# Hack to Hire- Data Science(4010)

Question-Answering Model using NLP

# Intro

This presentation comprises of the techniques used in creating the Question-Answering Model using Hugging Face for Quora question-answer dataset.

The project comprises of several basic steps using NLP techniques i.e.,

- Tokenization
- Data pre-processing
- Model Training
- Model Evaluation

# Overview of Dataset

- Quora Question- Answer Dataset
  - Features : ‘question’, ‘answer’
  - Training Data: 45,121 rows
  - Test Data: 11,281 rows
- I have split the dataset into Test and Training Set with ratio of 80:20

# Pre-processing

- This step of the model uses NLTK tools and libraries to pre-process the data to remove all the irrelevant information.
- In this process, I used basic regular expression library, to rid the data of any irrelevant information like special characters, URLs.
- After this, I used ‘punkt’ and ‘stopwords’ tokenization models.
- I have also used a combination of ‘stop_words’ and ‘PorterStemmer’ to achieve Stemming  while preprocessing the data.
- I have then used ‘word_tokenize’ to filter out common stop words from the data.

# Model Selection and Tokenization

- This step of the model uses ‘t5-small’, ‘bert-base-uncased’, ‘gpt2’ models to evaluate and train the dataset.
- I have made use of ‘AutoModelforQeustionAnswering’ and ‘AutoTokenizer’
- I have also made use of padding token if missing.
- After tokenization of the preprocessed datasheet, the structure for the dataset is as follows:
    - ‘Question’ , ‘answer’, ‘input_ids’, ‘attention_mask’, ‘start_position’, ‘end_position’.

# Model Training

- This step of the model training Arguments like
    - Batch size, epoch, logging steps
- I have also made use of ‘DefaultDataCollator’ for dynamic padding.
- Training of the models include setting up a Trainer claa for model, tokenizer, data colaltor, arguments and dataset.
- It then trainsthe models and saves them as well.

# Model Evaluation

- This step of the task includes evaluating the models.
- For this step, I have made use of F1 Score, BLEU score and Rouge score.

# Final Presentation

The final presentation will include the demo for the model.

GitHub Repository: https://github.com/Shagun0402/Hack-to-Hire-QA-Model

# Thanks!

Name: Shagun Paul
Email Id: shagun.paul0402@gmail.com
Phone No: 9625275670