**Guideline:**

1. You may use Python or R for the challenge. A notebook might be the best way to show your code and write your comments/thoughts for the reviewers to follow along.

2. Deliverables shall be a zip file contain of:

a. Notebook or Rmarkdown

b. Artifacts: Any models and external data source you have created or used.

c. README.md file to help interviewer understand how to execute your script

d. Helper: Any utility or helper functions you use in this challenge.

3. You can select any 2 out of the 3 questions.

**Q1.** *Text Summarization*

1. Refer to this article from Yahoo Finance - <https://finance.yahoo.com/news/musk-says-twitter-in-material-breach-of-its-merger-agreement-133732089.html> . You are required to:
2. Parse and clean text
3. Sentence Tokenization
4. Word tokenization
5. Word frequency table
6. Build an Extractive Text Summarization model
7. You are required to build an Abstractive Text Summarizer from your clean corpus above.

**Note:** You can use tools of your own choice for the tasks above.

**Q2.** *Text Classification*

Insurance industry is highly sensitive about Epidemic disease. A disease outbreak implying potential of high claims cases to insurer, hence resulting in financial loss. AIA being a major insurer in Asia pacific region, we have to always keep a close eye on epidemic trend and carry out pre-emptive action if any unusual trend been observed. Our claims department colleague deems social media is a good 3rd party source where we can obtain epidemic information. This data will complement and provide additional dimension for business decision making. However, to eyeball each single thread online and record down the topic category is time consuming and resource intensive, therefore, business user then reaches out to Data Science team and seek for a more effective solution. As a Data Scientist, you are required to:

1. Develop a text classifier model to help classify conversation topic from social media.
2. Include any interesting insight that you observe from the dataset based on your analysis.
3. Be mindful your target audience is business user; your final model output and business insight need to be easily understandable.

\*To ease your development, a covid 19 dataset (with label) (“Q2 Dataset”) will be provide. You shall use this dataset for your model development.

**Q3.** *Question-Answering*

1. There is a pdf file (“Q3 PDF”) about Covid 19 information in the folder. You are required to use this file and to build a Question-Answering model to answer the questions below
2. Which virus variants of covid 19 are detected in Malaysia?
3. What is the mean age of patient that Brought In Dead?
4. Which state has the highest number of patients that died in hospital?
5. Why foreigners have higher number of BID than Malaysians?
6. Convert answers from i and iv into speech or audio. You can save as audio file or embed into your notebook.

**Note:** You can use tools like PyPDF2, tools from Deepset.ai, Hugging Face, Google TTS, AWS Polly Or tools of your own choice for the tasks above.