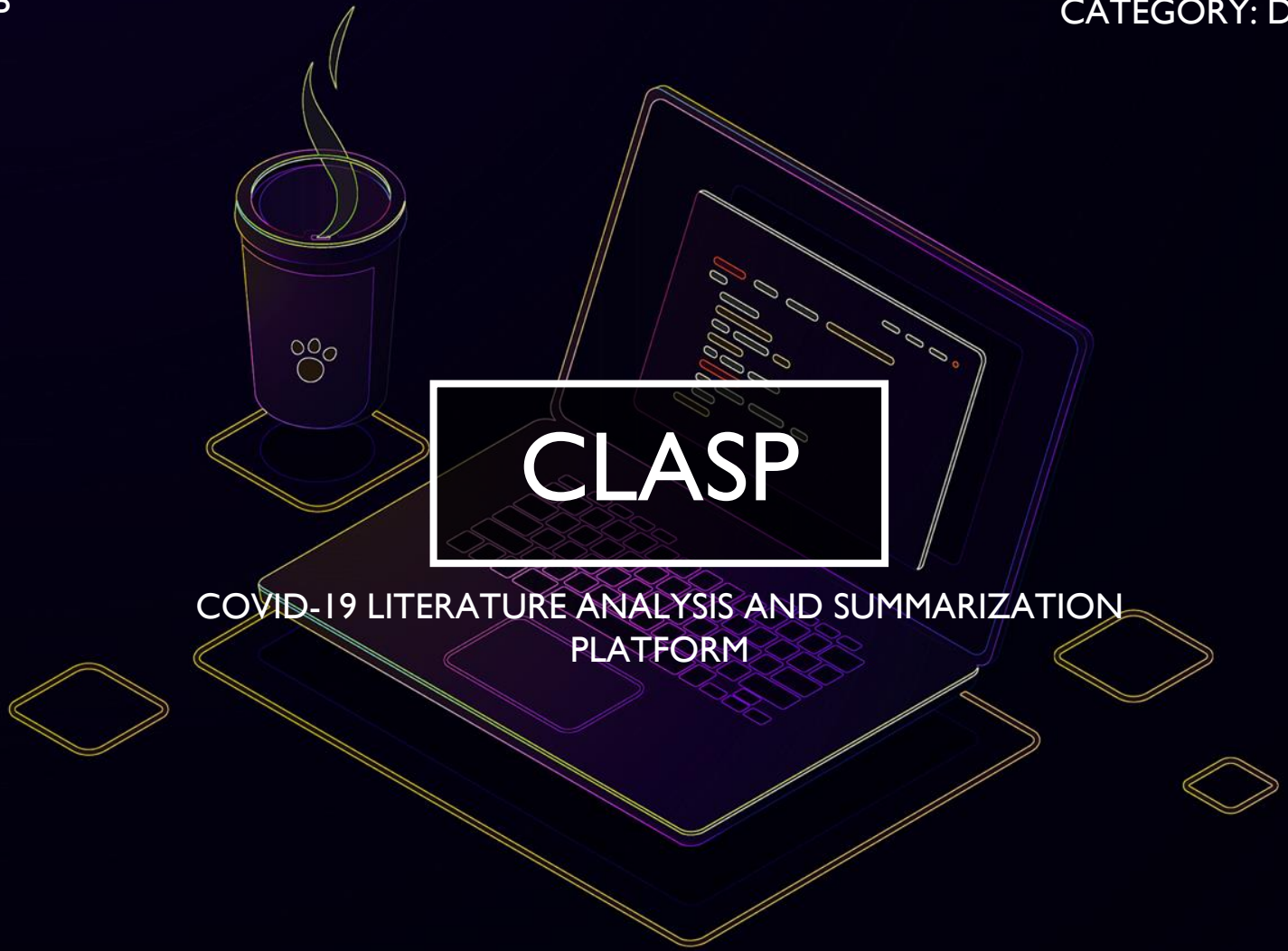
TEAM CLASP

CATEGORY: DATA MINING



CLASP

COVID-19 LITERATURE ANALYSIS AND SUMMARIZATION PLATFORM

Inspiration

- Since the advent of the pandemic, the number of research papers written about it, analysing its various aspects have shot up exponentially. As the world strives to create a vaccine, it is not possible to take into account every ounce of the data available, in its original, vast form. A lot of useful data remains unattended.

- The wildfire-like spread of the coronavirus has raised millions of questions worldwide. The common man and the researchers are in search of correct answers. Panic rises as the sources of data become more unreliable each passing day.

- The fact that no technology in 2020 was able to accurately predict the global impact of this virus has lead to its unprecedented spread and irrevocable damage to life and economy worldwide.

Target audience

- ➔ Our solution has been design in a robust manner such than it can be used efficiently by people looking to gain insights into the research on COVID-19.
- ➔ It enables one-click access to research updates on the pandemic for people across the globe.

Impact of the solution

- ➔ The use of unsupervised learning enables biomedical researchers to benefit from the vast amount of unlabelled data.
- ➔ Text summarisation helps in narrowing down the vast corpus of data to its most essential, key points. This ensures that scientists have every fact available and catalyses the development of a vaccine.
- ➔ Our solution enables the user to enter his/her query and have a verified source of data answer it. This diffuses panic considerably.
- ➔ Anomaly detection when specifically focussed on the 29 states in India can notify appropriate authorities to a potential anomaly before the onset of a pandemic.

About Us

1. Tanmay Pardeshi:

I am a sophomore studying in the Information Technology department at Pune Institute of Computer Technology. I have a keen interest in back-end and server side programming. I currently work with two python frameworks Django and Flask. I have worked on a few projects based on Django and REST APIs in my college. I also have knowledge about the basic concepts of Java. I am currently stepping my foot into areas involving cloud computing. Data structures and algorithms is one of interests as well.

2. Kaustubh Odak:

I am a sophomore doing my bachelors in Computer Engineering at Pune Institute of Computer Technology. I have experience in front-end development with ReactJS library and UI frameworks like Bootstrap, Material UI and MaterializeCSS. I also have experience with android app development in Kotlin and cross-platform app development in Flutter and Firebase Realtime NoSQL database. I have used this knowledge to make a few applications for college events.

About Us

3. Harsh Sakhrani:

I am a sophomore studying in the Information Technology Department at PICT. I have a budding interest in the field of Artificial Intelligence and NLP. I have completed several MOOCs to enhance my knowledge of Deep Learning. In the past I have worked on a few projects based on Statistical Machine Learning and Computer Vision. Competitive programming and Data Structures and Algorithms are few of my other interests.

4. Snigdha Singh:

I am studying Computer Science at Pune Institute of Computer Technology, currently in my sophomore year. I have knowledge of Machine Learning algorithms, Neural Networks and NLP having worked in the same. I have experience in development technologies such as Node.js, Express, Bootstrap and more. I have built multiple websites using the same. I have a keen interest in Data Structures and Algorithms.

About Us

5. Saloni Parekh:

I am a second year Information Technology student at PICT, I have a keen interest in Machine Learning and Deep Learning. I have completed the Deep Learning Specialisation on Coursera to enhance my knowledge of its techniques. I have worked on a Video-Based Dynamic Human Authentication System which had reached the Internal Hackathon held for the Smart India Hackathon, 2020. I am interested in working on the usage of ML and DL to solve problems encountered on a daily routine. I have also worked on a Django based project used to calculate Carbon Footprint.

Dataset Details

The text mining tool that we propose makes use of the CORD-19 dataset which comprises of over

128,000

scholarly articles including over

59,000

with full text, about COVID-19, SARS-CoV-2, and related coronaviruses

Limitations of existing work

A completely comprehensive data mining tool, solely dedicated to COVID-19 research, which is capable of taking new literature into consideration being mostly unsupervised in nature, isn't out there to the best of our knowledge. We provide a solution with the following features:

- ➔ Text summarization
- ➔ Query-based information retrieval engine
- ➔ Q and A system
- ➔ Anomaly detection
- ➔ Live research news section

Text Summarization - Model Details

- The text summarization technique that we have implemented makes use of pre trained Glove embeddings, cosine similarity averaged over sentences and the PageRank algorithm to rank sentences according to their conceptual relevance.
- The summary that our system produces comprises of all the conceptually relevant sentences which encapsulate the idea and the essence of the paper.
- Its size is approximately 20% of the actual paper.

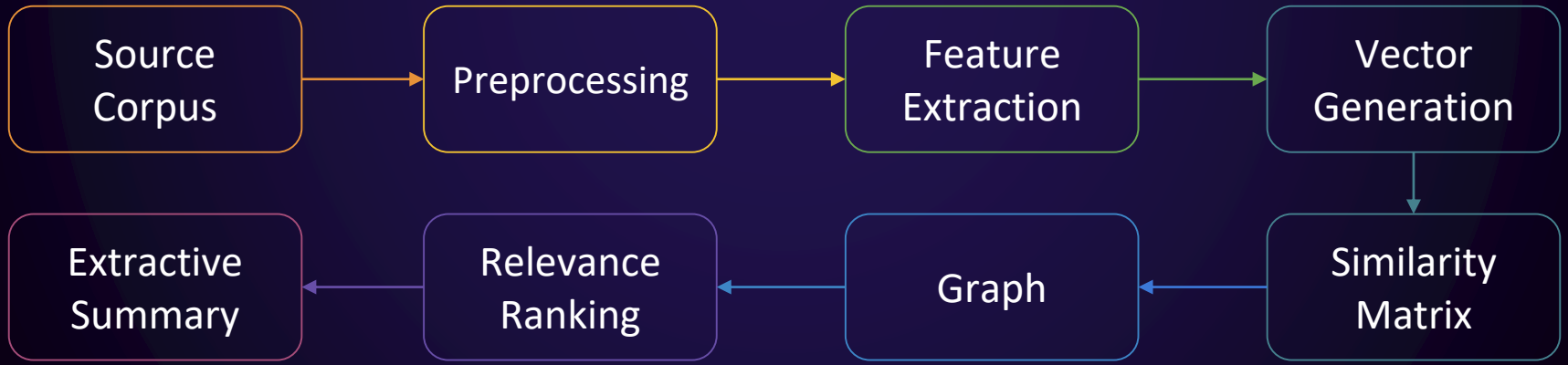
Results

Inferencing Time

~40s

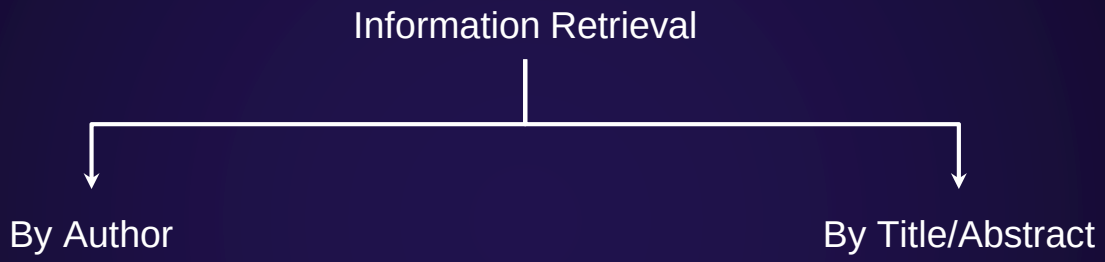
Text Summarization - Performance Details

- A number of metrics like Euclidean Distance, Jaccard Similarity and TF-IDF scores were used to calculate the similarity matrix.
- Cosine similarity generated the most contextually relevant results.



Information Retrieval Engine

MODEL DETAILS

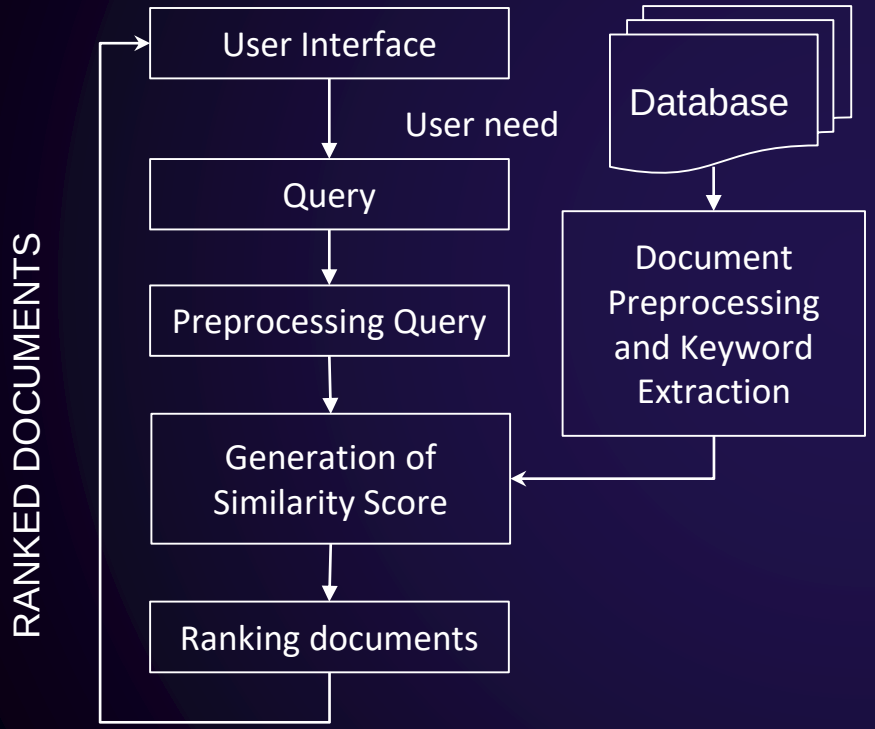


- ➔ The Information Retrieval tool that we built makes use of a combination of Keyword Indexing and Levenshtein Distance.

PERFORMANCE DETAILS

- ➔ A number of other similarity metrics like Cosine similarity and Jaccard similarity were taken into consideration to score the research papers on the basis of the query entered by the user.
- ➔ The results achieved by Levenshtein distance were more relevant in nature.

Information Retrieval Engine



Results

Precision

0.9404

Recall

0.2113

Inferencing Time

~30s

Q&A System - Model Details

- ➔ The Q&A system we built makes use of cdQA which is built on top of the HuggingFace library.
- ➔ The model was fine tuned using the CORD-19 dataset.
- ➔ The most probable articles in the dataset are selected using TF-IDF features and the cosine similarity between the question and each document.
- ➔ The system divides the most probable document into paragraphs and passes each paragraph and question to a pretrained model(BERT).
- ➔ The reader outputs an answer corresponding to each paragraph.
- ➔ The final layer compares the answers by an internal score function and outputs the most likely answer according to the scores.

Results

Training Time

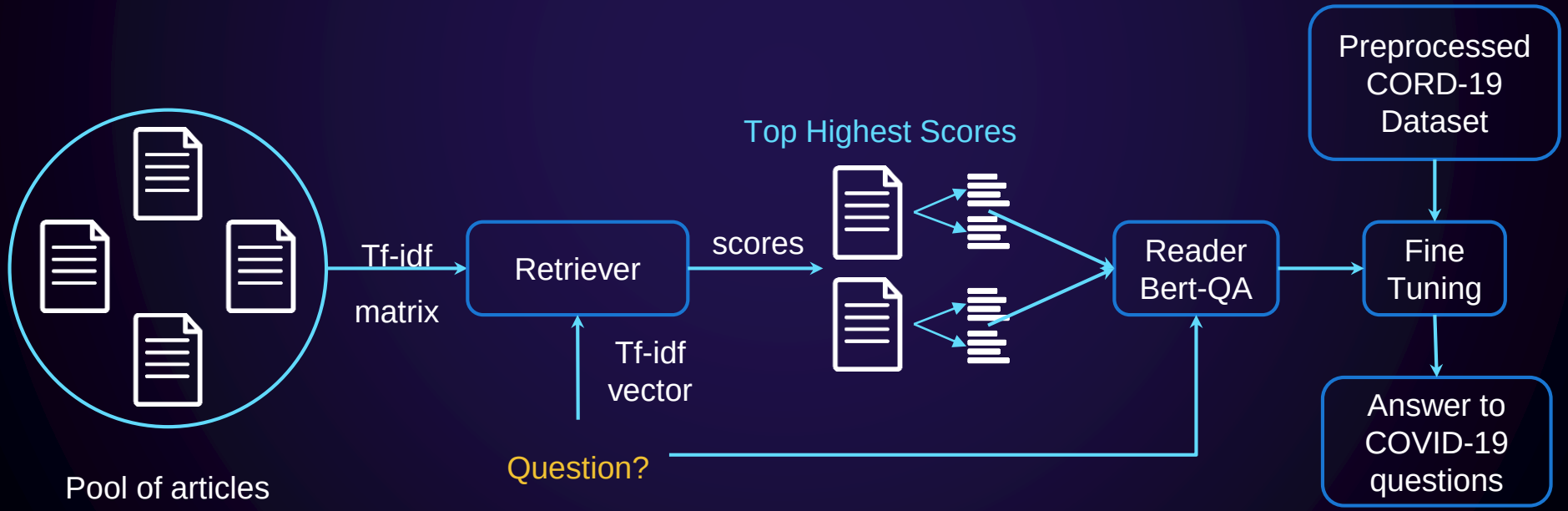
~123s

Inferencing Time

~20s

Q&A System - Performance Details

The pre-trained models from the transformers and SimpleTransformers library didn't give adequate results on our data. The model from the transformers library couldn't calculate the start and end scores correctly. Suitable data was unavailable to use the model from SimpleTransformers. These problems were easily resolved by fine tuning the cdQA model thus giving pretty accurate results.

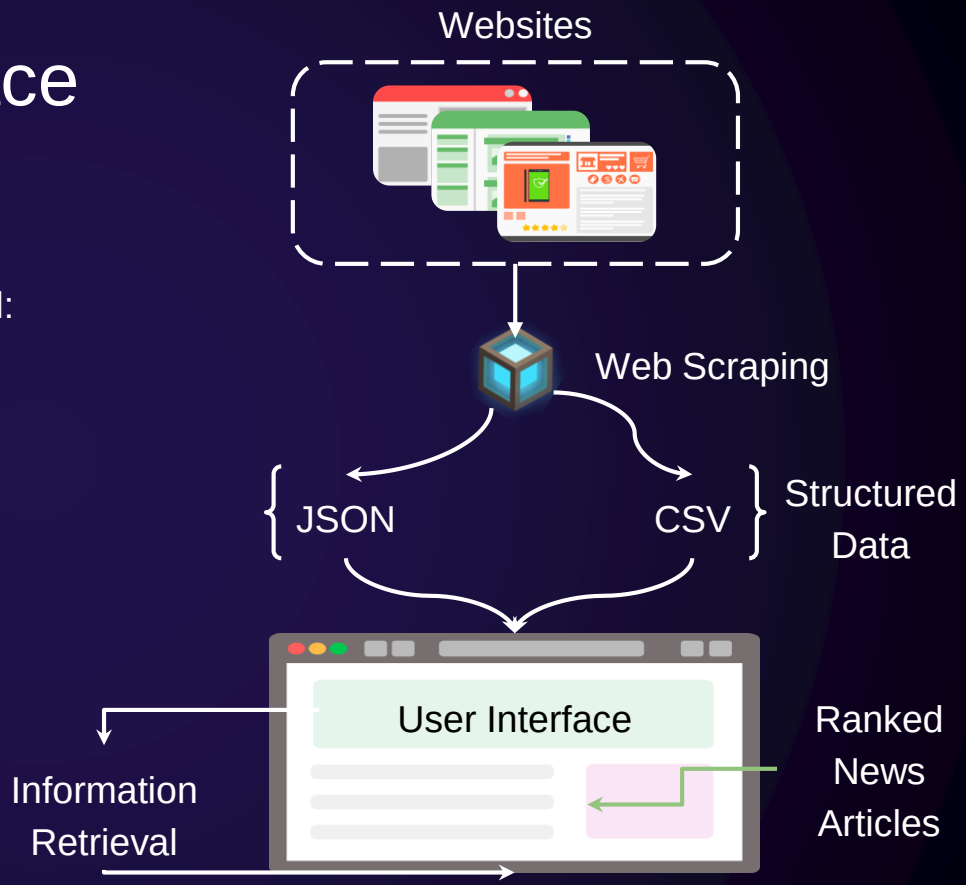


Live Research News Space

MODEL DETAILS:

Research News Updates with Information Retrieval:

The Live Research News Space makes use of Real Time scraping techniques to scrape large amounts of unstructured news and continuously updates the user about all the relevant research happenings all around the world with respect to COVID-19.



Live Research News Space

PERFORMANCE DETAILS:

News Search:

- ➔ The news search section makes use of pre-trained Glove embeddings along with cosine similarity to get the relevant news with respect to the query entered by the user.
- ➔ Jaccard similarity and Levenshtein Distance were also taken into consideration. The results achieved by cosine similarity were most relevant.

Results

Precision

0.78 - 0.89

according to the query

Recall

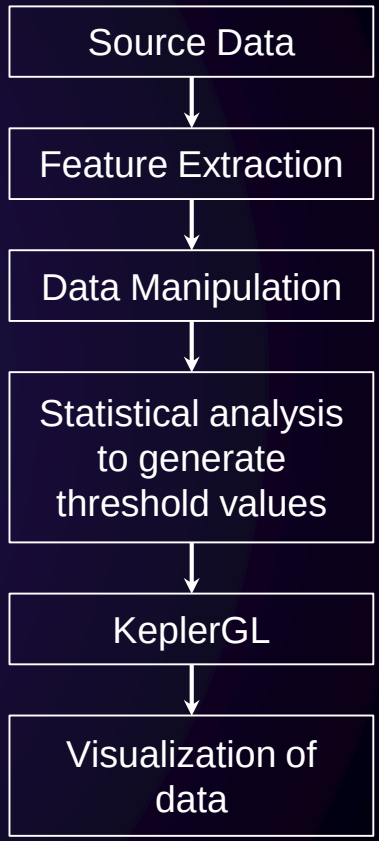
0.3655

Inferencing Time

~10s

Anomaly Detection and Alert System

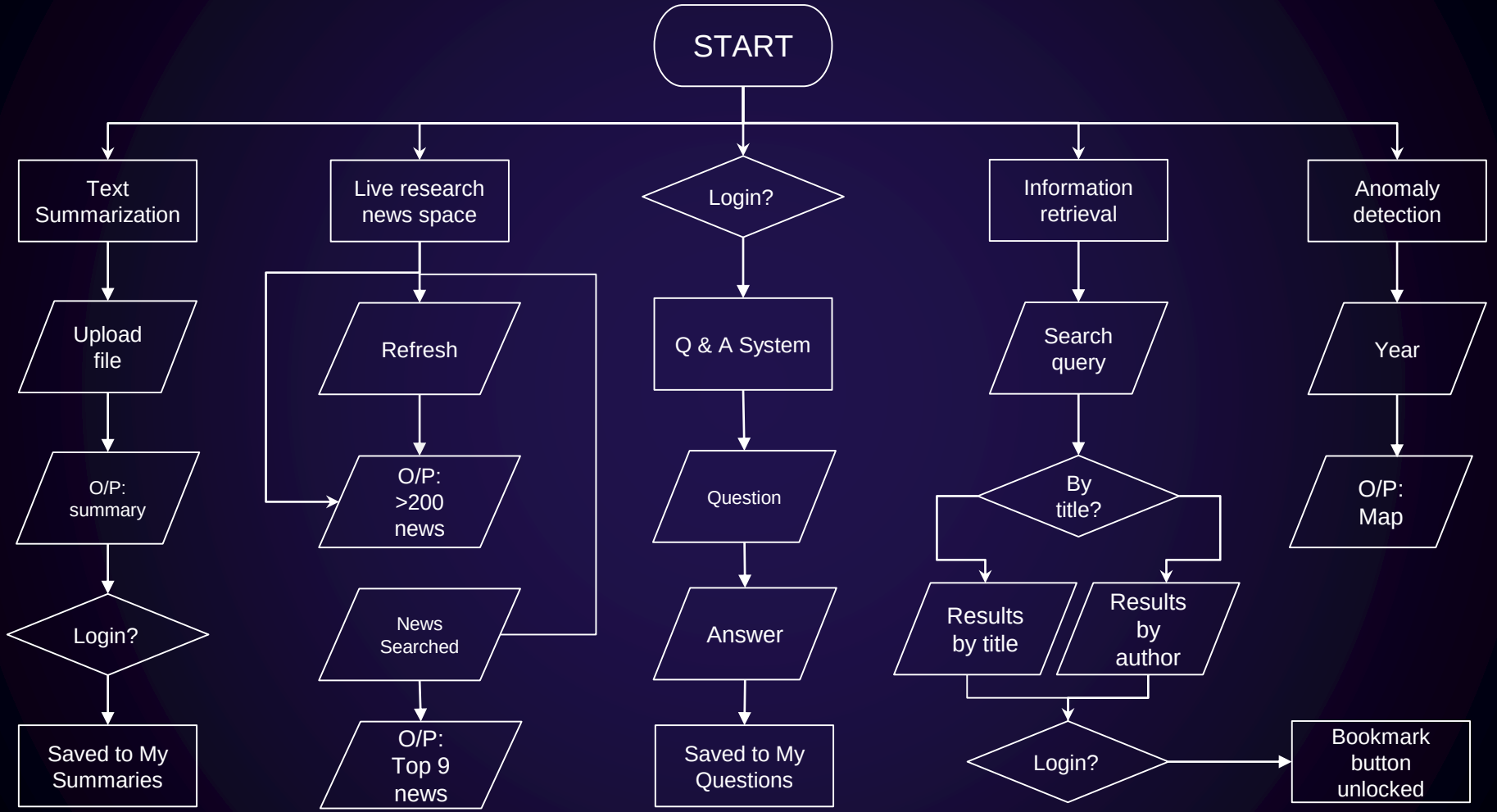
- ➔ Anomaly detection makes use of statistical trend analysis for more than 180 countries all around the world to detect abnormal behaviour with respect to parameters like GDP and Total Amount of Health Care Spend.
- ➔ We made use of KeplerGL to visualize the annual healthcare behaviour of all the major countries in the world.
- ➔ The objective of this feature is to alert the concerned authorities about the emergence of anomalies in various countries before it reaches epidemic proportions.
- ➔ Latitude and Longitude of each country has been taken into consideration for visualization.
- ➔ Statistical Trends were analyzed to calculate a unique threshold healthcare value for each country according to which the anomalous behaviour is recorded.



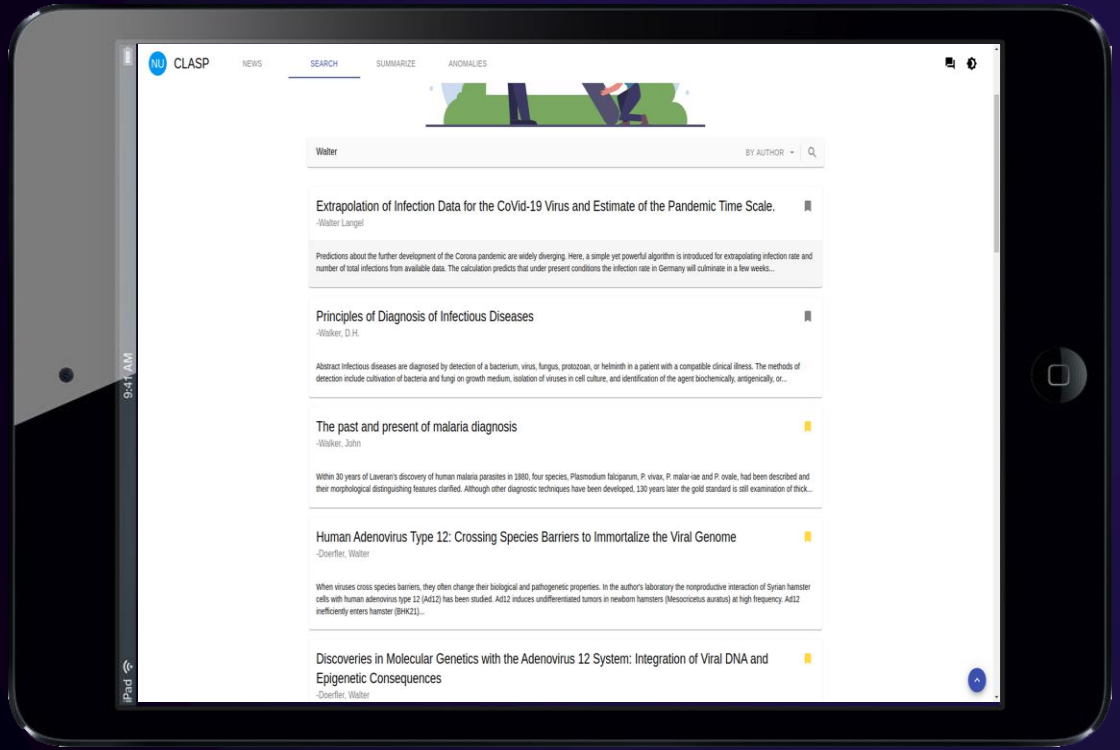
TEAM CLASP

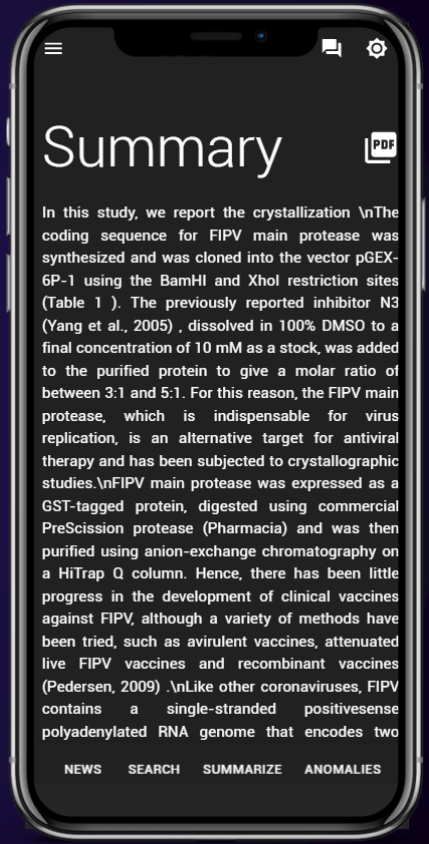
USER INTERACTION/FLOW

FINAL SOLUTION



User Interface





Accuracy

Features/Metrics	Precision	Recall
Text summarization	All the text summarization metrics out there (Bleu, Rouge, F1 score) require a Gold Standard summary to measure the performance*	
Information Retrieval	0.9404	0.2113
Q and A system	All the Q&A metrics out there (Bleu, Rouge, F1 score) require a Gold Standard Answer to measure the performance*	
Anomaly Detection	Not applicable	Not applicable
Live Research News	0.78 - 0.89	0.3655

**Gold Standard summaries or answers were not used, hence no specific performance metric has been evaluated.*

Performance Numbers

Features / Times	Training Times	Inferencing Times
Text summarization	-	~40 seconds
Information Retrieval	-	~30 seconds
Q and A system	~ 123 seconds	~ 20 seconds
Anomaly Detection	Not Applicable	Not Applicable
Live Research News	-	~10 seconds

CPU: *1.6 GHz Dual-core Intel Core i5*

GPU: *Intel UHD-617 1636 MB*

RAM: *16 GB 2133 Mhz LPDDR3*

TEAM CLASP

Tech Stack

New Learnings

- ➔ Making use of pre-trained models and fine tuning them
- ➔ Deploying a heavy product
- ➔ How scalability is important
- ➔ Usage of contextual similarity metrics
- ➔ Understanding how all the major pretrained models like BERT,GPT-2,ELMo work and their architectures
- ➔ Usage of open source visualization tools like KeplerGL
- ➔ Web scraping techniques
- ➔ React Context API and Material UI
- ➔ Flask REST APIs
- ➔ JWT Authentication
- ➔ Deployment on Linux servers using nginx and gunicorn servers

DGX Utilization And Interaction With Problem Statement Owners

If a machine like DGX is provided for more time we would definitely like to work with the problem statement owners and implement the following:

- Increase the size of the corpus for the Q&A system, thus making it more aware of the coronavirus literature out there.
- Also add AI to the Information Retrieval system we have already created to make it more robust in nature.
- If a machine like DGX is provided, there would be no deployment limitations whatsoever leading to a significant decrease in the inference time of all the features.

What We Can Do In The Future

- Work on the fine tuning process for the Q&A system, by increasing the size of the corpus,without worrying about the limitations of the deployment process.
- Explore the Anomaly Detection Space by also taking publicly available information like social media and news into consideration.

Scope For Improvement

Weekly data for every country indicating it’s Healthcare spend, Hospital Finance Reports, Healthcare and Wellbeing reports can help us predict anomalies in the same before it reaches pandemic proportions.