Topic Modelling of Swedish Newspaper Articles about Coronavirus:

a Case Study Using Latent Dirichlet Allocation

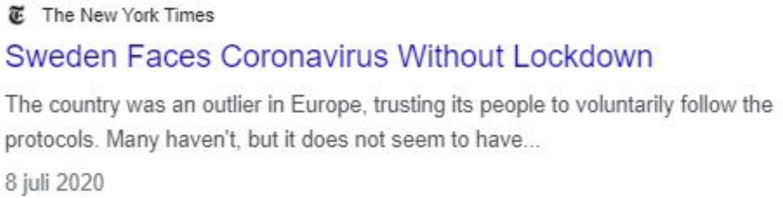
Bernadeta Griciute[1,2] Lifeng Han[3] Goran Nenadic[3]

University of Malta[1], Saarland University[2], University of Manchester[3]

HealthNLP2023: The 6th International Workshop on Health Natural Language Processing, with ICHI 2023
June 26th, 2023, Houston, Texas, USA [*https://www.healthnlp.info/*](https://www.healthnlp.info/)

Motivation

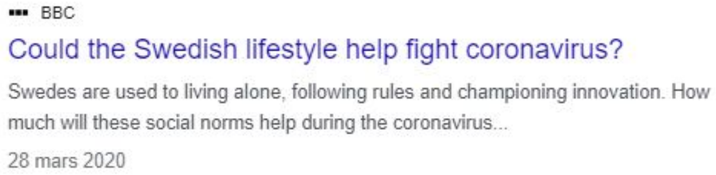
Sweden’s unconventional response to COVID-19


The New York Times

Sweden Faces Coronavirus Without Lockdown

The country was an outlier in Europe, trusting its people to voluntarily follow the protocols. Many haven't, but it does not seem to have...

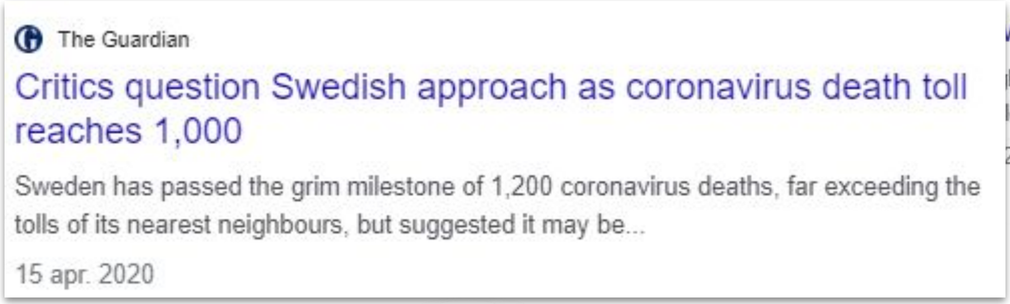
8 juli 2020


BBC

Could the Swedish lifestyle help fight coronavirus?

Swedes are used to living alone, following rules and championing innovation. How much will these social norms help during the coronavirus...

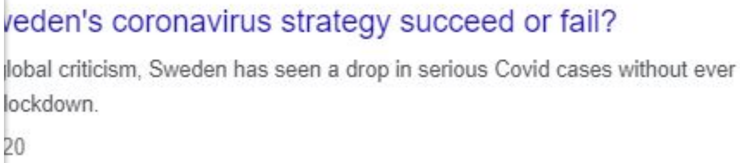
28 mars 2020


The Guardian

Critics question Swedish approach as coronavirus death toll reaches 1,000

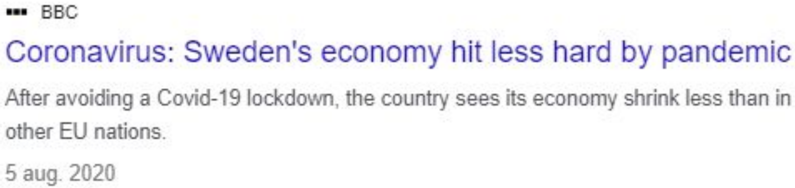
Sweden has passed the grim milestone of 1,200 coronavirus deaths, far exceeding the tolls of its nearest neighbours, but suggested it may be...

15 apr. 2020


veden's coronavirus strategy succeed or fail?

lobal criticism, Sweden has seen a drop in serious Covid cases without ever lockdown.

20


BBC

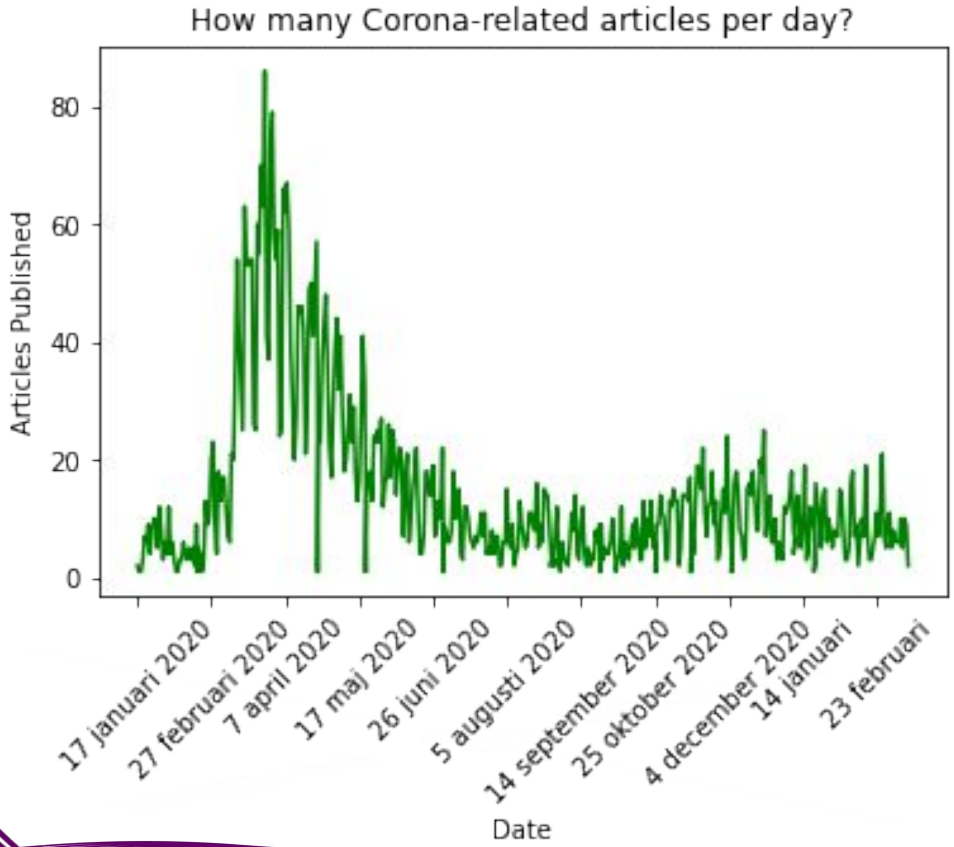
Coronavirus: Sweden's economy hit less hard by pandemic

After avoiding a Covid-19 lockdown, the country sees its economy shrink less than in other EU nations.

5 aug. 2020

Data

- Newspaper articles from SVT, Swedish national public television broadcaster;
- Articles with Covid as the main topic;
- 17/01/2020 - 17/01/2021;
- 6515 articles scraped, 2251 after filtering only nationwide and foreign news.



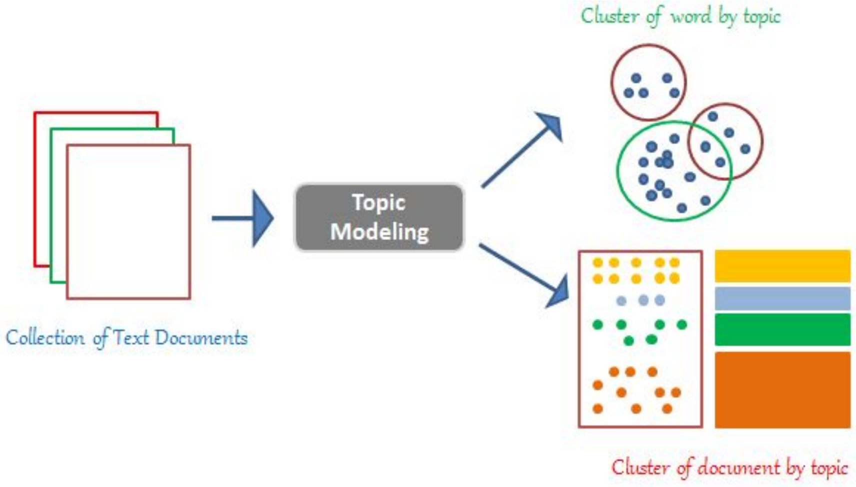
Methods I

Latent Dirichlet Allocation (LDA)

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})$$
$$= \Pi_{i=1}^{K} p(\beta_i) \Pi_{d=1}^{D} p(\theta_d)$$
$$(\Pi_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n}))$$

where the four main parameters β, θ, z, and w represent respectively the “topic distribution”, “topic proportion of document”, “topic assignment of document”, and the “observed words of document”.

Blei et al., 2003

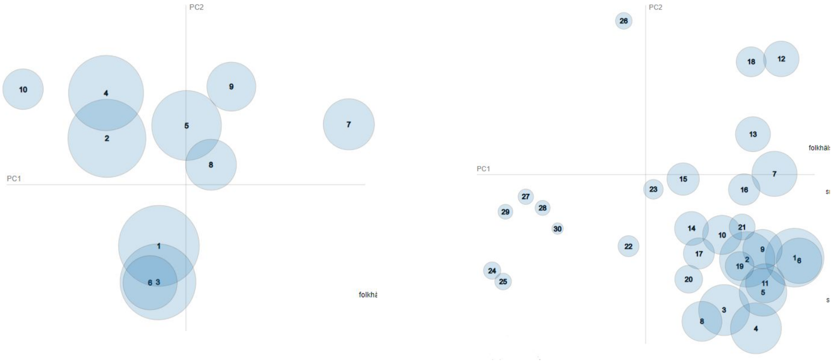


https://thinkinfi.com/latent-dirichlet-allocation-for-beginners-a-high-level-overview/

Methods II

Handpicking number of topics

tried various options between 10 to 50 topics, and it seemed that 15 to 25 topics are not too overlapping and meanwhile keep a good balance. (10, 30)=(left, right)



Dynamic Topic Modelling (DTM) - extension of LDA: evolution of topics over time.

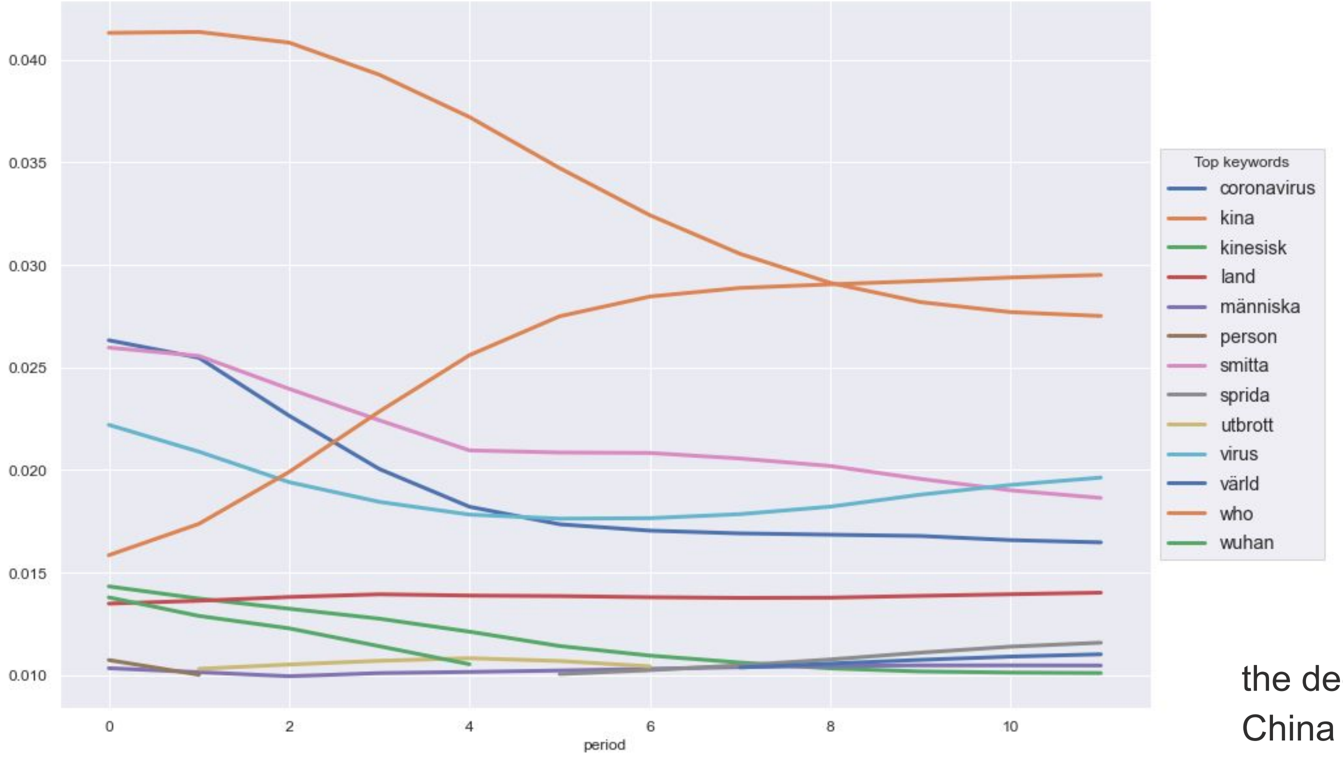
GENSIM library GENSIM topic modelling for humans

The most intense month consisted of 569 articles, while the smallest number was only 23 articles

Results

Read the pre-print paper for more details? https://arxiv.org/abs/2301.03029

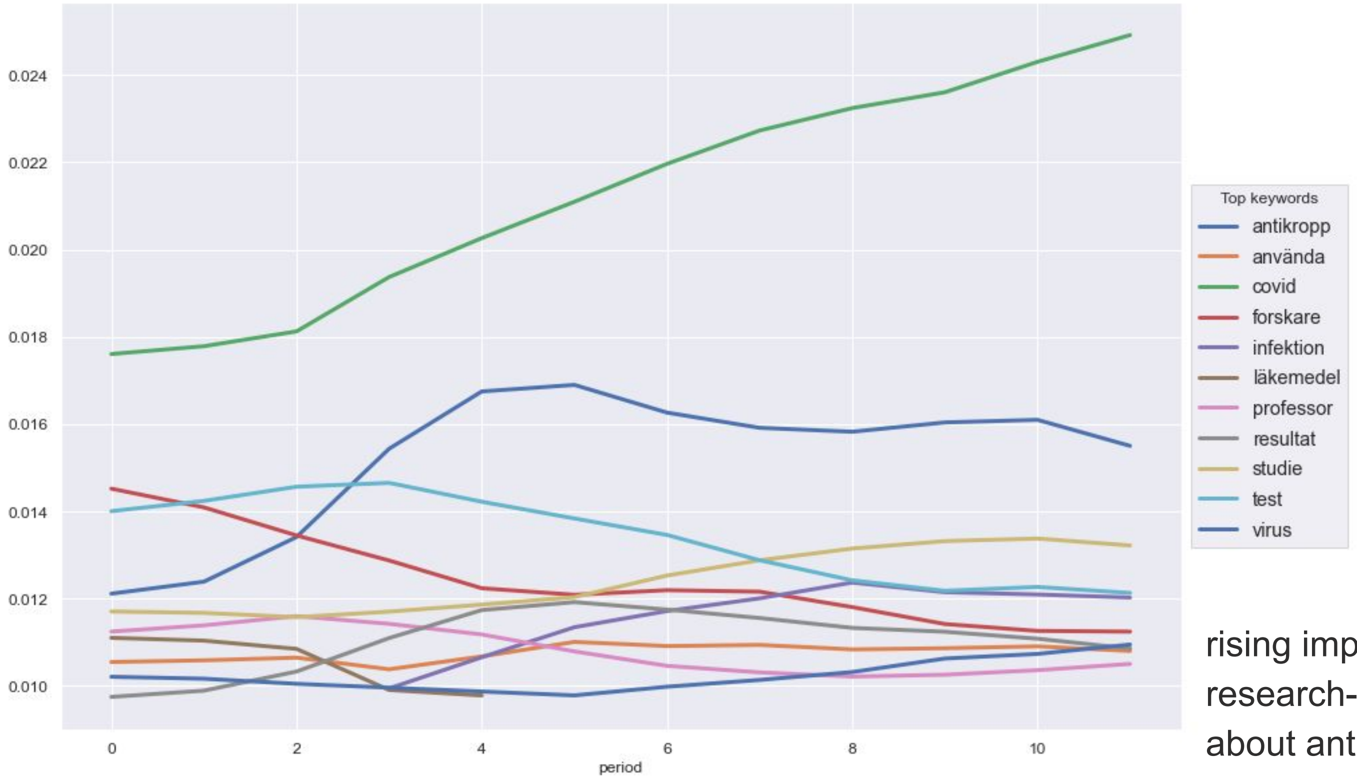
From Local to Global Issue



Coronavirus
China
Chinese
country
person
people
infect
spread
outbreak
virus
world
WHO
Wuhan

the decline in the graphs illustrating China and Wuhan and the rising curve of World Health Organisation

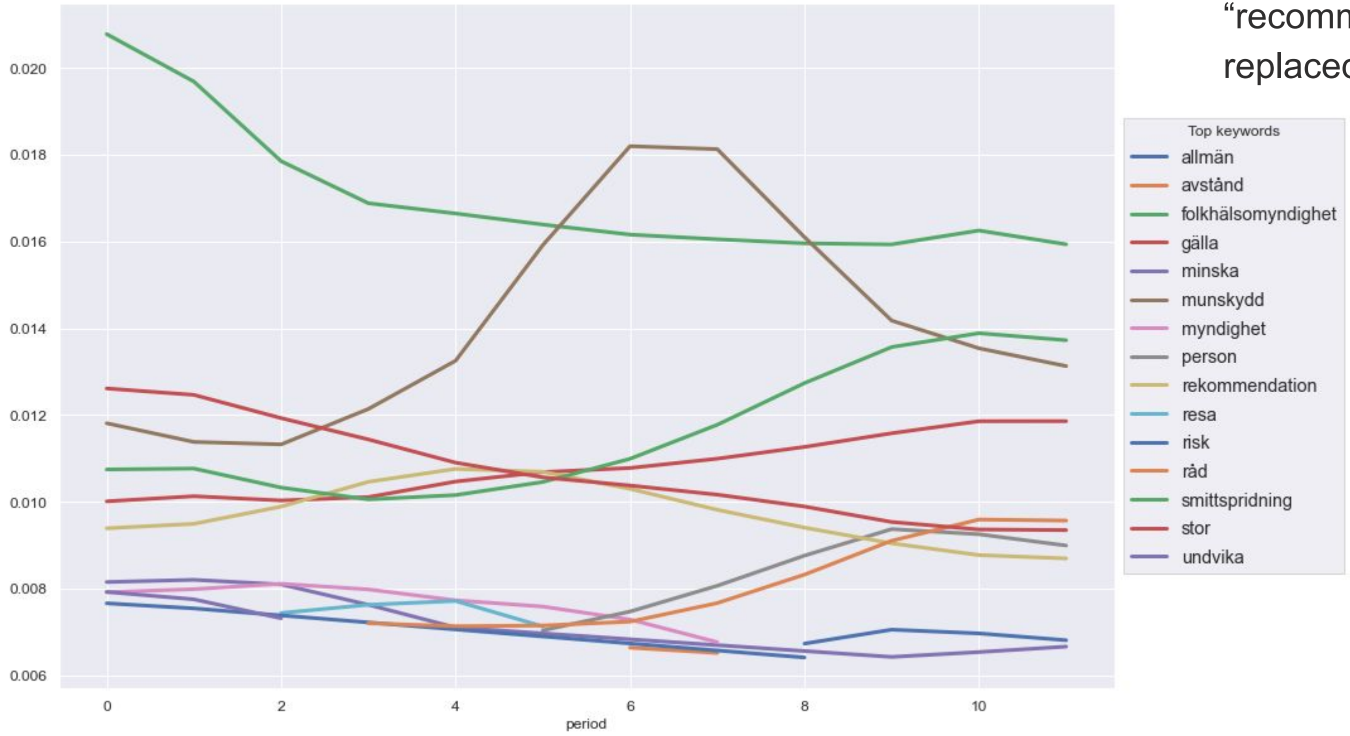
Covid Research



antibodies
use
Covid
researcher
infection
medicine
professor
results
study
test
virus

rising importance of the Covid research-related topic - discussions about antibodies

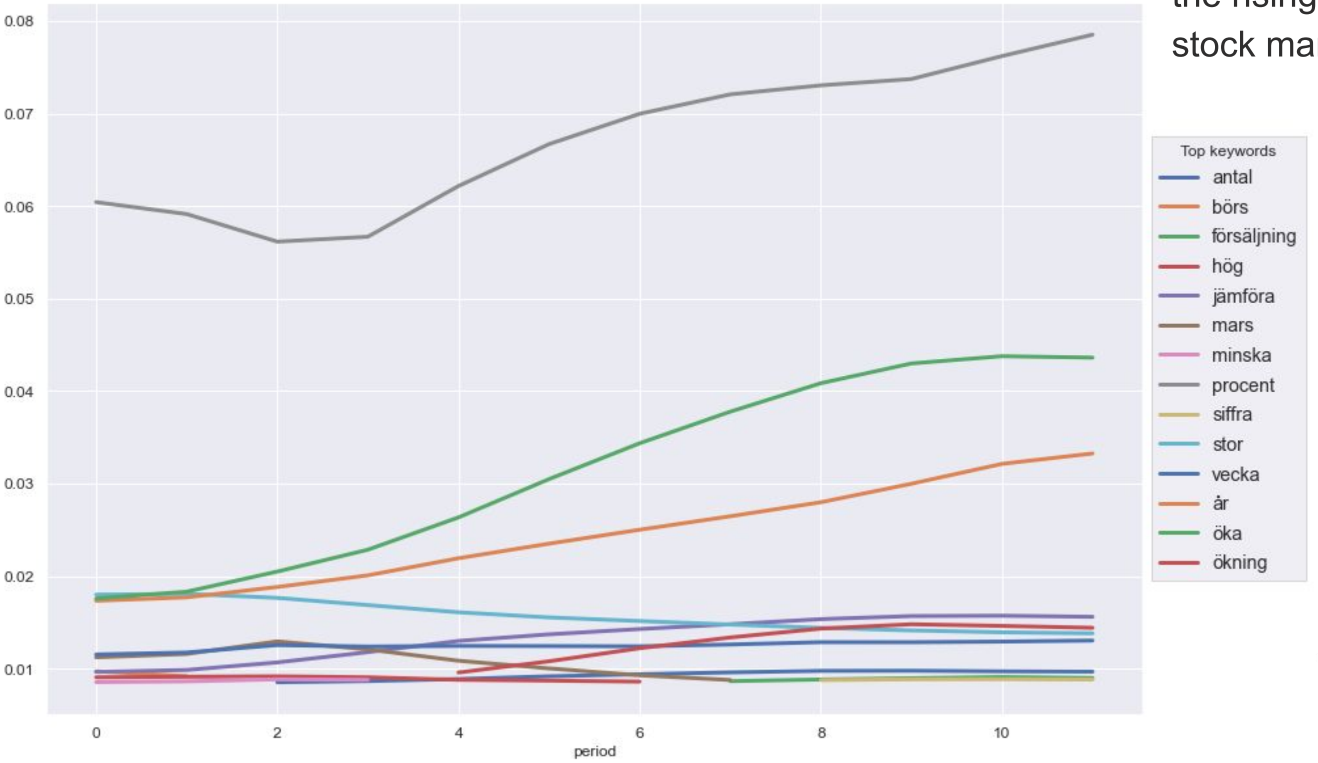
From Recommendations to Advices



the use of the word “recommendation” decreases, and is replaced by the “advice”

general
distance
Public Health Agency
apply to
decrease
face mask
authority
person
recommendation
travel
risk
advice
spread of infection
big
avoid

Economical Consequences



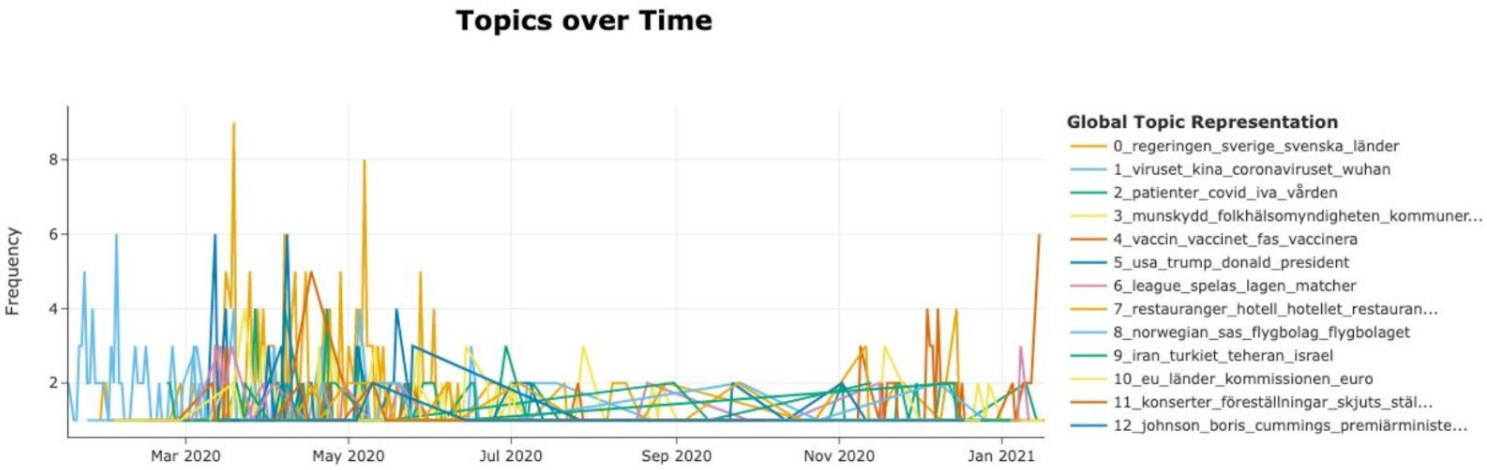
the rising concern about sales and stock market

number
stock market
sales
high
compare
March
decrease
percent
numbers
big
week
year
increase

Neural Models

comparing the importance of the topics themselves:

- in the beginning the China-related topic was dominant, but it soon got overtaken by all the other related topics;
- with Sweden and its government’s decisions getting the most audience, among them, the face-mask discussions.
- observe the rise of vaccine related topic, once it became available



government, Sweden, Swedish
virus, China, coronavirus
patients, Covid, Intensive Care Unit
face masks, Public Health Agency, communes
vaccine, the vaccine, phase
USA, Trump, Donald
League, play, teams
Restraurants, hotel, the hotel
Norwegian, SAS, airline
Iran, Turkey, Teheran
EU, countries, Commision
concerts, performances, postponed
Johnson, Boris, Cummings

BERTopic Output of Clustering using Minimum 10 Sentences per Cluster with Time Frame, upcoming work in collaboration with Hao Li

Conclusions

- Topic Modelling gives good insights on the topics discussed in the society during the global pandemic (local vs. global threat, changes in the tone of the government, social and economical impacts);
- An unsupervised approach is convenient to use in pressing issues like Covid; for the full potential, approaches where the number of topics is chosen without human intervention should be used.

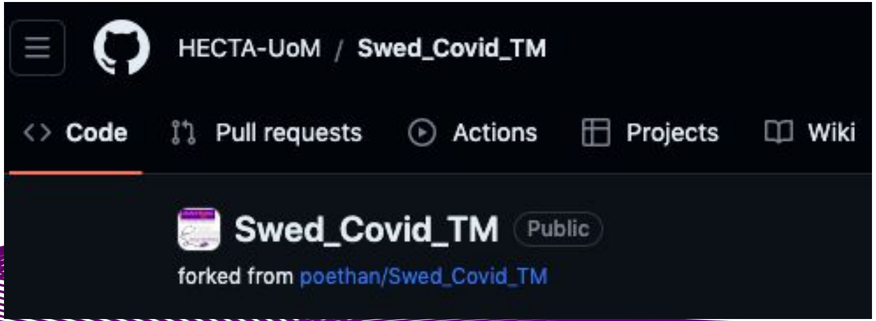
Future Work

- More data: other newspapers, social media, longer span
- More countries: other Scandinavian countries, countries with different policies (China)
- More methods: BERTopic vs LDA in details

Bernadeta Griciūtė, Lifeng Han, Hao Li, Goran Nenadic. Topic Modelling of News Articles on Covid19: Investigation using Statistical and Neural Methods. HealTAC 2023: HEALTHCARE TEXT ANALYTICS CONFERENCE 2023 MANCHESTER, JUNE 14-16, 2023

Open-source Project

- Corpus collected, stop-word lists, codes used via Colab
- Different outputs using various parameters, e.g. number of sentences for BERT-topic.
- https://github.com/HECTA-UoM/Swed_Covid_TM



HECTA-UoM / Swed_Covid_TM

Swed_Covid_TM Public

forked from poethan/Swed_Covid_TM

Resources

1. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.
2. Blei, David M; Lafferty, John D (2006). *Dynamic topic models*. *Proceedings of the ICML*. ICML'06. pp. 113–120.
3. Naskar, A. Latent Dirichlet Allocation for Beginners: A high level overview. Retrieved 03/05/2023 from https://thinkinfi.com/latent-dirichlet-allocation-for-beginners-a-high-level-overview/ .
4. This pre-print paper link: https://arxiv.org/abs/2301.03029

Thank you!

Get in touch?

Griciute.Bernadeta[AT]gmail.com

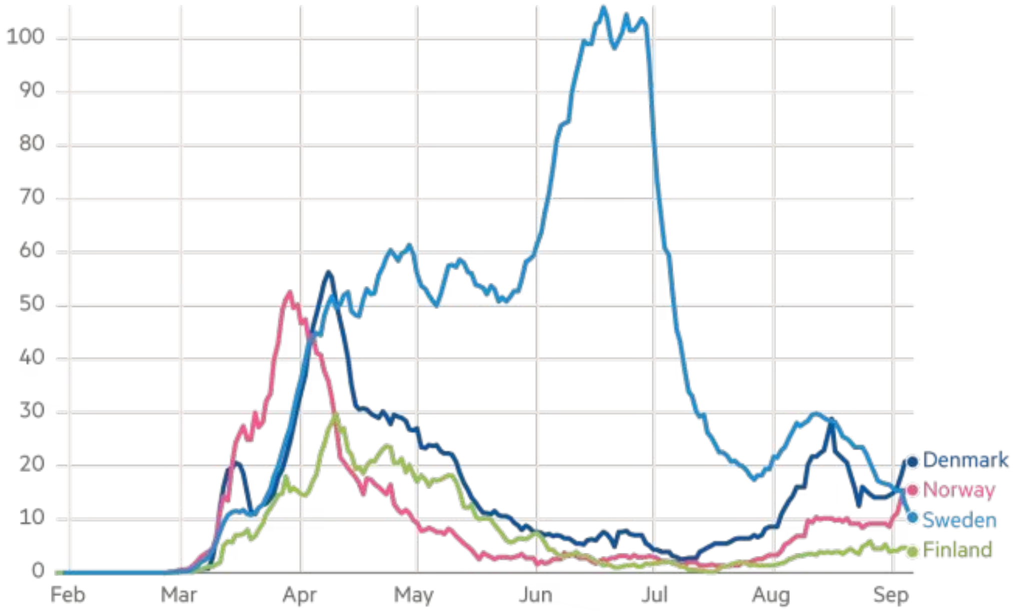
{G.Nenadic,

Lifeng.Han}[AT]manchester.ac.uk

Appendices

The profile of Sweden's pandemic differs radically from those of its neighbours

New confirmed cases of Covid-19, seven-day rolling average of new cases (per million)



Source: FT analysis of data from the European Centre for Disease Prevention and Control, the Covid Tracking Project
Data updated Sep 8 at 1pm BST. Interactive version: ft.com/covid19
© *FT*