Masters Thesis Overview

AI-Powered COVID-19 Research Cockpit Presentation

Supplement – CORD-19 Initial Analysis (W.I.P.)

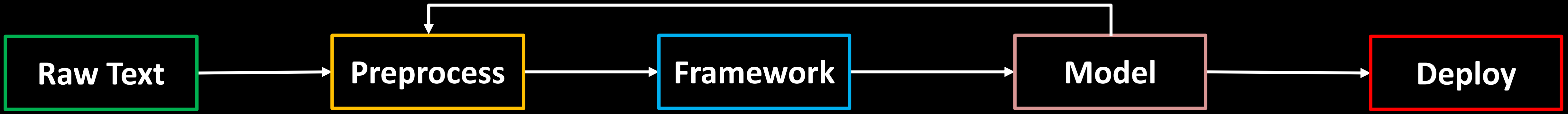
Jon-Paul Boyd

3rd September 2020

DE MONTFORT UNIVERSITY LEICESTER

Pipeline From Development To Deployment

Goal help medical researchers find answers to their questions from literature, for example searching for better treatments & policy decisions. Interactive analysis of texts with data driven and NLP methods supported by AI techniques.



Raw Text

Preprocess

Framework

Model

Deploy

Metadata.csv
title, author, affiliations, publish date, links to full body text

PMC_JSON & PDF_JSON
full body text including abstract, results, conclusions

* Only PMC is peer-reviewed

# Original Metadata Articles
237207*

# Original PMC Articles
72916

# Original PDF Articles
99336

Cleansing
Comparison titles/body text/abstract (PMC preferred as peer-reviewed)

Duplicate Articles
5

Documents Per Language

en	95796
de	1107
es	803
fr	652
it	351

95796 recs remaining after cleansing

* Difference indicates articles with abstracts only ☹

Unsupervised LDA (Latent Dirichlet Allocation)*
Topic Modelling With Ktrain

* Other modelling approaches to be tested & benchmarked, including variety of LDA implementations & alternatives e.g. lda2vec

Serverless Backend

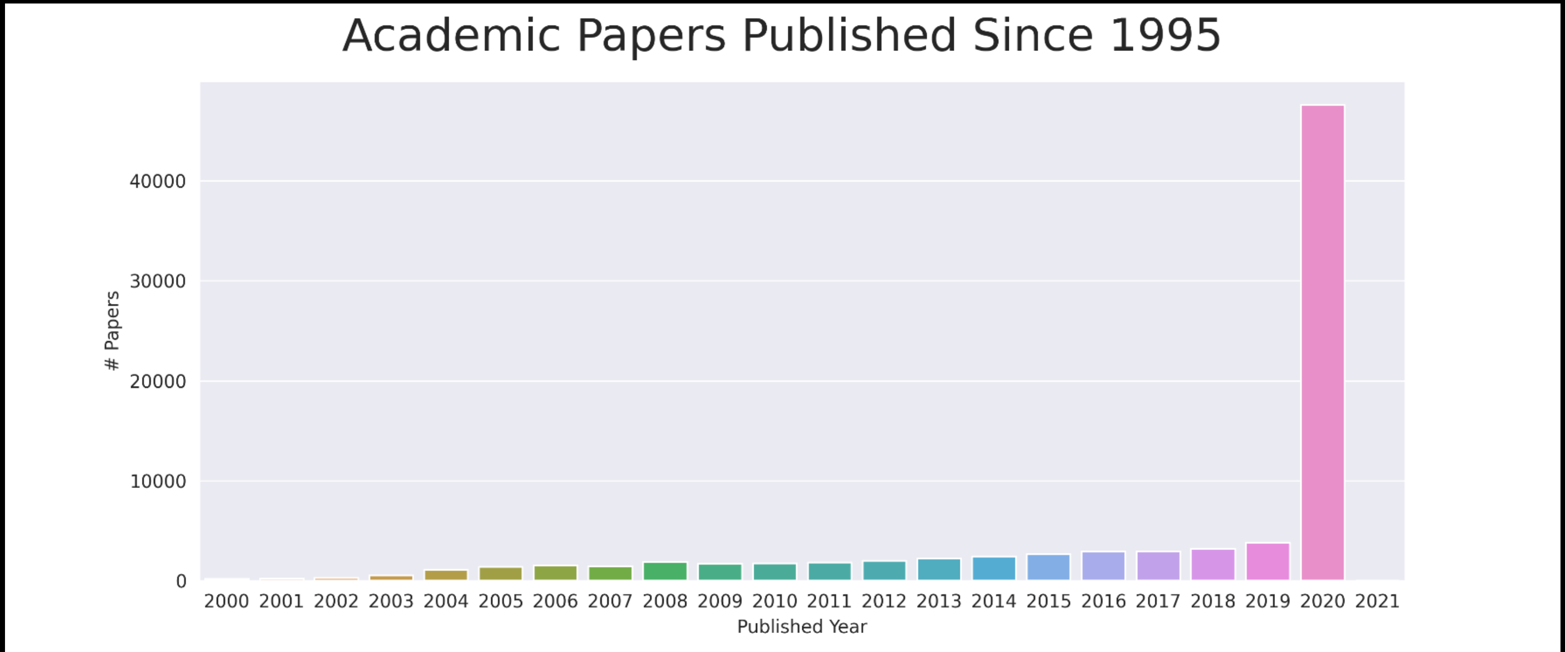
FaaS (Functions As A Service)

NoSQL Database

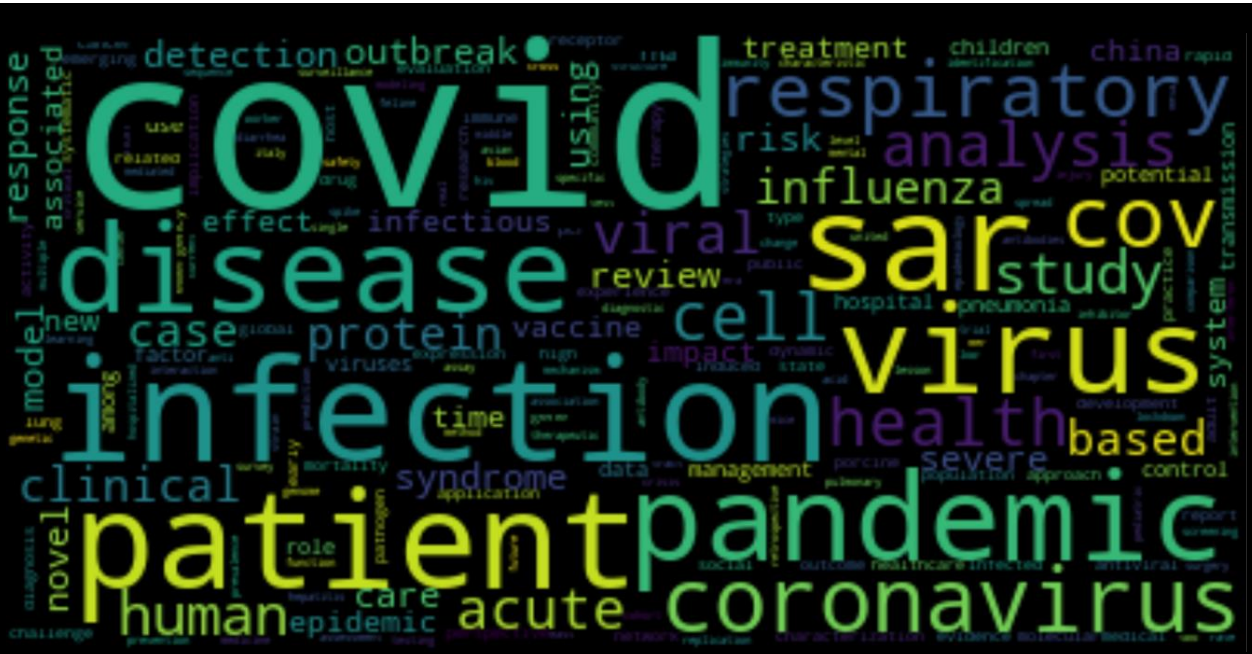
Web Server

* Estimated cloud costs (5K USD) total to May 2020, to be refined following further prototyping

Data Analysis Coronavirus Publication Explosion & Word Significance



Number of coronavirus-related academic publications increased tremendously in 2020

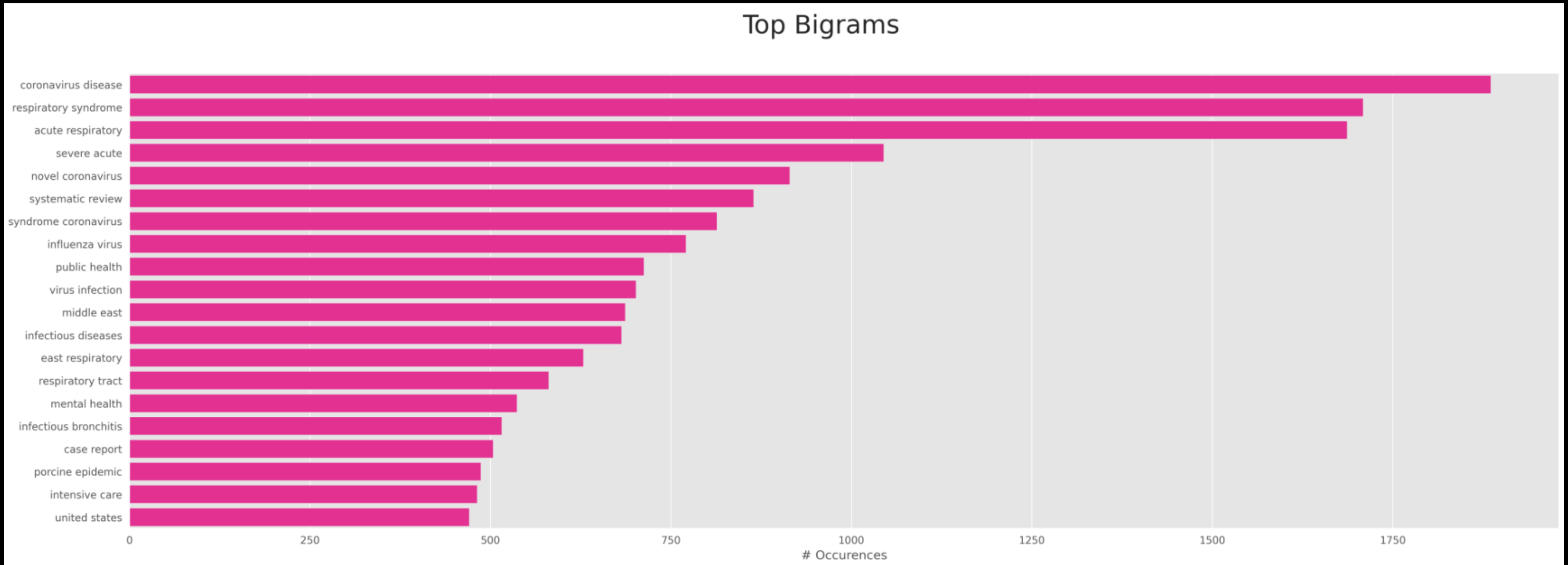


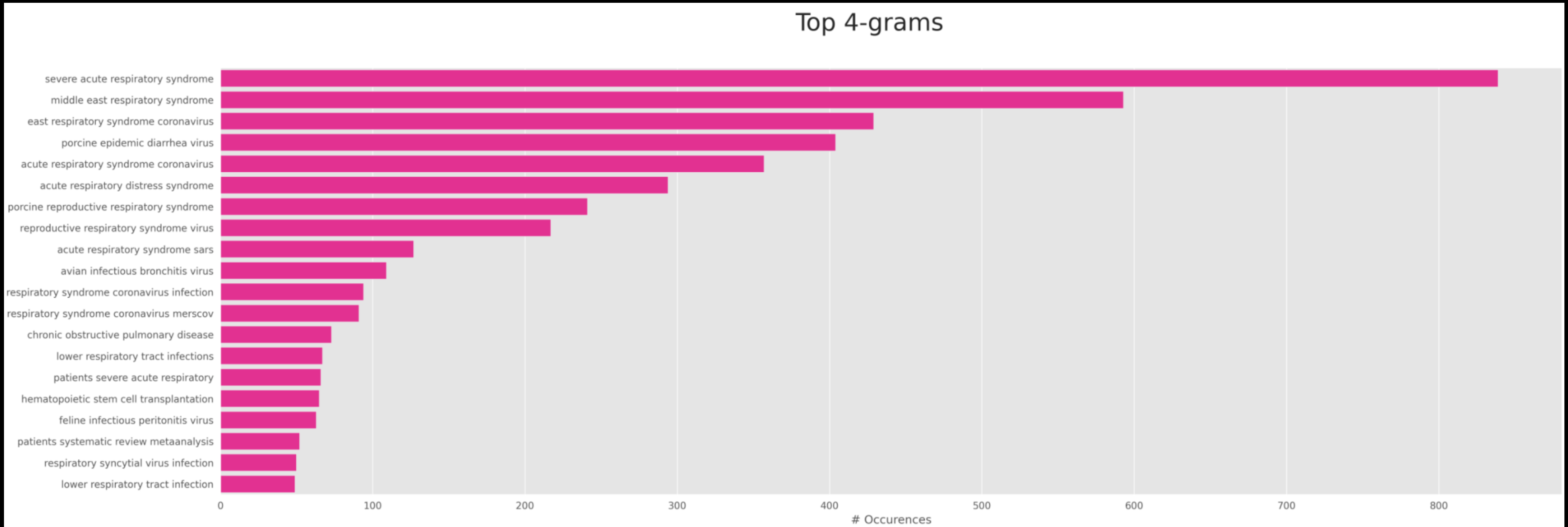
Word cloud representing word frequency over 95K articles

Data Analysis N-Gram Multiword Token Examples

Retain words appearing frequently enough together to include as complete entities.

Can be used for keyword token search, topic association, document categorisation & filtering.

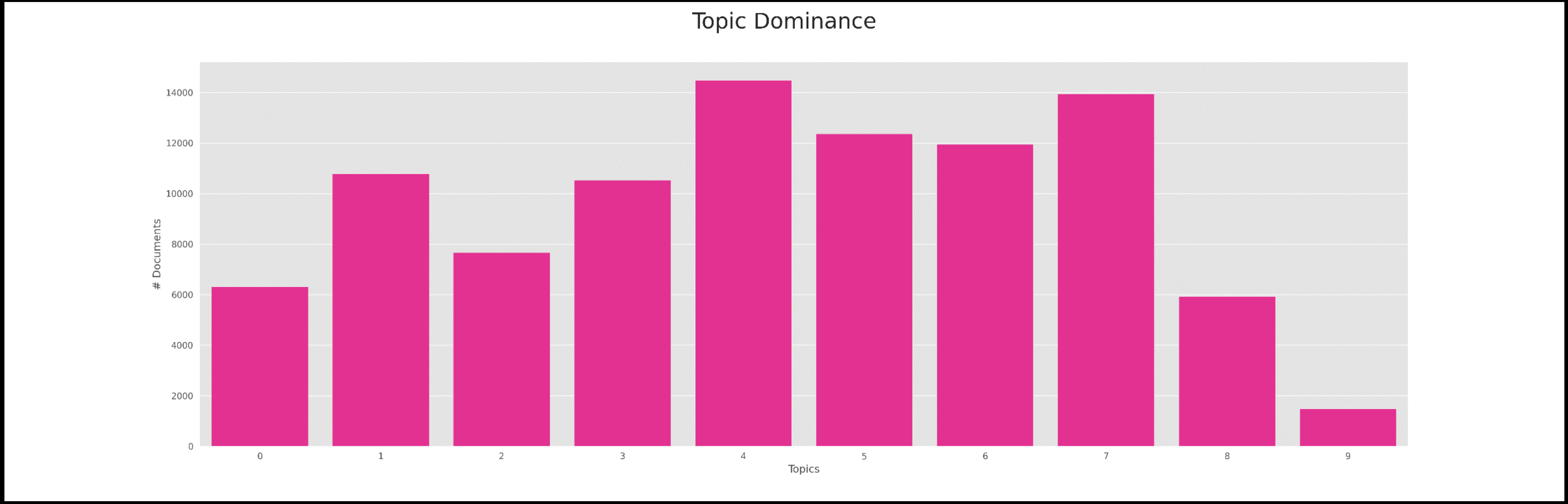




Topic Modelling **Dominance** *

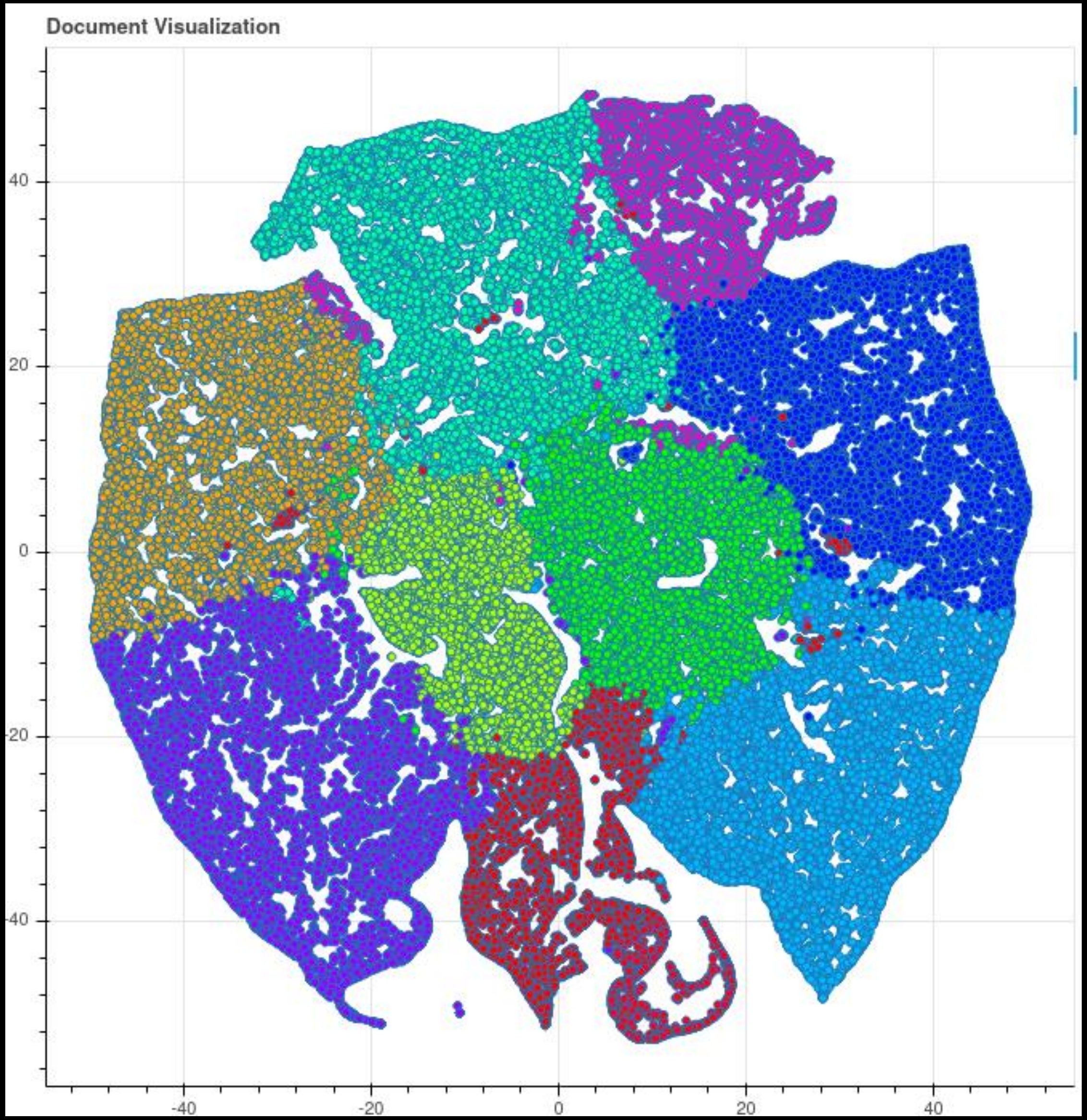
Topic modelling is **fully automated** & **extracts underlying semantic structure of documents** to help shape their meaning. Each topic is a collection of words. **Words belonging to topics have weights**, so can be used to infer dominance & possibly a single subfield of study a given article can be associated with e.g. virology, immunology, genetics etc. Topic models **can be trained on any problem domain corpus of text**. Once trained, **a TM can topic classify a new document without further training**. In this example, training the TM took 37 minutes, processing 95K CORD-19 full article texts on a 6-core desktop CPU.

Topic 4	Topic 7	Topic 5	Topic 6	Topic 1	Topic 3	Topic 2	Topic 0	Topic 8	Topic 9
14483 docs	13949 docs	12362 docs	11946 docs	10781 docs	10534 docs	7661 docs	6314 docs	5924 docs	1490 docs
covid-19	research	protein	cells	care	viruses	preprint	model	blood	use
children	social	rna	cell	covid-19	samples	covid-19	models	group	package
symptoms	public	proteins	expression	patient	vaccine	license	fig	patient	document
age	people	fig	mice	medical	species	medrxiv	set	therapy	end
risk	information	binding	immune	pandemic	influenza	population	method	lung	begin
sars-cov-2	work	sequence	response	risk	strains	model	network	liver	minimal
pneumonia	countries	activity	protein	hospital	animals	epidemic	information	normal	amssymb
group	global	acid	levels	healthcare	infections	infected	values	diagnosis	amsfonts
days	risk	dna	fig	testing	detected	doi	methods	cancer	documentclass
mortality	development	sequences	activation	staff	infected	copyright	value	years	amsmath
									amsbsy



* W.I.P. Additional stopword & lemmatization work required to remove & group words (e.g. remove fig, group cell/cells & protein/proteins). Also consider including n-grams e.g. blood group/liver cancer/lung cancer

Topic Modelling Visualisation



PCA (Principal Component Analysis)

There are 10 topics. Each is represented by a different colour.

Each dot is a corpus document, coloured to indicate the assigned topic.

PCA reduces the 10-dimensional topic word vectors to 2 dimensions to support visualisation.

This dimensionality reduction facilitates measuring distance between document vectors, for example by using Euclidean geometry, to indicate a given document's nearest neighbours, even if they belong to a different topic.

It means a search engine can enhance a query result hitlist by including additional neighbouring documents of a document included in search results, even if that neighbour does not include the query keywords e.g. include nearby red topic documents when orange topic documents rank highly in search results (grid ref x -30, y +5)

This helps support discovery and new insights.

Topic Modelling Recommender Example

Ktrain topic modelling also provides a recommendation engine that can recommend documents from the corpus that are semantically related to the question being asked. It surfaces scored documents from each topic with index allowing retrieval of original text.

Question

What do we know about therapeutics, interventions, and clinical studies?

Answer

--- Topic id 1 (care covid-19 patient medical pandemic risk hospital healthcare testing staff)

Number of topic docs 10781
Most relevant document index 32074
Score 0.9737986869576298
Text: The source of infection is a person with SARS-CoV-2 infection. The number of patients attending healthcare facilities should be minimized by (1) advising persons with mild symptoms to be tested safely and then isolate, monitor their condition, and only seek in-person care if symptoms worsen; and (2) using telemedicine to provide care for patients whose medical needs can be addressed remotely. For

--- Topic id 3 (viruses samples vaccine species influenza strains animals infections detected infected)

Number of topic docs 10534
Most relevant document index 79628
Score 0.981071403441446
Text: Wildlife diseases may represent a potential threat not only to local wildlife populations but also to domestic animals and humans. Various studies have been carried out to analyse the prevalence of pathogens in wild boar populations and the role of these populations as reservoir for pathogens or a source of infection for domestic pigs (Kaden et al. 2009). Wild boar (Sus scrofa) populations are fou

Further Thoughts

Q. How do you see the potential of NLP-based systems in helping your medical experts?

How successful are any of your current solutions in helping retrieve the right papers or answering questions correctly?

Q. What user experience are you looking for?

For example, a simple search bar with document retrieval ordered from most to least relevant? Or something further supported with information extraction and visualisation to explain ranking and generate insight? How important is including a snippet of text to directly answer the query?

Q. What unstructured data sources do you have access to?

Internal/external? Format? Volumes? Datasets available now? Licenses to academic publication servers? Do you extract from social media?

Q. Would you recommend including non-peer reviewed sources?

Should there be a weighting to prefer reviewed sources?

Q. How best to assess paper quality?

This is a tough problem that is also dangerous where using WHO trials, citations, social media critique could introduce bias. Thoughts?

Q. How should such NLP-based systems be measured for accuracy?

In typical classification & regression, performance evaluation is straight-forward by calculating loss from the ground truth. Given number of dimensions in unstructured text, how best to measure the effectiveness of a Q&A system? Manual evaluation requires experts and very time consuming? A golden ranking benchmark still requires manual assembly. Existing benchmarks (e.g. Stanford Question Answering Dataset, Machine Reading Comprehension Dataset, General Language Understanding Evaluation) relate to different, non-medical domains.

Q. What are your thoughts on our relationship and expected deliverables going forward?