# DistilBERT Question Answering System

Assignment-2

Harihar Panda(umds20003)

**Objective :-** To develop a small Question-Answering system using pretrained DistilBert model.

**Introduction :-** In simple transformers there are numerous types of model used for QA tasks. DistilBERT is one of the popular type of model built on Wikipedia (2500M words) and Book Corpus (800M words). In general BERT stands for Bidirectional Encoder Representation of Transformer. It learns information from both the left and the right side of a token's context during training. It is based out of 12 layers (transformer blocks), 768 hidden layers, 12 attention models and 110M parameters. It is pretained on Masked Language Modelling and Next Sentence Prediction.

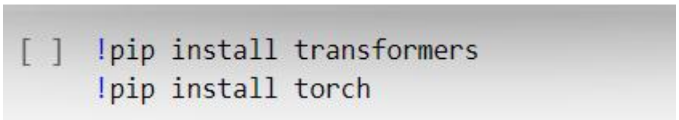
The biggest challenge especially for academic researchers and small companies is how to put these models in production under low latency constraints? Well, the potential solution found as Knowledge Distillation. A technique to compress a large model, called the teacher, into a smaller model, called the student.

**Approach :-** In the teacher-student training, a student network is trained to mimic the full output distribution of the teacher network. DistilBERT trains the student to generalize the same way as the teacher by matching the output distribution. Using the teacher signal, a smaller language model, called DistilBERT was built, from the supervision of BERT.

DistilBERT’s 66 million parameters make is 40% smaller and 60% faster than BERT-base, all while retaining more than 95% of BERT’s performance.

**Demo :-** Hugging Face is a python based library that exposes an API to use many well known transformers architectures, such as BERT, RoBERTa, GPT-2 or DistilBERT, that obtain state-of-the-art results on a variety of NLP tasks like text classification, information extraction, question answering and text generation.

1. Getting started with Transformers only requires to install the pip package. The library has seen super-fast growth in PyTorch.

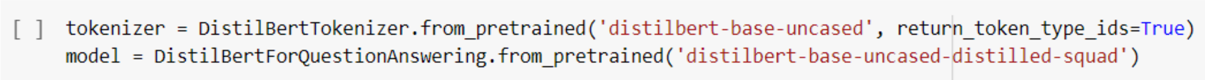


```
[ ]  !pip install transformers
     !pip install torch
```

2. From the transformers the ‘DistilBertTokenizer’ and ‘DistilBertQuestionAnswering’ has been imported.

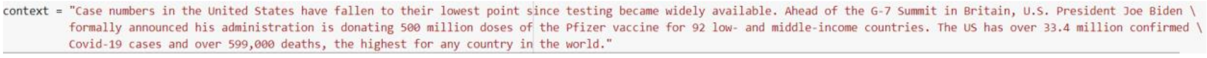
3. We are using the tokenizer created by ‘DistilBertTokenizer’ from the pretrained model called as the ‘distilbert-base-uncased’. This is the student model from the BERT base. Then we are

calling the actual trained model ‘distilbert-base-uncased-distilled-squad’.

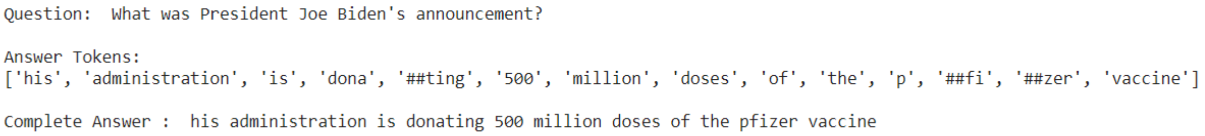


```
[ ]  tokenizer = DistilBertTokenizer.from_pretrained('distilbert-base-uncased', return_token_type_ids=True)
     model = DistilBertForQuestionAnswering.from_pretrained('distilbert-base-uncased-distilled-squad')
```

4. We have the baseline knowledge from the pretrained model. Now we give the context based on which we ask certain questions. After that encoder has been called which uses the tokenizer to get tokenize in a correct way post which it starts scoring by using the scores that is already there in the trained model and then it converts to the related tokens. Finally, it aggregates the related tokens to get the complete answer.



```
context = "Case numbers in the United States have fallen to their lowest point since testing became widely available. Ahead of the G-7 Summit in Britain, U.S. President Joe Biden \
          formally announced his administration is donating 500 million doses of the Pfizer vaccine for 92 low- and middle-income countries. The US has over 33.4 million confirmed \
          Covid-19 cases and over 599,000 deaths, the highest for any country in the world."
```



```
Question:  What was President Joe Biden's announcement?

Answer Tokens:
['his', 'administration', 'is', 'dona', '##ting', '500', 'million', 'doses', 'of', 'the', 'p', '##fi', '##zer', 'vaccine']

Complete Answer :  his administration is donating 500 million doses of the pfizer vaccine
```