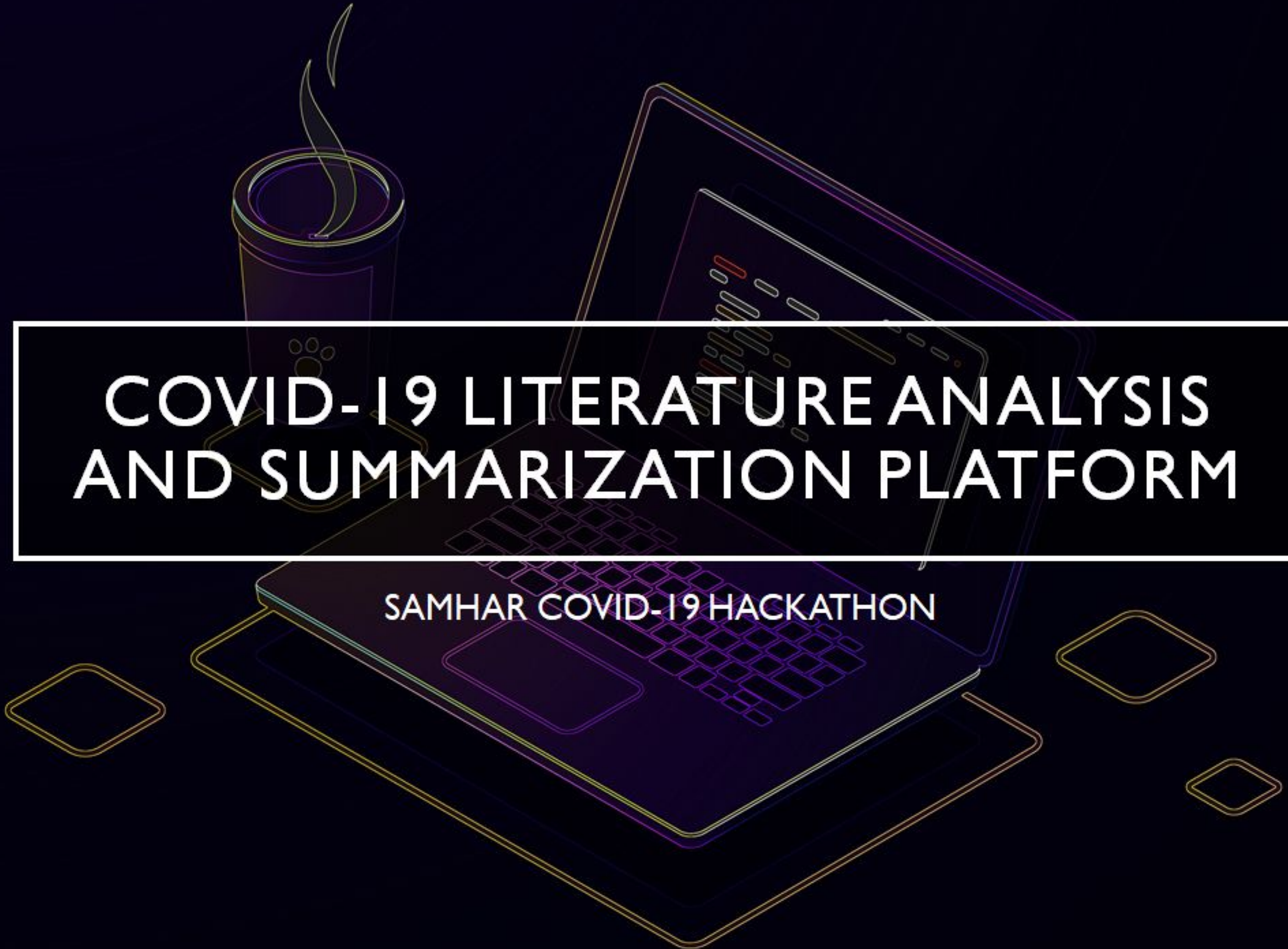


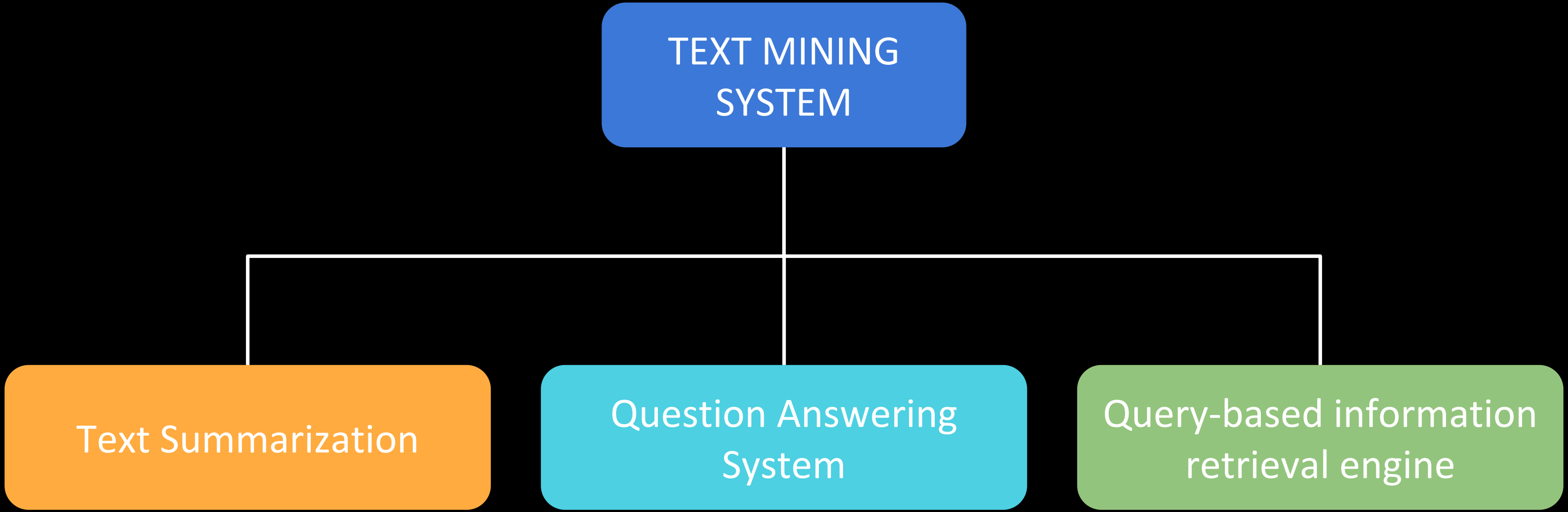
COVID-19 LITERATURE ANALYSIS AND SUMMARIZATION PLATFORM

SAMHAR COVID-19 HACKATHON

OUR UNDERSTANDING OF THE PROBLEM

- Owning to the pandemic, the acceleration in biomedical research has been breathtaking. As of now there are more than 50,000 Research articles already published and growing. If a text mining tool is provided to the researchers, it will be of great help.
- The system will be unsupervised in nature. Research is an exponentially growing forte deeming it impossible to label.
- Though there are some text mining tools in the market, hardly any of them are biomedical centric in nature and most of them make heavy use of supervised learning.
- We propose a solution which makes use of unsupervised learning to carry out text mining in fields which are extremely dynamic in nature and continuously growing.

PRODUCT FLOW



TEXT MINING SYSTEM

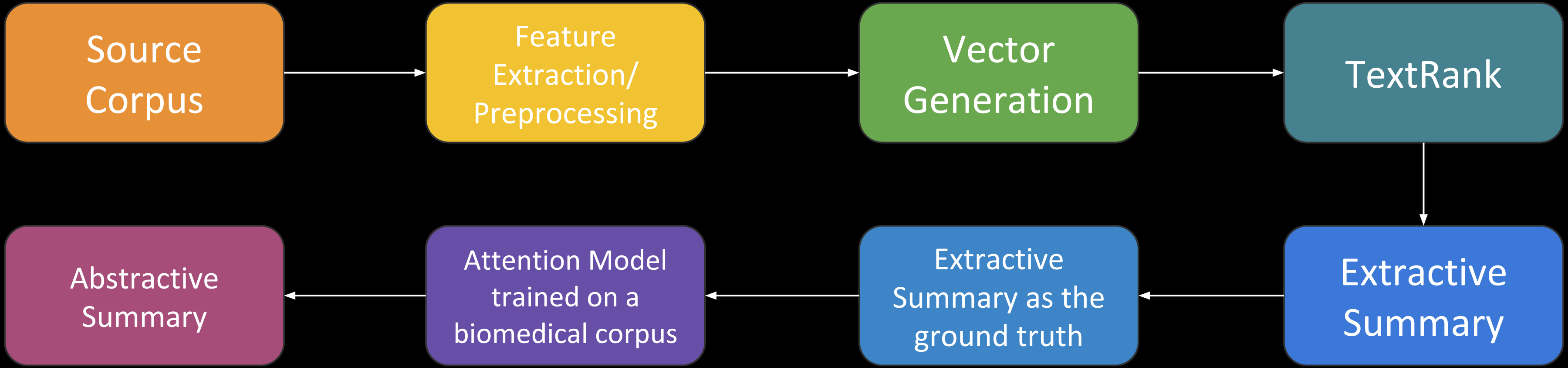
Text Summarization

Question Answering System

Query-based information retrieval engine

TEXT SUMMARIZATION SYSTEM

- We developed an unorthodox approach that makes use of extractive text summarisation.
- A python script is used to preprocess the data and extract only the necessary features from the corpus.



Source Corpus

Feature Extraction/ Preprocessing

Vector Generation

TextRank

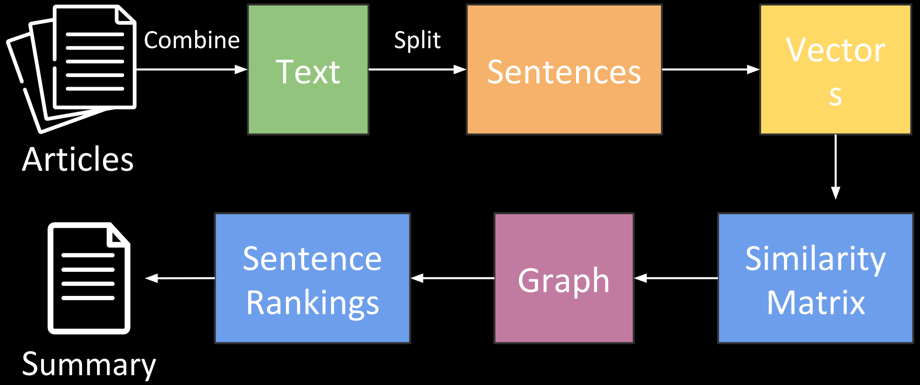
Extractive Summary

Extractive Summary as the ground truth

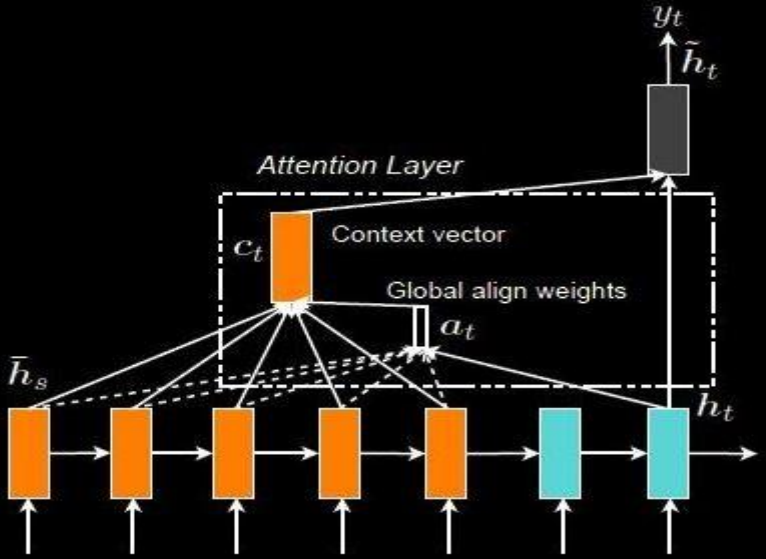
Attention Model trained on a biomedical corpus

Abstractive Summary

- For Extractive summarization, we make use of Unsupervised Graph based TextRank which is used as Ground Truth for the supervised Abstractive approach.
- Abstractive summarization makes use of a fine tuned Attention model, pretrained on a biomedical corpus.



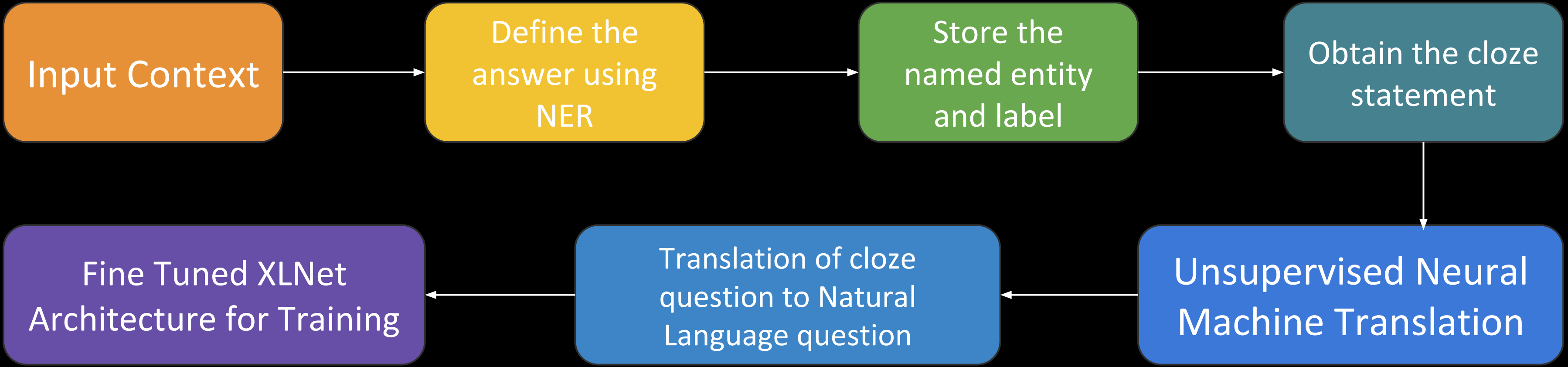
TextRank Architecture



Global Attention Mechanism

Q&A SYSTEM

We propose an Unsupervised Question Answering system based on Cloze Translation rather than the common Supervised Question-Answer-Context triplet Architecture.

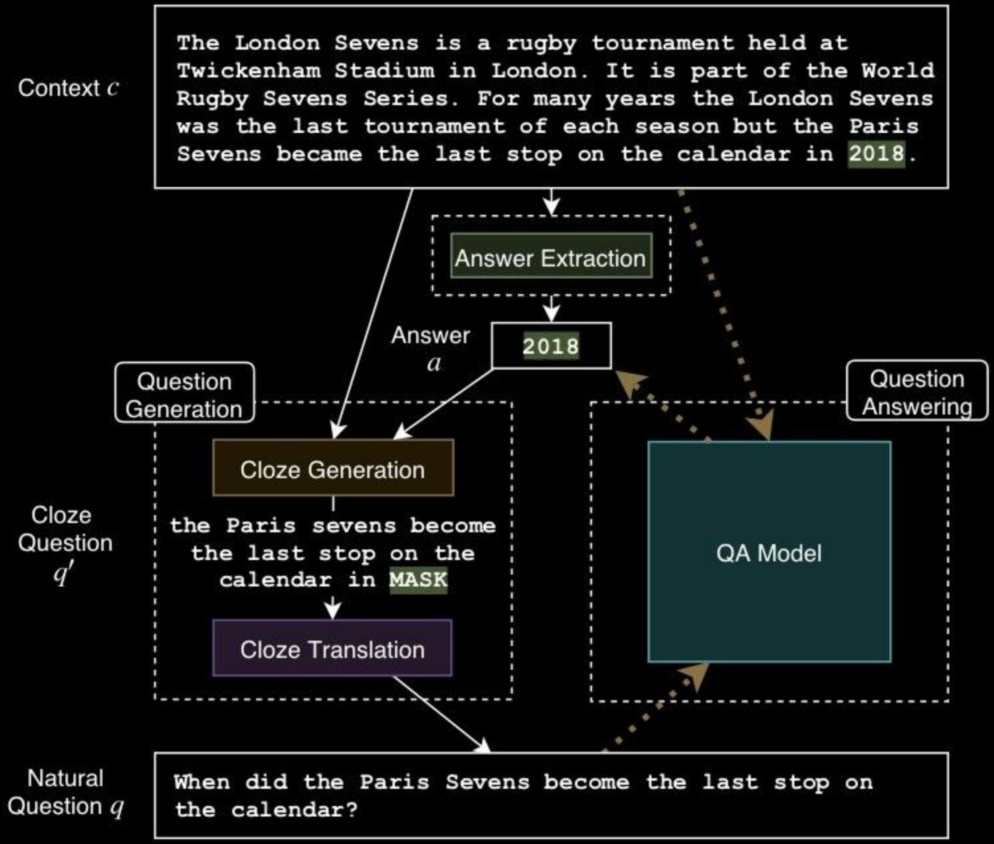

Input Context
Define the answer using NER
Store the named entity and label
Obtain the cloze statement
Unsupervised Neural Machine Translation
Translation of cloze question to Natural Language question
Fine Tuned XLNet Architecture for Training

1) We first sample a context in our target domain-BioMedical Research papers in our case.

2) Sampling is done from a set of candidate answers within that context, using pre trained components (Named Entity Recognition) to identify such candidates.

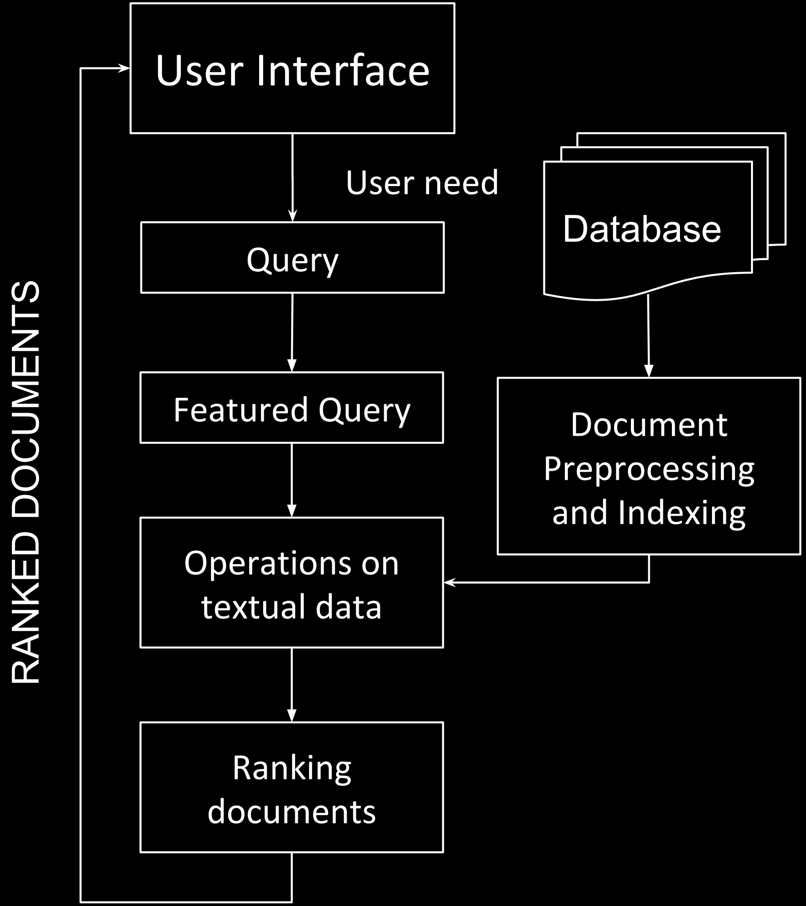
3) These require supervision, but no aligned (question, answer) or (question, context) data. Given a candidate answer and context, we can extract “fill-the-blank” cloze questions.

4) Lastly, we convert cloze questions into natural questions using an unsupervised cloze-to-natural question translator.

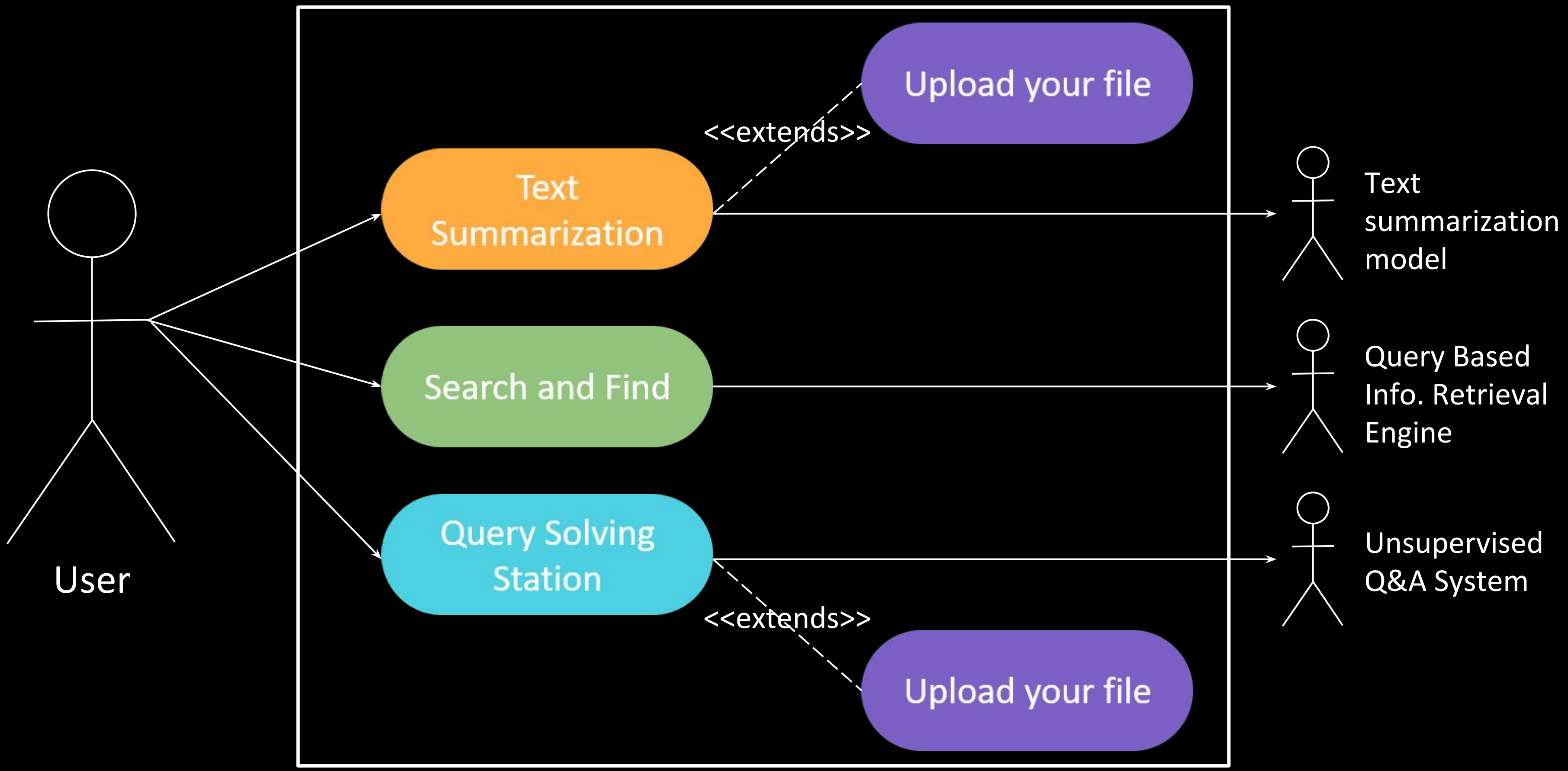

Context c
The London Sevens is a rugby tournament held at Twickenham Stadium in London. It is part of the World Rugby Sevens Series. For many years the London Sevens was the last tournament of each season but the Paris Sevens became the last stop on the calendar in 2018.
Answer Extraction
Answer a
2018
Question Generation
Question Answering
Cloze Generation
Cloze Question q'
the Paris sevens become the last stop on the calendar in MASK
Cloze Translation
QA Model
Natural Question q
When did the Paris Sevens become the last stop on the calendar?

INFORMATION RETRIEVAL

- For efficient retrieval of information, the first and rather essential step is to organize the information by indexing.
- We take notes of key signals - including keywords and other signals - and keep track of it all in the search index.
- Next we identify the relevant documents pertaining to the query and rank them.
- This can be done using a language model.
- In a LM (Language Model) approach to IR (Information Retrieval), we attempt to model the query generation process.
- Then, we rank the documents on the basis of the similarity score between the query and the documents.

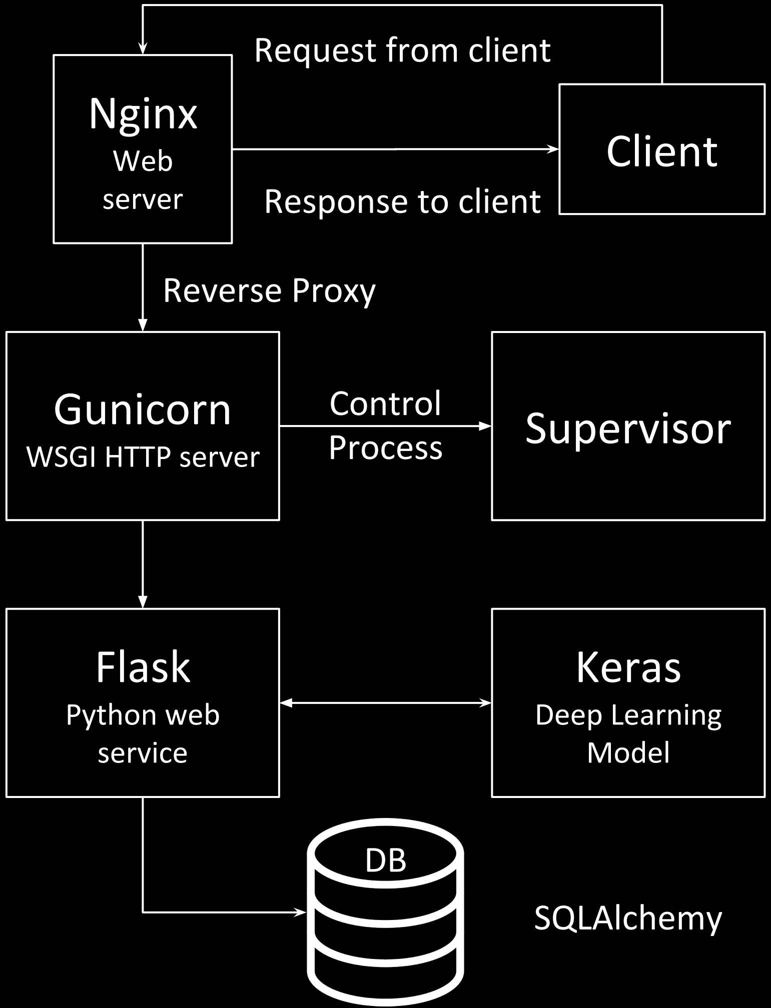


USE-CASE DIAGRAM



DEPLOYMENT DETAILS

- The backend of the model designed in the Flask framework will be served using gunicorn which is a Python WSGI HTTP server.
- The frontend designed using React and Bootstrap will be served on Nginx which is a free, open-source high functioning server.
- The database connectivity during development and testing stages will be done using SQLite. The Flask SQLAlchemy toolkit will be used for database operations.
- Communication i.e. Requests and Responses between the React and Flask servers will be using REST APIs and authentication will be done using JWTs.
- The project will be served on a virtual machine hosted on Microsoft Azure.



TECHNOLOGY STACK

Keras

TensorFlow

NLTK

spaCy

NGINX
Part of F5

gunicorn

SQLite

Flask

React

tesseract.js

Bootstrap

Azure

REFERENCES

- Text Summarisation : https://www.cs.utexas.edu/~asaran/reports/summarization.pdf
- Biomedical Text Summarisation : http://cs229.stanford.edu/proj2019spr/report/77.pdf
- Extractive Query Based Summarisation : https://www.aclweb.org/anthology/W18-5604.pdf
- Information Retrieval : https://www.cse.iitb.ac.in/~soumen/readings/papers/BergerL1999xlate.pdf
- Statistical Language Modelling : http://ciir.cs.umass.edu/pubfiles/ir-318.pdf
- Unsupervised Question Answering : https://research.fb.com/wp-content/uploads/2019/07/Unsupervised-Question-Answering-by-Cloze-Translation.pdf
- The BioMedical Corpus which was used to train the Attention model for Text Summarisation was provided by BioASQ which comprises of BioMedical contexts along with their Gold Standard Summary.