A CASE STUDY

Trends in COVID-19 News Coverage:
An Introduction to Topic Models Using Text and Data Mining

Princeton University Researcher Gavin Cook explains how he used TDM Studio, ProQuest's text and data mining service, to analyze a corpus of 700,000 newspaper articles related to COVID-19 — and how creating topic models from that analysis can give us unprecedented insights on the areas of greatest interest.

proquest.com

To talk to the sales department, contact us at
1-800-779-0137 or **sales@proquest.com**.

ProQuest®

Gavin G. Cook is a PhD student in the Department of Sociology affiliated with the Paul and Marcia Wythes Center on Contemporary China. His research interests include the "science of science", cultural sociology with an emphasis on discipline and organization, and how the world views China. Gavin graduated from Princeton University in 2015 with a BA in East Asian Studies.

Introduction

At the time of this writing, the United States — and much if not most of the rest of the world — is in the throes of an outbreak of COVID-19. Schools and businesses have been shuttered, and offices across the country are empty as people obey shelter-in-place orders and work from home. COVID-19 has quickly become a decade-defining issue, and analyzing ProQuest's newspaper corpus with text and data mining (TDM) can help shed light on how the world has reacted to it.

There may be no more water coolers to talk around, but that certainly hasn't stopped people from pontificating on the virus. What makes the coronavirus so interesting from a social scientific perspective is that everybody has their own take on the reality of the disease and an additional opinion on what everybody else thinks. Humans are gossip machines, and we care about what the rest of our species thinks — especially people in positions of power.

A lot of opinion about opinion, however, is based on hearsay. Fortunately, we can use data to turn rumor into research. Because people develop their opinions in part from media, we can turn to newspapers to get a sense of what information people draw from when forming their own views of the world. Using TDM to analyze ProQuest's newspaper corpus allows scholars to generate high-level insights about how the media reports on key issues. We can supplement speculation with statistics to get a much better picture of what the media says and how they say it.

How the media has covered COVID-19 is an empirical question — one we can get a more accurate picture of with statistical techniques for natural language processing. The detail of human analysis is impossible to replicate with machines, but we can supplement human depth with machine breadth. TDM lets us process many thousands more document than any human could at once.

The Data

For this project, I analyzed a ProQuest dataset that includes nearly one million articles related to the coronavirus, spanning multiple countries and multiple languages. For now, we will focus on the roughly 70% of the articles in the corpus that are in English. This gives us a final sample of 782,662 articles.

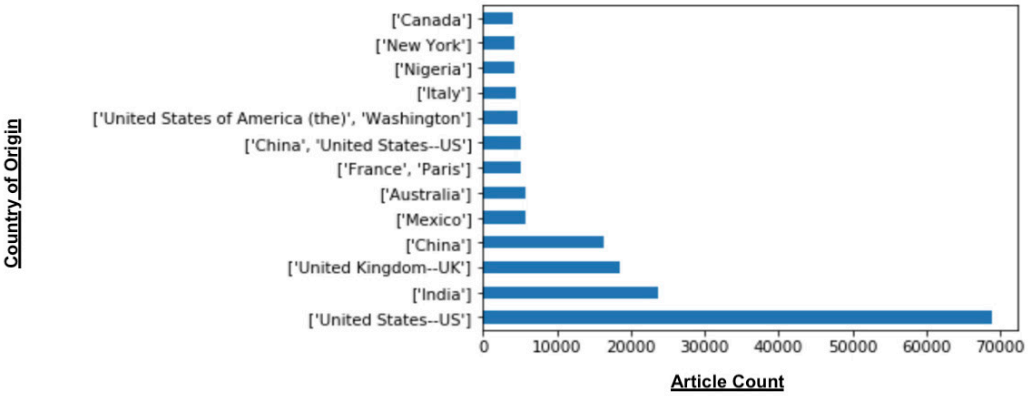


Figure 1: National Origin of Articles in the ProQuest Corpus. This figure shows the distribution of the country of publication of every English-language article in the ProQuest Coronavrus Corpu. The most common country of origin for articles in the corpus from outlets based in the United States. The second most common country of origin is India.

The corpus includes English-language articles from the United States, the UK and Australia, but, interestingly, there are more English-language articles in the corpus from India than the UK. The Times of India is one of the biggest and most important newspapers in the world, and it is English.

This corpus is, as we can see, enormous — far too big for us to read every article it contains, let alone for us to summarize the main themes of the articles succinctly or accurately. We can, however, train a machine to read and summarize the corpus for us. One of the most tried-and-true methods to do so is with a topic model.

The Process: Topic Models Made Simple

Topic models are one of many ways of summarizing information from a collection of texts. In basic terms, topic models posit that each document in a corpus represents a collection of topics, and that the general themes of all documents in a corpus can be summarized with topics. The output of a topic model is a set of topics, and the words associated with each one.

One of the earliest and most famous topic models is the latent Dirichlet allocation model made famous in a 2003 paper by David Blei, Andrew Ng, and Michael I. Jordan (Blei et al. 2003). The original paper, published in the Journal of Machine Learning Research, has more than 30,000 citations. The paper itself is very technical, as you can see in the example below.

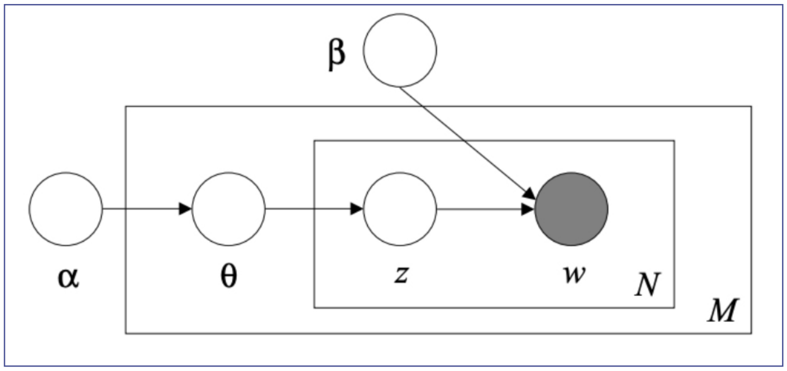
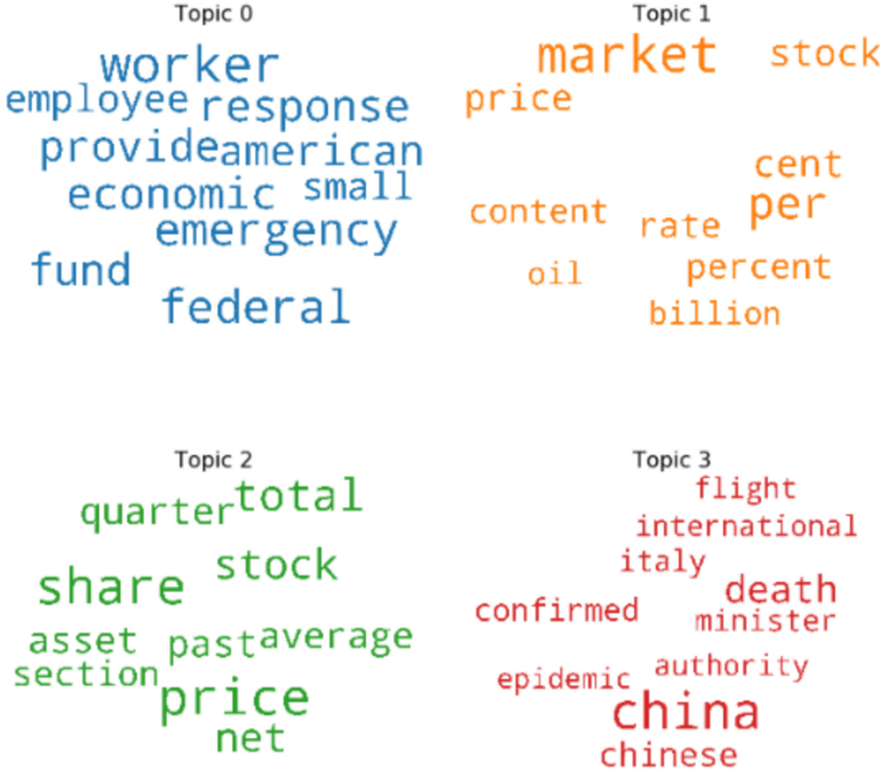


Figure 2: Graphical model representation of LDA. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeted choice of topics and words within a document. (Blei et al. 2003).

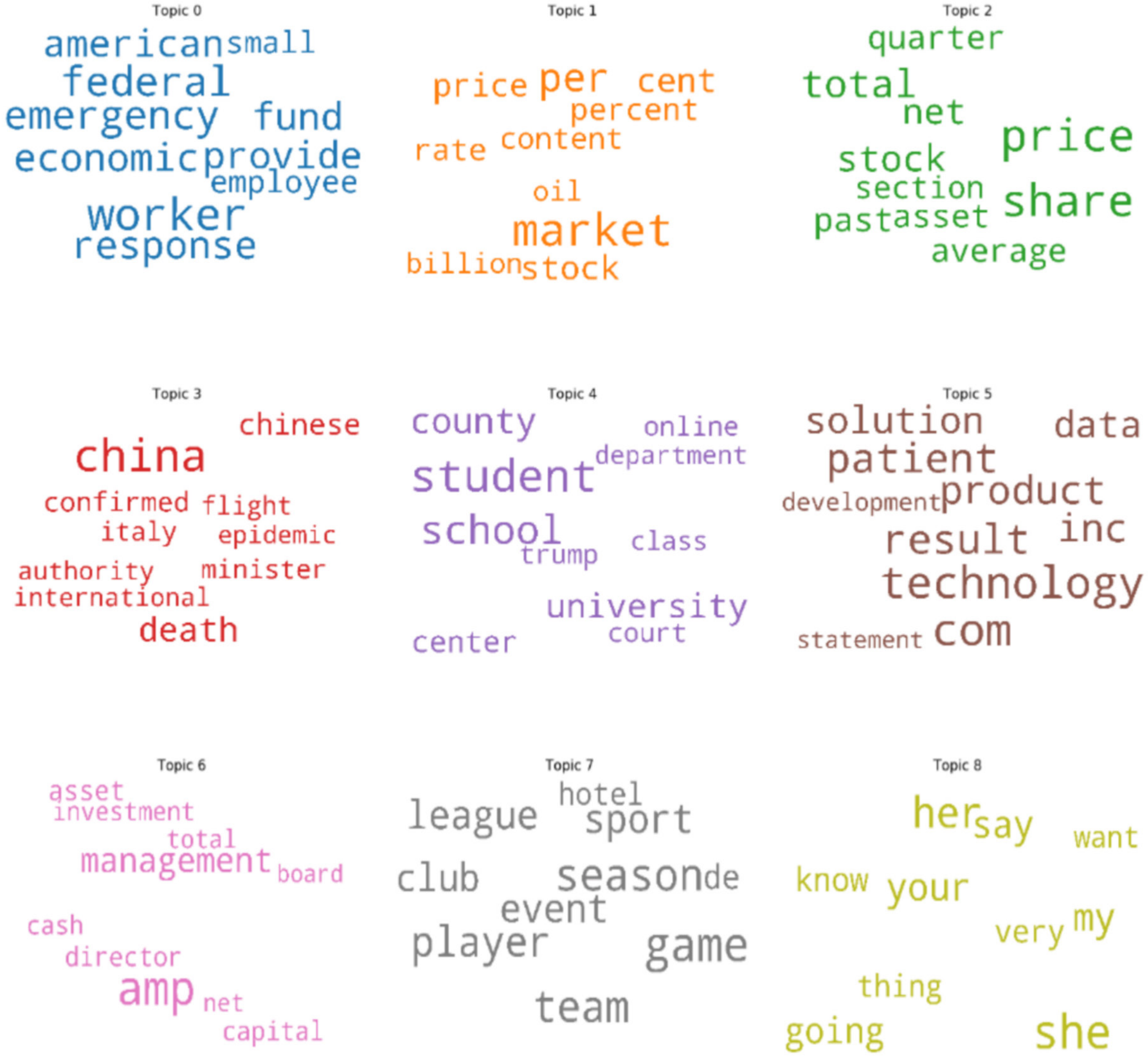
And this is not even one of the many equations in the paper. For a more in-depth but still accessible introduction to how the LDA algorithm works, see: http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/. ***[Note:** this blog post was recommended by the author of Gensim, a famous topic-modeling package in Python!]*

While running a topic model is straight-up science, interpreting topic models is not and requires both art and artifice. Brandon Stewart, a sociology professor at Princeton, argues that while some fear that natural language processing may render qualitative research useless, the true conceit of natural language processing is to *bring qualitative research back into quantitative research*.

To get a sense of how we might bring the qualitative into the quantitative, we can look at some word clouds from the topic model. This is not very helpful on its own, but we can use it to get a sense of how we might go about interpreting this model. We must read between the lines and intuit the higher-level meaning behind the keywords. This will let us assign a working label to each topic. The image below shows the word clouds I generated using TDM Studio and the ProQuest newspaper dataset.

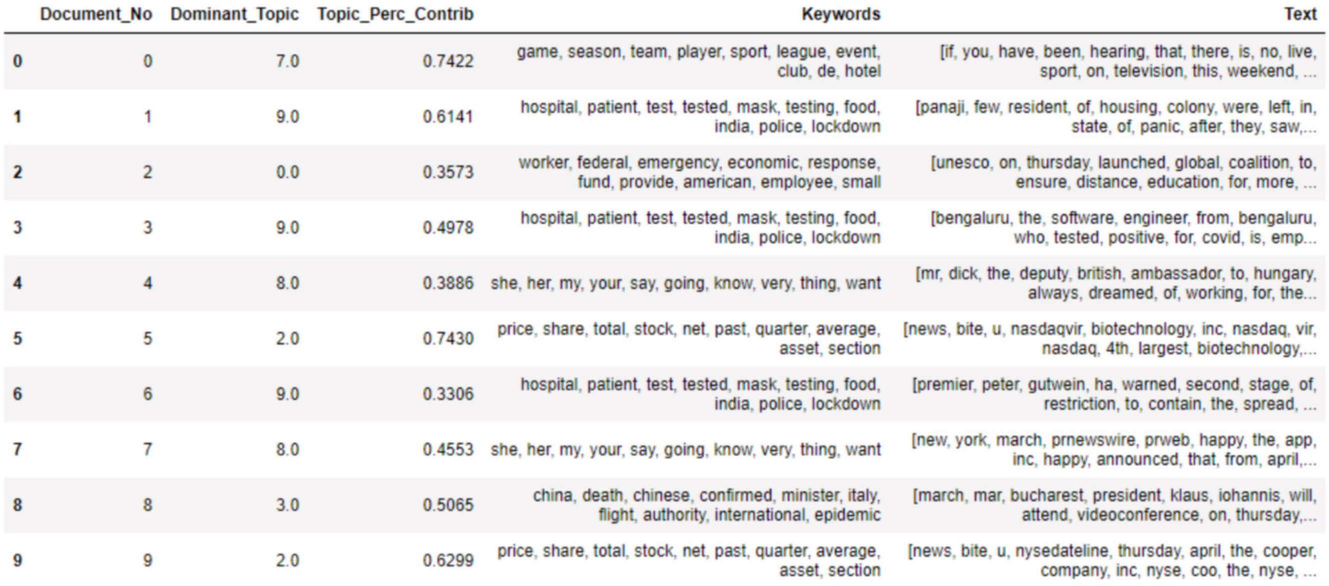


These are the top four topics. Let's see what happens when we look at a few more:



Do any words jump out at you? Do you see any themes that emerge from these topics?

For a deeper pass at the data, we can take a closer look at the first 10 documents in our corpus and see what topic they most strongly represent.



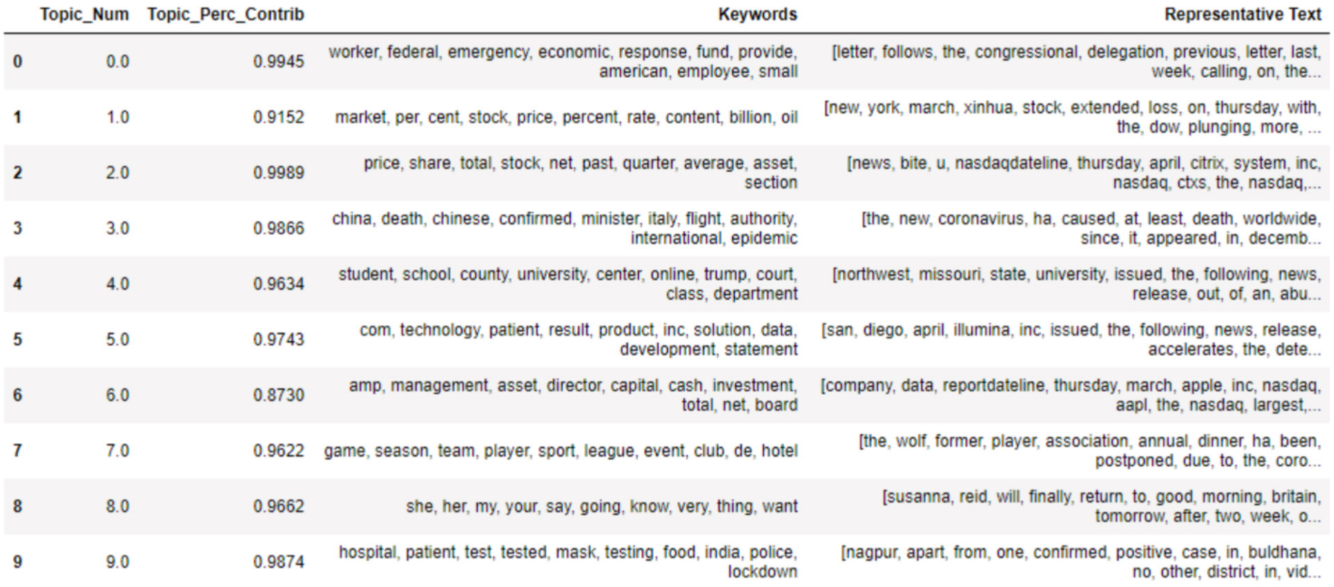
	Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Text
0	0	7.0	0.7422	game, season, team, player, sport, league, event, club, de, hotel	[if, you, have, been, hearing, that, there, is, no, live, sport, on, television, this, weekend, ...
1	1	9.0	0.6141	hospital, patient, test, tested, mask, testing, food, india, police, lockdown	[panaji, few, resident, of, housing, colony, were, left, in, state, of, panic, after, they, saw,...
2	2	0.0	0.3573	worker, federal, emergency, economic, response, fund, provide, american, employee, small	[unesco, on, thursday, launched, global, coalition, to, ensure, distance, education, for, more, ...
3	3	9.0	0.4978	hospital, patient, test, tested, mask, testing, food, india, police, lockdown	[bengaluru, the, software, engineer, from, bengaluru, who, tested, positive, for, covid, is, emp...
4	4	8.0	0.3886	she, her, my, your, say, going, know, very, thing, want	[mr, dick, the, deputy, british, ambassador, to, hungary, always, dreamed, of, working, for, the...
5	5	2.0	0.7430	price, share, total, stock, net, past, quarter, average, asset, section	[news, bite, u, nasdaqvir, biotechnology, inc, nasdaq, vir, nasdaq, 4th, largest, biotechnology,...
6	6	9.0	0.3306	hospital, patient, test, tested, mask, testing, food, india, police, lockdown	[premier, peter, gutwein, ha, warned, second, stage, of, restriction, to, contain, the, spread, ...
7	7	8.0	0.4553	she, her, my, your, say, going, know, very, thing, want	[new, york, march, prnewswire, prweb, happy, the, app, inc, happy, announced, that, from, april,...
8	8	3.0	0.5065	china, death, chinese, confirmed, minister, italy, flight, authority, international, epidemic	[march, mar, bucharest, president, klaus, iohannis, will, attend, videoconference, on, thursday,...
9	9	2.0	0.6299	price, share, total, stock, net, past, quarter, average, asset, section	[news, bite, u, nysedateline, thursday, april, the, cooper, company, inc, nyse, coo, the, nyse, ...

The columns are, in order:

Document_No: This is the number of the document in our corpus. Python starts counting from zero, so that's why document 1 is labeled as document zero.

- **Dominant_Topic:** This is the strongest topic in a given document. LDA models each document as a mixture of topics, so every document has more than one topic attached to it. This output just gives us the topic that is strongest in a given document.
- **Topic_Perc_Contrib:** This tells us how heavily the dominant topic in a given document contributes to that document. Because LDA models each document as a *mixture* of topics, each document can map to more than one topic.
- **Keywords:** These are the keywords associated with a given topic.
- **Text**: This is the beginning of the text. You may notice that some of the words are truncated. This is because they have been lemmatized as when the corpus was preprocessing for analysis. This makes it easier for algorithms to work on the text by rendering each word in a more succinct form. Some very common English words have also been removed.

This gives tells us how our documents map to topics. What if we want to know the reverse: how topics map to documents? We can look at documents that strongly represent a given a topic. This is a list of topics with the beginning of the single text in the corpus that most strongly represents that topic:



	Topic_Num	Topic_Perc_Contrib	Keywords	Representative Text
0	0.0	0.9945	worker, federal, emergency, economic, response, fund, provide, american, employee, small	[letter, follows, the, congressional, delegation, previous, letter, last, week, calling, on, the...
1	1.0	0.9152	market, per, cent, stock, price, percent, rate, content, billion, oil	[new, york, march, xinhua, stock, extended, loss, on, thursday, with, the, dow, plunging, more, ...
2	2.0	0.9989	price, share, total, stock, net, past, quarter, average, asset, section	[news, bite, u, nasdaqdateline, thursday, april, citrix, system, inc, nasdaq, ctxs, the, nasdaq,...
3	3.0	0.9866	china, death, chinese, confirmed, minister, italy, flight, authority, international, epidemic	[the, new, coronavirus, ha, caused, at, least, death, worldwide, since, it, appeared, in, decemb...
4	4.0	0.9634	student, school, county, university, center, online, trump, court, class, department	[northwest, missouri, state, university, issued, the, following, news, release, out, of, an, abu...
5	5.0	0.9743	com, technology, patient, result, product, inc, solution, data, development, statement	[san, diego, april, illumina, inc, issued, the, following, news, release, accelerates, the, dete...
6	6.0	0.8730	amp, management, asset, director, capital, cash, investment, total, net, board	[company, data, reportdateline, thursday, march, apple, inc, nasdaq, aapl, the, nasdaq, largest,...
7	7.0	0.9622	game, season, team, player, sport, league, event, club, de, hotel	[the, wolf, former, player, association, annual, dinner, ha, been, postponed, due, to, the, coro...
8	8.0	0.9662	she, her, my, your, say, going, know, very, thing, want	[susanna, reid, will, finally, return, to, good, morning, britain, tomorrow, after, two, week, o...
9	9.0	0.9874	hospital, patient, test, tested, mask, testing, food, india, police, lockdown	[nagpur, apart, from, one, confirmed, positive, case, in, buldhana, no, other, district, in, vid...

The columns are similar to the raw output above. You may notice that the "Topic_Perc_Contrib" values are all very high, and this is because each text is the most representative text for a given topic in the corpus. This is by design — those values better be close to 1 (a.k.a. 100%)!

Topic 0, which includes the keywords "worker," "employee," "small" and "economic," seems to generally cover words related to the economic plights of working-class Americans and small businesses. Topic 1 is about markets and commodities, as is Topic 6, and Topic 2 is similar but more specifically about financial markets. Topic 3 is interesting in that it covers both China and Italy. More generally, it seems to capture governments and their responses to the coronavirus in hard-hit areas outside the United States. Topic 4 is about schools and

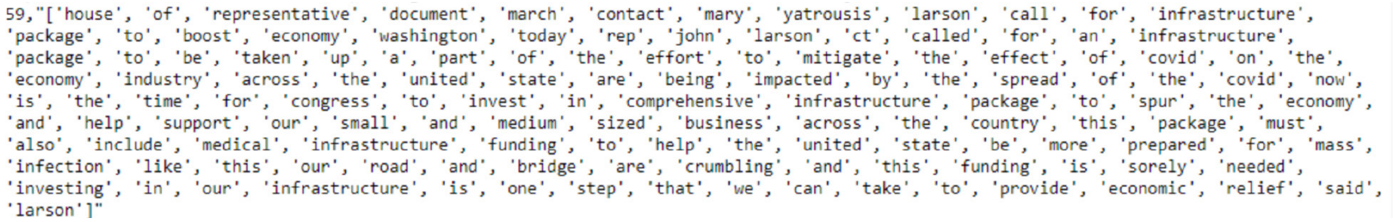
> *"COVID-19 has quickly become a decade-defining issue, and analyzing ProQuest's newspaper corpus with TDM can help shed light on how the world has reacted to it. "*

moving online, and Topic 5 relates to vaccine development. Topic 7 is about everything that has unfortunately been cancelled by the virus: sports, hotels, and more. We could label the first 10 topics:

1. American Workers
2. Markets and Commodities
3. Stock Markets
4. International Responses
5. Students and Schools
6. Solutions and Developments
7. Company Drama
8. Canceled Events
9. Common Words
10. India's Response

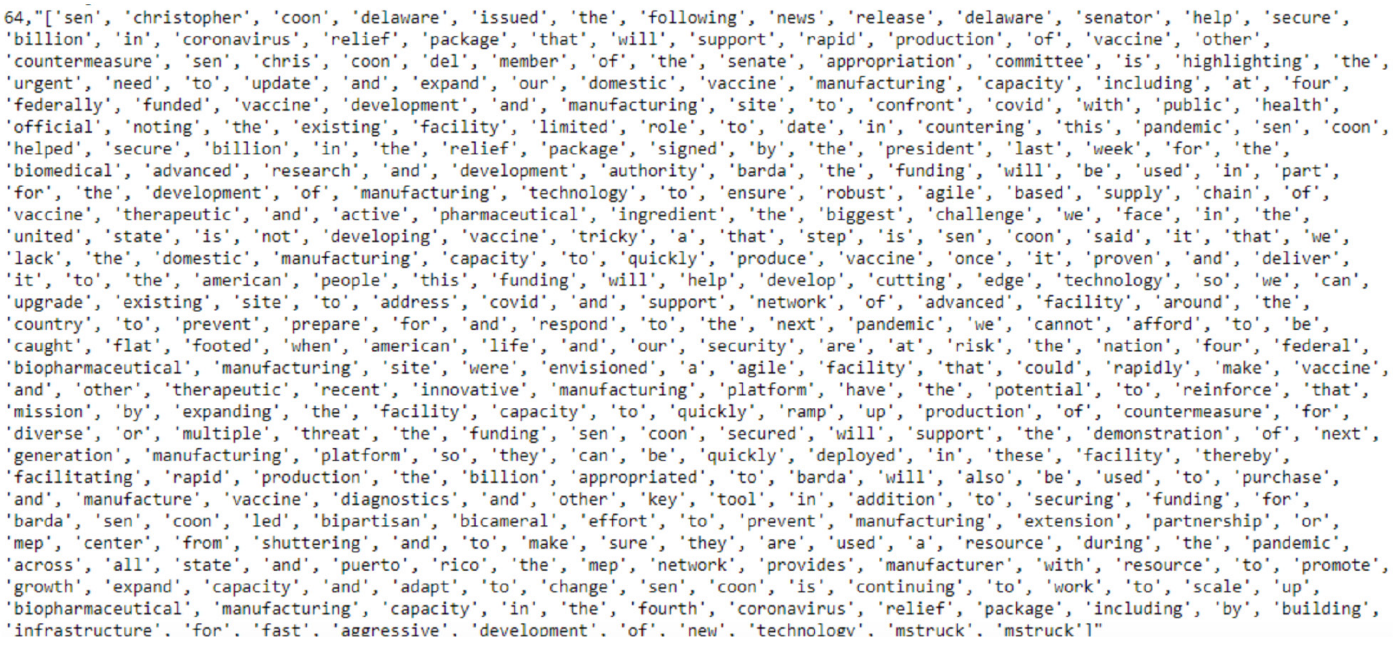
Now that we know all the topics, we have a high-level view of what the various documents in our corpus talk about when they report on the coronavirus. As a final exercise, we can read a selection of documents that have been clustered into a single topic at greater length.

Let us take a brief look at some documents that have been clustered under Topic 1: American Workers. For example:



```
59,"['house', 'of', 'representative', 'document', 'march', 'contact', 'mary', 'yatrousis', 'larson', 'call', 'for', 'infrastructure',
'package', 'to', 'boost', 'economy', 'washington', 'today', 'rep', 'john', 'larson', 'ct', 'called', 'for', 'an', 'infrastructure',
'package', 'to', 'be', 'taken', 'up', 'a', 'part', 'of', 'the', 'effort', 'to', 'mitigate', 'the', 'effect', 'of', 'covid', 'on', 'the',
'economy', 'industry', 'across', 'the', 'united', 'state', 'are', 'being', 'impacted', 'by', 'the', 'spread', 'of', 'the', 'covid', 'now',
'is', 'the', 'time', 'for', 'congress', 'to', 'invest', 'in', 'comprehensive', 'infrastructure', 'package', 'to', 'spur', 'the', 'economy',
'and', 'help', 'support', 'our', 'small', 'and', 'medium', 'sized', 'business', 'across', 'the', 'country', 'this', 'package', 'must',
'also', 'include', 'medical', 'infrastructure', 'funding', 'to', 'help', 'the', 'united', 'state', 'be', 'more', 'prepared', 'for', 'mass',
'infection', 'like', 'this', 'our', 'road', 'and', 'bridge', 'are', 'crumbling', 'and', 'this', 'funding', 'is', 'sorely', 'needed',
'investing', 'in', 'our', 'infrastructure', 'is', 'one', 'step', 'that', 'we', 'can', 'take', 'to', 'provide', 'economic', 'relief', 'said',
'larson']"
```

We see that most words of the document in question of the model have been truncated. The words in the document have been lemmatized, and some common words have been removed from the output. Wading through this might not be easy, but we can still make out most of the original document. What jumps out to you? In isolation, this is not very interesting, but we can look at another document in the same topic for more insight:



```
64,"['sen', 'christopher', 'coon', 'delaware', 'issued', 'the', 'following', 'news', 'release', 'delaware', 'senator', 'help', 'secure',
'billion', 'in', 'coronavirus', 'relief', 'package', 'that', 'will', 'support', 'rapid', 'production', 'of', 'vaccine', 'other',
'countermeasure', 'sen', 'chris', 'coon', 'del', 'member', 'of', 'the', 'senate', 'appropriation', 'committee', 'is', 'highlighting', 'the',
'urgent', 'need', 'to', 'update', 'and', 'expand', 'our', 'domestic', 'vaccine', 'manufacturing', 'capacity', 'including', 'at', 'four',
'federally', 'funded', 'vaccine', 'development', 'and', 'manufacturing', 'site', 'to', 'confront', 'covid', 'with', 'public', 'health',
'official', 'noting', 'the', 'existing', 'facility', 'limited', 'role', 'to', 'date', 'in', 'countering', 'this', 'pandemic', 'sen', 'coon',
'helped', 'secure', 'billion', 'in', 'the', 'relief', 'package', 'signed', 'by', 'the', 'president', 'last', 'week', 'for', 'the',
'biomedical', 'advanced', 'research', 'and', 'development', 'authority', 'barda', 'the', 'funding', 'will', 'be', 'used', 'in', 'part',
'for', 'the', 'development', 'of', 'manufacturing', 'technology', 'to', 'ensure', 'robust', 'agile', 'based', 'supply', 'chain', 'of',
'vaccine', 'therapeutic', 'and', 'active', 'pharmaceutical', 'ingredient', 'the', 'biggest', 'challenge', 'we', 'face', 'in', 'the',
'united', 'state', 'is', 'not', 'developing', 'vaccine', 'tricky', 'a', 'that', 'step', 'is', 'sen', 'coon', 'said', 'it', 'that', 'we',
'lack', 'the', 'domestic', 'manufacturing', 'capacity', 'to', 'quickly', 'produce', 'vaccine', 'once', 'it', 'proven', 'and', 'deliver',
'it', 'to', 'the', 'american', 'people', 'this', 'funding', 'will', 'help', 'develop', 'cutting', 'edge', 'technology', 'so', 'we', 'can',
'upgrade', 'existing', 'site', 'to', 'address', 'covid', 'and', 'support', 'network', 'of', 'advanced', 'facility', 'around', 'the',
'country', 'to', 'prevent', 'prepare', 'for', 'and', 'respond', 'to', 'the', 'next', 'pandemic', 'we', 'cannot', 'afford', 'to', 'be',
'caught', 'flat', 'footed', 'when', 'american', 'life', 'and', 'our', 'security', 'are', 'at', 'risk', 'the', 'nation', 'four', 'federal',
'biopharmaceutical', 'manufacturing', 'site', 'were', 'envisioned', 'a', 'agile', 'facility', 'that', 'could', 'rapidly', 'make', 'vaccine',
'and', 'other', 'therapeutic', 'recent', 'innovative', 'manufacturing', 'platform', 'have', 'the', 'potential', 'to', 'reinforce', 'that',
'mission', 'by', 'expanding', 'the', 'facility', 'capacity', 'to', 'quickly', 'ramp', 'up', 'production', 'of', 'countermeasure', 'for',
'diverse', 'or', 'multiple', 'threat', 'the', 'funding', 'sen', 'coon', 'secured', 'will', 'support', 'the', 'demonstration', 'of', 'next',
'generation', 'manufacturing', 'platform', 'so', 'they', 'can', 'be', 'quickly', 'deployed', 'in', 'these', 'facility', 'thereby',
'facilitating', 'rapid', 'production', 'the', 'billion', 'appropriated', 'to', 'barda', 'will', 'also', 'be', 'used', 'to', 'purchase',
'and', 'manufacture', 'vaccine', 'diagnostics', 'and', 'other', 'key', 'tool', 'in', 'addition', 'to', 'securing', 'funding', 'for',
'barda', 'sen', 'coon', 'led', 'bipartisan', 'bicameral', 'effort', 'to', 'prevent', 'manufacturing', 'extension', 'partnership', 'or',
'mep', 'center', 'from', 'shuttering', 'and', 'to', 'make', 'sure', 'they', 'are', 'used', 'a', 'resource', 'during', 'the', 'pandemic',
'across', 'all', 'state', 'and', 'puerto', 'rico', 'the', 'mep', 'network', 'provides', 'manufacturer', 'with', 'resource', 'to', 'promote',
'growth', 'expand', 'capacity', 'and', 'adapt', 'to', 'change', 'sen', 'coon', 'is', 'continuing', 'to', 'work', 'to', 'scale', 'up',
'biopharmaceutical', 'manufacturing', 'capacity', 'in', 'the', 'fourth', 'coronavirus', 'relief', 'package', 'including', 'by', 'building',
'infrastructure', 'for', 'fast', 'aggressive', 'development', 'of', 'new', 'technology', 'mstruck', 'mstruck']"
```

> *"While running a topic model is straight-up science, interpreting topic models is not and requires both art and artifice."*

This article is about vaccine production. We see that articles on infrastructure development and vaccines are clustered into the same topic. This is interesting! It suggests that there is a latent similarity between the ways that these two seemingly disparate things are discussed in the news. Vaccines and infrastructure — who knew? One could surmise that these two things both relate to development and saving America from the various impacts of the virus.

The keywords for the American Workers topic are certainly interesting in isolation, and the topic model reveals additional insights when we look at the documents for a given topic in more detail. We could repeat this with more documents and see what other higher-order patterns emerge.

Conclusion

Using TDM to create topic models offers powerful insight into how a vast corpus of news discusses the novel coronavirus. It shows us that news outlets are talking about — very broadly — the people, institutions and events affected by the outbreak. In future papers, I'll explore more detailed topic models and additional dimensions of this dataset.

About TDM Studio

ProQuest's workflow solution for text and data mining is designed for research, teaching and learning. TDM Studio provides access to sought-after content including current and historical newspapers, primary sources, scholarly journals, and dissertations and theses. It empowers researchers, students and faculty to analyze documents by uncovering connections and patterns that lead to career-defining discoveries.

Learn more at www.proquest.com/go/tdm-studio
or contact your ProQuest representative today.

References

Bird, Steven, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. 1st ed. Beijing ; Cambridge [Mass.]: O'Reilly, 2009.

Blei, David M., Andrew Y. Ng, Michael I. Jordan, and John Lafferty. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3, no. 4/5 (May 15, 2003): 993–1022.

Henrich, Joseph. *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species and Making Us Smarter*. Princeton Oxford: Princeton University Press, 2016.

Řehůřek, Radim, and Petr Sojka. "Software Framework for Topic Modelling with Large Corpora." In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA, 2010.

Selva Prabhakaran. "Topic Modeling Visualization – How to Present the Results of LDA Models?" *Machine Learning Plus* (blog), n.d. https://www.machinelearningplus.com/nlp/topic-modeling-visualization-how-to-present-results-lda-models/.

Sievert, Carson, and Kenneth Shirley. "LDAvis: A Method for Visualizing and Interpreting Topics." In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 63–70. Baltimore, Maryland, USA: Association for Computational Linguistics, 2014. https://doi.org/10.3115/v1/W14-3110.