Modern data analytics - COVID-19 project

Sebastiaan Van den Broeck

18/08/2022

There is no doubt about the enormous impact of COVID-19 on our world. That is to say, it is one of the most disastrous events of our lifetime yet. Despite the critical importance of this information, not much is known about it. The information has become wrapped up in politics and misinformation. The appearance of a pandemic is a random event. It is therefore not possible to predict when the next one will occur. However, it is a safe assumption that one will happen again in the future. Any information we can uncover now can surely be of use later (Relman, 2020). To that end, a data-driven analysis has been applied.

1 Methodology

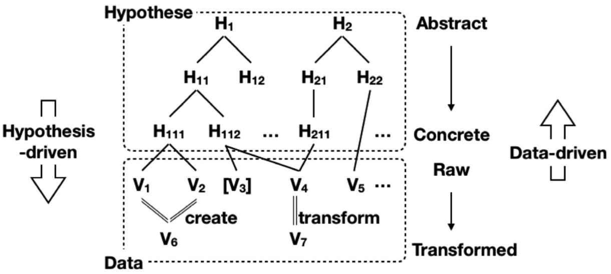

Figure 1: *The relationship between the hypothesis space, data space, a hypothesis-driven analysis and a data-driven analysis. Taken from Matsumuro & Miwa (2019).*

The scientific process can be described as a search in one of two spaces. On the one hand, there exists the well-known hypothesis space. It contains all hypothesis that are organized from abstract to specific. First, the researcher will select a general hypothesis (H_1). Next, a more specific hypothesis will have to be formulated (H_{111}). Afterwards, the researcher can reach the variables contained in the data space. This method therefore has a high depth. It can start from a very general hypothesis and reach specific variables.

On the other hand, the data space has gotten more attention recently, because of a higher availability of data. It contains all relevant information with respect to the analysis. This method therefore has a wide reach, but not a lot of depth. A sound understanding of the data is therefore needed to analyze it. To that end, plots and tables have been used to visualize the data. However, datasets often contain inconsistencies and noise. It is therefore paramount to any data science undertaking to clean the data (Sarker, 2021). The problem with a hypothesis-driven analysis has now become more clear. The researchers can only select one hypothesis (for example, H_{111}) and reach only a specific amount of variables. However, much more information is available in the data space. The data-driven analysis, then, tries to exploit this by including as much of the data space as possible. The question that remains is how one should reach a relevant hypothesis from the data (Matsumuro & Miwa, 2019).

2 Data sources

Before starting the analysis we will discuss the data. The data treated in this report comes from various sources. First, there is the COVID-19 dataset (The New York Times, 2021). The report takes a data-driven approach instead of a traditional hypothesis-driven approach. This dataset therefore fulfills a principal, directional role. It provides information on the number of cases on the national level, state level and county level in the United States. Moreover, there is information on mask use and the number of cases in prisons and colleges. Second, the previous dataset has been complemented with economical information (Bureau of Economic Analysis, 2022). It provides information on the GDP of each state per industry. The information will allow us to not only examine the human impact of the pandemic, but also take a closer look at the economical impact. Third, there is no doubt air transportation has played an important role in the pandemic. The crowdsourced air traffic dataset has therefore been added to the analysis (Strohmeier et al., 2021). It is a transformed and cleaned version from the OpenSky dataset and contains important information regarding the development of air traffic before and during the pandemic. Fourth, the CORD-19 dataset has been added to the analysis (Lucy et al., 2020). It contains a collection of more than 140000 scientific papers and related research.

3 Graph mining

The second part of the project has been devoted to graph mining. Graph mining is a general term that means mining information from a dataset that is represented by a graph structure (Rehman et al., 2012). In other words, the data is represented by nodes, relationships and properties. We will tap into a new part of the data space in the hopes of reaching an interesting hypothesis. To that end, different data sources can be combined. Modelling the spread of a disease using a graph approach is not new concept. For example, Manriquez et al. (2021) used an edge-weighted graph to study the dissemination of an epidemic. Specifically, centrality measures have been used to find out the importance of airports. To that end, Neo4J can be used as a graph database. We can consider the airports as nodes and the flights as relationships between the airports. Moreover, an object-oriented programming approach has been taken.

The identification of which nodes in a network are the most important has been an important issue in network analysis (Opsahl et al., 2010). Indeed, different interpretations of "importance" will lead to different approaches. To that end, three important interpretations of centrality have been formulated. First, the centrality of a node depends on the connectedness through immediate connections to other nodes. Second, the centrality of a node depends on the frequency with which the node falls on paths between other nodes. Third, the centrality of a graph depends on the degree to which it is close to the rest of the graph (Klein, 2010). The first interpretation corresponds to the degree centrality. This metric tries to assess the importance of a node using the number of direct connections it has to other nodes. Within the context of our analysis it is certainly important to know how many connections an airport has to other ones. However, this centrality measure does not take into account the global structure of the network, but only the direct connections. The second interpretation denotes the betweenness of a node. The betweenness centrality measures the extent to which a node lies on the shortest path between other nodes. There is no doubt that this interpretation is also important for our analysis. The third interpretation is called the closeness centrality. It simply measures the inverse sum of the shortest distances to all the other nodes. However, we are not working with a weighted graph. The distance metric is therefore simply the number of hops between two airports. The interpretation after calculating this metric may be difficult. Another interesting approach could be constructing a graph where the weights are the distances between airports (Opsahl et al., 2020).

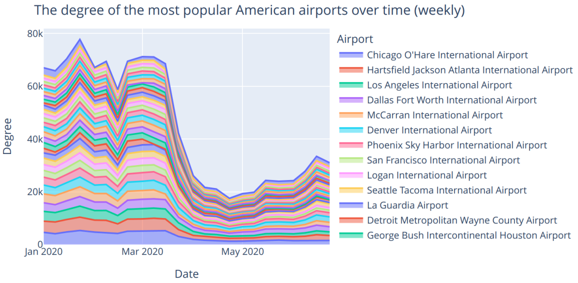

Figure 2: *The degree of the most popular American airports took a sharp dive after the WHO declared COVID-19 a pandemic and after Trump declared a national emergency.*

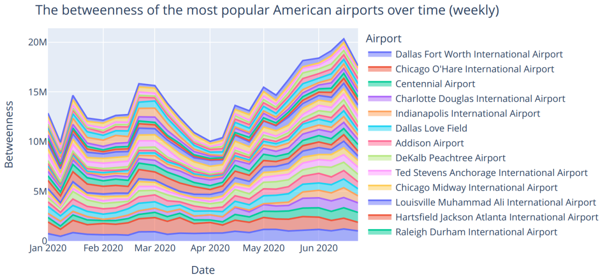

Figure 3: *A similar trend can be detected for the betweenness centrality of American airports.*

4 Text mining

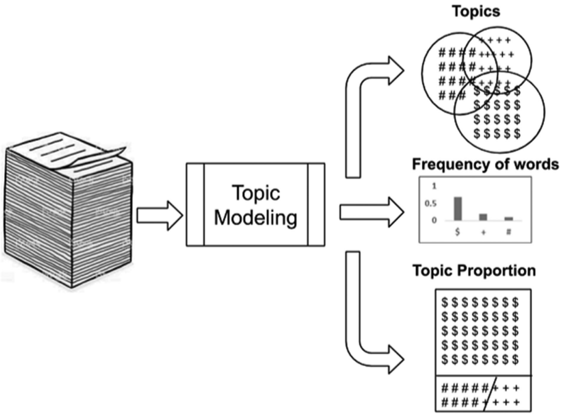

Figure 4: *The area of topic modelling is concerned with making inferences from a huge collection of documents in terms of topics, frequency of words and topic proportions. Taken from Chauchan & Shah (2022).*

Third, another non-tabular data source that often goes unnoticed is text. Indeed, natural language processing, or text mining, is a family of techniques that tries to extract meaning from text documents. Topic modelling is one area of NLP that can be used to provide an overview from a very large collection of documents. Topic models, then, are statistical models that can be used to show the hidden structure of text data. In other words, they learn the latent semantic structures of a collection of text documents (Chauhan & Shah, 2022). The technique allows inferences to be made in terms of the whole collection, individual documents or relationships between the documents. Specifically, latent dirichlet allocation will be applied. The technique models a topic by word probabilities. Therefore, one can look at the words with the highest probabilities for each topic to get an idea of what the topic is about (Jelodar et al., 2019).

The dataset at hand is the COVID-19 open research dataset (CORD-19). It provides a collection of more than 140000 scientific papers and related research. The goal of the dataset is to use machine learning to discover relevant information. Therefore, it has been specifically designed for text mining purposes. However, limited preprocessing of the text has been done. We will therefore go through these steps ourselves (Lucy et al., 2020).

Table 1: *Topics before pandemic*

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**	**Topic 6**	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**
cell	health	patient	cov	cell	infect	immun	vaccin	calv	der
protein	diseas	respiratori	detect	protein	respiratori	vaccin	viru	protein	die
viru	develop	studi	viru	activ	viral	cell	infect	viru	und
viral	model	infect	sequenc	lung	group	respons	influenza	structur	model
infect	public	associ	human	express	studi	protein	studi	studi	ein
activ	emerg	children	respiratori	induc	viru	antigen	result	result	network
express	data	clinic	virus	airwai	associ	antibodi	ibv	effect	von
rna	infecti	influenza	pcr	result	virus	specif	virus	test	mit
virus	studi	case	infect	increas	caus	express	protect	differ	bei
host	risk	hospit	sars	mice	patient	infect	effect	domain	cell

Next, we will look at the clusters of documents before and during the pandemic. A more detailed analysis has been given in the notebook. The table shown above denotes the 10 most important words per topic before the pandemic. First, we may notice that some improvements can still be made in the preprocessing of the documents. For example, the words *viru*, *viral* and *virus* all mean the same thing. They should therefore be stemmed in a way that they become one token. Another problem becomes apparent when taking a closer look at topic 10. Apparently, one or more German papers have made their way into the dataset and have formed their own cluster. This may actually indicate that the LDA is working well, since these papers should clearly be in a cluster of their own. Second, inferences can be made regarding the research itself. On the one hand, we can detect some clusters that have to do with the inner workings of the disease. For example, for the topic 1 and 5 the words *cell*, *protein* and *rna* are important. On the other hand, some other topics appear to deal with the patient and the sociatal impact. For example, the words *health*, *disease*, *public* and *model* are important for topic 2. Topic 3, then, has the words *patient*, *respiratori* and *studi* associated with it. It is also interesting to note that there is a cluster of documents that deals with vaccination. Topic 8 has the words *vaccin*, *infect* and *influenza* associated with it.

Next, we will take a look at the interpretation of the LDA that is trained on documents from after the start of the pandemic. First, we may notice that there is a new cluster compared to the previous model. Cluster 9 has the words *mask*, *effect* and *aerosol* associated with it. In other words, this is a cluster about the effectiveness of masks for the spread of COVID-19. Second, there is again a cluster with German words. This gives an indication that still some work can be done on the preprocessing steps. On the other hand, it can indicate that the LDA is working well, because it can isolate this cluster. Third, it is clear that the word *covid* appears a lot in the clusters and is not very informative. It could therefore be removed in the preprocessing steps. More interpretation regarding the topics and their relationships to each other has been given in the notebook.

Table 2: *Topics during pandemic*

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**	**Topic 6**	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**
health	patient	covid	test	patient	cov	protein	health	mask	der
covid	covid	model	covid	cell	sars	cov	studi	effect	die
studi	hospit	vaccin	model	covid	covid	cell	mental	covid	und
research	risk	case	data	diseas	infect	sars	depress	cov	text
data	studi	studi	method	treatment	patient	ace	associ	sars	studi
care	clinic	infect	detect	immun	antibodi	viral	anxieti	aerosol	infect
pandem	sever	number	perform	activ	respiratori	bind	women	patient	activ
provid	associ	effect	result	studi	coronaviru	covid	social	studi	gene
develop	group	data	studi	clinic	sever	viru	covid	infect	effect
social	patients	time	sampl	associ	studi	drug	effect	risk	model

5 Conclusion

In this paper I have provided a fresh perspective on the COVID-19 pandemic. First, the New York Times COVID-19 dataset was used as the guiding force. In other words, it played a very important role in the exploratory data analysis. Second, economical information was used to provide more robust explanations. Third, the COVID-19 OpenSky dataset was used for graph mining. Fourth, the CORD-19 dataset was added to the mix to perform topic modelling. The results of the analysis indicate a detrimental effect on all centrality measures of the airport-flights graph. We uncovered a sharp drop in the degree centrality after the WHO declared COVID-19 a pandemic and after Trump declared COVID-19 a national emergency. Moreover, we took a look at the locations of international airports and first COVID-19 cases. As could be expected, the locations of international airports are close to the first COVID-19 cases. Second, text mining was applied to detect the change in scientific literature before and after the pandemic. The results indicate a higher importance of the treatment of individuals, vaccination and the effectiveness of masks.

6 Bibliography

Brandes, U. (2008). On variants of shortest-path betweenness centrality and their generic computation. Social Networks, 30(2), 136–145. https://doi.org/10.1016/j.socnet.2007.11.001

Bureau of Economic Analysis. (2022). Annual GDP per state [Dataset]. https://apps.bea.gov/regional/downloadzip.cfm

Chauhan, U., & Shah, A. (2022). Topic Modeling Using Latent Dirichlet allocation: A Survey. ACM Computing Surveys, 54(7), 1–35. https://doi.org/10.1145/3462478

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. Multimedia Tools and Applications, 78(11), 15169–15211. https://doi.org/10.1007/s11042-018-6894-4

Klein, D. J. (2010). Centrality measure in graphs. Journal of Mathematical Chemistry, 47(4), 1209–1223. https://doi.org/10.1007/s10910-009-9635-0

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, et al.. 2020. CORD-19: The COVID-19 Open Research Dataset. In Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020, Online. Association for Computational Linguistics.

Manríquez, R., Guerrero-Nancuante, C., Martínez, F., & Taramasco, C. (2021). Spread of Epidemic Disease on Edge-Weighted Graphs from a Database: A Case Study of COVID-19. International Journal of Environmental Research and Public Health, 18(9), 4432. https://doi.org/10.3390/ijerph18094432

Matsumuro, M., & Miwa, K. (2019). Model for Data Analysis Process and Its Relationship to the Hypothesis-Driven and Data-Driven Research Approaches. In A. Coy, Y. Hayashi, & M. Chang (Eds.), Intelligent Tutoring Systems (Vol. 11528, pp. 123–132). Springer International Publishing. https://doi.org/10.1007/978-3-030-22244-4_16

Opsahl, T., Agneessens, F., & Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. Social Networks, 32(3), 245–251. https://doi.org/10.1016/j.socnet.2010.03.006

Rehman, S. U., Khan, A. U., & Fong, S. (2012). Graph mining: A survey of graph mining techniques. Seventh International Conference on Digital Information Management (ICDIM 2012), 88–92. https://doi.org/10.1109/ICDIM.2012.6360146

Relman, D. A. (2020). To stop the next pandemic, we need to unravel the origins of COVID-19. Proceedings of the National Academy of Sciences, 117(47), 29246–29248. https://doi.org/10.1073/pnas.2021133117

Sarker, I. H. (2021). Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective. SN Computer Science, 2(5), 377. https://doi.org/10.1007/s42979-021-00765-8

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, 63–70. https://doi.org/10.3115/v1/W14-3110

Strohmeier, M., Olive, X., Lübbe, J., Schäfer, M., & Lenders, V. (2021). Crowdsourced air traffic data from the OpenSky Network 2019–2020. Earth System Science Data, 13(2), 357–366. https://doi.org/10.5194/essd-13-357-2021

Stroustrup, B. (1988). What is object-oriented programming? IEEE Software, 5(3), 10–20. https://doi.org/10.1109/52.2020

The New York Times. (2021). Coronavirus (Covid-19) Data in the United States. Retrieved from https://github.com/nytimes/covid-19-data.