Big Data Platforms

Final Project
November 30, 2022

Manish Kumar
Ekansh Trivedi
Iliyana Staneva
Medha Yadav

Agenda

Opportunity

Data

Methodology & Result

Applications

Deployment

Challenges & Future Work

Opportunity

NEWS | 05 May 2021

COVID research: a year of scientific milestones

COVID-19

Coronavirus and the risks of 'speed science'

Published: 07 May 2021

Research on Covid-19: a disruptive phenomenon for

NEWS · 28 AUGUST 2020

COVID-19 research update: How many pandemic papers have been published?

A briefing

Published: 23 June 2021

Publication patterns' changes due to the COVID-19 pandemic: a longitudinal and short-term scientometric

A publishing pandemic du

become?

Impact c

dynamics and non-COVID-19 researc

Feb 23, 2021

More than 87,000 scientific papers on coronavirus since pandemic

Study finds "astonishing" growth even as partnerships shrink

SCIENCEINSIDER | SCIENTIFIC COMMUNITY

Scientists are drowning in C

new tool

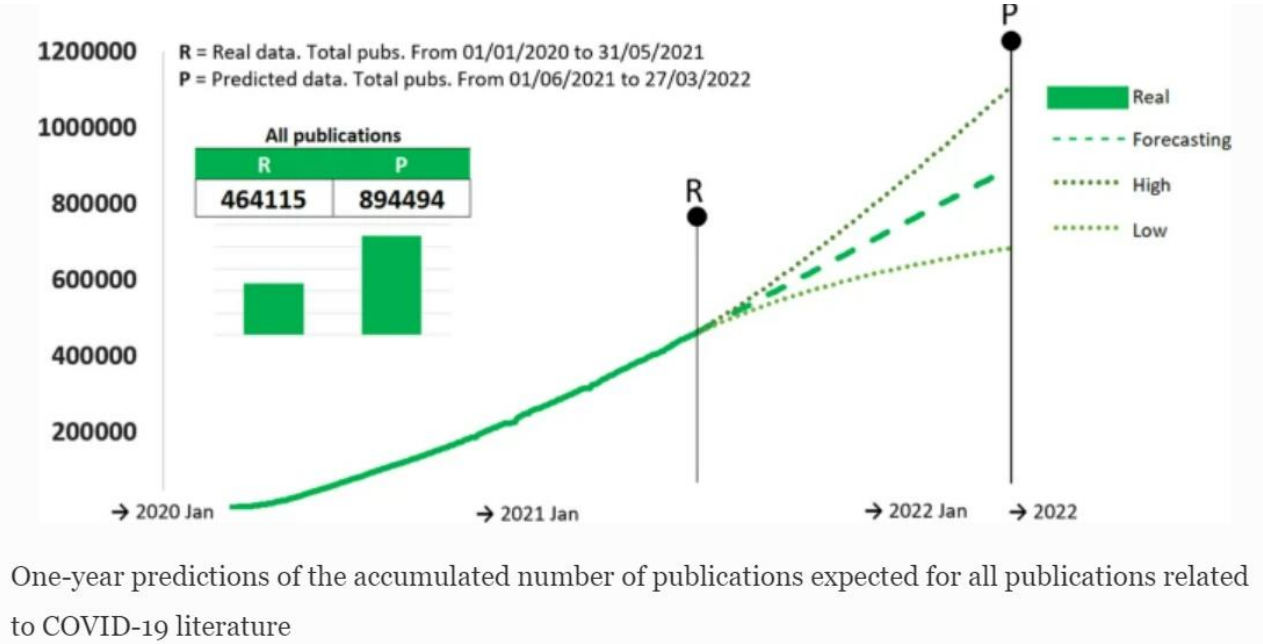
The hunt is on f

The Pandemic of Publications: Are We Sacrificing Quality for Quantity?

The rapid, massive growth of COVID-19 authors in the scientific literature

How many publications are we looking at?

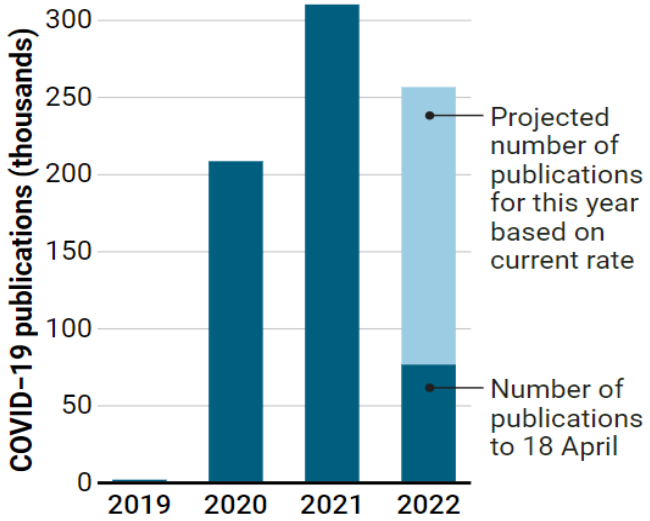
1.1 Million



One-year predictions of the accumulated number of publications expected for all publications related to COVID-19 literature

https://link.springer.com/article/10.1007/s11192-022-04536-x

0.75 Million



(GRAPHIC) K. FRANKLIN/*SCIENCE*; (DATA) DIMENSIONS DATA USED BY PHILIP SHAPIRA, BIORXIV, 2020.12.06.413682

https://www.science.org/content/article/pivot-covid-19-research-eases-publishing-surge-starts-level

1.8 Million papers annually

https://www.smithsonianmag.com/smart-news/half-academic-studies-are-never-read-more-three-people-180950222/

A novice researcher's everyday

Innumerable publications

Labor-intensive manual read-throughs

Limited domain knowledge

Lost time with negligible results

What can we do?

Input paper → **CLOUD BASED SPARK ML SOLUTION** → Predict the topic/theme of the paper

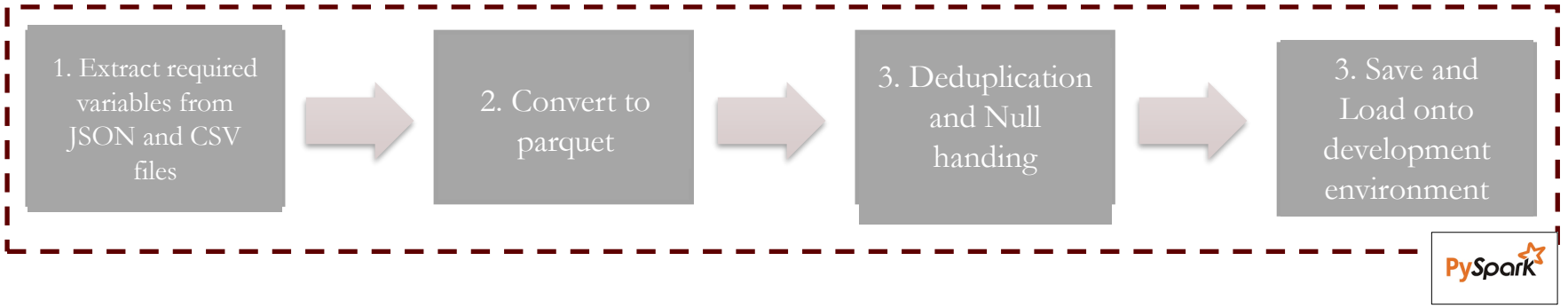
Data

Data Profile and Preprocessing

Data Profile

- **Source:**
  Allen Institute for AI (https://ai2-semanticscholar-cord-19.s3-us-west-2.amazonaws.com/historical_releases.html)
- **Format:**
  Zipped JSON and CSV files
- **Size:**
  ~75 GB; Published Papers in JSON format
- **Raw Data Variables:**
  title, paper id, metadata, abstract, body text, bibliography, references, back matter

Data Preprocessing



Heavily nested JSON raw input

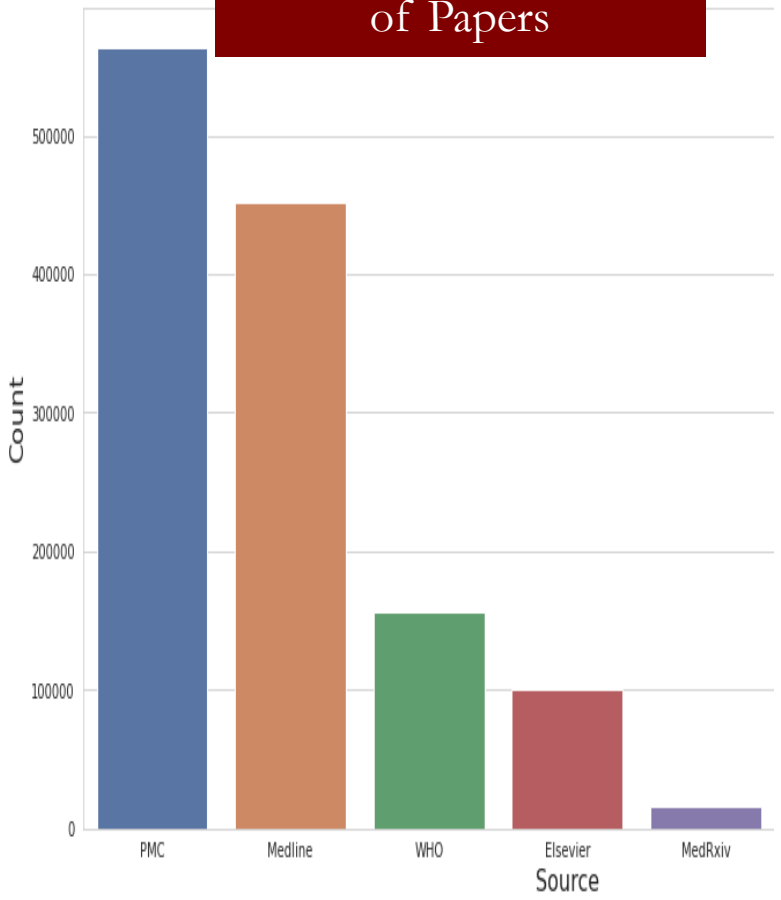

```
{
    "paper_id": "000b88a3a342f55aab834668b790458e1d4bab43",
    "metadata": {
    "abstract": [
    "body_text": [
    "bib_entries": {
    "ref_entries": {
    "back_matter": [
}
```


Preprocessed CSV

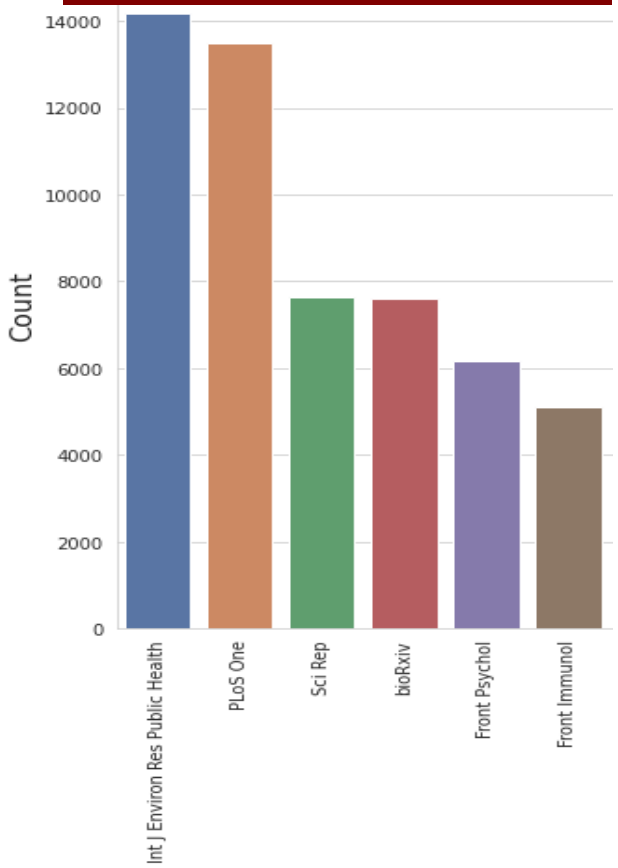
paper_id	title	authors	abstract	body_text
0000028b5cc154f68b8a269f6578f21e31f62977	"Multi-faceted" COVID-19: Russian experience			According to current live statistics at the time of editing this letter, Russia has been the third country in the world to be affected by COVID-19
0000b6da665726420ab8ac9246d526f2f44d5943	The cell phone vibration test: A telemedicine substitute for the tuning fork test	Robert; J Lewis; Nora Watson; Charles A Riley; Anthony M Tolisano	for differentiating between conductive and sensorineural hearing loss remains elusive. Our goal	COVID-19 pandemic, an acceleration in the implementation of telemedicine occurred in order to better triage patients while
0000b93c66f991236db92dc16fa6db119b27ca12	Infections in Hematopoietic Stem Cell Transplantation (HSCT) Patients 24	Biju George; Sanjay Bhattacharya		[illegible] are important causes of morbidity, mortality, hospital stay, intensive care unit admissions, and healthcare cost besides healthcare resource utilization in the setting of hematopoietic stem cell
000122a9a774ec76fa35ec0c0f6734e7e8d0c541	ST-segment elevation myocardial infarction care. The Spanish experience	Leor; BelÃ© N Cid-Ã Lvarez; Armando PÃ© Rez De Prado; Xavier Rossello;	objectives: The COVID-19 outbreak has had an unclear impact on the treatment and outcomes of	COVID-19 outbreak has had an unclear impact on the treatment and outcomes of patients with ST-segment elevation myocardial

Exploratory Data Analysis

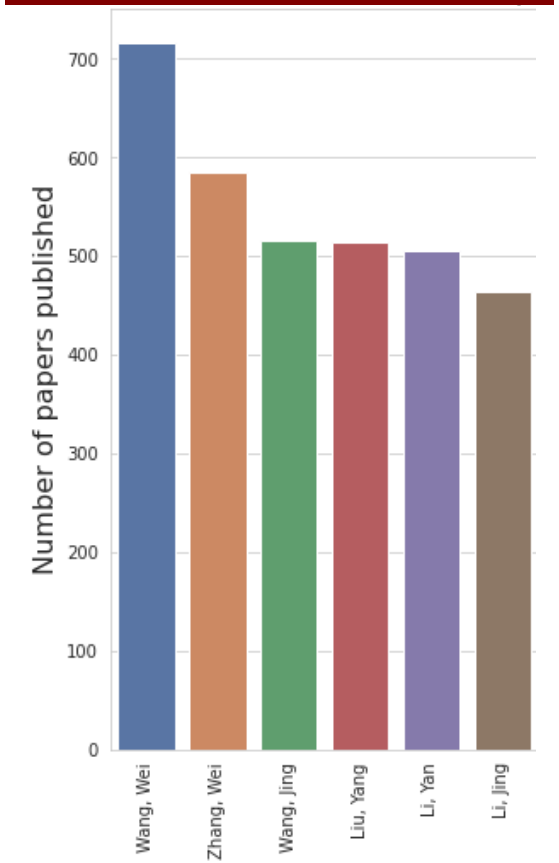
Most Common Sources of Papers



Top 5 Journals publishing COVID papers

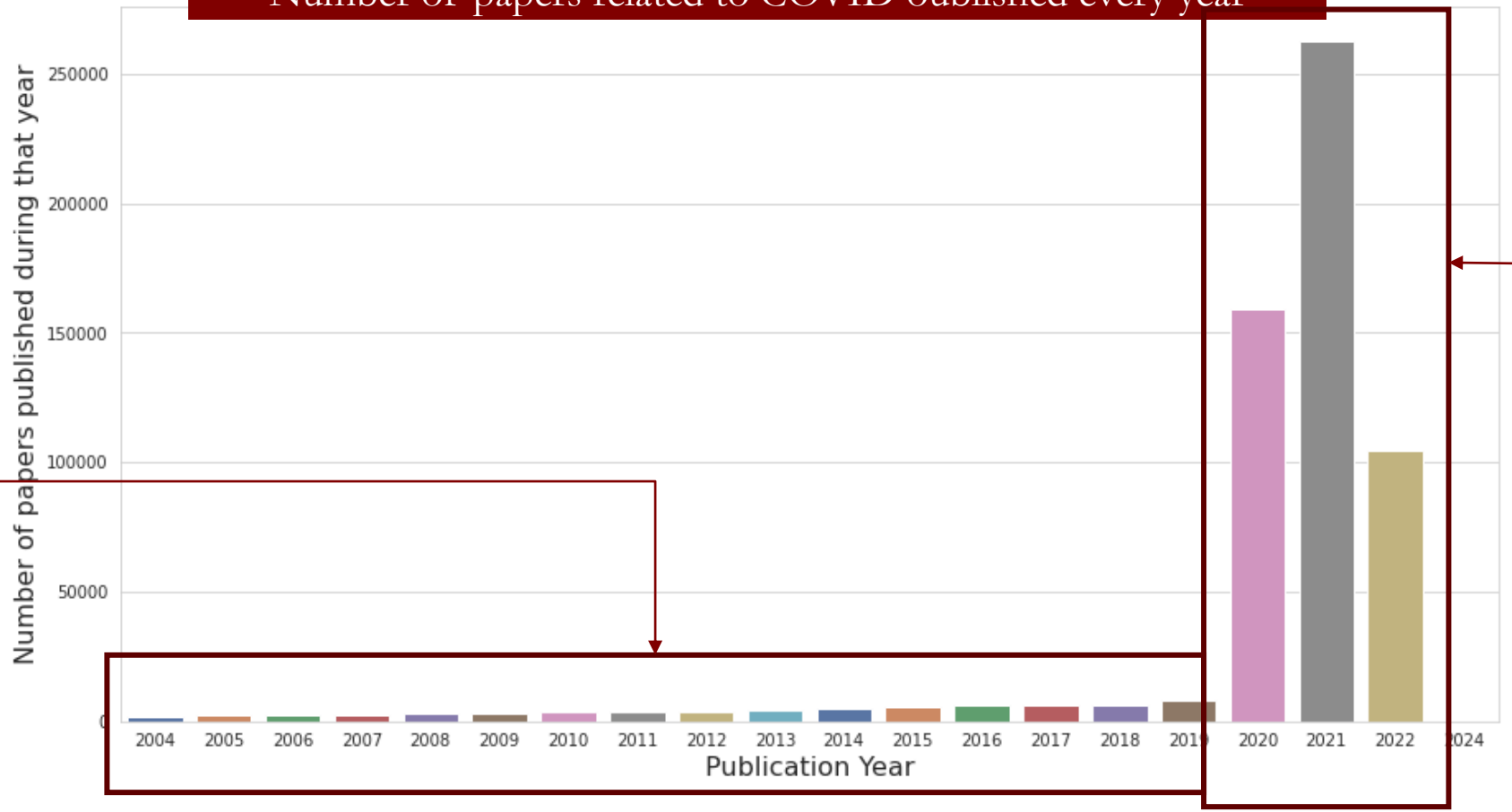


Top 5 authors publishing COVID papers



Exploratory Data Analysis

Number of papers related to COVID oublished every year



Research on Sars-CoV was being done and published even before COVID hit

Research on COVID increased exponentially in 202, and peaked in 2021

Methodology & Results

Methodology : Topic Modeling

- Abstract Data
- Body of the article

- Null Handling
- Case Handling
- Stop word Removal
- Numeric word removal
- Punctuation Removal
- Lemmatization

- TF IDF
- Word2vec
- BERT

- Latent Dirichlet allocation (LDA)
- KMeans
- BERT QA

Selected Architecture

Input

Title + Abstract data

Data Cleanup

Stop word + lemmatization + removal of numeric and punctuation

Feature Extraction

TF IDF
Word2Vec
BERT Embeddings

Model

LDA, KMeans, BERT QA

LDA: How well is the model doing?

Evaluation Metric

Key characteristics – **human interpretability** or **semantic interpretability** of topics.

Human judgment
- Observation-based
- Interpretation-based

Quantitative metrics
- Perplexity
- Coherence

Hybrid approach for evaluation –
- Human judgement – Observation based
- Quantitative metrics – Coherence Calculation

Model Performance

Model Parameters – 10 max iterations and 5 topics
Perplexity Score – 6.56

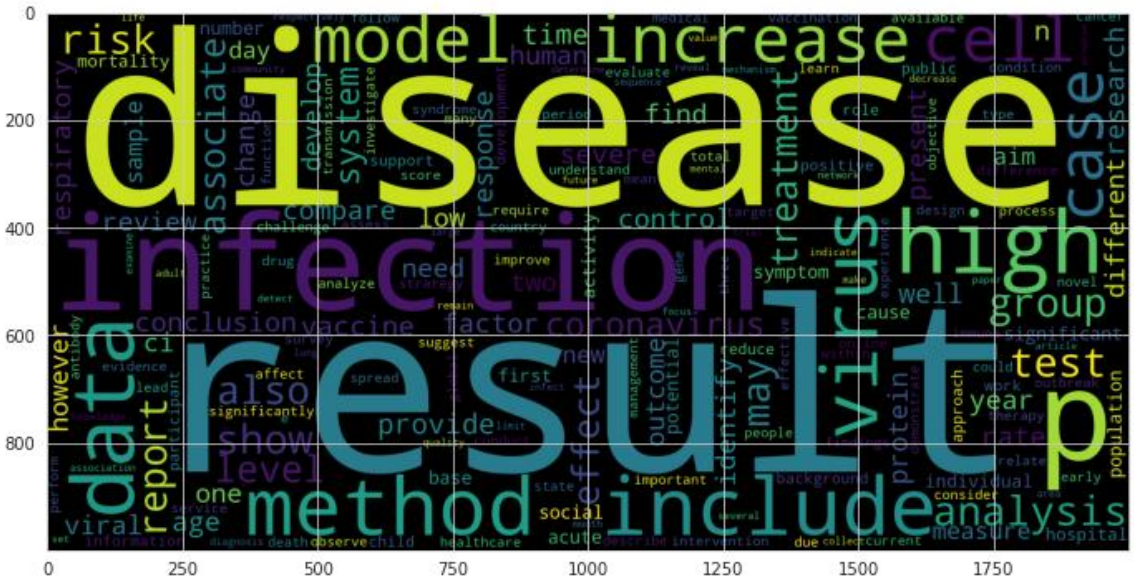
Top 3 topics post training

Topic 1	Topic 2	Topic 3
cell	social	virus
antibody	mental	model
la	research	viral
vaccine	service	protein
food	stress	rna
immune	anxiety	sequence
response	experience	human
mask	support	infection
virus	behavior	detection
expression	work	drug

Better words when using TF-IDF tokenization

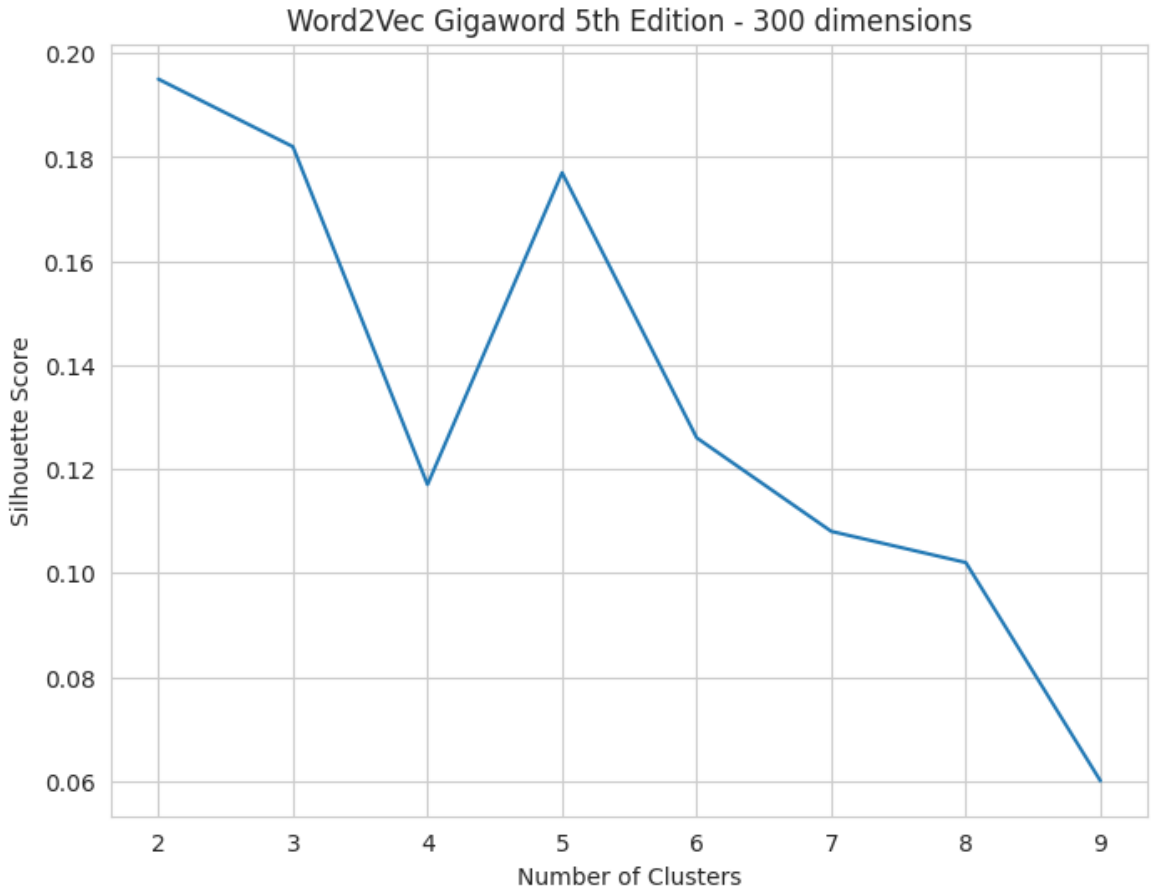
TF-IDF

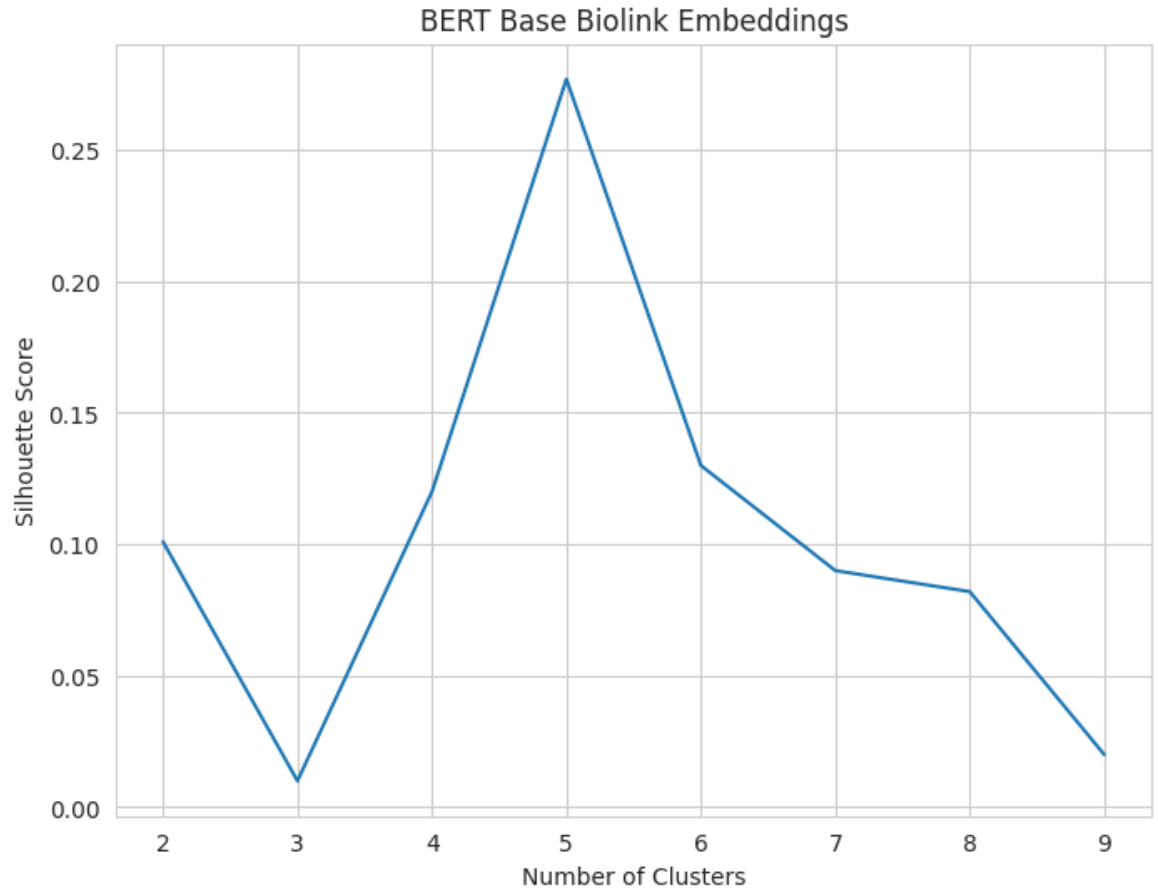
Word count



KMeans: How well is the model doing?

- We tried two different embeddings
- Biolink produced a better clusters with cluster sie of 5





Applications

Application I : Question Answering Model – Overview

A simple question answering model can help find quick insights from vast number of papers

1

Bert For Question Answering Cased model

Source - Callmenicky

A fine-tuned version of bert-base-cased model

2

Deberta For Question Answering model

Source - nbroad

Uses disentangled attention and enhanced mask decoder on top of BERT

3

Longformer For Question Answering model

Source - allenai

A transformer model for long documents

4

Roberta For Question Answering

Source - deepset

A Roberta-base model, fine-tuned using the SQuAD2.0 and COVID-19 QA dataset

Application I : Question Answering Model – Results

Results Summary

- ✓ First three models only got the first (and also the easiest) question right
- ✓ The third model returned answers to every questions even if they were incorrect, unlike the first two which returned empty for the remaining 4 questions
- ✓ Last model got two answers while returning incorrect answer for one of them
- ✓ For simpler questions, the last model also provided additional details related to the answers in contrast to the concise answers by other models

This work is shown in Modeling-QA.ipynb notebook file

Sample Outputs

Sample Questions

```
What is the most common cause of atypical pneumonia?
What is the main reason why NO production is regulated?
What is the name of the molecule that binds to a pulmonary pathogen?
What factors can cause the development of pulmonary fibrosis?
How does pneumoviruses enter respiratory epithelial cells?
```

Bert For Question Answering Cased model

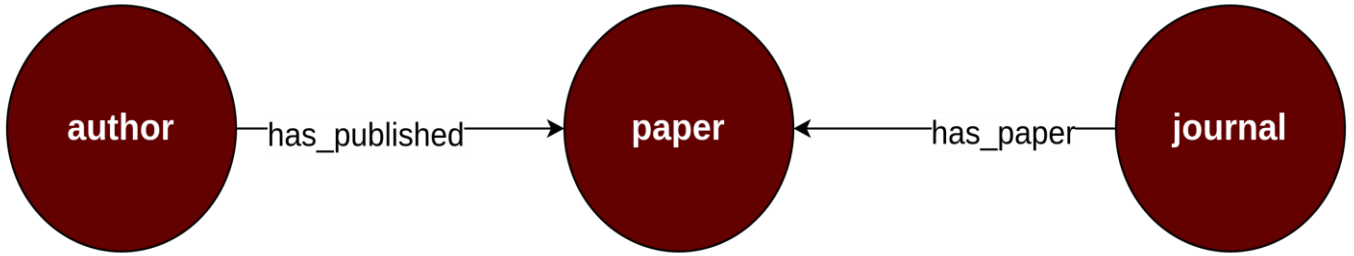
```
What is the most common cause of atypical pneumonia?              |[Mycoplasma pneumoniae]
What is the main reason why NO production is regulated?           |[]
What is the name of the molecule that binds to a pulmonary pathogen?|[]
What factors can cause the development of pulmonary fibrosis?     |[]
How does pneumoviruses enter respiratory epithelial cells?        |[]
```

Roberta For Question Answering

```
|What is the most common cause of atypical pneumonia?              |[mycoplasma pneumoniae is a commo
n cause of upper and lower respiratory tract infections . it remains one of the most frequent causes of
atypical pneumonia particularly among young adults]
|
|What is the main reason why NO production is regulated?           |[the formation of such rns is tho
ught to be the prime reason why no • can in many cases contribute to the etiology of inflammatory lung
disease]
```

Application II : Graph Databases – Overview

- ✓ Graph databases provide a way to generate and visualize relationships between entities
- ✓ Both Pyspark GraphFrame and neo4j can achieve graph-based data storage. We explored both the tools
- ✓ Each author, paper, and journal acts as a node
- ✓ All nodes are connected as per relationships – "has_published" or "has_paper"
- ✓ Data was prepared using python to make it ready to import to neo4j
- ✓ Docker was used to install the neo4j (neo4j version 5.2.0)
- ✓ Bash script (start_neo4j.sh) starts the docker container, neo4j server and imports the data



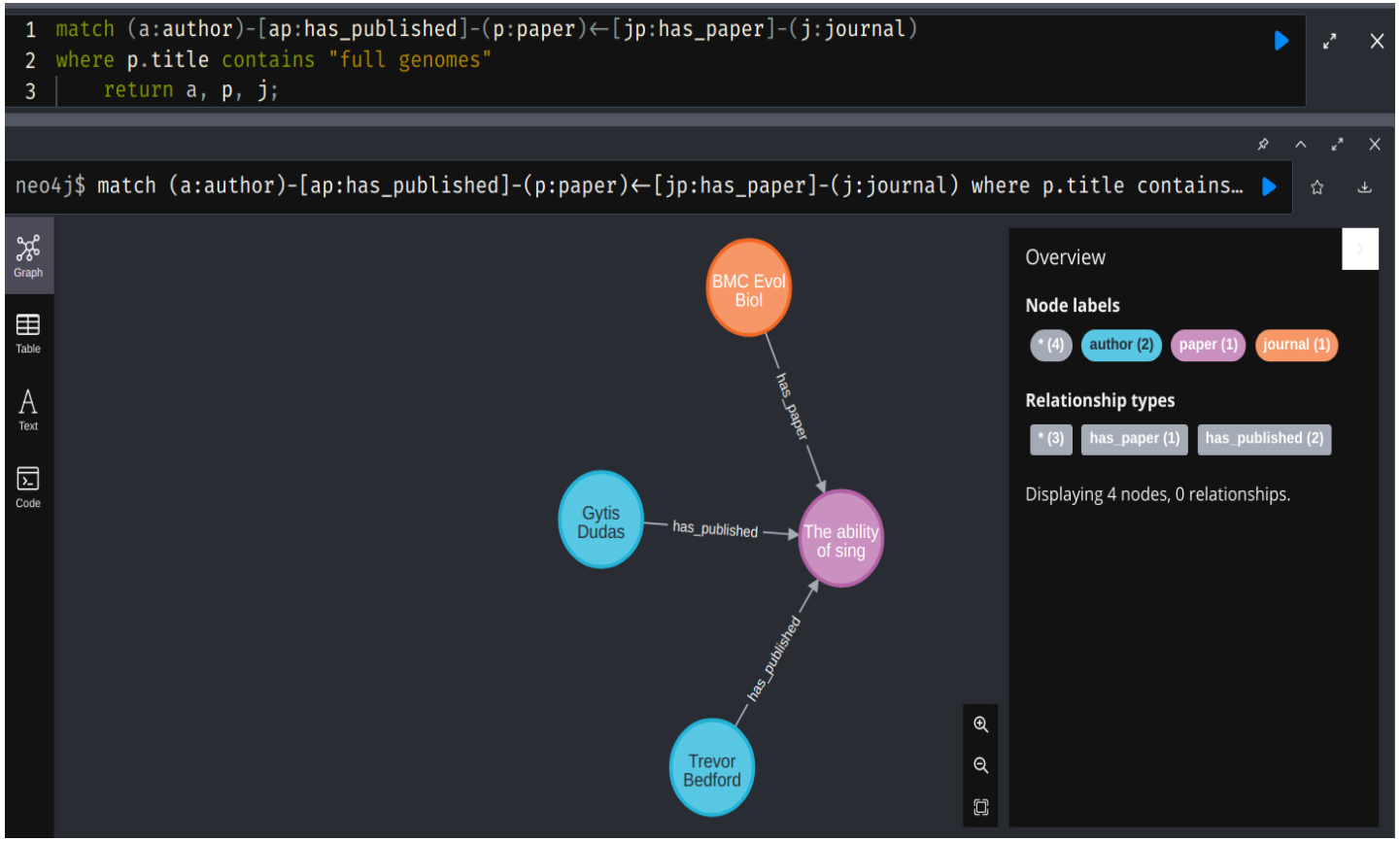
This work is shown in Prepare-neo4j-data.ipynb notebook file

Application II : Graph Databases – Sample Output

Pyspark GraphFrame

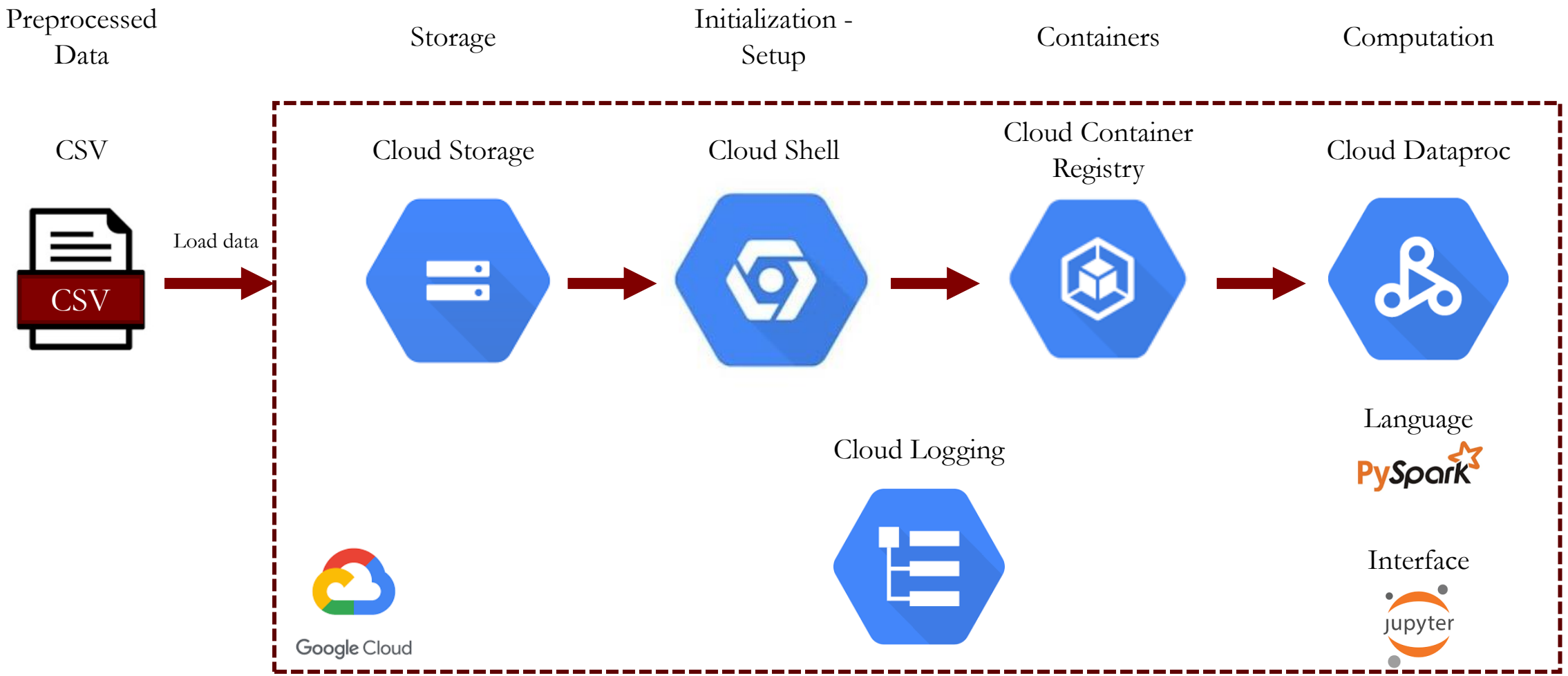
src	dst	relationship
patt_debra_a	PMC8202122	published
wilfong_lalan	PMC8202122	published
toth_sara	PMC8202122	published
broussard_stephanie	PMC8202122	published
kanipe_kristen	PMC8202122	published
hammonds_jason	PMC8202122	published
allen_victoria	PMC8202122	published
mautner_beatrice	PMC8202122	published
campbell_nakedra	PMC8202122	published
dubey_ajay_k	PMC8202122	published
wu_nini	PMC8202122	published

Neo4j



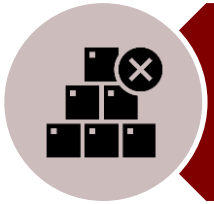
Deployment

Productionizing on GCP

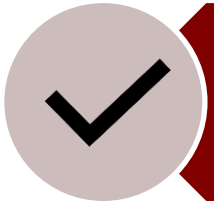


Challenges & Future Work

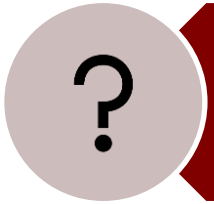
Data Challenges



Deduplication –
How to find the latest or the correct version of a paper if there are multiple versions?



Data Integrity –
Do we trust the information from the metadata file or the JSON files, if there's a conflict?



Handling Missing Data –
Abstract not provided separately for PMC papers

Similar Data Values –
Difficult to differentiate between two authors with same names

Potential future work

Research Browser

- A ReactJS based research browser can be built that can connect to the neo4j database using GraphQL queries
- We can index the data (title, abstract, full paper text, author names, and journal names) in neo4j using its inbuilt procedure methods (APOC)
- Data indexing can enable free text search to complete the research browser search functionality

Automation

- From starting the GCP cluster to data preprcessing, modeling, and neo4j database, all these steps can be automated using Airflow

Increased Compute

- For most of our tasks, we used the paper abstract as a source of information due to memory and storage constraints
- Full body text of the paper can provide a lot more information provided additional resources

Thank You

Presented by:

Manish Kumar
mnis@uchicago.edu

Medha Yadav
medhaydv@uchicago.edu

Iliyana Staneva
iliyanastaneva@uchicago.edu

Ekansh Trivedi
ekansh@uchicago.edu

References

- https://www.science.org/content/article/scientists-are-drowning-covid-19-papers-can-new-tools-keep-them-afloat
- https://news.osu.edu/more-than-87000-scientific-papers-on-coronavirus-since-pandemic/
- https://royalsocietypublishing.org/doi/10.1098/rsos.210389
- https://link.springer.com/article/10.1007/s11192-021-04059-x
- https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov
- https://www.weforum.org/agenda/2020/03/speed-science-coronavirus-covid19-research-academic
- https://www.science.org/content/article/pivot-covid-19-research-eases-publishing-surge-starts-level
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7230425/
- https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-021-01404-9
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7528741/
- https://link.springer.com/article/10.1007/s11192-021-03989-w
- https://www.nature.com/nature-index/news-blog/how-coronavirus-is-changing-research-practices-and-publishing
- https://www.nature.com/articles/d41586-020-00502-w
- https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0
- https://towardsdatascience.com/6-tips-to-optimize-an-nlp-topic-model-for-interpretability-20742f3047e2
- https://highdemandskills.com/topic-model-evaluation/
- https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0