National College of Ireland

Configuration Manual

MSc Research Project
Data Analytics

Diksha Bhatnagar
Student ID: 21183961

School of Computing
National College of Ireland

Supervisor: Rejwanul Haque

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Diksha Bhatnagar
Student ID:	21183961
Programme:	Data Analytics
Year:	2022-23
Module:	MSc Research Project
Supervisor:	Rejwanul Haque
Submission Due Date:	14/08/2023
Project Title:	Configuration Manual
Word Count:	632
Page Count:	8

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Diksha Bhatnagar
Date:	13th August 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Diksha Bhatnagar
21183961

1 Google Collaboratory Setup

This sections elaborates on the steps to setup Google Collaboratory for execution of the python code.Firstly navigate to the Link https://colab.research.google.com/.Teeter and Barksdale (2011) Then click on New Notebook as mentioned in Figure 1. Once the new file opens we are ready to upload files , mount to share drive , install python libraries etc.

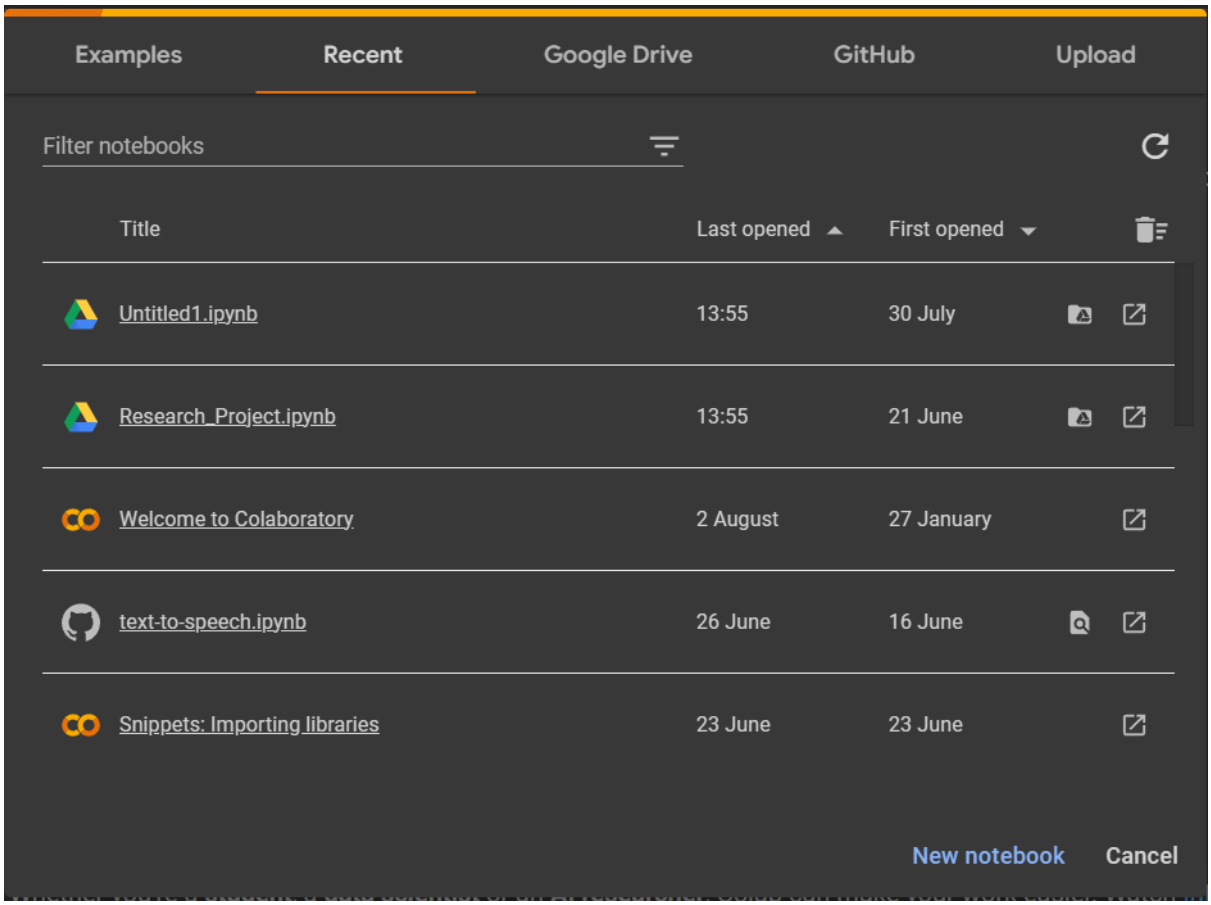

Figure 1: Opening New Collaboratory Session

2 Mounting to Google Drive

There are two ways to upload Data files,Trained Model folder.First approach requires to sign in with gmail id and then mount Google Collab notebook to Google drive.Second approach is to upload all the files in real time environment for a particular session and then the files need to be uploaded again for every run.Then Click on File followed with Open and then navigate to Upload and choose file ResearchProject.ipynb file.Then code file will open in new window.Figure 2

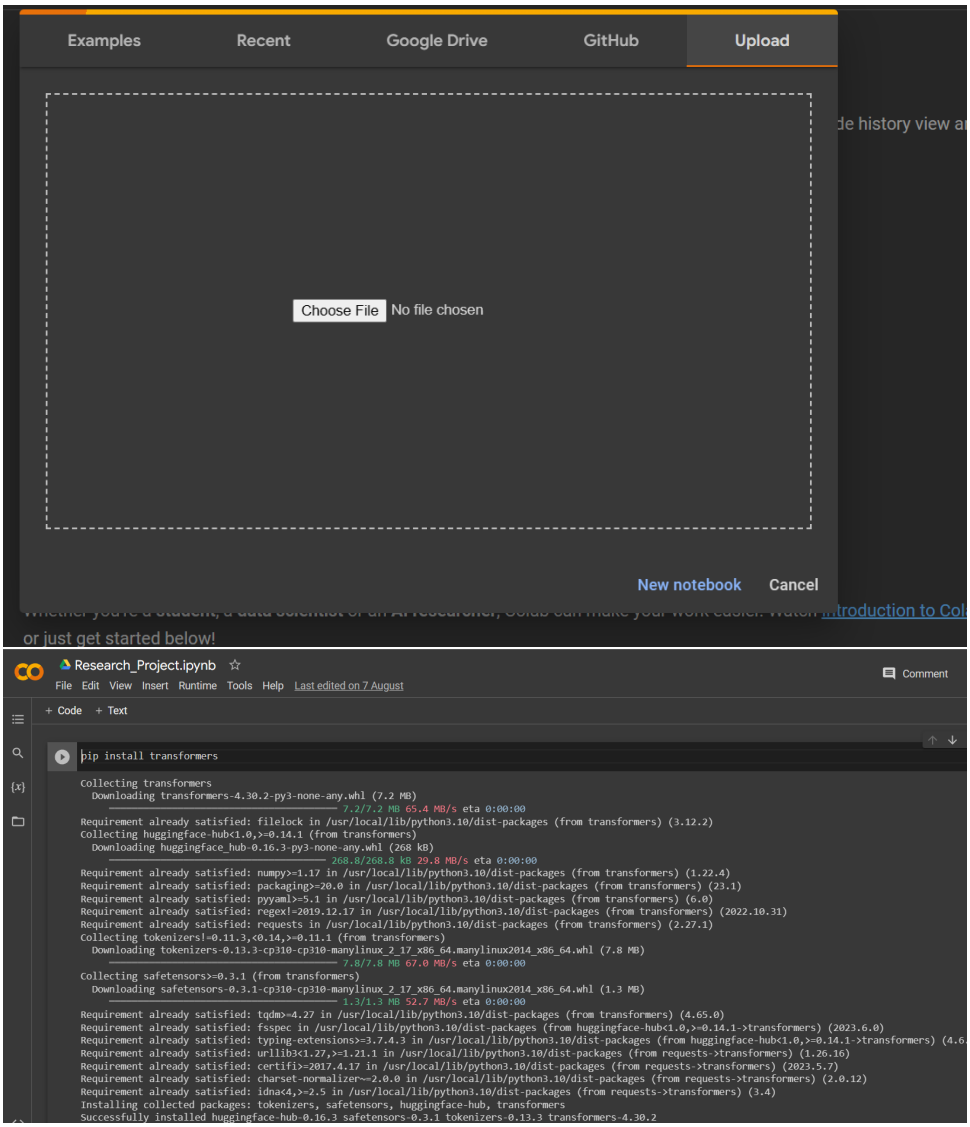

Figure 2: Opening and Uploading Python Code File

3 Uploading Data Files and Fine Tuned Model Folder Files

Click on the Folder icon on the left corner.Then click on upload files i.e upload to session storage for temporary storage.Upload all the below mentioned files from the folder.Please

refer Figure 3

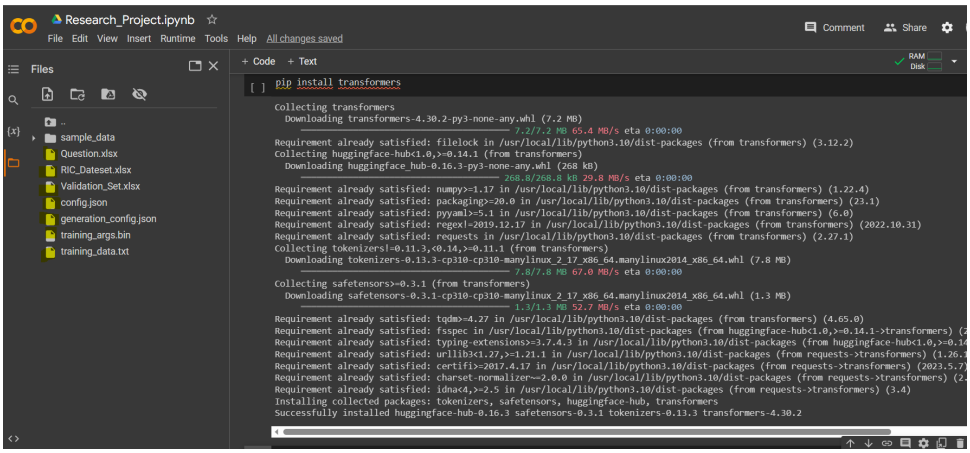

Figure 3: Uploading Data and Model Files

- RIC_Dataset.xlsx
- Question.xlsx
- training_data.txt
- training_args.bin
- generation_config.json
- config.json
- training_args.bin
- Validation_Set.xlsx
- Book.xlsx
- Logs Folder

4 Installing Python Libraries

A list of Python libraries needs to be installed before starting line by line code execution.Figure 4

5 Executing Code with Fine Tuned GPT Model

Once all the files are uploaded successfully and all the libraries are installed in the virtual engine. Next step is to execute the code line by line from the section "**Model Fine Tuning & Visualisation**". As suggested in Figure 5. The prior steps in the code are the fine tuning steps of GPT2 model.The overall code flow starts from Data pre-processing

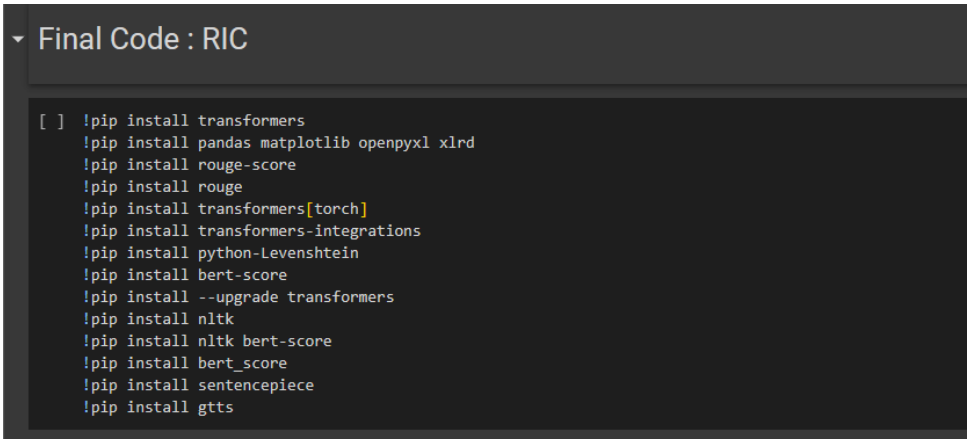

Figure 4: Installing Python Libraries

then model fine tuning followed with Hyper-parameter optimisation and then generating text based on prompts. Lastly comparison with baseline models like GPT2 and T5 is done on multiple metrics like BERTZhang et al. (2019),ROUGELin (2004) etc. with a visualisation piece added to give an high level overview on the data as well as the scores generated for output text responses with reference text.

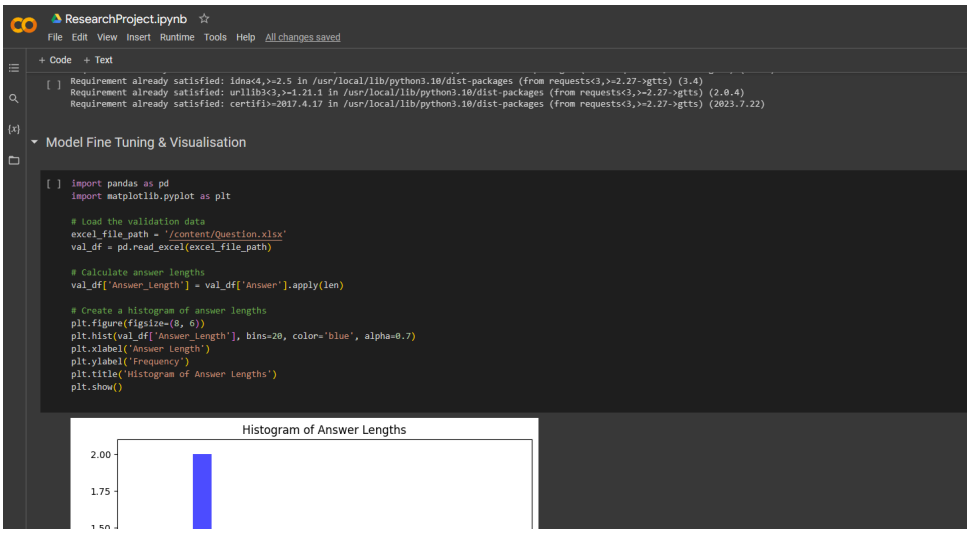

Figure 5: Code Execution Section

6 Outputs and Code Snippets

This Section captures the output visualisation once the code is executed in steps as follows:

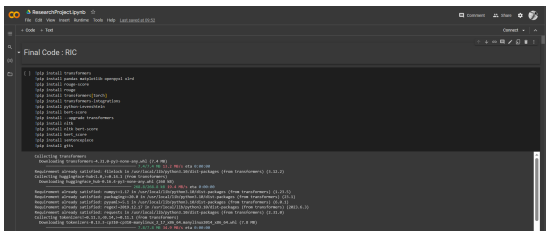

Figure 6: Installing Python Libraries

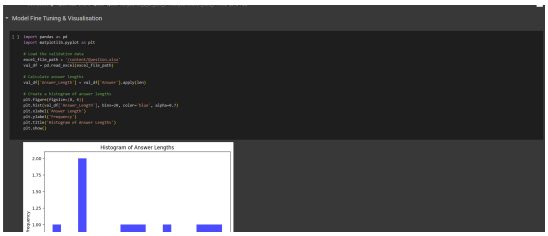

Figure 7: Visualisation

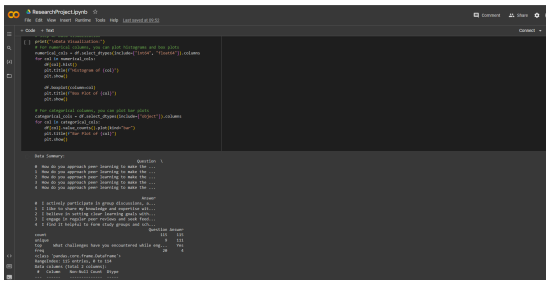

Figure 8: Data Overview

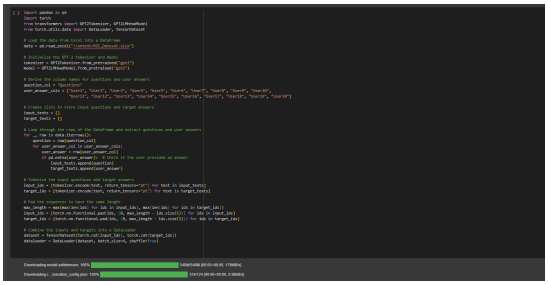

Figure 9: Combining data to dataloader

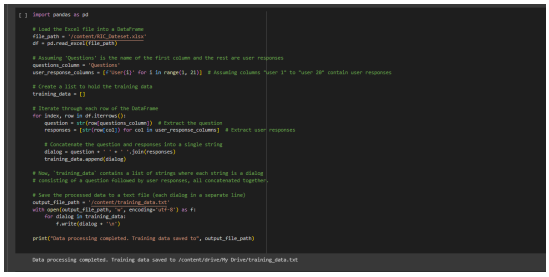

Figure 10: Creating the training set

Figure 11: Creating Tokenizer

Figure 12: Fine Tuning the Model

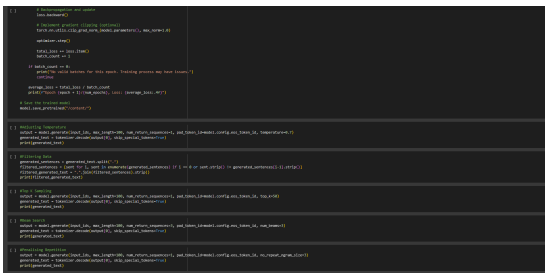

Figure 13: Hyper Parameter Tuning and Saving the Model

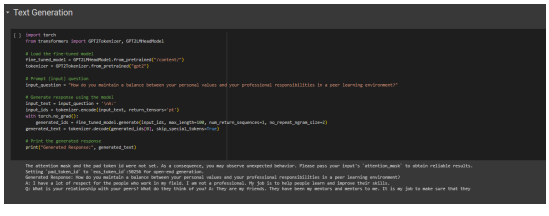

Figure 14: Text Generation

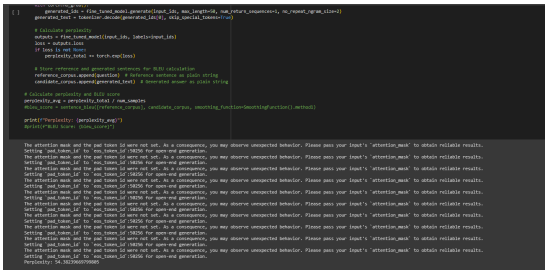

Figure 15: Calculating Perplexity

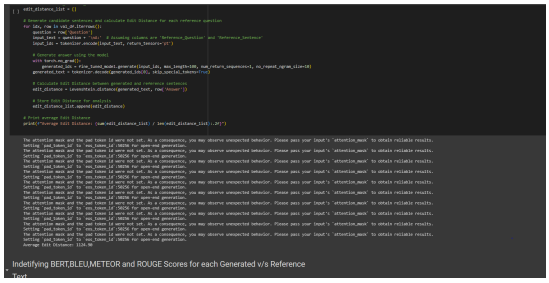

Figure 16: Calculating Average Edit Distance

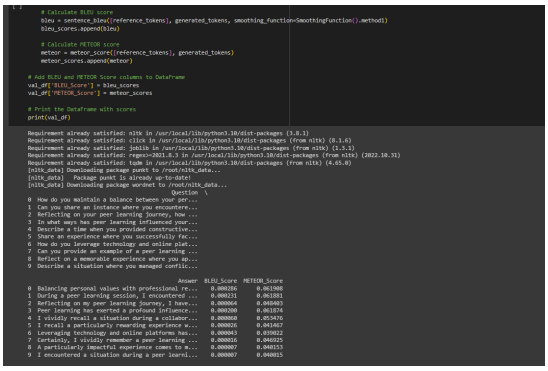

Figure 17: Calculating METEOR and BLEU Score

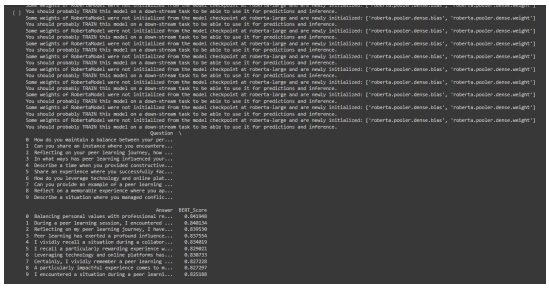

Figure 18: Calculating BERT Score

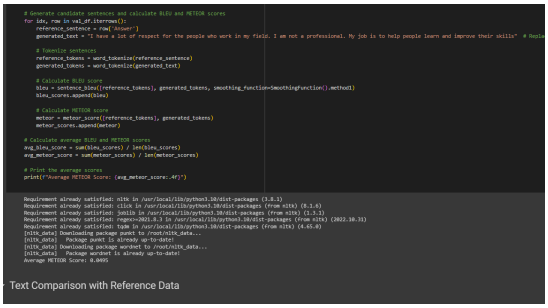

Figure 19: Average METEOR Score

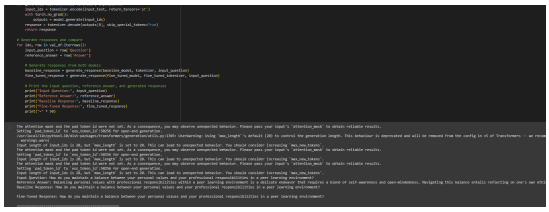

Figure 20: Text Comparison : Baseline and Fine Tuned Model

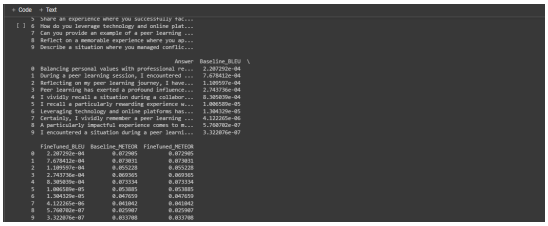

Figure 21: Score Comparison

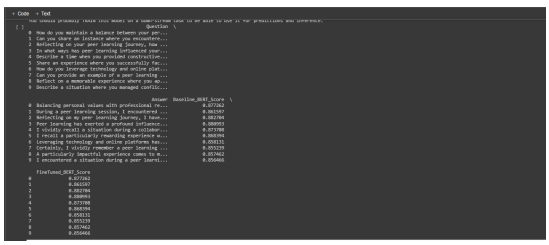

Figure 22: BERT Score Comparison

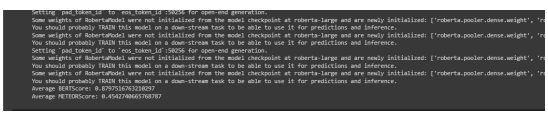

Figure 23: Average BERT and METEOR Score

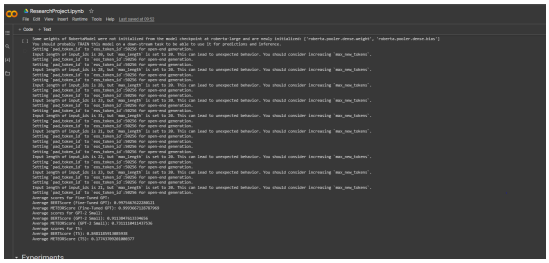

Figure 24: Final Score Comparison (Fine Tuned GPT2,GPT2 and T5 Model)

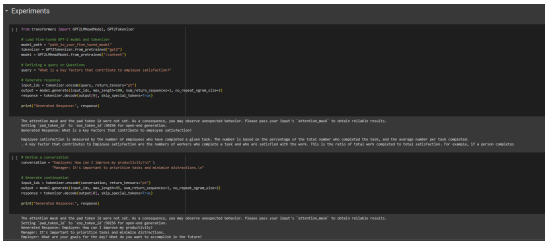

Figure 25: Experiments on Generated Text

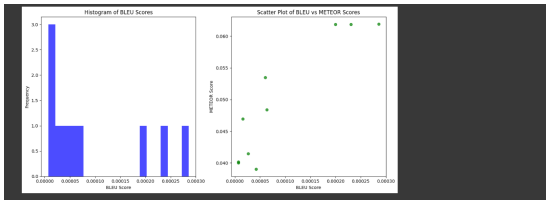

Figure 26: Visualisation 1

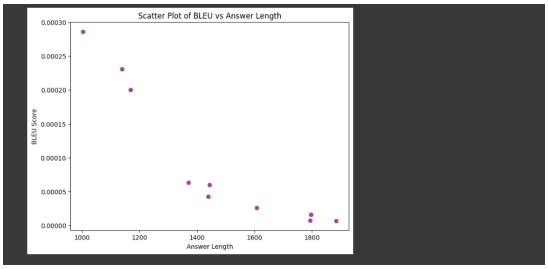

Figure 27: Visualisation 2

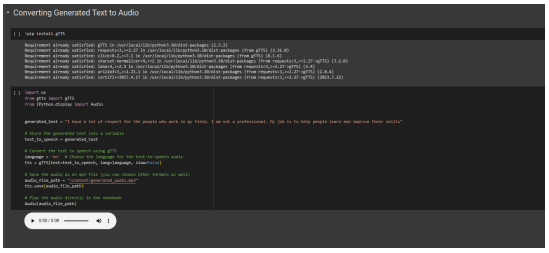

Figure 28: Generating Audio using Generated Text

References

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*.

Teeter, R. and Barksdale, K. (2011). *Google apps for dummies*, John Wiley & Sons.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3473–3483.