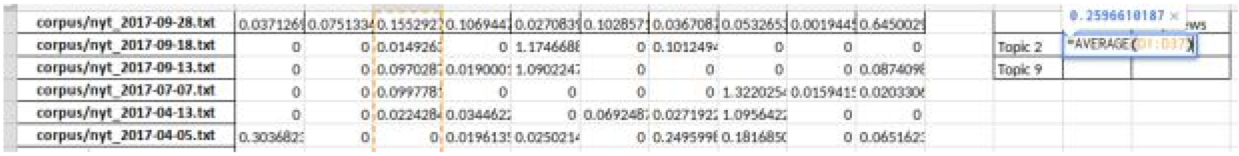
corpus/nyt_2017-01-19.txt
corpus/nyt_2017-01-22.txt
corpus/nyt_2017-02-04.txt 0.0618

Topic Presence

To analyze how much of each topic is present in each corpus, we first create a new table off to the side of the 'docs' sheet.

	NYT	Daily News
Topic 2		
Topic 9		

We want to populate it with the average presence of each topic in each sub-corpus (separated by source). To do this, you can use the average function in excel. While selecting the cell to populate, type “=AVERAGE(D1:D37)” to populate the cell with the average of cells between D1 and D37.



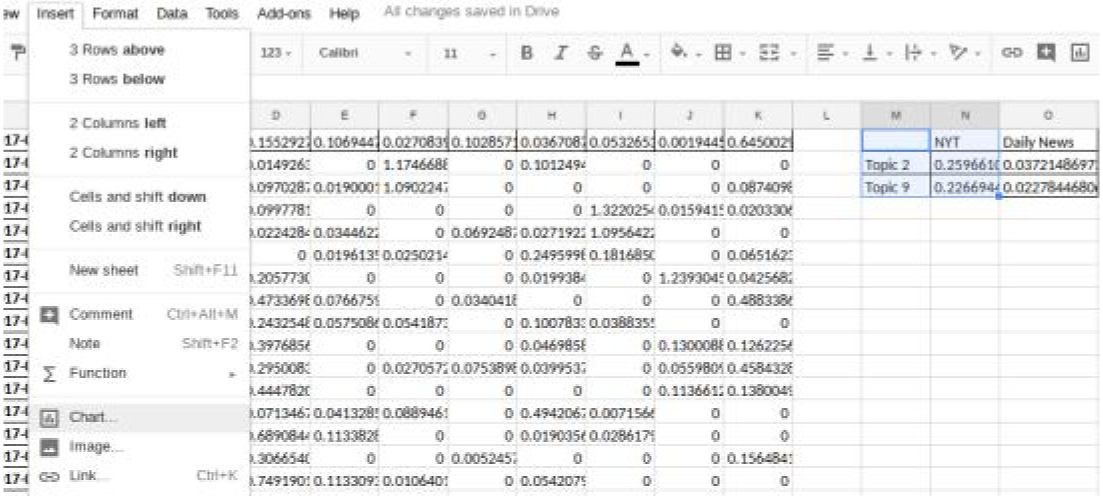
corpus/nyt_2017-09-28.txt	0.0371269	0.0751334	0.1552927	0.1069441	0.0270839	0.1028571	0.0367081	0.0532655	0.0019445	0.6450029
corpus/nyt_2017-09-18.txt	0	0	0.0149263	0	1.1746688	0	0.1012494	0	0	0
corpus/nyt_2017-09-13.txt	0	0	0.0970287	0.0190001	1.0902247	0	0	0	0	0.0874098
corpus/nyt_2017-07-07.txt	0	0	0.0997781	0	0	0	0	1.3220254	0.0159415	0.0203306
corpus/nyt_2017-04-13.txt	0	0	0.0224284	0.0344622	0	0.0692487	0.0271922	1.0956422	0	0
corpus/nyt_2017-04-05.txt	0.3036823	0	0	0.0196135	0.0250214	0	0.2495998	0.1816850	0	0.0651623

		ws
Topic 2	=AVERAGE(D1:D37)	
Topic 9		

Once you populate the table, it should look like this.

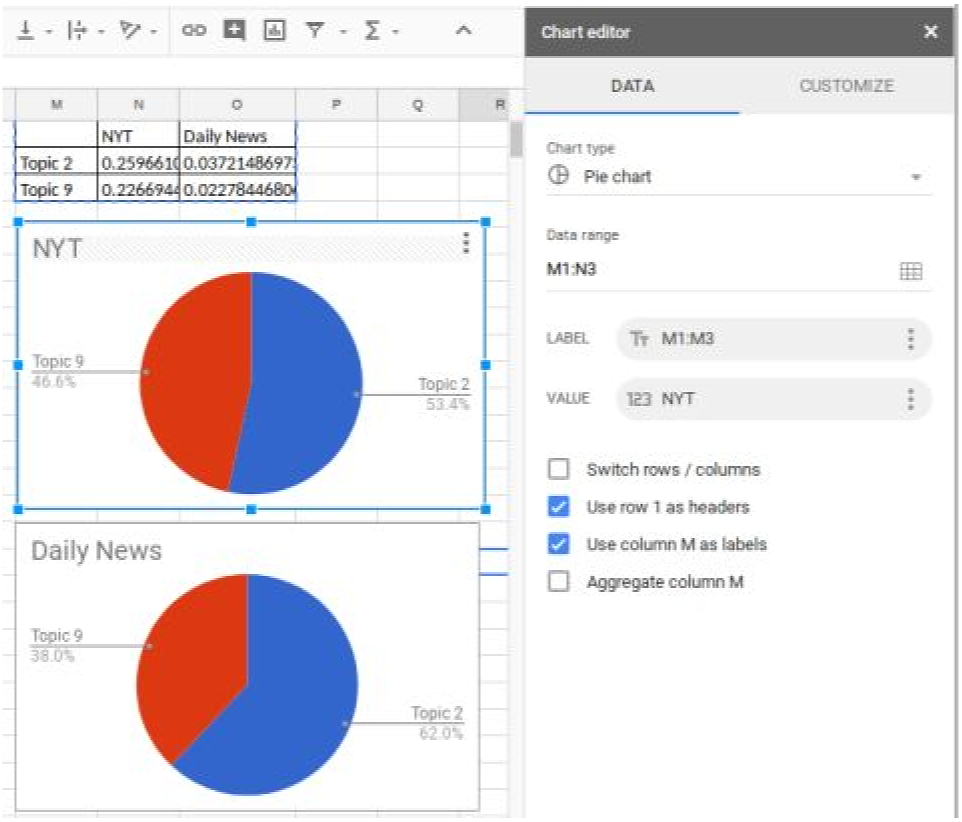
	NYT	Daily News
Topic 2	0.259	0.0372
Topic 9	0.226	0.0227

You can then generate a pie chart from this new table. In this case, I’ll compare the relative topic presence in each of the sources independently. Create the NYT chart by selecting the following.



	D	E	F	G	H	I	J	K	L	M	N	O
17-	.1552927	0.1069441	0.0270839	0.1028571	0.0367081	0.0532655	0.0019445	0.6450029			NYT	Daily News
17-	.0149263	0	1.1746688	0	0.1012494	0	0	0		Topic 2	0.2596610	0.0372148697
17-	.0970287	0.0190001	1.0902247	0	0	0	0	0.0874098		Topic 9	0.2266944	0.0227844680
17-	.0997781	0	0	0	0	1.3220254	0.0159415	0.0203306				
17-	.0224284	0.0344622	0	0.0692487	0.0271922	1.0956422	0	0				
17-	0	0.0196135	0.0250214	0	0.2495998	0.1816850	0	0.0651623				
17-	.2057730	0	0	0	0.0199384	0	1.2393045	0.0425682				
17-	.4733698	0.0766759	0	0.0340418	0	0	0	0.4883386				
17-	.2432548	0.0575086	0.0541873	0	0.1007833	0.0388355	0	0				
17-	.3976856	0	0	0	0.0469858	0	0.1300088	0.1262256				
17-	.2950083	0	0.0270571	0.0753898	0.0399537	0	0.0559809	0.4584328				
17-	.4447820	0	0	0	0	0	0.1136612	0.1380049				
17-	.0713467	0.0413285	0.0889461	0	0.4942067	0.0071566	0	0				
17-	.6890844	0.1133828	0	0	0.0190356	0.0286179	0	0				
17-	.3066540	0	0	0.0052457	0	0	0	0.1564841				
17-	.7491901	0.1133093	0.0106401	0	0.0542079	0	0	0				

Switch the chart type to ‘pie chart’, then repeat for the daily news.



M	N	O
	NYT	Daily News
Topic 2	0.2596610	0.0372148697
Topic 9	0.2266944	0.0227844680

Chart editor

DATA CUSTOMIZE

Chart type: Pie chart

Data range: M1:N3

LABEL: M1:M3

VALUE: NYT

☐ Switch rows / columns

☑ Use row 1 as headers

☑ Use column M as labels

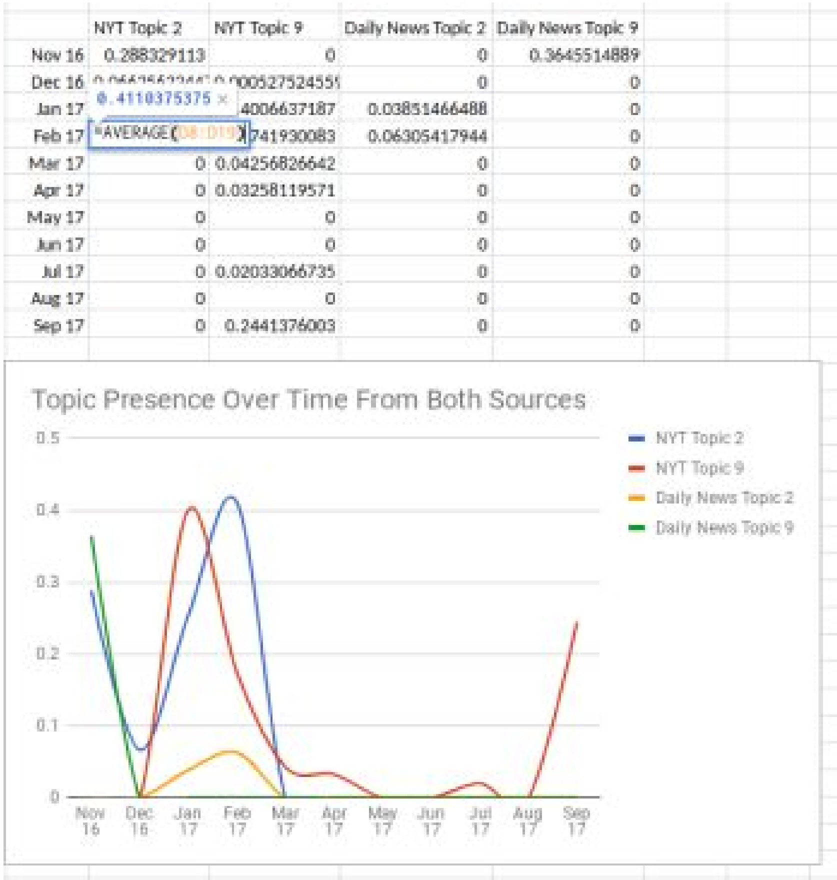
☐ Aggregate column M

From this we can conclude that the new york times talked about Topic 9 relatively more than the Daily News. Feel free to play around with these modes of analysis. Alternatively we could have analyzed how the topics were split among the sources. In this case, it wouldn't be that useful because these two topics belong mostly to NYT articles. Something to consider might happen in your own analysis.

A more advanced analysis would include groupings of topics which are related to a specific content or style of discourse. Think carefully about how you formulate your charts and how you can use it for analysis.

Topic Presence Timeline

To get a better idea of how the topic presence evolves over time, you'll need to create a new table with document averages for each time period. Here I combined manually-selected monthly time chunks into a single table with the two topics and two sources. Then I simply selected the new table, added a chart, and adjusted the settings until it showed time on the x-axis and content on the y-axis with the column labels as the legend.



	NYT Topic 2	NYT Topic 9	Daily News Topic 2	Daily News Topic 9
Nov 16	0.288329113	0	0	0.3645514889
Dec 16	0.06625622444	0.00052752455	0	0
Jan 17	0.4110375375 ×	4006637187	0.03851466488	0
Feb 17	=AVERAGE(D8:D19)	741930083	0.06305417944	0
Mar 17	0	0.04256826642	0	0
Apr 17	0	0.03258119571	0	0
May 17	0	0	0	0
Jun 17	0	0	0	0
Jul 17	0	0.02033066735	0	0
Aug 17	0	0	0	0
Sep 17	0	0.2441376003	0	0

Topic Presence Over Time From Both Sources

This chart shows that in the New York Times topic 9 seemed to spike first, followed by discussion about topic 2. The Daily News saw a spike of topic 9 early on and a small spike of topic 2 around February 2017. The temporal evolution of news data is an important aspect to capture if one wants to understand discourse around public figures. Be sure to think back to the topic descriptions you wrote to understand this evolution.

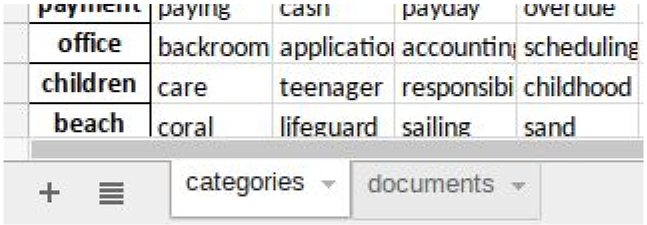
Sentiment Analysis

Sentiment analysis is a powerful way to understand the emotions being evoked in the documents under examination. The method used here is a simple word-bank. For instance, if we have a set of words that convey positive sentiment and a set of words that convey negative sentiment, we can calculate scores for both positive and negative by counting the number of appearances of words in either category in each document. Note that most psychology research on sentiment shows that positive and negative are two independent axes of salience. Texts can be both highly positive and negative, and the two don't cancel each other out.

Our method of sentiment analysis was introduced in this paper. In a sense, it is a set of manually pre-selected topics which are then run over your set of documents.

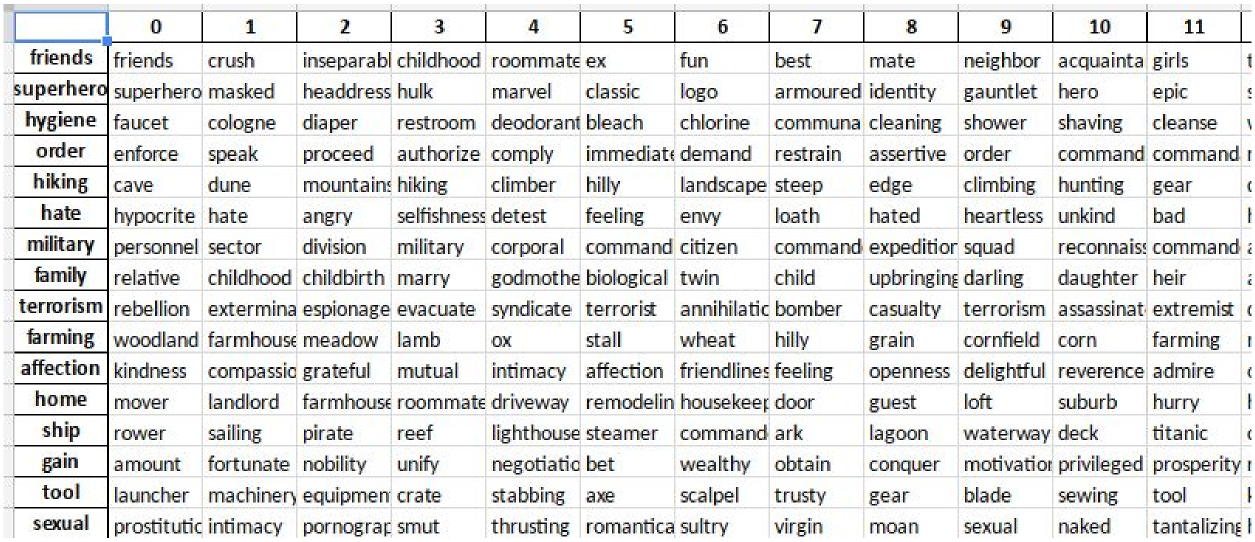
Results

Much like the topic models, the sentiment output also contains two sheets in the excel file.



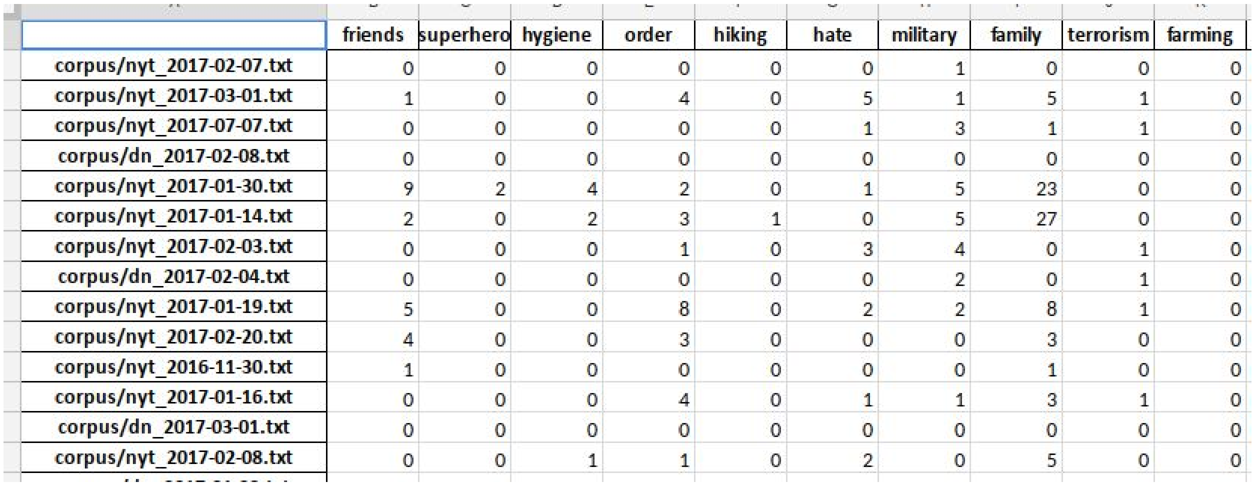
payment	paying	cash	payday	overdue
office	backroom	applicatio	accountin	scheduling
children	care	teenager	responsibi	childhood
beach	coral	lifeguard	sailing	sand

The first sheet is simply a list of the pre-defined categories - similar to topics, but with predefined meanings. You may be most interested in the categories “positive_emotion” and “negative_emotion”, depending on what you decide to do. As you can see, each row is a pre-defined category, and each non-ordered column is a word that belongs to that category.



	0	1	2	3	4	5	6	7	8	9	10	11
friends	friends	crush	inseparabl	childhood	roommate	ex	fun	best	mate	neighbor	acquainta	girls
superhero	superhero	masked	headdress	hulk	marvel	classic	logo	armoured	identity	gauntlet	hero	epic
hygiene	faucet	cologne	diaper	restroom	deodorant	bleach	chlorine	communa	cleaning	shower	shaving	cleanse
order	enforce	speak	proceed	authorize	comply	immediate	demand	restrain	assertive	order	command	command
hiking	cave	dune	mountains	hiking	climber	hilly	landscape	steep	edge	climbing	hunting	gear
hate	hypocrite	hate	angry	selfishness	detest	feeling	envy	loath	hated	heartless	unkind	bad
military	personnel	sector	division	military	corporal	command	citizen	command	expeditior	squad	reconnaiss	command
family	relative	childhood	childbirth	marry	godmothe	biological	twin	child	upbringing	darling	daughter	heir
terrorism	rebellion	extermina	espionage	evacuate	syndicate	terrorist	annihilatic	bomber	casualty	terrorism	assassinat	extremist
farming	woodland	farmhouse	meadow	lamb	ox	stall	wheat	hilly	grain	cornfield	corn	farming
affection	kindness	compassic	grateful	mutual	intimacy	affection	friendlines	feeling	openness	delightful	reverence	admire
home	mover	landlord	farmhouse	roommate	driveway	remodelin	housekeep	door	guest	loft	suburb	hurry
ship	rower	sailing	pirate	reef	lighthouse	steamer	command	ark	lagoon	waterway	deck	titanic
gain	amount	fortunate	nobility	unify	negotiatic	bet	wealthy	obtain	conquer	motivatior	privileged	prosperity
tool	launcher	machinery	equipmen	crate	stabbing	axe	scalpel	trusty	gear	blade	sewing	tool
sexual	prostitutic	intimacy	pornograp	smut	thrusting	romantica	sultry	virgin	moan	sexual	naked	tantalizing

The “documents” sheet relates these categories to documents in a fashion very similar to the topic models. In this sheet, each row is a document and each column is one of the pre-defined categories.



	friends	superhero	hygiene	order	hiking	hate	military	family	terrorism	farming
corpus/nyt_2017-02-07.txt	0	0	0	0	0	0	1	0	0	0
corpus/nyt_2017-03-01.txt	1	0	0	4	0	5	1	5	1	0
corpus/nyt_2017-07-07.txt	0	0	0	0	0	1	3	1	1	0
corpus/dn_2017-02-08.txt	0	0	0	0	0	0	0	0	0	0
corpus/nyt_2017-01-30.txt	9	2	4	2	0	1	5	23	0	0
corpus/nyt_2017-01-14.txt	2	0	2	3	1	0	5	27	0	0
corpus/nyt_2017-02-03.txt	0	0	0	1	0	3	4	0	1	0
corpus/dn_2017-02-04.txt	0	0	0	0	0	0	2	0	1	0
corpus/nyt_2017-01-19.txt	5	0	0	8	0	2	2	8	1	0
corpus/nyt_2017-02-20.txt	4	0	0	3	0	0	0	3	0	0
corpus/nyt_2016-11-30.txt	1	0	0	0	0	0	0	1	0	0
corpus/nyt_2017-01-16.txt	0	0	0	4	0	1	1	3	1	0
corpus/dn_2017-03-01.txt	0	0	0	0	0	0	0	0	0	0
corpus/nyt_2017-02-08.txt	0	0	1	1	0	2	0	5	0	0

Analysis

Because the topic modeling and sentiment spreadsheets are so similar, many of the analyses can be done on both. See methods in the topic modeling section to see which modes of analysis may be most important to your corpus.

Category Meaning

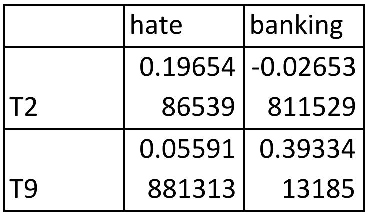
While the sentiment categories were selected based on some preconceptions of what they hope to find, you still need to perform the same 'close reading' that you used on the topic models to understand what each of the categories means in *your* specific corpus. Sort by word frequency and manually examine the documents which contain these types of sentiment. You can also use your word clouds to examine documents at a glance or convey document contents to the reader of your paper. Your paper should include a detailed analysis of any sentiment categories you hope to use based on the 'close readings'.

Sentiment Compositions and Timelines

Sentiment categories can be plotted both in terms of composition and composition over time by chunking documents based on month or year. See the analysis section in the topic modeling guide in this document for more information on that.

Sentiment-Topic Correlation

You can also combine results from both sentiment and topic modeling. We can compare how much of each topic is related to the sentiment categories. For example, here we show correlations between two sentiment categories, hate and banking, with topic 2 and 9. The following table shows that while topic two is weakly related to hate, it seems to be independent from baking. On the other hand, topic 9 seems to be unrelated to hate, while it is highly related to banking. From this we might learn something about the topic: topic 2 seems to carry more emotional weight than topic 9, while topic 9 seems to contain much more topical information about banking.



	hate	banking
T2	0.1965486539	-0.02653811529
T9	0.05591881313	0.3933413185

Use these tools to learn more about the topics and the content from a more sentiment-based perspective.

Conclusions

Always begin the analysis with some idea of what you are interested in finding in your corpus. Write a detailed description of each of the topics or sentiment categories of interest as they relate to *your* corpus, and then analyze it quantitatively to answer questions about frequency and changes over time. Finally, try to formulate a cohesive argument with your results; use the data to tell the story of your analysis.