|  |  |
| --- | --- |
|  | Tech Immersion and Placement Programme  Applied Artificial Intelligence  6 January – 3 April 2020  Project Capstone Report  **Submitted By:**  Eugin Lee  (email@somewhere.com)  Koay Seng Tian  ([sengtian@yahoo.com](mailto:sengtian@yahoo.com))  **Supervisors:**  Nvidia:  Timothy Liu ([timothyl@nvidia.com](mailto:timothyl@nvidia.com))  Lai Zhang Sheng ([zlai@nvidia.com](mailto:zlai@nvidia.com))  Republic Polytechnic:  Tan Poh Keam ([tan\_poh\_keam@rp.edu.sg](mailto:tan_poh_keam@rp.edu.sg)) |

Table of Contents

[1 Background 2](#_Toc36451878)

[1.1 Goal 2](#_Toc36451879)

[2 Methodology and Design 2](#_Toc36451880)

[2.1 Software Data Flow 2](#_Toc36451881)

[2.2 Collecting User Feedback 3](#_Toc36451882)

[3 Applications 3](#_Toc36451883)

[4 Insights 4](#_Toc36451884)

[4.1 BERT Models Are Not Equal 4](#_Toc36451885)

[4.2 ROUGE and not BLEU 4](#_Toc36451886)

[5 Results and conclusions 5](#_Toc36451887)

[6 Recommendations 5](#_Toc36451888)

[7 Appendices 6](#_Toc36451889)

[7.1 GitHub 6](#_Toc36451890)

[7.2 Development Environment 6](#_Toc36451891)

[7.3 Extractive Summarizer 7](#_Toc36451892)

[7.4 Running the Summarizer 8](#_Toc36451893)

[7.5 Library Versions 9](#_Toc36451894)

[8 References 10](#_Toc36451895)

# Background

This project is a collaboration between Nvidia and Republic Polytechnic (RP) as a capstone project for the Tech Immersion and Placement Programme (TIPP) programme by Infocomm Media Development Authority (IMDA). The project aims to provide a real-world work environment for the students to apply artificial intelligence techniques taught in the course. Supervisors in return receive the intellectual property of the final product.

## Goal

The goal of this project is to provide an annotation tool, to differentiate key takeaways in a corpus of text. The summarisation technique will be powered by Bidirectional Encoder Representations from Transformers (BERT) models.

# Methodology and Design

Extractive Summariser using BERT transformer model.

The main entry to the application is via a webpage where user enters a URL as an input. This BERT model summariser aims to extract key feature sentences of the main corpus.

These extracted features will then be presented as an annotated form, together with the main corpus as an output document.

Using a client-server model, the web application provides seamless transition between server (flask), user interface (Streamlit) and the underlying python code. The software is developed using open- source software, libraries and modules.

For evaluation, we will be using the ROUGE scoring system which is optimised to calculate distances/ similarities between summarised articles. By using ROUGE and comparing between BERT generated summaries against human generated summaries, we calculate the similarity between the text and recommended ways to further improve the results.

## Software Data Flow

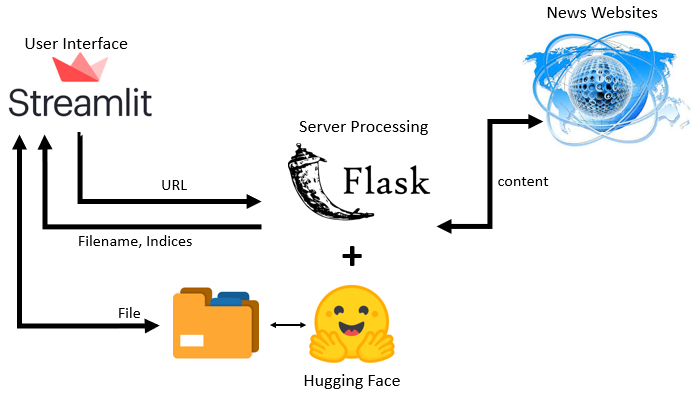


Figure 1 Software Data Flow

The software is developed using available open-source application frameworks (Streamlit, Flask etc.) and Hugging Face BERT/Transformer model. Python scripting language is used mostly.

Streamlit architecture is based on the ability to write web application the same way a plain Python script is written. Streamlit applications have a unique data flow. Any time something must be updated on the screen (for example, when the application needs to response to a button when pressed), Streamlit will just rerun the entire Python script from top to bottom.

This will pose a challenge for the application/UI developer because it is not implemented as a 'call back', like most web applications will perform. Some of these quirks can be modified using Streamlit's cache decorator (streamlit@cache) which allows developers to skip certain costly computations when the application reruns. However, such technique, as we have observed, may create stability issues.

## Collecting User Feedback

User can feedback or suggest user-defined summary by checking or unchecking the returned check boxes. The original and enhanced summaries are saved as CSV files for future model fine tuning and improvements.

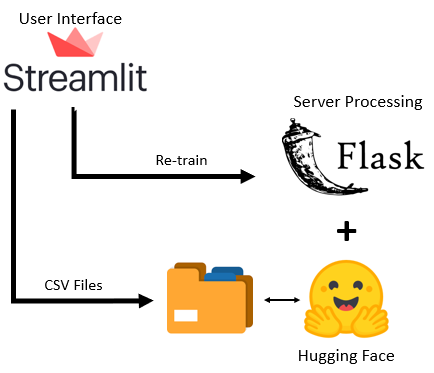


Figure 2 Capturing User Feedbacks

# Applications

In this project we utilised these libraries. The versions used in current development will be documented in later section.

|  |  |  |
| --- | --- | --- |
| **Python Library** | **Usage** | **Installation** |
| Newspaper3k | Scraping tool | pip install newspaper |
| Pandas | Data manipulation | conda install pandas |
| summarizer | BERT summariser | pip install summarizer |
| datetime | Time stamp | - |
| Rouge | ROUGE scoring | pip install rouge |
| NLTK | Tokenising words | pip install NLTK |
| Flask | Server | pip install flask |
| Streamlit | User Interface | pip install streamlit |

# Insights

In this project we discovered two major takeaways that arise due to our experiments and research. We discovered that (1) model selection plays a big part of in operational efficiency, (2) and

## BERT Models Are Not Equal

When we first started out, we utilised the “stock” BERT model (bert-base-uncased) that was embedded in the summarizer library. While testing, we quickly realised that we may face into some usability issues as the waiting time for the extractive summary takes around 17 seconds to complete.

According to recent research on attention spans, current attention spans range between 8 to 12 seconds. This means that users might regard the seventeen seconds wait to be “too long” and may not even use the product even if it was a perfect product. We need to find a compromise between speed and performance.

To achieve this, we look at the different BERT models available in the market. We eliminated the “large” BERT models at first cut, because they contain more parameters which will mean longer processing time. This narrows our focus to the base BERT model. To evaluate the performance, we run a loop test, controlling all parameters except a model change. In this experiment, we recorded the performance in the form of processing time and text length.

At this point, we are using just quantitative measures, such as (1) processing speed and (2) text length as our selection.

Within the BERT models, we chose DistillBERT as its processing speed falls within the sweet spot between 8-12 seconds and the summarized text length seems to be within an acceptable ballpark area. In addition, to artificially improve on the run time, we created a time illusion (to make time seem to pass faster), by inserting loading animation and stimulating an active program during the processing runtime.

|  |  |
| --- | --- |
|  |  |

Figure 3 Where You Got this from

## ROUGE and not BLEU

We selected ROUGE as our main technique to measure model accuracy. The main reason is because ROUGE provides more insights into the scoring mechanism. In ROUGE, the precision and recall scores are also provided which gives an understanding in the event of a low score, whether the machine generated text, is too short or too long, or is it not relevant. With this knowledge, we could either tweak the parameters correctly, or if the summary is poor, we can fine-tune it with a longer relevant corpus. With BLEU, another popular technique, only the final BLEU score (between 0 to 1) is given. This gives us no direction to improve.

ROUGE and BLEU are both popular methods in summaries evaluation. They use similar calculation methods surrounding n-grams and variations of recall and precision. When testing, we discovered that BLEU does not seem to work very well when there are differences in corpus length (between BERT and human generated). This could be due to the brevity penalty effect. The documentation recommended a smoothing technique to overcome this, but we did not find the value to be as intuitively accurate as ROUGE.

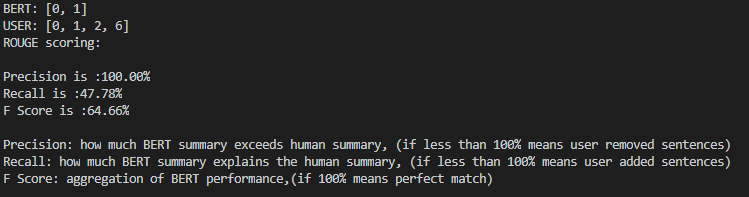


Figure 4 Rouge Score

# Results and conclusions

We successfully created and deployed a web application that allows a user to (1) enter a URL and received a full text with annotated key points (2) user can provide feedback, which can be used to provide a quantitative scoring and (3) we could improve the BERT model based on the scoring.

Our average runtime for the model is around 8-10 seconds. The model for some reason, seems to run slightly faster after a few iterations.

Project managers from Nvidia are extremely satisfied with the progress we made.

# Recommendations

Given that the code is designed to be modular based and robust, the engineer will be able to switch out part of the code further customised it. For example, as our parser is optimised for newspaper articles, it may not work as well for PDF documents. In which case, the engineer could switch in a PDF parser instead.

The Streamlit application works perfectly for a proof-of-concept user interface; to quickly unlock the power of the underlying model and test out concept feasibility. However, the UI content has some limitations such as feedback functionality like a HTML, and limited fine-tuning options, which limits our ability to improve the product.

In Natural Language Processing (NLP) literature, the computer-generated summary is always compared against a “gold standard” summary to determine its accuracy. However, in our case, the BERT summary is compared against to a random user feedback. This user may have its own biasness and may have an inconsistent proficiency of the subject. This could affect our evaluation score of the model. To avoid this issue, it might be useful to collect an aggregation of user responses, before tweaking the model.

# Appendices

## GitHub

The source codes, documentation and the test data for the project are hosted in Github.

Repository URL: <https://github.com/koayst/rp_capstone>

If you have issue accessing the GitHub, please contact any of us.

Github uses Git as an open-source version control system. The purpose is to keep the revisions straight, storing modification in a central repository. This allows us, as a developer, to easily collaborate, as we can download a new version of the software, make changes and upload the newest version, after we make the modifications.

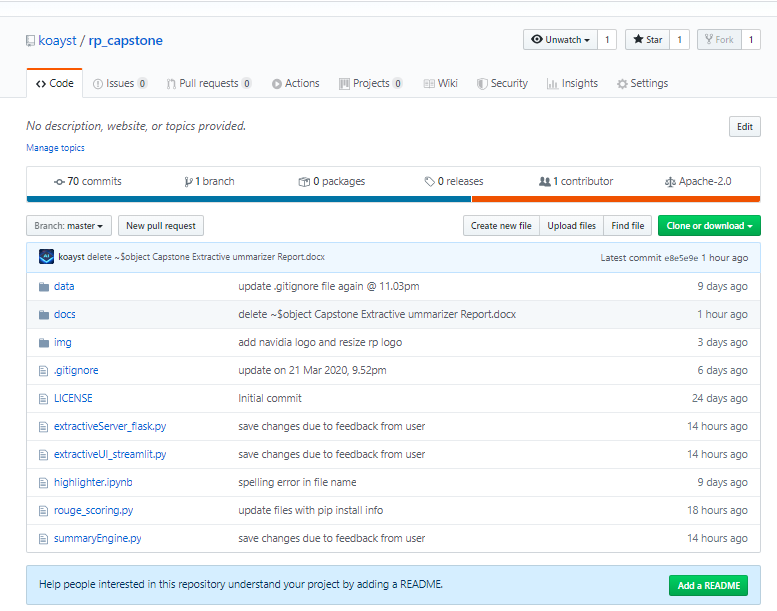


Figure 5 GitHub Repository Screen Shot

## Development Environment

We are using Microsoft Windows Operating System (Windows 10 and Windows 8) as our development environment.

Anaconda Individual Edition (2019-10) was installed and then relevant modules are installed/updated as required by the project.

## Extractive Summarizer

To run the system, it is advised to create a separate virtual environment. Use Git to download the source codes by cloning it from GitHub.

Install the required libraries and modules as indicated in section 3.

Open two terminals on a local machine (command prompt for Windows OS) i.e. one to run the Flask server and the other one to run the Streamlit webapp.

* + Flask server: *python extractiveServer\_flask.py*
    - Running on http://127.0.0.1:5500/ (Press CTRL+C to quit)
  + Streamlit: *streamlit run extractiveUI\_streamlit.py*
    - Local URL: <http://localhost:8501>
* Once Streamlit is running, you will notice the webapp is running in your browser. If the Streamlit web app is not running in your browser, you can bring up the Streamlit web app by ‘http://localhost:8501’. We are using Chrome browser for development and testing.

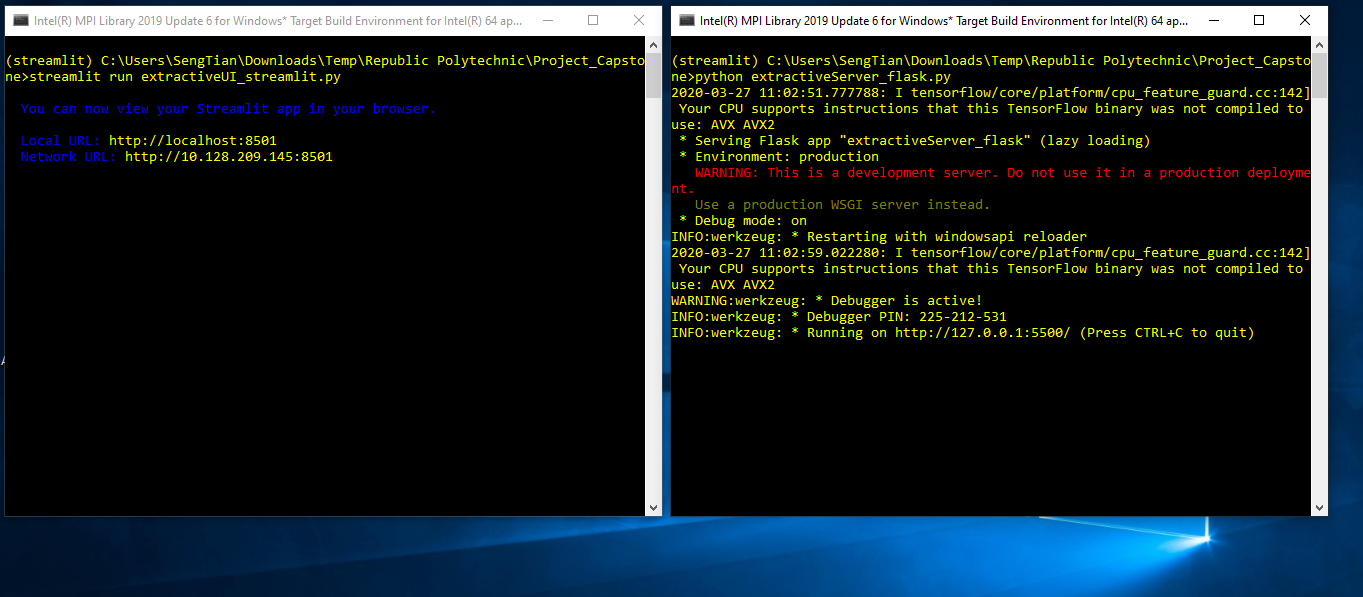


Figure 6 Screen shots for running the summarizer

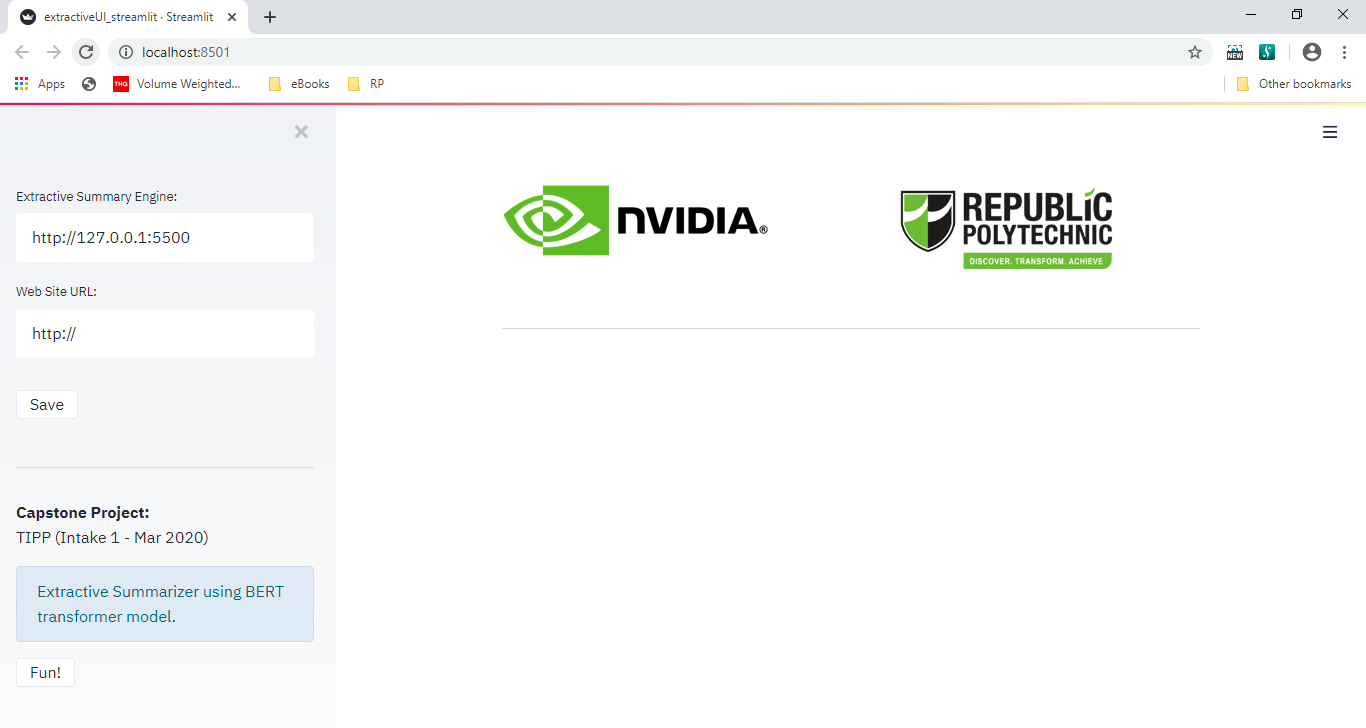


Figure 7 Streamlit webapp screen shot

## Running the Summarizer

Go to Singapore’s Channel New Site (<https://www.channelnewsasia.com/>) and select a news article you want to summarize. Copy the URL and paste it to the Streamlit webapp. Press ‘Enter’ to start the summarizer engine.

The output will show a list of checkboxes. User can enhance the modelling by checking and unchecking the checkboxes.

User’s feedback can be saved by clicking the ‘Save’ button.

Two CSV (Comma-Separated Values) files are stored in the Data directory. It is highlighted as shown below

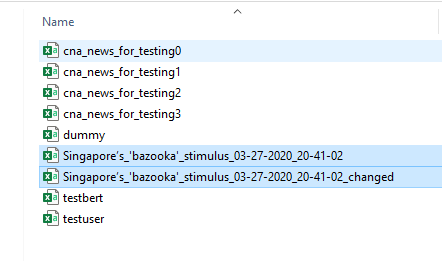


Figure 8 Two CSV files

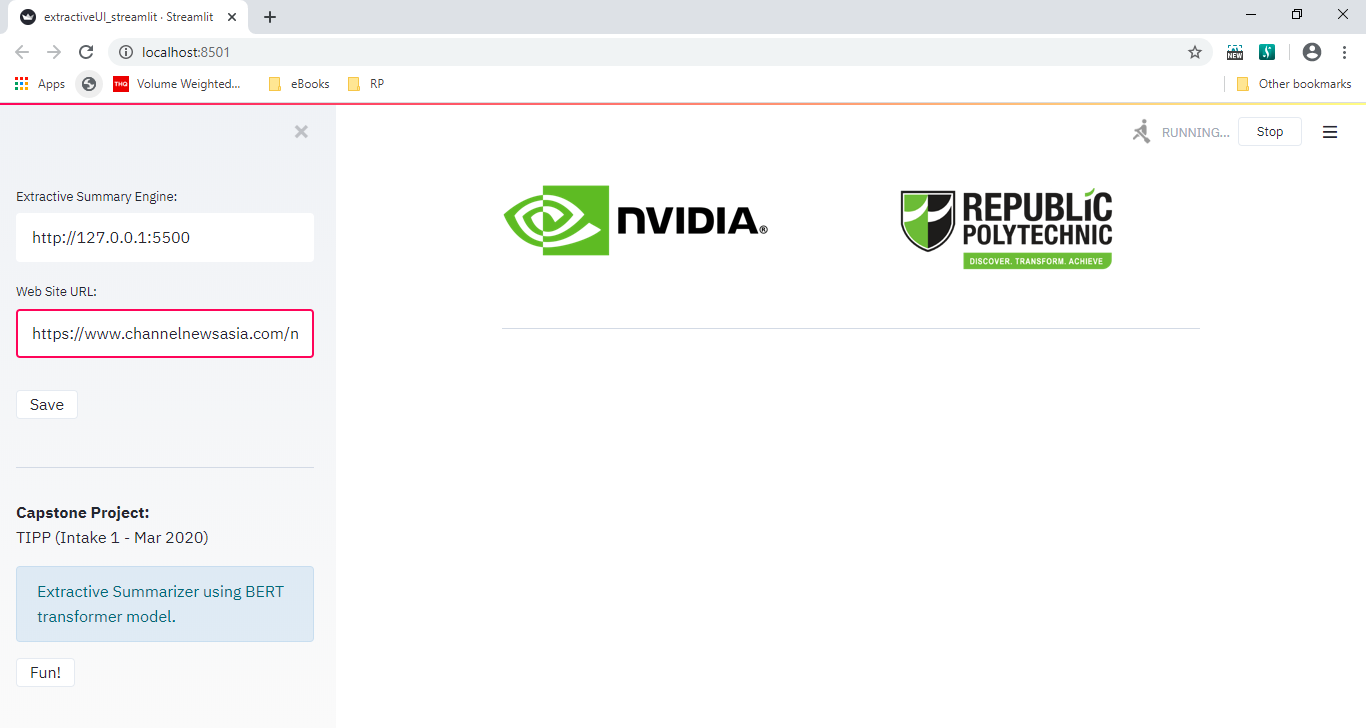


Figure 9 Input URL and Press ‘Enter’

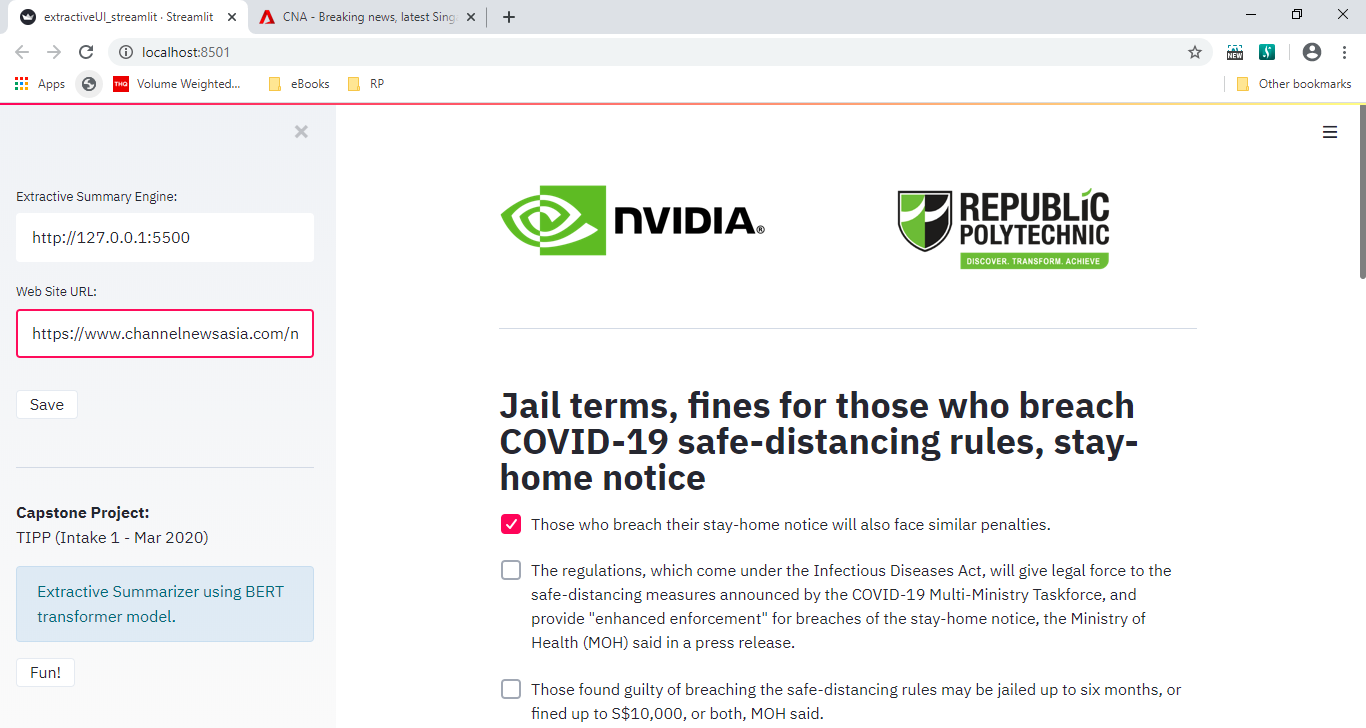


Figure 10 Output of Summarizer

## Library Versions

|  |  |
| --- | --- |
| **Name** | **Version** |
| Python | 3.7.6 |
| Flask | 1.1.1 |
| Keras-applications | 1.0.8 |
| Keras-preprocessing | 1.1.0 |
| Matplotlib | 3.1.3 |
| Newspaper3k | 0.2.8 |
| Numpy | 1.18.1 |
| Pandas | 1.0.1 |
| Pytorch | 1.4.0 |
| Regex | 2020.2.20 |
| Request | 2.23.0 |
| Rouge | 1.0.0 |
| Streamlit | 0.56.0 |
| Tensorflow | 2.0.0 |
| Tokenizers | 0.5.2 |
| Torchvision | 0.5.0 |
| Transformers | 2.2.0 |
| Urllib3 | 1.2.58 |

# References

* Streamlit – <https://www.streamlit.io>
* Flask - <https://flask.palletsprojects.com>
* Eugin: Please update your reference here