From Bats to Masks: Change of Topics in Swedish Articles about Coronavirus

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

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

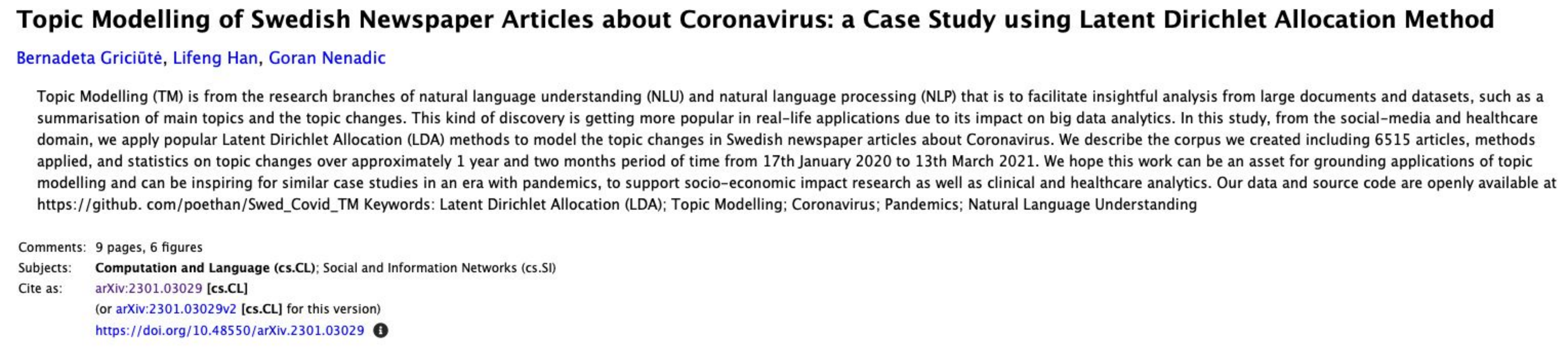
Content

Motivation

Corpus preparation

Experimental work using LDA tools

Findings



Topic Modelling of Swedish Newspaper Articles about Coronavirus: a Case Study using Latent Dirichlet Allocation Method

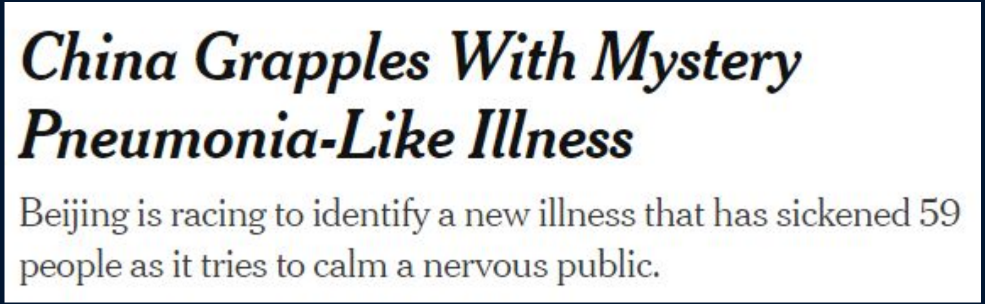
Bernadeta Griciūtė, Lifeng Han, Goran Nenadic

Topic Modelling (TM) is from the research branches of natural language understanding (NLU) and natural language processing (NLP) that is to facilitate insightful analysis from large documents and datasets, such as a summarisation of main topics and the topic changes. This kind of discovery is getting more popular in real-life applications due to its impact on big data analytics. In this study, from the social-media and healthcare domain, we apply popular Latent Dirichlet Allocation (LDA) methods to model the topic changes in Swedish newspaper articles about Coronavirus. We describe the corpus we created including 6515 articles, methods applied, and statistics on topic changes over approximately 1 year and two months period of time from 17th January 2020 to 13th March 2021. We hope this work can be an asset for grounding applications of topic modelling and can be inspiring for similar case studies in an era with pandemics, to support socio-economic impact research as well as clinical and healthcare analytics. Our data and source code are openly available at https://github. com/poethan/Swed_Covid_TM Keywords: Latent Dirichlet Allocation (LDA); Topic Modelling; Coronavirus; Pandemics; Natural Language Understanding

Comments: 9 pages, 6 figures
Subjects: **Computation and Language (cs.CL)**; Social and Information Networks (cs.SI)
Cite as: arXiv:2301.03029 **[cs.CL]**
(or arXiv:2301.03029v2 **[cs.CL]** for this version)
https://doi.org/10.48550/arXiv.2301.03029

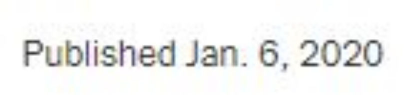
https://doi.org/10.48550/arXiv.2301.03029

Motivation

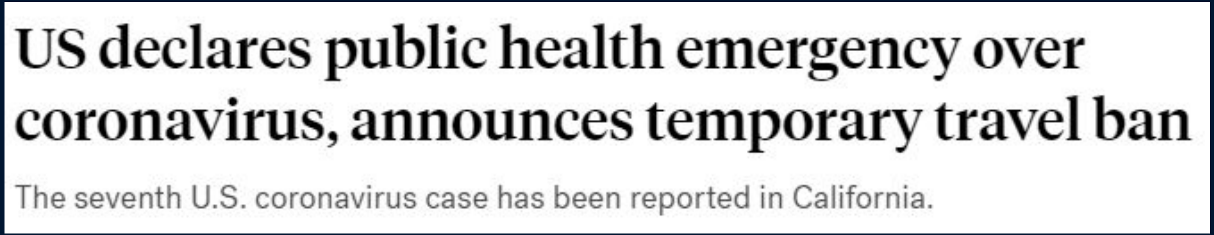


China Grapples With Mystery Pneumonia-Like Illness

Beijing is racing to identify a new illness that has sickened 59 people as it tries to calm a nervous public.

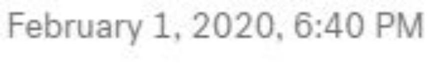


Published Jan. 6, 2020



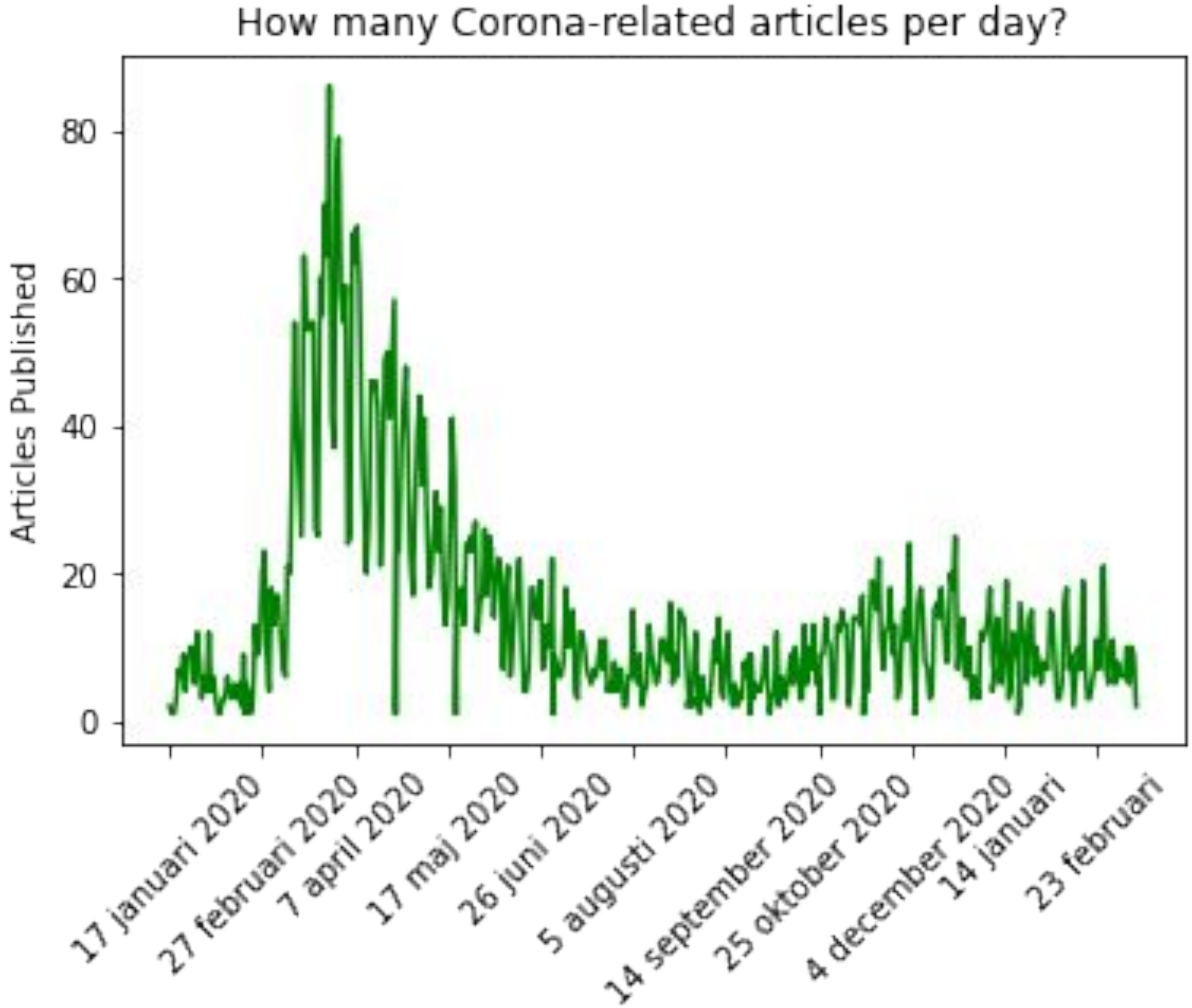
US declares public health emergency over coronavirus, announces temporary travel ban

The seventh U.S. coronavirus case has been reported in California.



February 1, 2020, 6:40 PM

COVID AND MEDIA



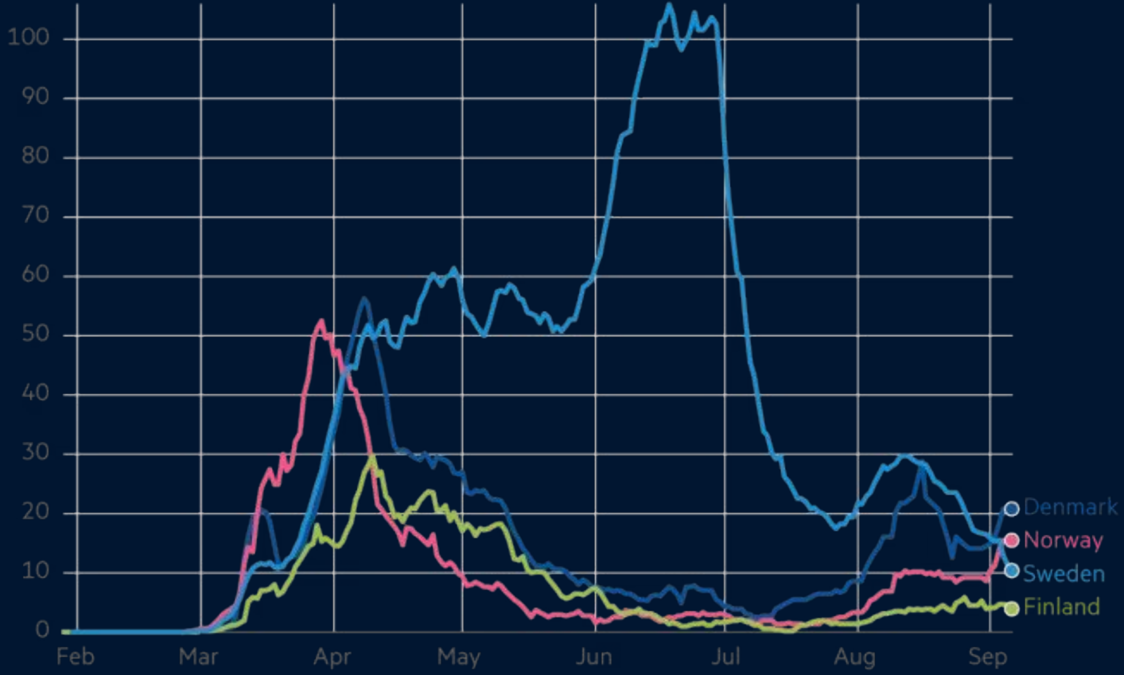
SVT, Sweden

Different approaches

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

New confirmed cases of Covid-19 (per million)

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

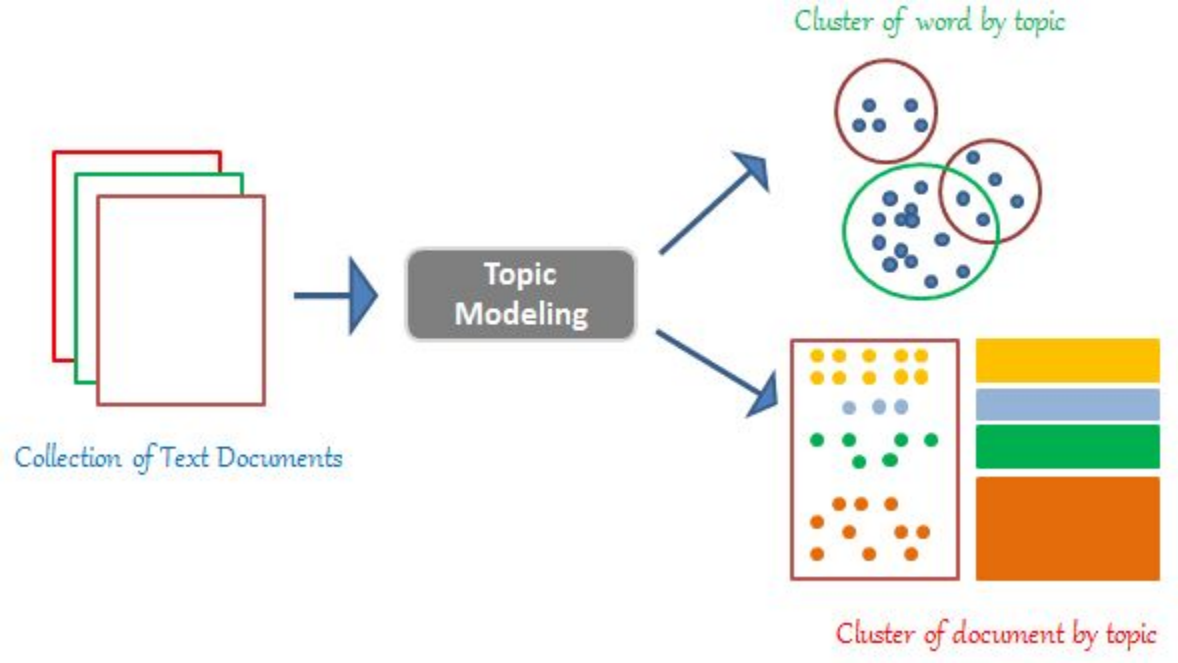


Source: FT analysis of data from the European Centre for Disease Prevention and Control, the Covid Tracking Project

Data updated Sep 8 at 1pm BST. Interactive version: ft.com/covid19

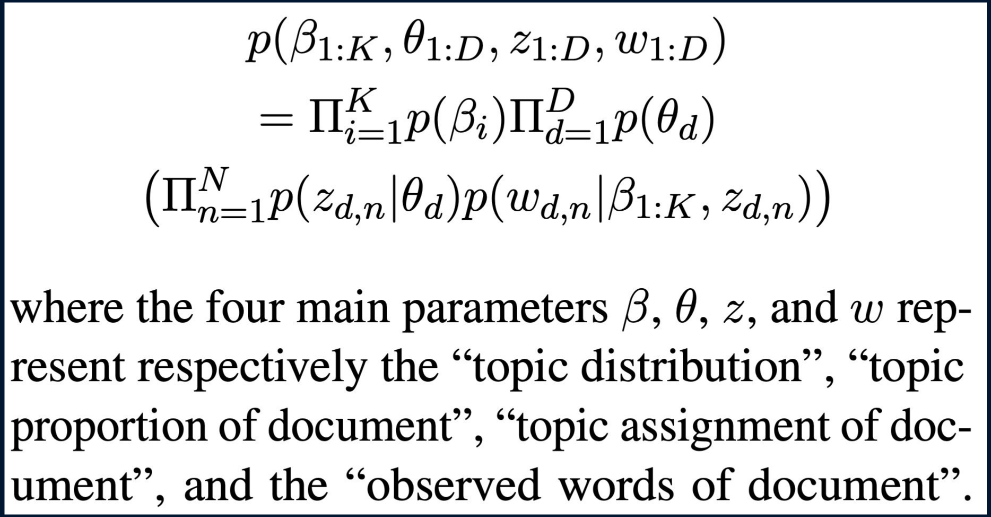
© FT

Topic Modelling.
Latent Dirichlet Allocation (LDA)



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

More LDA



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

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; Blei, 2012)

Read our paper for more detailed interpretations https://arxiv.org/abs/2301.03029

DTM

In comparison to “statistical assumptions of a **static** topic model, such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003).”

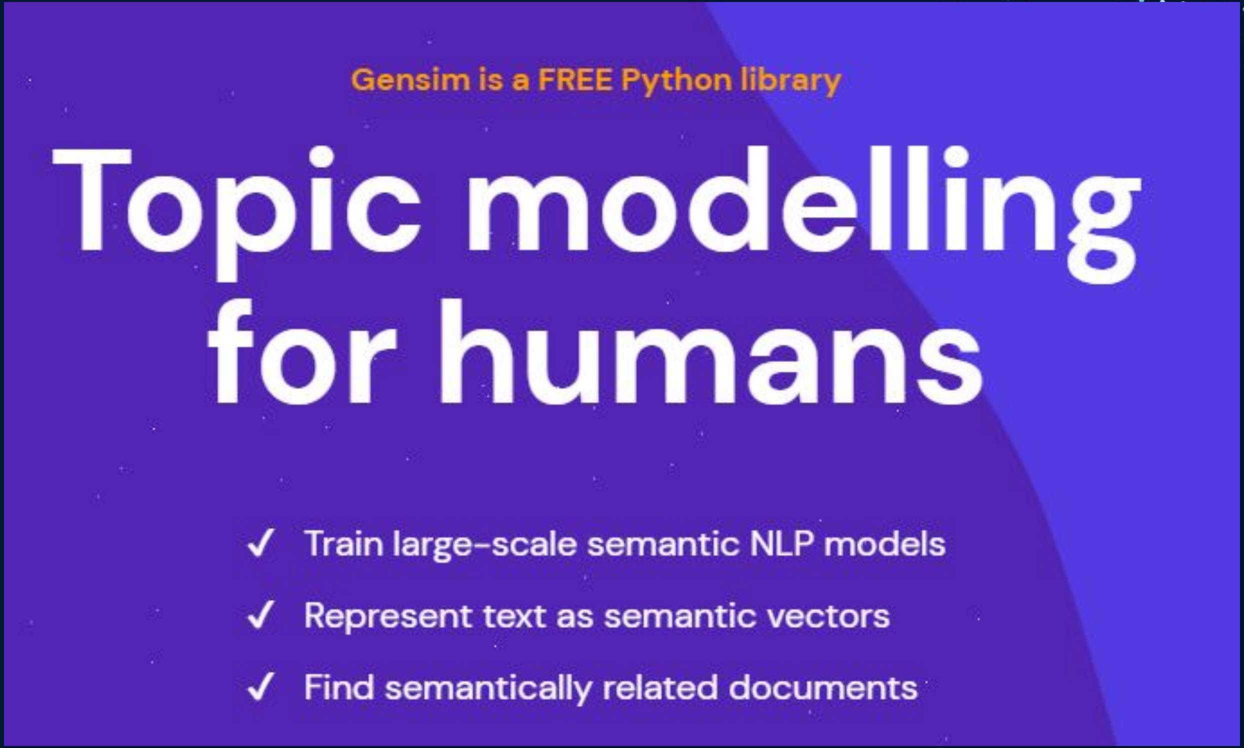
- LDA assumes the documents are drawn exchangeably form the same set of topics.
- However, the order of some collections reflects an evolving set of topics

To address this, DTM approaches the task by dividing data by time slice, e.g. day/month/year

- Model the documents with k-component topic model
- Topics associated with slit t evolve forms the topics associated with slice ‘t-1’
- Read Blei and Lafferty (2006) for more math implementation, included in **Gensim toolkit**

Dynamic Topic Models (DTM), introduced by Blei and Lafferty (2006); Blei (2012)

Gensim

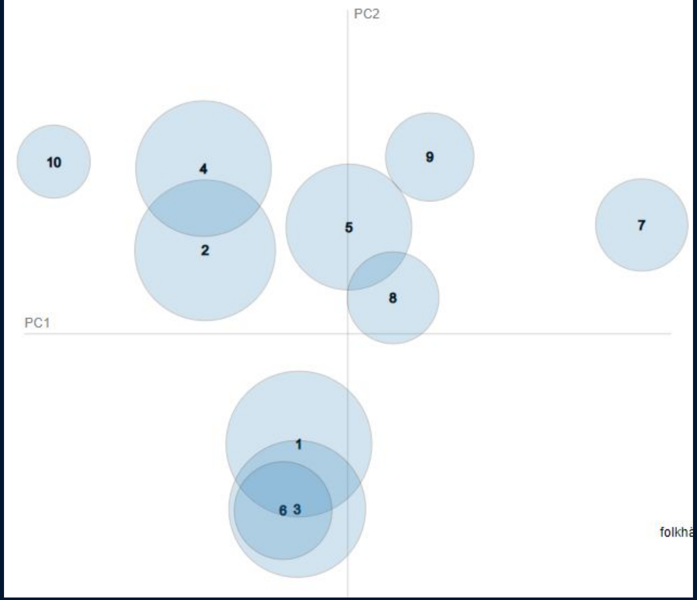


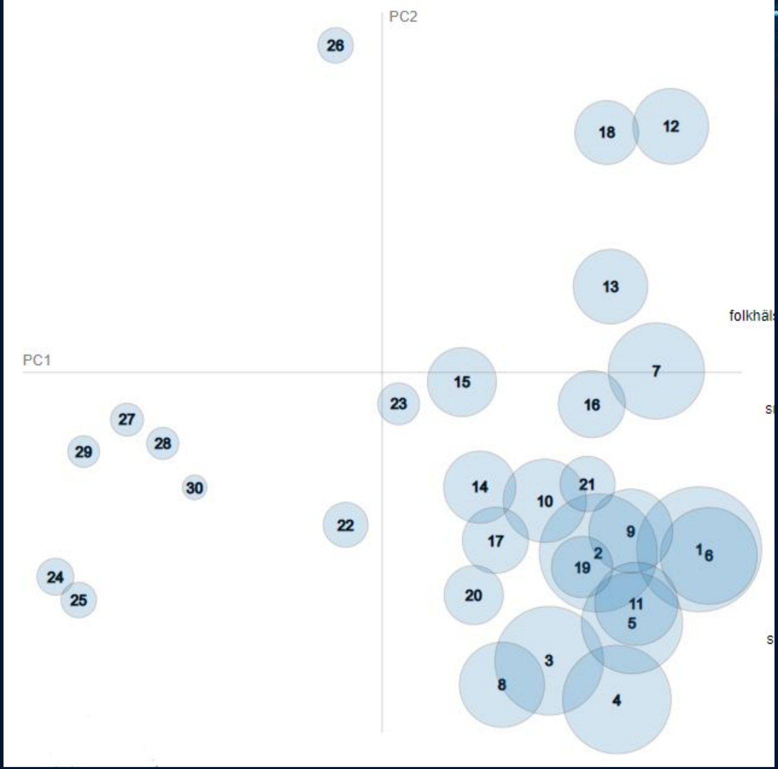
Gensim is a FREE Python library

Topic modelling for humans

- ✓ Train large-scale semantic NLP models
- ✓ Represent text as semantic vectors
- ✓ Find semantically related documents

Choosing Number of Topics





Dataset

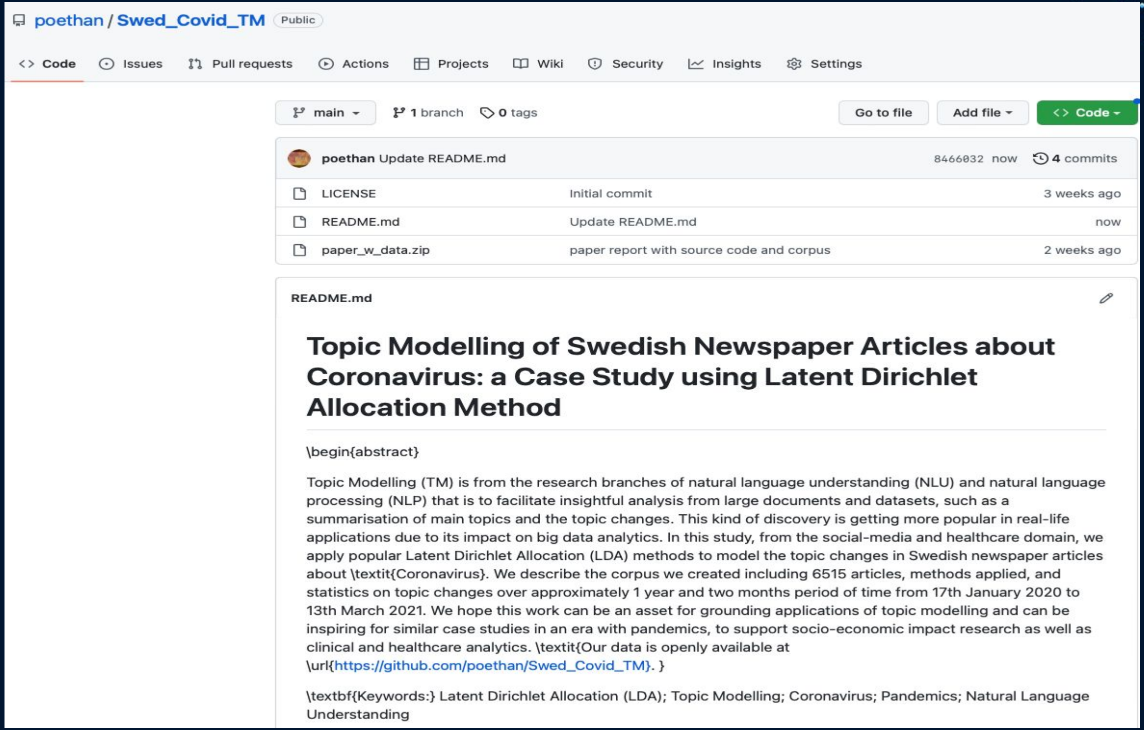
Newspaper articles having Covid as one of the main topics

SVT (Sveriges Television) - Sweden's national public television broadcaster

2 251 article

One year from the first article: 2020/01/17 - 2021/01/17

Dataset: how to get it?



poethan Update README.md 8466032 now 4 commits

LICENSE	Initial commit	3 weeks ago
README.md	Update README.md	now
paper_w_data.zip	paper report with source code and corpus	2 weeks ago

README.md

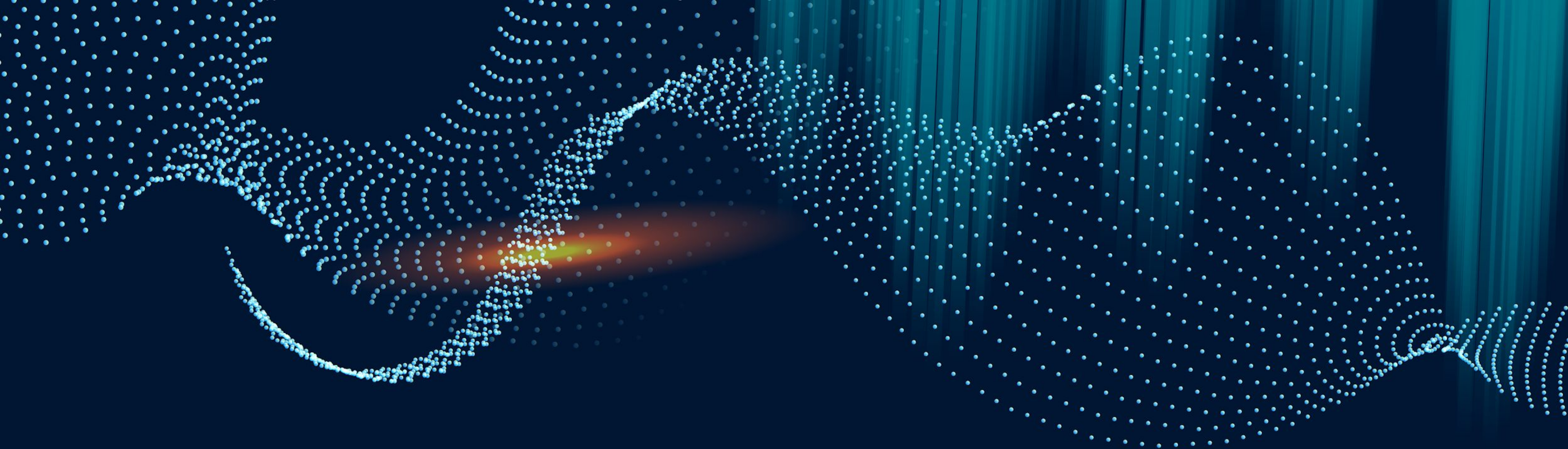
Topic Modelling of Swedish Newspaper Articles about Coronavirus: a Case Study using Latent Dirichlet Allocation Method

\begin{abstract}

Topic Modelling (TM) is from the research branches of natural language understanding (NLU) and natural language processing (NLP) that is to facilitate insightful analysis from large documents and datasets, such as a summarisation of main topics and the topic changes. This kind of discovery is getting more popular in real-life applications due to its impact on big data analytics. In this study, from the social-media and healthcare domain, we apply popular Latent Dirichlet Allocation (LDA) methods to model the topic changes in Swedish newspaper articles about \textit{Coronavirus}. We describe the corpus we created including 6515 articles, methods applied, and statistics on topic changes over approximately 1 year and two months period of time from 17th January 2020 to 13th March 2021. We hope this work can be an asset for grounding applications of topic modelling and can be inspiring for similar case studies in an era with pandemics, to support socio-economic impact research as well as clinical and healthcare analytics. \textit{Our data is openly available at \url{https://github.com/poethan/Swed_Covid_TM}. }

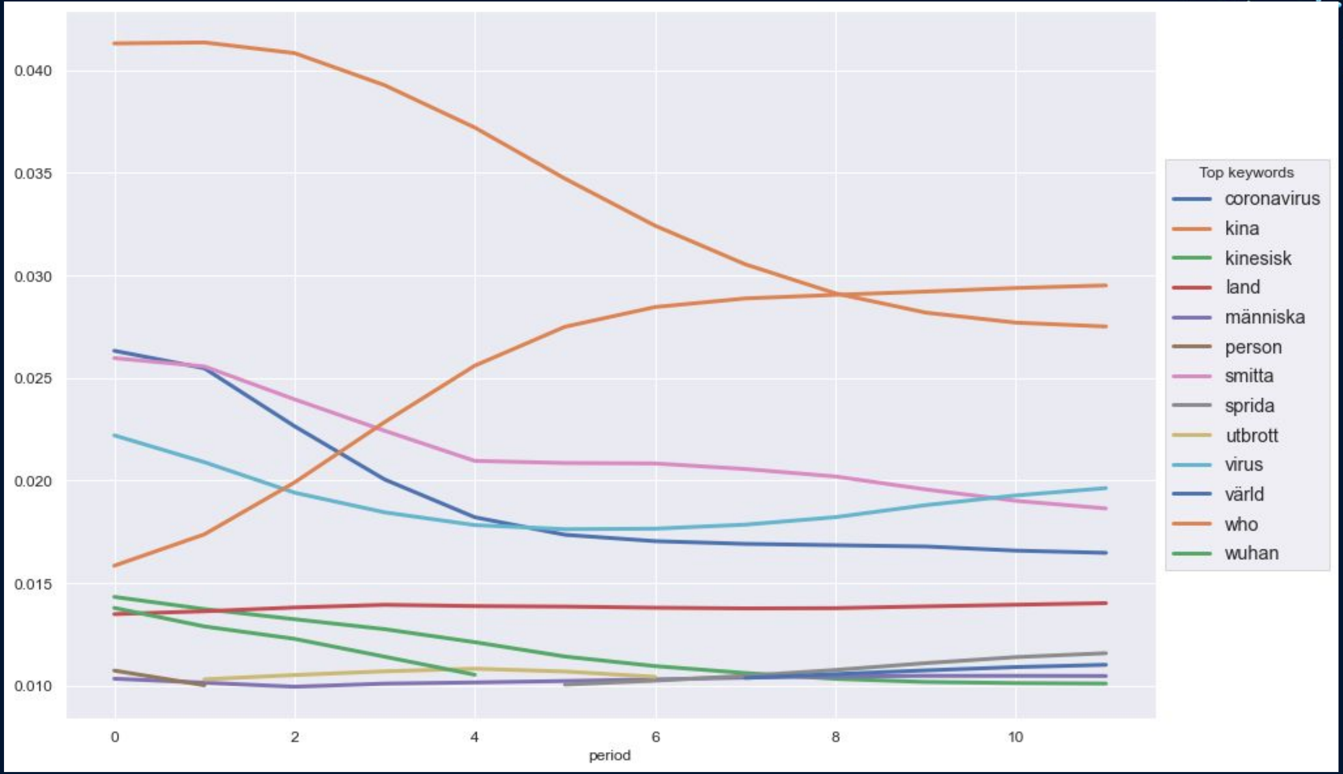
\textbf{Keywords:} Latent Dirichlet Allocation (LDA); Topic Modelling; Coronavirus; Pandemics; Natural Language Understanding

https://github.com/poethan/Swed_Covid_TM



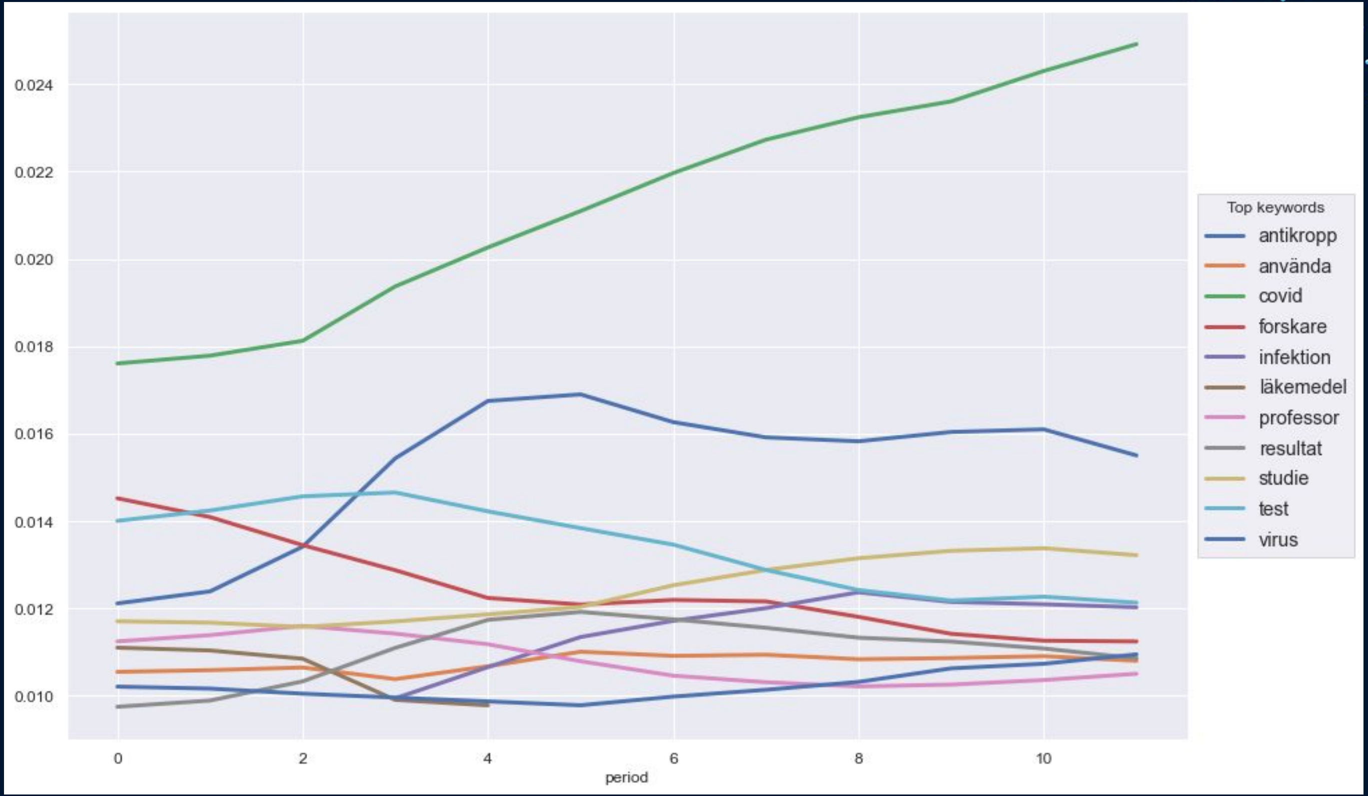
Findings

From Local to Global Issue



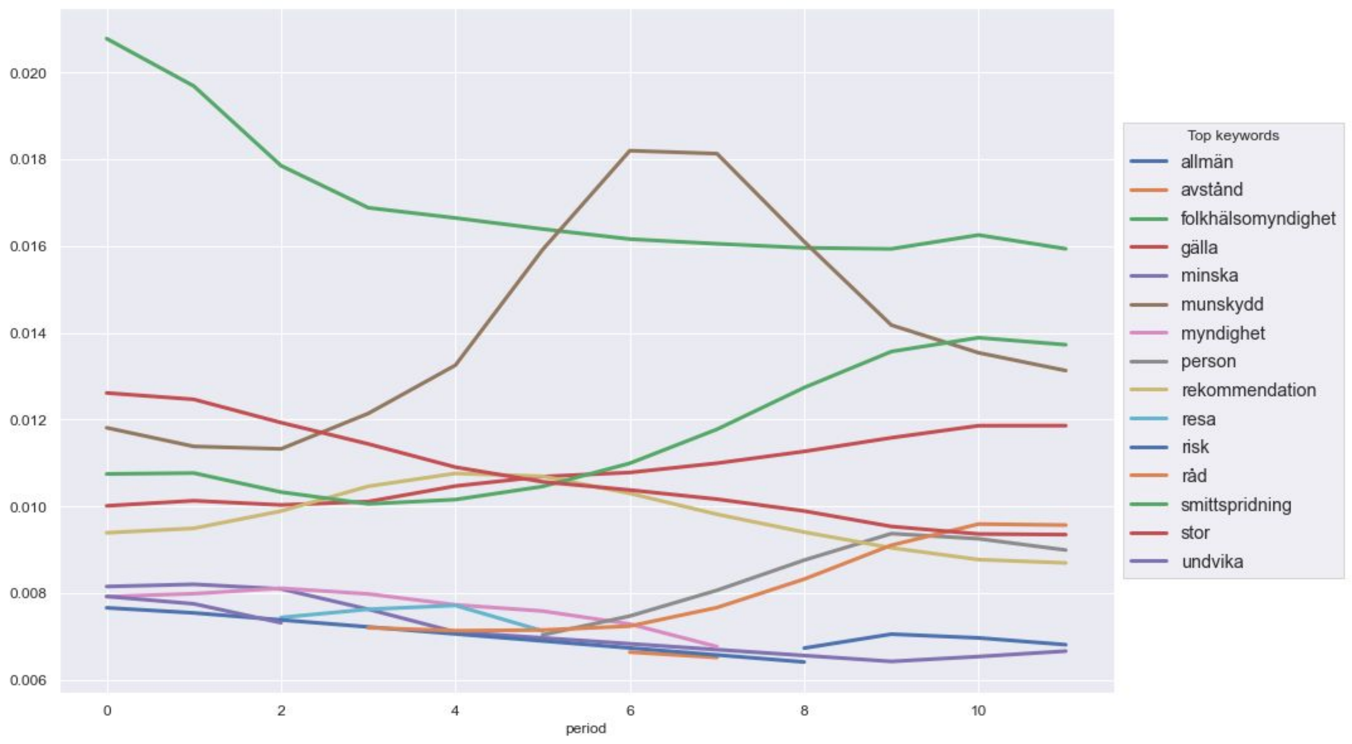
Coronavirus
China
Chinese
country
person
people
infect
spread
outbreak
virus
world
WHO
Wuhan

Factual Information



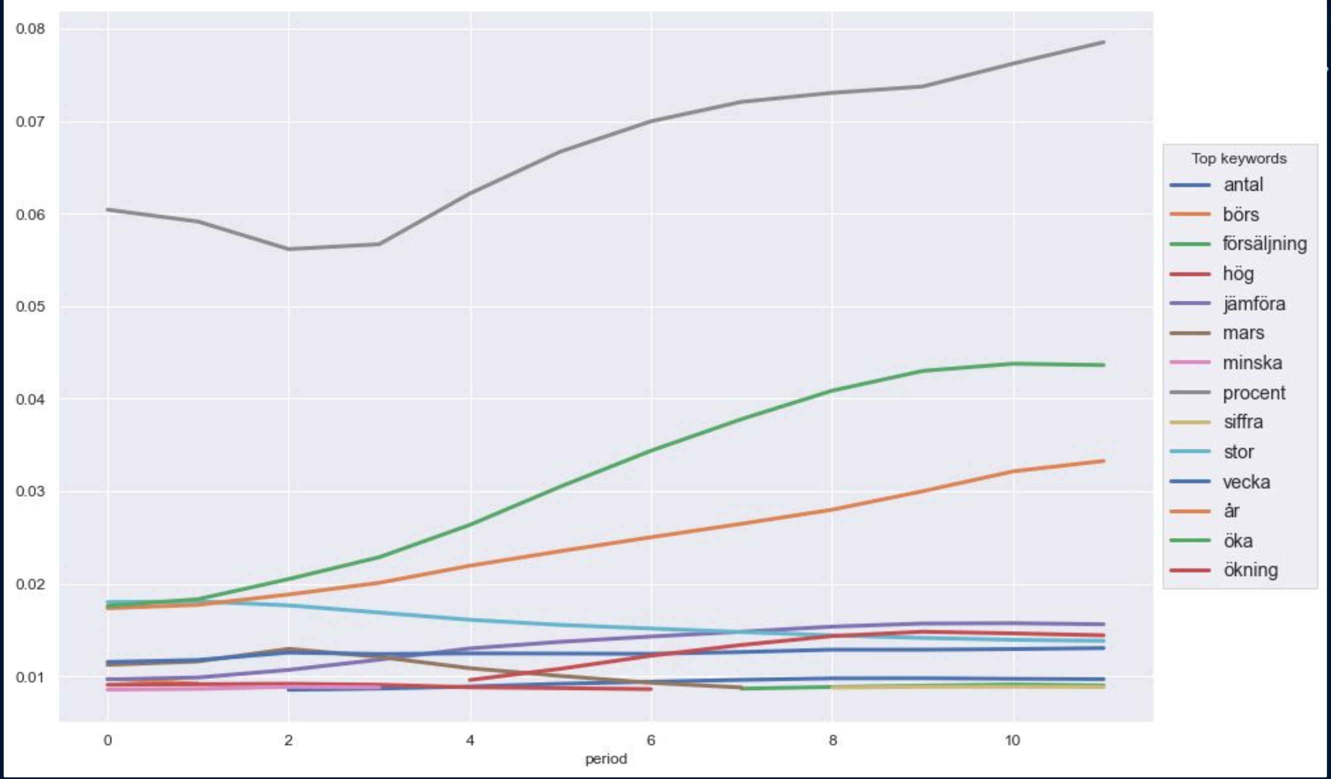
antibodies
use
Covid
researcher
infection
medicine
professor
results
study
test
virus

From Recommendations to Advices



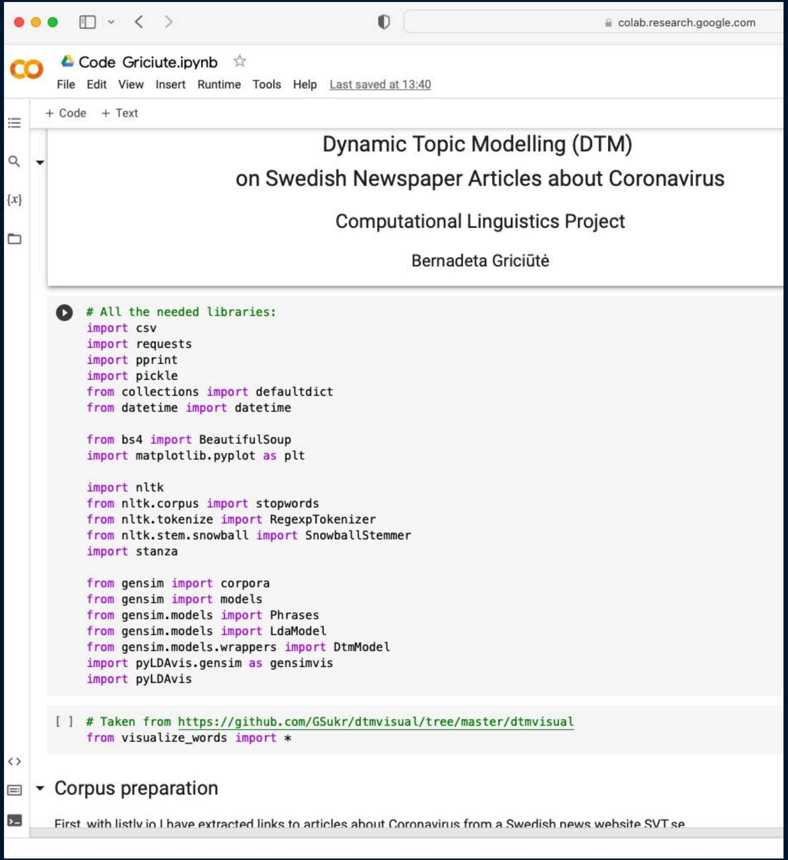
general
distance
Public Health Agency
apply to
decrease
face mask
authority
person
recommendation
travel
risk
advice
spread of infection
big
avoid

Consequences for Economy



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

Data with code:


colab.research.google.com
CO
Code_Griciute.ipynb
File Edit View Insert Runtime Tools Help Last saved at 13:40
+ Code + Text

Dynamic Topic Modelling (DTM)

on Swedish Newspaper Articles about Coronavirus

Computational Linguistics Project

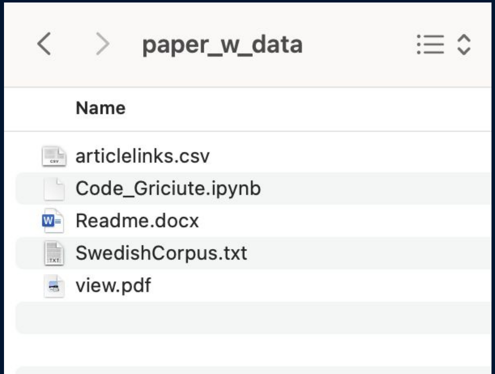
Bernadeta Griciūtė

```
# All the needed libraries:
import csv
import requests
import pprint
import pickle
from collections import defaultdict
from datetime import datetime

from bs4 import BeautifulSoup
import matplotlib.pyplot as plt

import nltk
from nltk.corpus import stopwords
from nltk.tokenize import RegexpTokenizer
from nltk.stem.snowball import SnowballStemmer
import stanza

from gensim import corpora
from gensim import models
from gensim.models import Phrases
from gensim.models import LdaModel
from gensim.models.wrappers import DtmModel
import pyLDAvis.gensim as gensimvis
import pyLDAvis
```

```
[ ] # Taken from https://github.com/GSukr/dtmvisual/tree/master/dtmvisual
    from visualize_words import *
```

Corpus preparation

First, with listly.io I have extracted links to articles about Coronavirus from a Swedish news website SVT.se



articlelinks.csv
Code_Griciute.ipynb
Readme.docx
SwedishCorpus.txt
view.pdf

https://github.com/poethan/Swed_Covid_TM

Future Work

Expand the data frame of the data

Use more varied data sources

Compare to other languages

Thank you!

Time for questions and suggestions :)