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.

PubMed Central

medRxiv THE PREPRINT SERVER FOR HEALTH SCIENCES

bioRxiv THE PREPRINT SERVER FOR BIOLOGY

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

PMC_JSON & PDF_JSON
title, authors, full body text with abstract, conclusions

* Only PMC peer-reviewed

 # Metadata Articles
237207*

 # PMC Articles
72916

 # 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

* Difference indicates articles with abstracts only ☹

python

Keras

scikit learn

Unsupervised LDA (Latent Dirichlet Allocation)
Topic Modelling With Ktrain

Azure

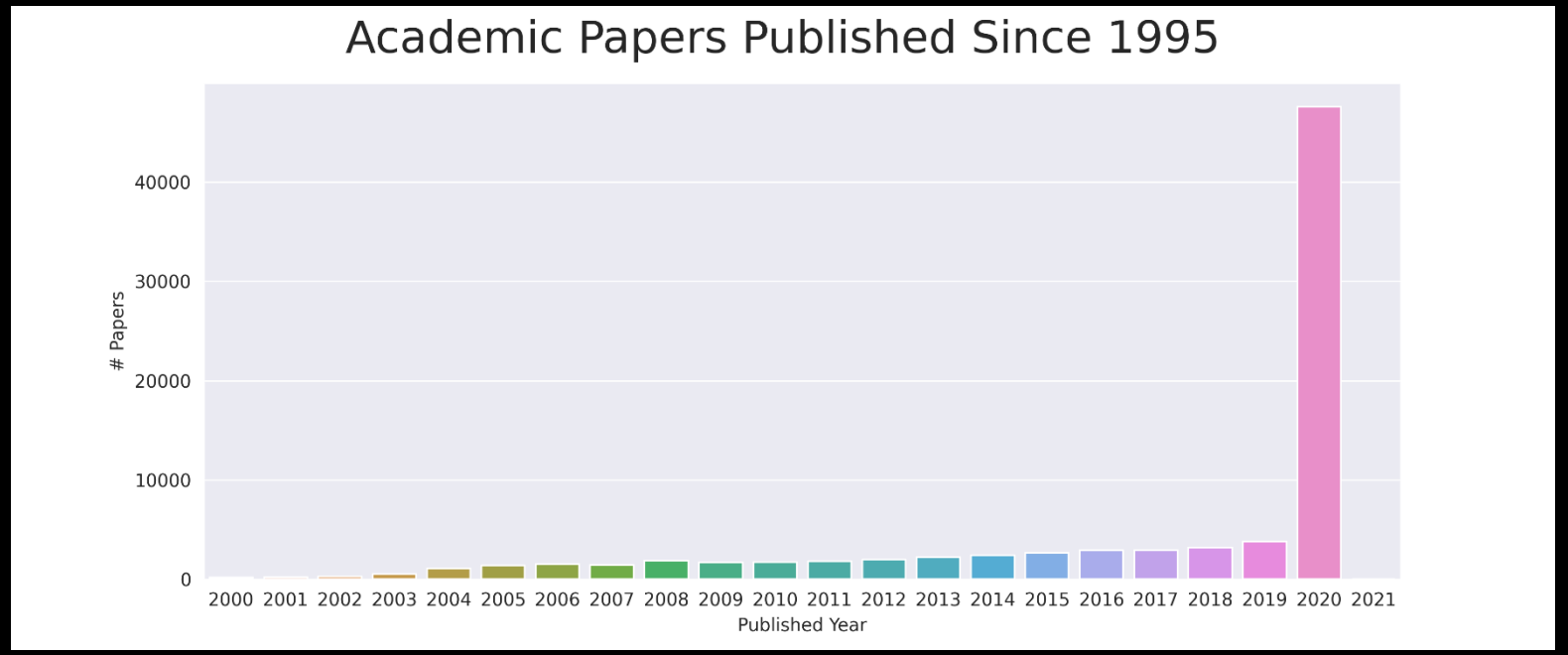
Serverless Backend

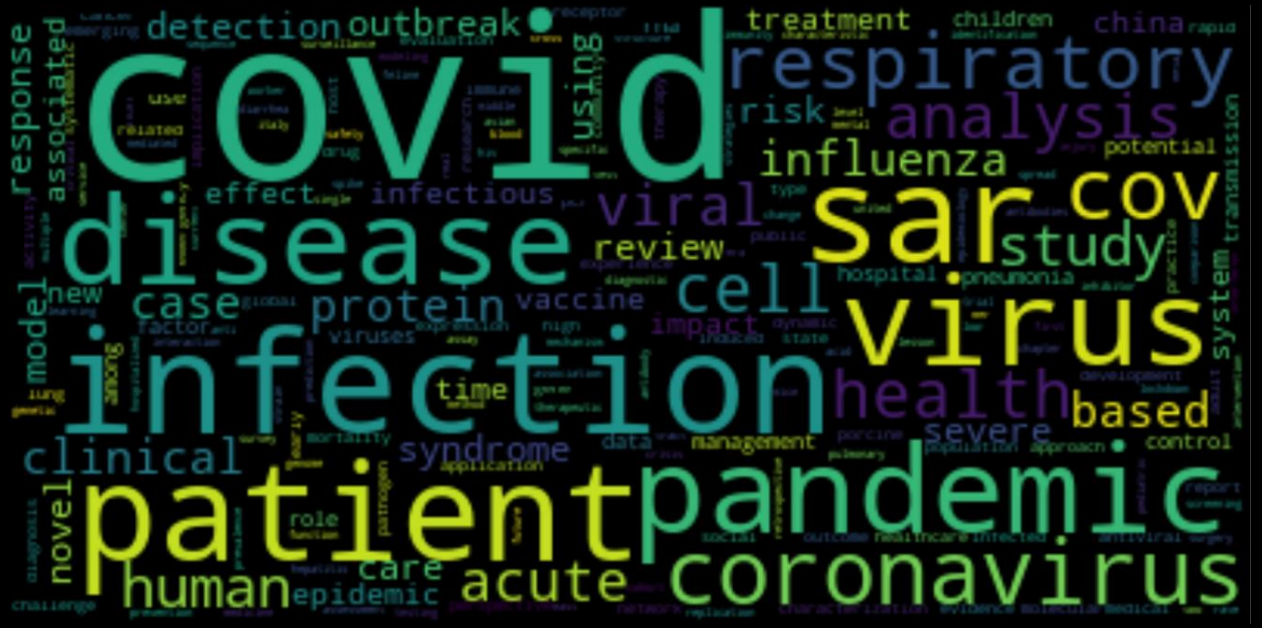
FaaS (Functions As A Service)

NoSQL Database

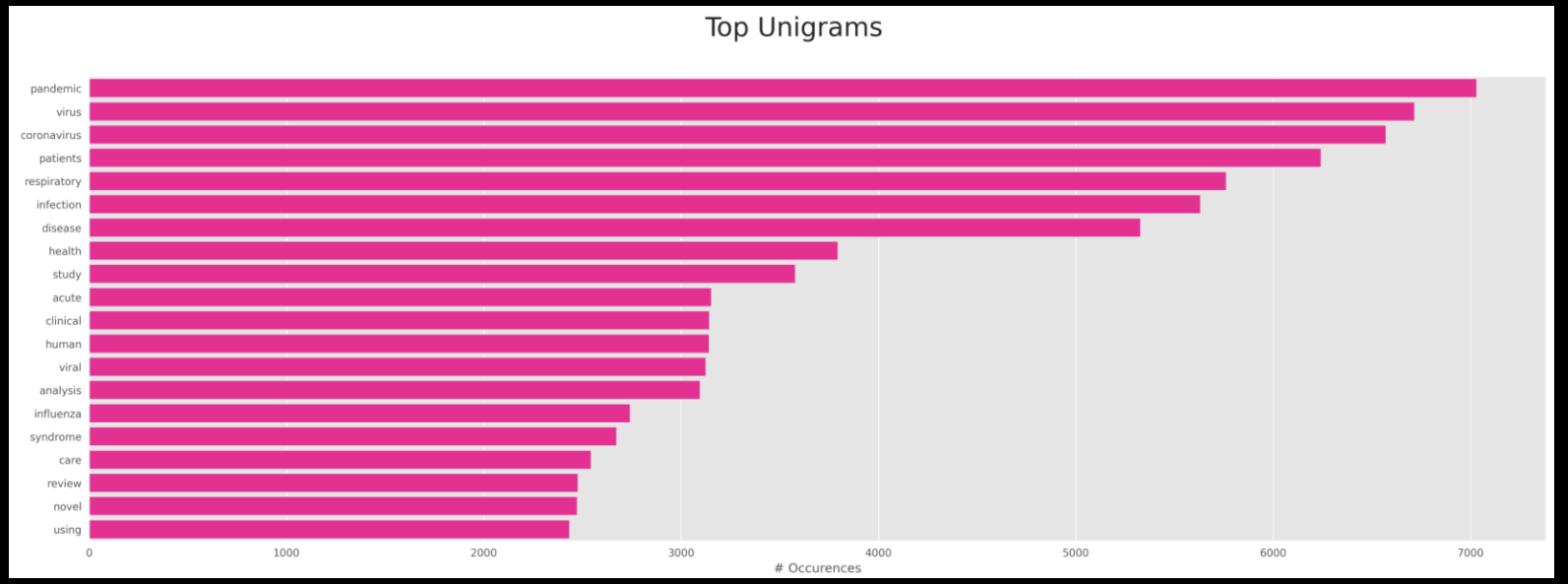
Web Server

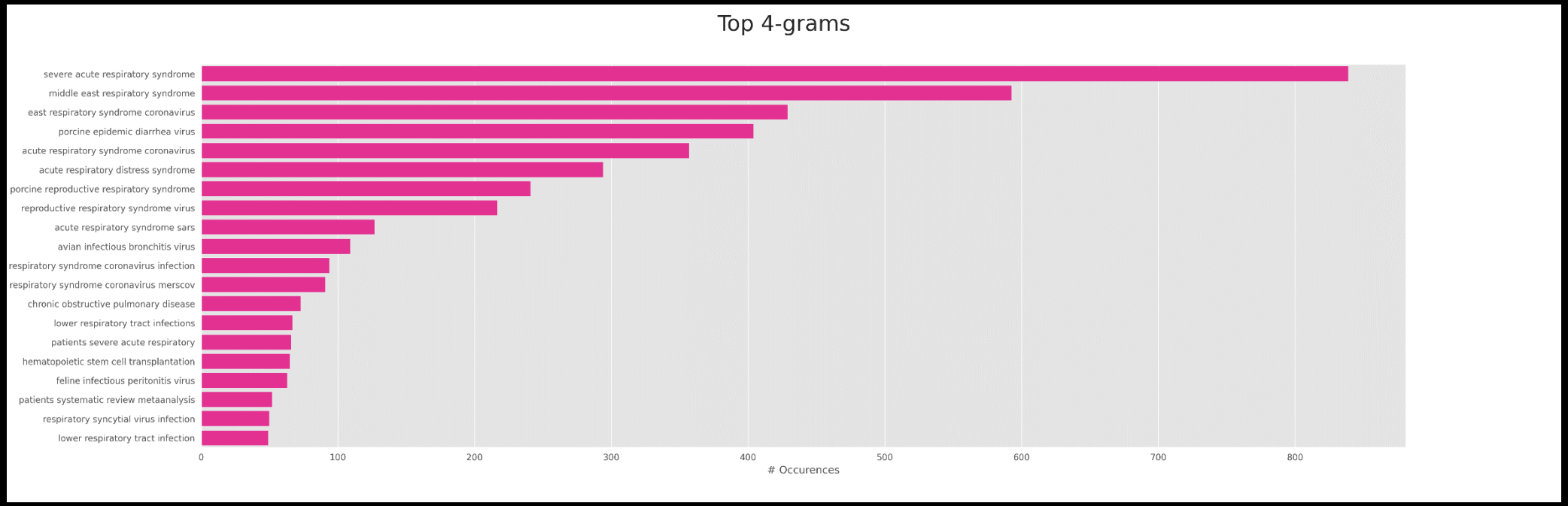
Data Analysis Coronavirus Publication Explosion & Word Significance





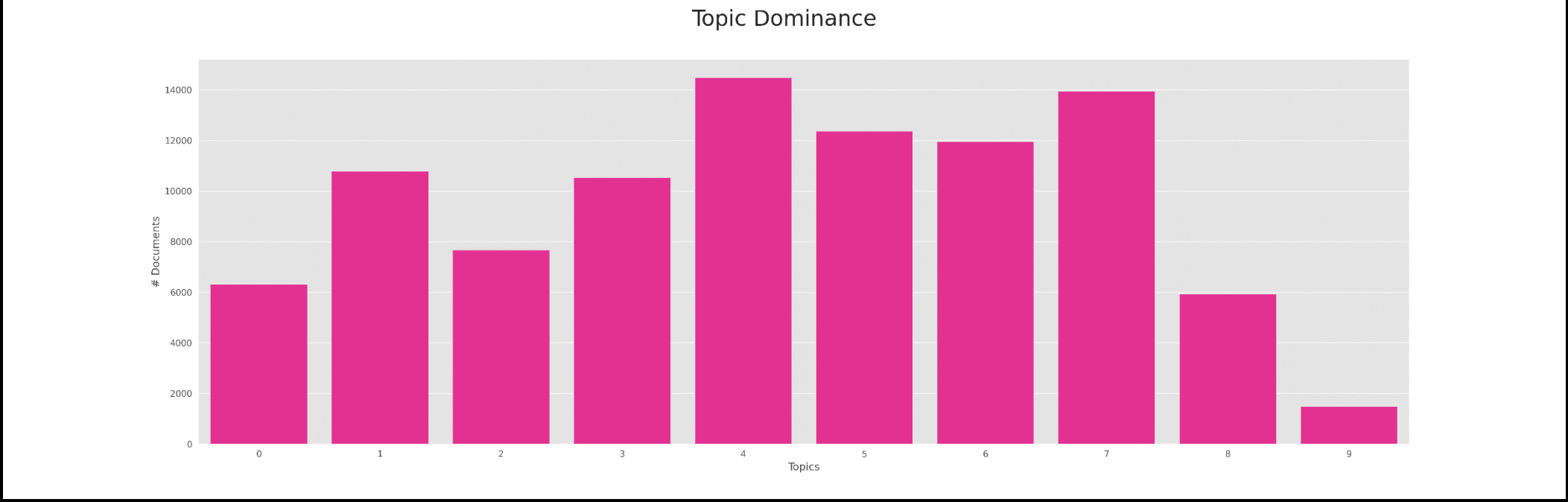
Data Analysis N-Gram Multiword Token Examples



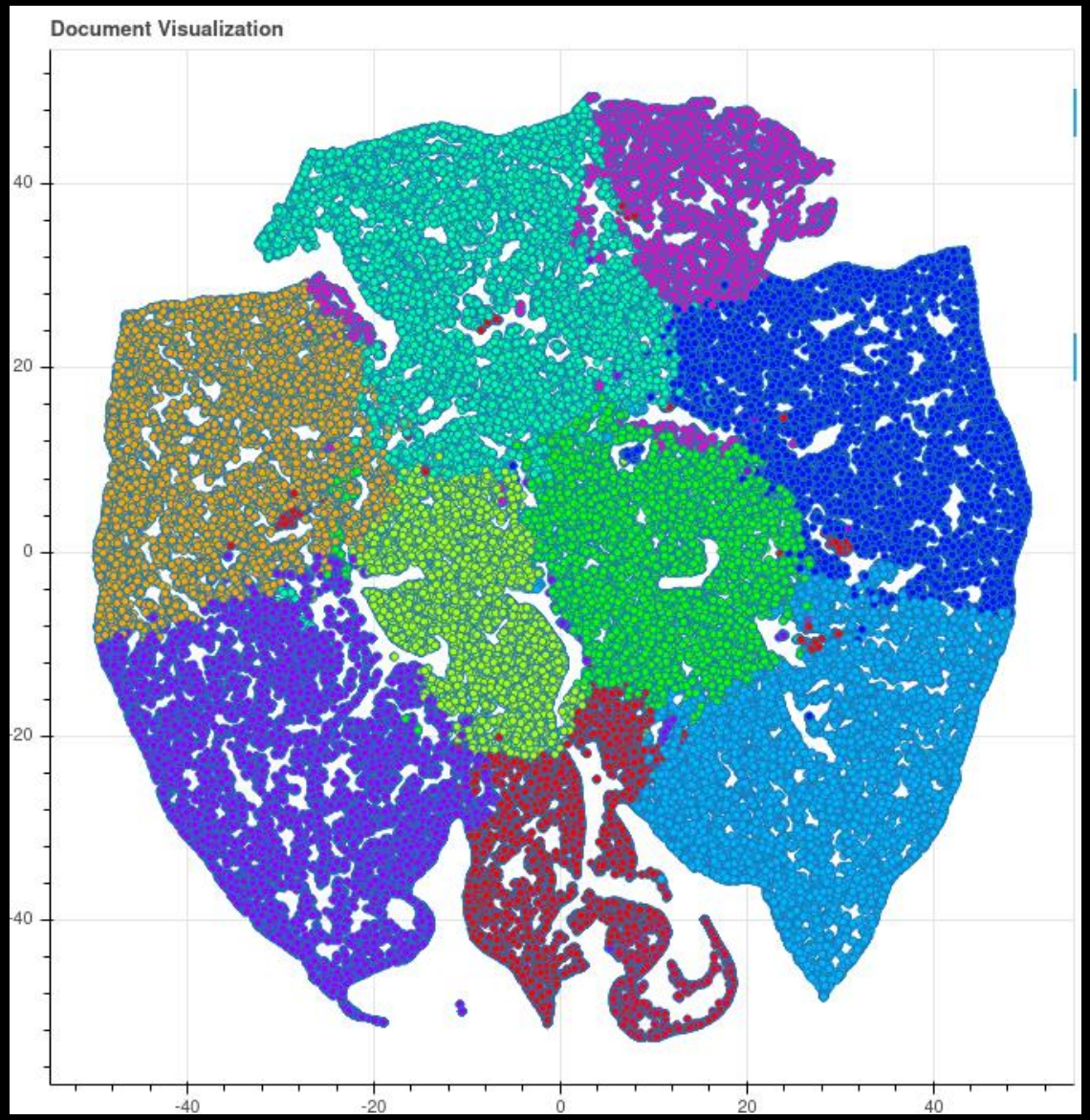


Topic Modelling Dominance

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



Topic Modelling Visualisation



PCA (Principal Component Analysis)

Reduces topic word vectors to 2 dimensions to visualize and facilitate measuring distance between feature vectors, for example by using Euclidean geometry

Topic Modelling Recommender

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 sources do you have access to?
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.