Topic modelling of scientific articles

an example using CORD-19 database

UNIGE's Data Science Day

September 2021

Erol Orel

Semi-automated literature search engine

Text Mining

- **Web Scrapping:** Retrieve papers with available metadata and full text from the web
  Currently: arXiv, biorXiv, JSTOR, medXriv, Paperity, PubMed, PubMedCentral

- **Text Processing:** Clean, tokenize, lemmatize, stop words, etc…

➢ Quantify frequencies: From text to numerical matrix

Topic Modelling

- **Clustering:** Reduce dimension and classify the articles by similar topic

- **Extract Information:** Define topics' subjects and relevant information from each

Case Study: COVID-19 transmission at school

CORD-19: The Covid-19 Open Research Dataset

On March 16, 2020, the Allen Institute for AI (AI2), in collaboration with The White House Office of Science and Technology Policy (OSTP), the National Library of Medicine (NLM), the Chan Zuckerburg Initiative (CZI), Microsoft Research, and Kaggle, coordinated by Georgetown University's Center for Security and Emerging Technology (CSET), released the first version.

« COVID-19 » AND « Children »

- Publication date > 2019-07-01

- Keywords:

```
searchfor3 = ['children',
              'child',
              'infant',
              'infants',
              'pediatric',
              'paediatric',
              'nursery',
              'kindergarten',
              'kid',
              'kids',
              'toddler',
              'toddlers',
              'adolescent',
              'adolescents',
              'babies',
              'baby']
```

➢ **Result of the search: 7'290 articles with full text**

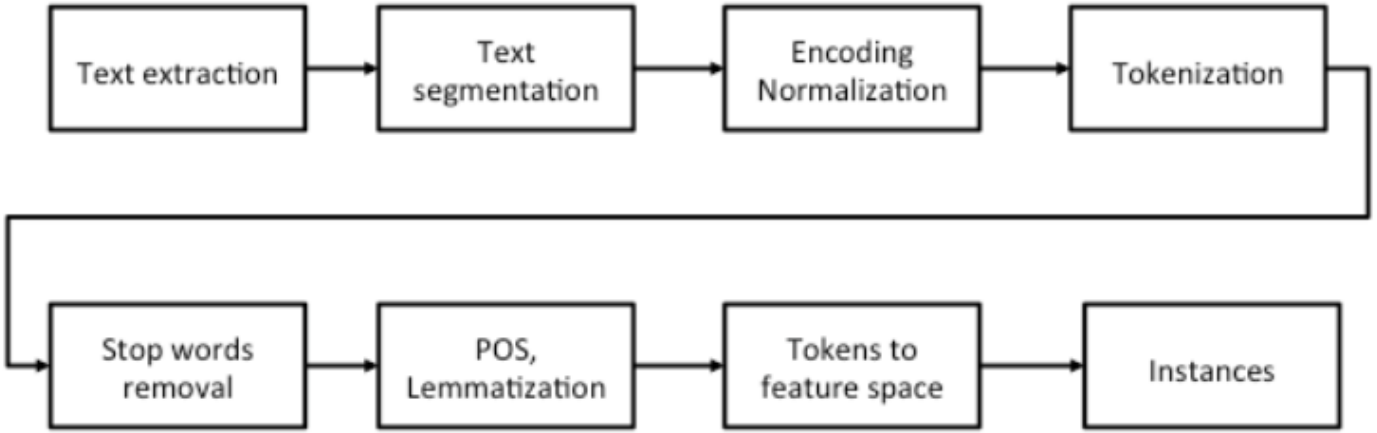
CORD-19 Database

[2]:

	Title	Date	DOI	PMCID	Authors	Abstract	Full_text	URL	Journal	cord_uid
0	Caring for Pediatric Patients with Diabetes am...	2020-05-05	10.1016/j.jpeds.2020.04.067	PMC7199676	Cindy, Ho; Beng Hui, Ng Nicholas; Seng, Lee Yung		Coronavirus disease 2019 (COVID-19) caused by ...	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7...	J Pediatr	kd7g7v3k
1	Response to "Visualization of Putative Coronav...	2020-05-16	10.1016/j.kint.2020.05.007	PMC7229727	Su, Hua; Gao, Ding; Yang, Hai-Chun; Fogo, Agne...		Dear Editors, We have carefully read and consi...	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7...	Kidney Int	15a2avvk
2	Correspondence regarding recently published ed...	2020-05-19	10.1183/13993003.01601-2020	PMC7241110	Ebmeier, Stefan; Cunnington, Aubrey J.	When individuals without prior immunity are co...	authors proposed a number of possible reasons ...	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7...	Eur Respir J	bfq66x1c
3	Addendum to: Risk Stratification and Personal ...	2020-04-22	10.1097/mpg.0000000000002762	PMC7273941	Say, Daphne S.; de Lorimier, Arthur; Lammers, ...		1. We continue to perform all previously sched...	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7...	J Pediatr Gastroenterol Nutr	eb2guuz4
4	Pediatric Coronavirus Disease-2019–Associated ...	2020-05-22	10.1093/jpids/piaa062	PMC7313948	Shulman, Stanford T		In the waning days of April 2020, reports from...	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7...	J Pediatric Infect Dis Soc	9tawlbms

```
Current dataframe shape (7290, 12)
```

Natural Language Processing



Instances: TF-IDF: How important a word is to a document in a corpus

- **Occurrence vs Importance**

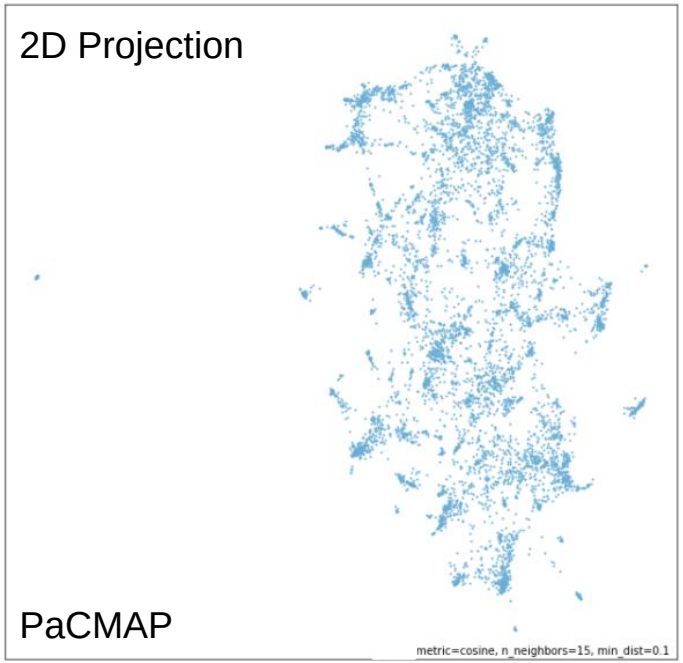
TF-IDF Matrix

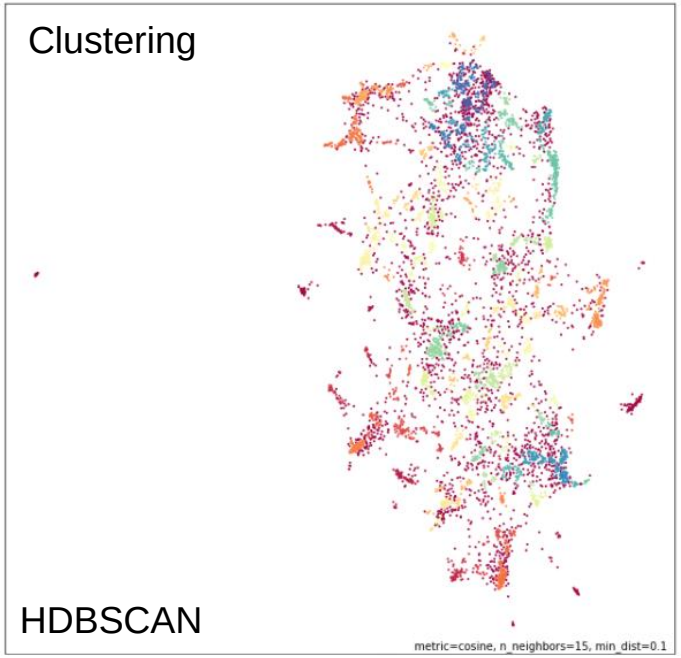
[4]:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
mining_quarrying	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.00000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
apologize_confusion	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.00000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
wetland	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.00000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
chocolate_factory	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.00000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
importer	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.00000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
parent	0.050197	0.000000	0.0	0.000000	0.000000	0.041008	0.00000	0.025075	0.0	0.000000	0.000000	0.000000	0.000000	0.065661	0.162708
school	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.00000	0.000000	0.0	0.000000	0.012412	0.000000	0.000000	0.089294	0.217016
aki	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.00000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
pediatric	0.024848	0.000000	0.0	0.029884	0.100708	0.040599	0.00000	0.024825	0.0	0.071896	0.012048	0.072717	0.036635	0.000000	0.012391
cell	0.023527	0.052658	0.0	0.000000	0.000000	0.000000	0.14258	0.070517	0.0	0.000000	0.045629	0.005738	0.000000	0.000000	0.000000

35843 rows × 7147 columns

Dimension Reduction and Clustering



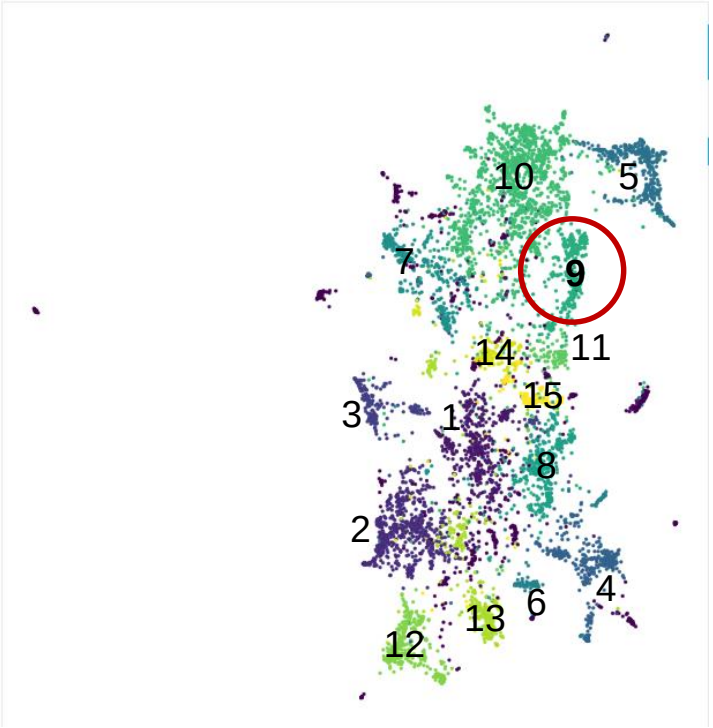


“COVID-19” AND “Children”

Up until 13th of April 2021: 7’241 articles

15 Topics: Most relevant keywords

Cluster 1: mortality, icu, model, admission, hypertension, cohort, factor, chronic
Cluster 2: cell, protein, expression, viral, lung, tissue, receptor, human, immune, cov
Cluster 3: vaccine, vaccination, influenza, country, immunization, coverage, dose
Cluster 4: mis, syndrome, acute, inflammatory, cardiac, cell, pediatric, fever, kawasaki
Cluster 5: woman, mother, pregnant, pregnancy, birth, maternal, delivery, milk
Cluster 6: brain, neurological, cell, seizure, acute, viral, syndrome, blood, epilepsy, stroke
Cluster 7: surgery, pediatric, surgical, procedure, emergency, mask, paediatric
Cluster 8: pediatric, pneumonia, asymptomatic, chest, lung, fever, viral, respiratory
Cluster 9: school, student, contact, transmission, parent, education, social, teacher
Cluster 10: parent, family, social, work, health, mental, anxiety, home, anxiety, service
Cluster 11: household, contact, transmission, index, estimate, sar, asymptomatic
Cluster 12: aki, kidney, renal, dialysis, mortality, acute, injury
Cluster 13: transplant, trial, kidney, dose, recipient, drug, therapy
Cluster 14: license, perpetuity, holder, post, funder, display, international
Cluster 15: sample, antibody, viral, assay, post, license, testing



Extract Relevant Information

379 articles in cluster 9 or only **141** when relevant articles are given

- Check for correct classification

- Extract relevant information from the short-listed articles

➢ **Done manually by the researchers**

or

➢ **Citizen Science**

Support and promote the collaboration of academic scientists and the general public (“citizens”)

UNIVERSITÉ DE GENÈVE

FACULTÉ DE MÉDECINE
Institut de santé globale

Thanks !