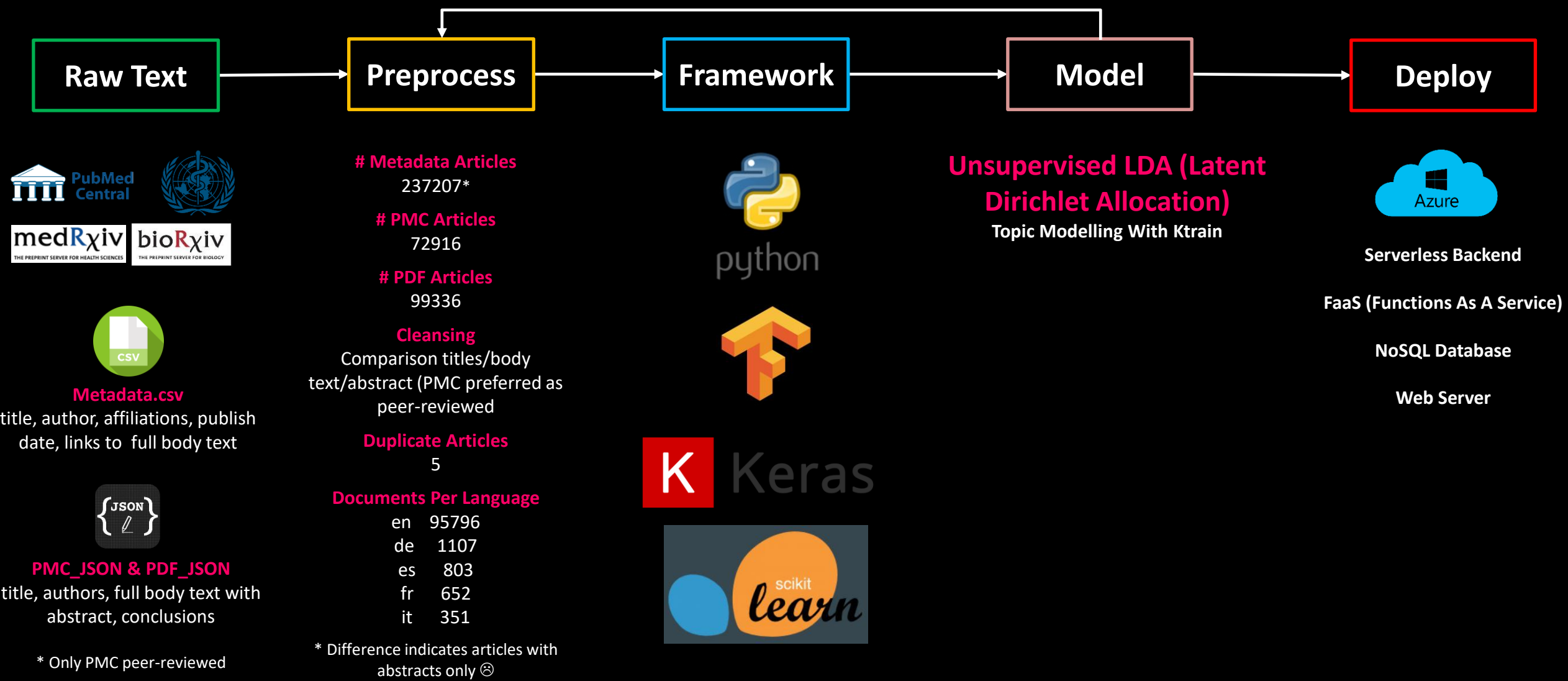
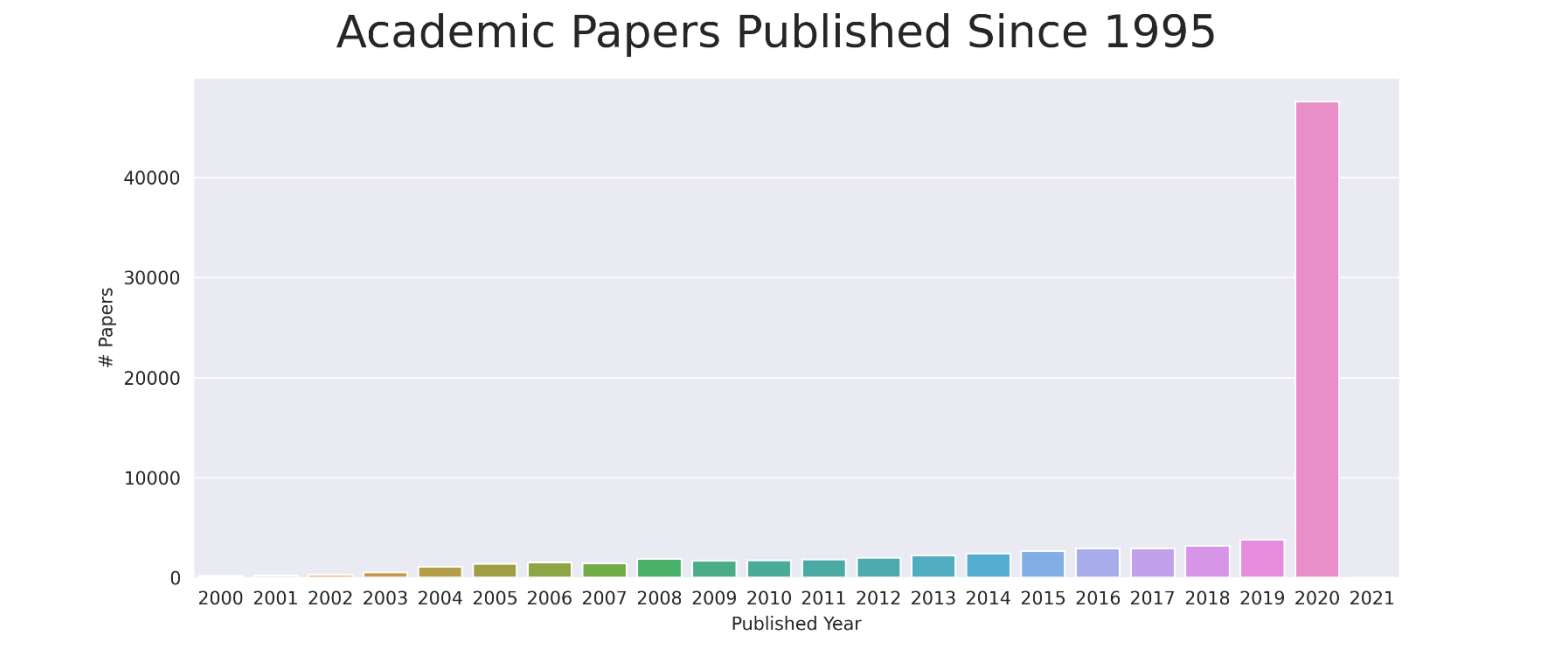


# Pipeline From Development To Deployment

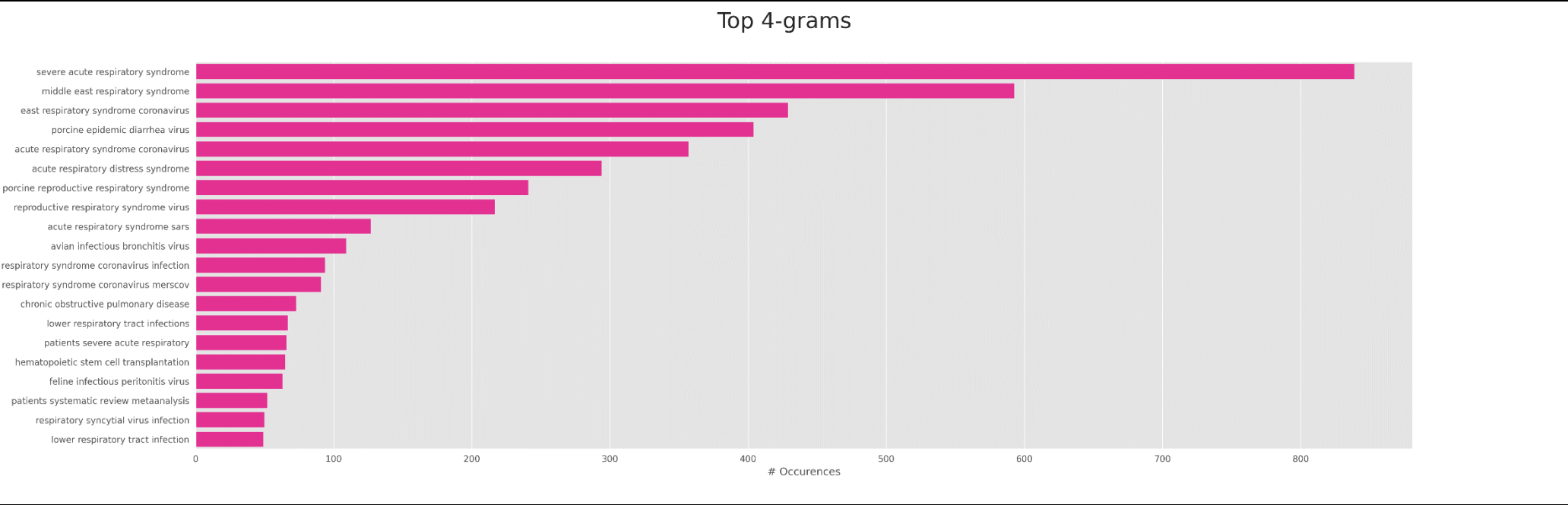
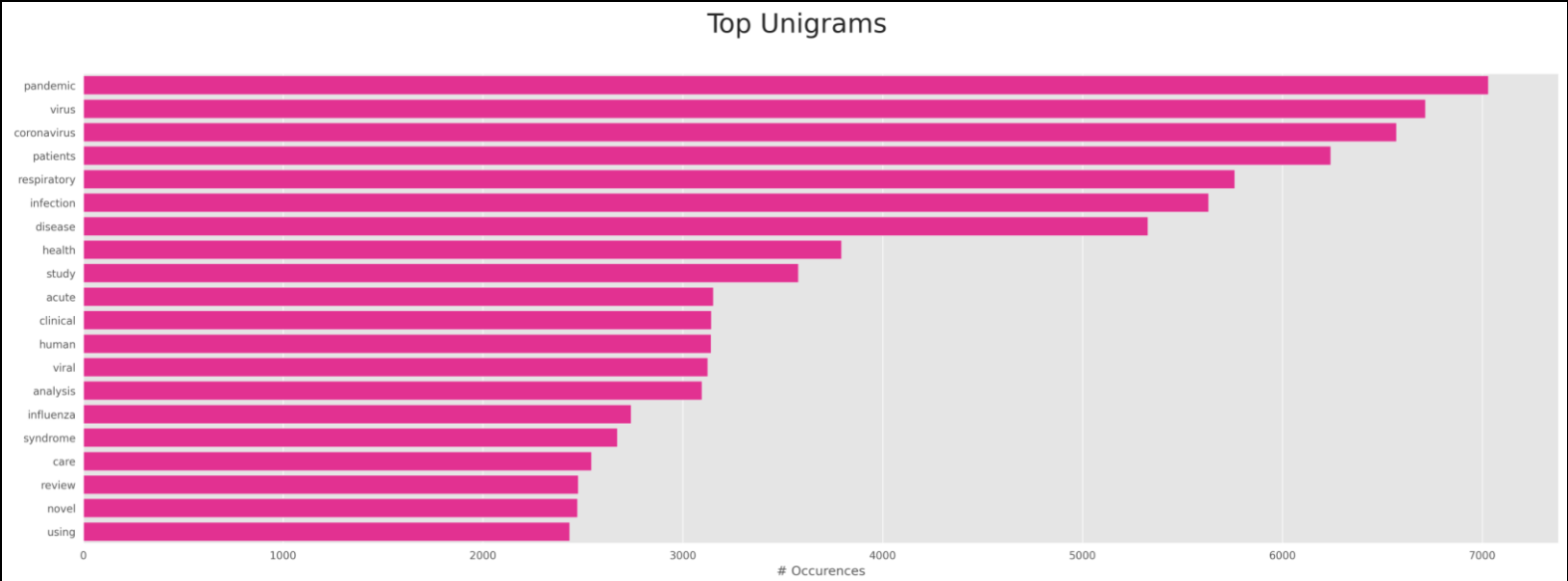
**Goal** help medical researchers find answers to their questions from literature, for example searching for better treatments & policy decisions. Interactive analysis of texts with data driven and NLP methods supported by AI techniques.



# Data Analysis Coronavirus Publication Explosion & Word Significance

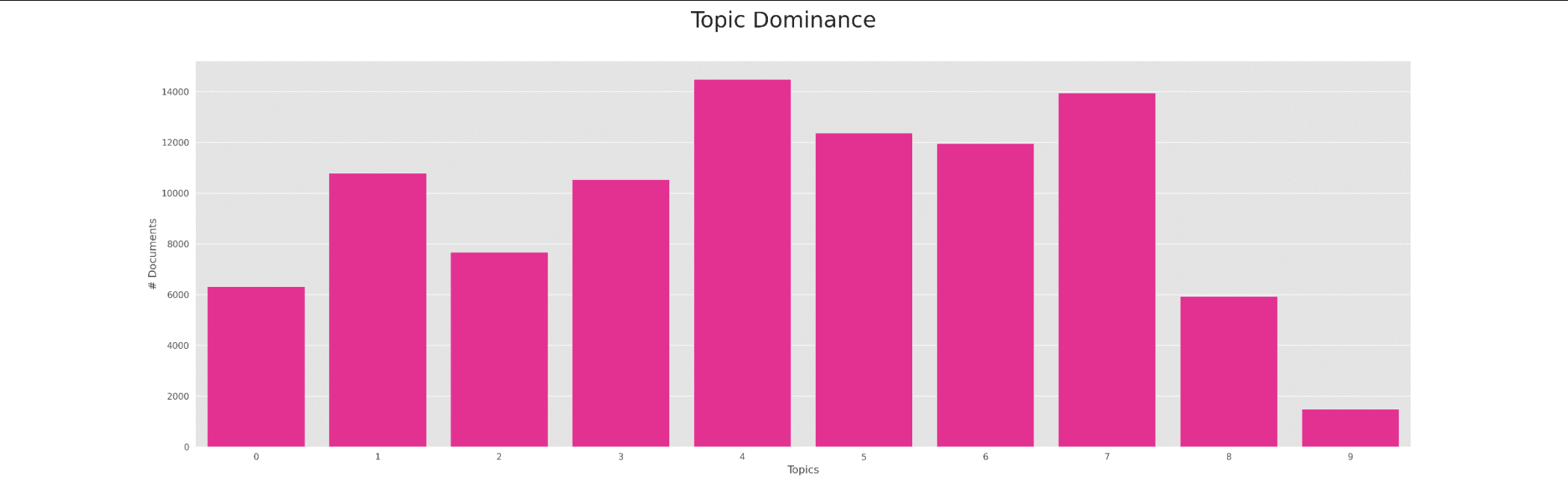


# Data Analysis N-Gram Examples



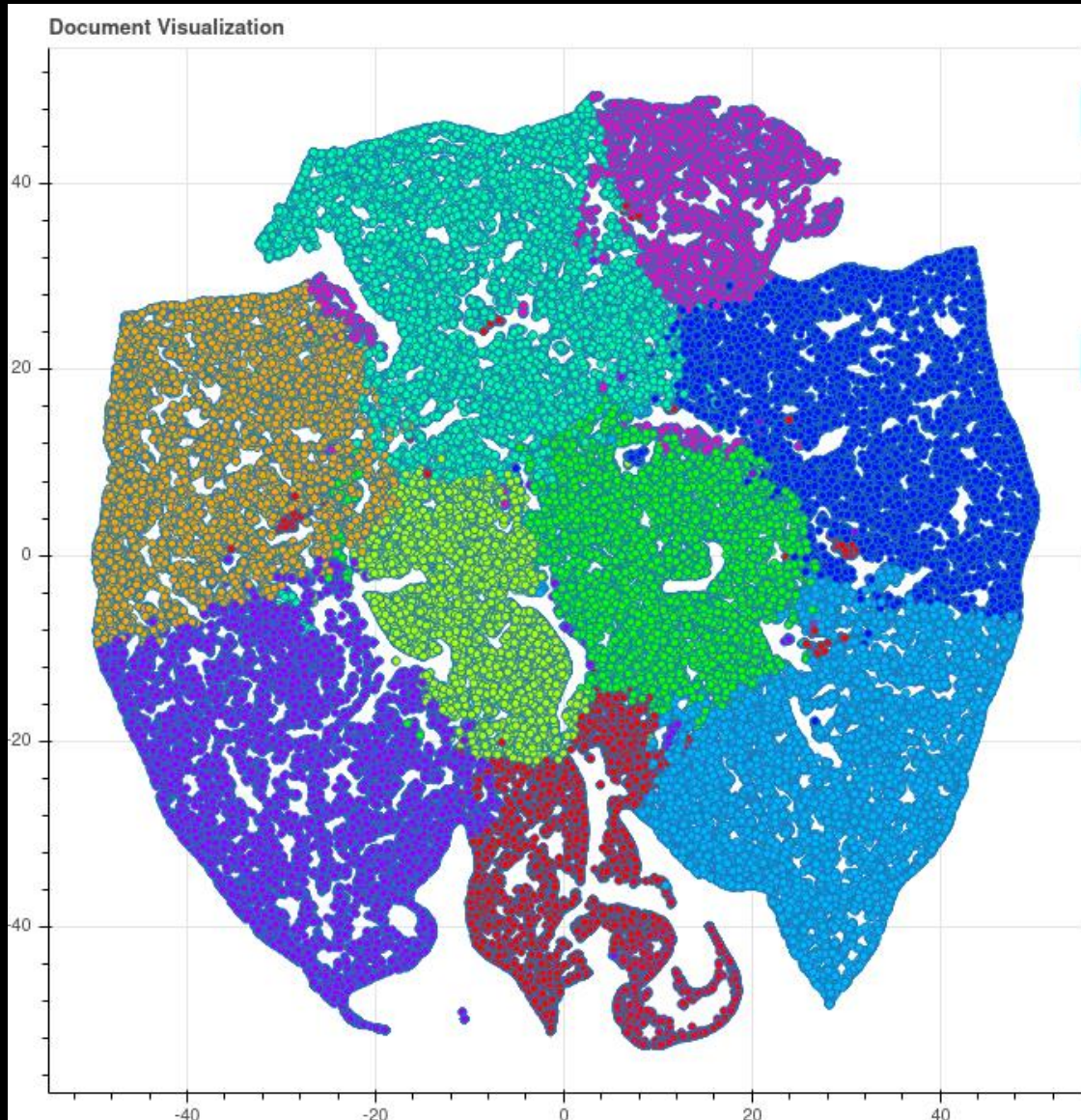
# Topic Modelling Dominance

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