Summer 2022

QUERAI – A Smart Quiz Generator

Elton da Silva
CCT College Dublin

Fernando Aires da Silva
CCT College Dublin

Kim Jang Womg
CCT College Dublin

Tai Teei Ho
CCT College Dublin

Recommended Citation

da Silva, Elton; Aires da Silva, Fernando; Womg, Kim Jang; and Ho, Tai Teei, "QUERAI – A Smart Quiz Generator" (2022). *ICT*. 26.
https://arc.cct.ie/ict/26

CCT College Dublin

Assessment Cover Page

Module Title:	Problem-Solving for Industry
Assessment Title:	QUERAI – A Smart Quiz Generator
Lecturer Name:	Dr Muhammad Iqbal
Student Full Name & Student Number:	Elton da Silva 2018322 & Fernando Aires da Silva 2017243 & Kim Jang Wong 2017300 & Tai Teei Ho 2018249
Video link:	https://youtu.be/XhYWqi7n59Q
Github link:	https://github.com/eltonsilvamtm/QUERAI
Project link:	https://querai-app.nw.r.appspot.com/
Assessment Due Date:	16th May 2022
Date of Submission:	16th May 2022

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other sources.
I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Table of Contents

List of Figures

List of Tables

Abstract

QUERAI is a website powered by an Artificial Intelligence Question & Answer quiz generator model aiming to enhance students' learning experience and improve teachers' qualitative work by giving them more time to deal with other activities such as assignment correction, general grading, and class preparation.

KEYWORDS: Natural Language Processing, NLP, Artificial Intelligence, Machine Learning, Web application

1. Introduction

Learning something new is, at most times, challenging and frustrating. According to (Schawbel, 2022), it takes approximately 20 hours to learn something new. Why are people constantly giving up learning a new skill if it takes 20 hours only?

Many factors cause an individual to give up on learning something new, and college students find it very difficult to keep up with lectures in the first few semesters of an undergraduate course.

Research conducted by (Richards and Frankland, 2017) states that the human brain absorbs somewhere between 10% to 20% of all the information that is received daily, meaning that the goal of human memory is not just to store information accurately but to "optimise decision-making" in a chaotic and quickly changing environment.

This natural phenomenon is known as "the forgetting curve", which is the idea that information disappears at an exponential rate (Cloke, 2022). A highly successful learning approach can be achieved by using a technique called "Spaced Repetition" (Tamm, 2022).

To maximise the amount of information to be retained by the brain, we need to reinforce such information periodically, as shown in the image below.

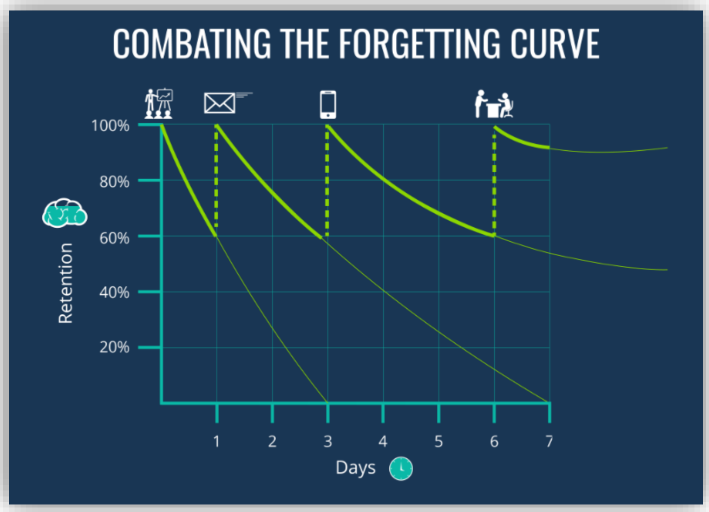


Figure 1: Combating the forgetting curve (Cloke, 2022)

2. Problem Analysis

Research conducted by (Hanson, 2021) affirms that 66% of males and 71% of females around the world aged between 16 to 24 enrolled in college in 2018, followed by an average 40% dropout rate with around 10% finishing the degree at a different institution.

Between 60% and 80% of Computing and Engineering Irish students dropped out of college in the first year from 2008 to 2011, with national rates at about 55% (O'Brien, 2022).

With little chance of completing college in time or at all being relatively small, a handful of factors might cause this issue: lack of preparation from students, course curriculum not updated to the market, lack of tools for professors to understand the class needs, and a teaching methodology including the spaced repetition technique mentioned above.

In an attempt to mitigate this problem, a solution is being proposed in the next section, followed by documentation that will walk through the development of a product intends to answer the following questions:

- Can spaced repetition improve a student's outcome?
- Will colleges accept that the newer generations need a more interactive way of teaching?
- Is the traditional way of teaching not working anymore?
- Is memorisation the key to learning?

3. Solution Proposal

Intending to avoid enormous dropout rates during the first year after seeing some other alumni deciding to give up their studies because they could not fully understand the fundamentals of Computer Science, this idea came to light.

With the help of AI algorithms, deep learning techniques, and neural network concepts, this project will create a website with a quiz generator capability. Combining it with the spaced-repetition approach transforms the teaching and learning experience into a more fun activity for learners and tutors.

These questions could be multiple-choice, fill in the blanks, true or false, or binary style questions, which will be analysed and returned to the professor in a feedback format, showcasing which areas the students are struggling to understand.

4. Methodologies

The methodology is an area of study on methods to systematically develop a product, paper, or any type of project in a structured manner. It consists of breaking the development process into phases or applying a cyclical approach where the project is constantly floating around sections (Jansen and Warren, 2020).

4.1 CRISP-DM

The Cross-Industry Process for Data Mining (CRISP-DM) methodology provides a structured approach to planning a data science project consisting of six steps that help the team plan, organise, and implement the project (Agues, 2020).

Step 1 - Business Understanding: Understanding how learning habits incentivise and shape a more enjoyable learning and teaching experience.

Step 2 - Data Understanding: by fully understanding the information available to the business, goals can be polished, creating a better final product.

Step 3 - Data Preparation: preparing the information is crucial to developing a good product; the goal is to use the most reliable and practical techniques to guarantee the final product's success.

Step 4 - Modelling: putting everything discussed together by turning the initial idea into a viable, commercialised product.

Step 5 - Evaluation: in conjunction with the modelling step, this step manages the quality of the product created in the previous step. It is prevalent for a development team to do the evaluation and modelling steps several times to guarantee the product hits its maximum potential.

Step 6 - Deployment: the final step happens when the final product is ready and distributed to schools and universities.

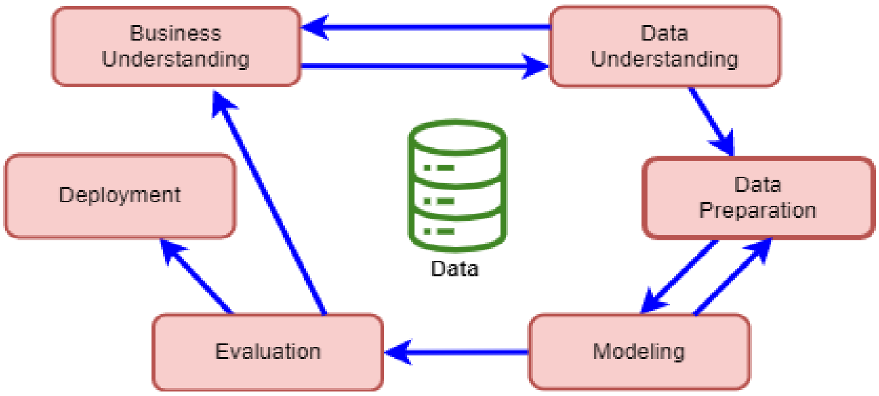


Figure 2: CRISP-DM Diagram (Data Science Process Alliance, 2022)

5. Business Understanding

This chapter provides a walkthrough of the business approach toward the strategic decisions taken to ensure that the business infrastructure is consistent and the final product meets commercialisation standards.

5.1 Business Strategy

There's no question that it is easy for a professional in the teaching area to have a 'heavy loaded' workflow with a wide range of teaching needs in their class. As preceding our introduction, many students do not progress after their first year in college, which is one of the many challenges in the teaching area.

Nonetheless, according to the list of 10 challenges faced by professionals in teaching (Gurudrasil, 2022), we observed that from the list, one of the major concerns was that it is time-consuming to prepare the teaching materials and other inconvenient scenes.

Further research concludes that the two major Learning Management systems (LMS), Moodle and Blackboard, lack automation of processes, leading lecturers and students to suffer from burnout and demotivation (Paradiso eLearning Blog, 2022).

The image below describes a high-level overview of the strategic decisions taken to ensure that the company's goals are achievable by fulfiling the client's expectations.

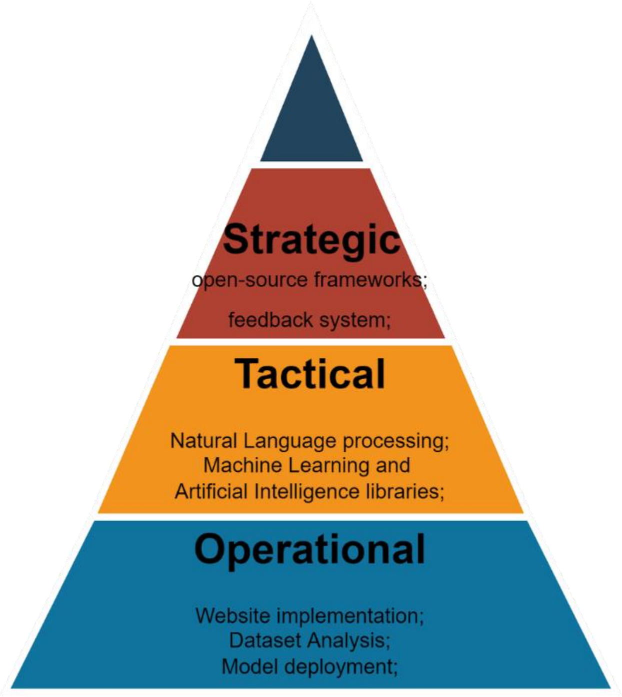


Figure 3: Business decisions Pyramide overview

5.1.1 Strategic decisions

- Use open-source and industry-standard frameworks, libraries, and other resources to reduce the deployment cost of the application;
- Apply a reward system to motivate students and lecturers to use the platform consistently;
- Simulate a more personal teaching approach with the assistance of analytic tools;
- Develop a web application once most universities and colleges have a website-based approach to dealing with students' work;
- Work with libraries and toolkits well known by the community, trying to reduce the time troubleshooting;

5.1.2 Who will most benefit from the product?

- ***Students:*** by increasing the student's content absorption, it is possible to make them more interested in the subject, therefore, less likely to be another dropout;
- ***Lecturers:*** by taking less time to prepare questionnaires, the lecture can focus on the development of more complex activities;
- ***Mentors:*** focus on the group's weaknesses when reviewing content by using the class breakdown performance analytics;

- ***Academic Organisation:*** higher retention of freshers by consistently applying the concepts of spaced repetition through the learning platform, which will result in an increase in profits for the organisation;

5.2 Business Analysis

According to (GoIreland, 2018), the Irish higher education system consists of 6 private colleges, 9 Institutes of Technology, and 8 Public Universities. However, further research concludes that there is no similar product in the Irish education market, which gives us the leverage and the challenges of being the first to offer such innovation.

5.2.1 SWOT

SWOT (MindTools, 2022) is an industry-standard analysis tool used to identify Strengths, Weaknesses, Opportunities, and Threats concerning a business strategy. It consists of analysing an organisation's current impact (positive/negative) regarding the internal and external environments through the aspects that compose the word SWOT summarised below.

The table below describes the SWOT Analysis practically:

SWOT Analysis			
Strengths	**Weaknesses**	**Opportunities**	**Threats**
The team	Project management	Huge market education	Political interference in education
Supportive mentors (Professors)	Technical Knowledge	Scarce competition	A better software might appear
Strong and meaningful idea	Short time frame	Online learning has grown exponentially	

Table 1: SWOT analysis

5.2.2 MosCoW

MosCoW (Korolev, 2021) is a priority-based project management business approach that defines the importance of certain features of a project/product by improving the quality of rapid app development (RAD) processes.

Those processes are divided into four priority categories, with the table below describing those priorities concerning the QUERAI web application:

MoSCow Analysis			
Must-Have	**Should Have**	**Could Have**	**Won't Have**
Upload and read the contents of a file/URL	Login/Logout Option	Moodle Integration	Real-time module adjustment tool
AI algorithm to interpret the contents of a file/URL	Stores feedback into a database	Feedback analysis of a group of students	Text-to-speech AI to read out the questions
Generate a Quiz	Improvement recommendations tool	Reward system for quizzes competition	
Quiz feedback			

Table 2: MoSCoW implementation

5.3 Use Case

CCT College Dublin is experiencing a high number of dropouts on the Computing faculty and is looking for ways to facilitate students' overall college experience, therefore, minimising the number of students dropping out, eventually resulting in more revenue for the organisation.

After long research, CCT coordinators got interested in the spaced repetition methodology. They decided to implement the QUERAI website capabilities of auto quizzes generation to help their students and professors be more productive, resulting in overall increased satisfaction with the teaching/learning experience.

The product (website) and the proper usage of the spaced repetition technique is a great way to keep the students motivated and keep track of their progress and where they need more attention, creating the sense of a more personal learning approach.

In the following use case scenario, the professor automatically generates weekly quizzes using the QUERAI website, and the students will attempt to answer them. The professor will receive their performance analytics towards the activities, which can be further analysed, resulting in a more personal teaching/learning experience.

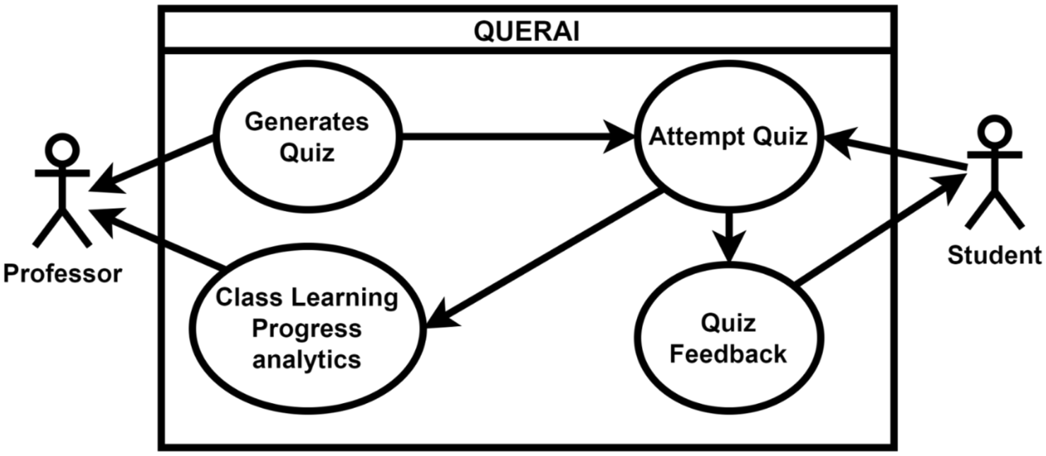
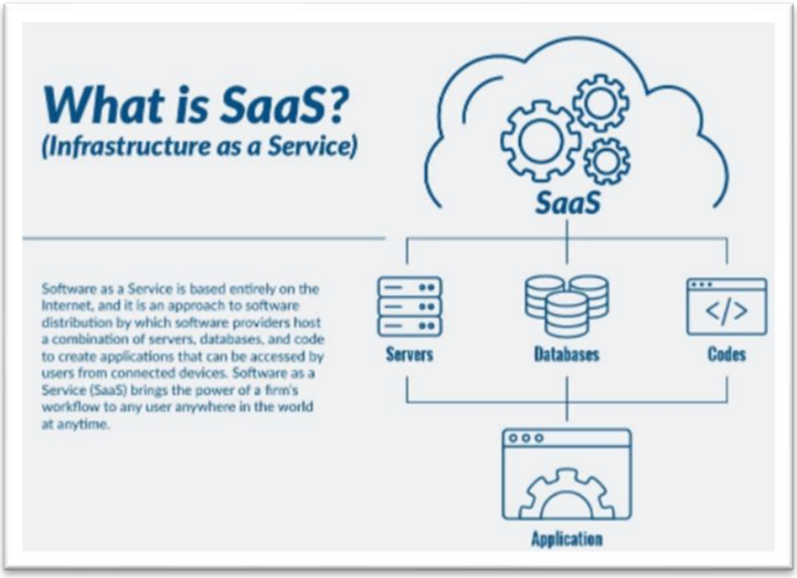


Figure 4: Use Case

5.4 Pricing

Following the big tech companies' trends (Square, 2020), charging customers in a subscription manner is the best alternative for this product. If the product is a one-time buy, customers will not buy the item as often.

A monthly subscription would also imply that institutions would require higher volumes of usage of the platform to make sense of the money spent. Software as a Service (SaaS) is the most feasible way of extracting maximum revenue from customers and encouraging the QUERAI team to constantly work on services upgrades by giving customers the freedom to leave at any moment (Matteo Duò, 2020).



Software as a Service is based entirely on the Internet, and it is an approach to software distribution by which software providers host a combination of servers, databases, and code to create applications that can be accessed by users from connected devices. Software as a Service (SaaS) brings the power of a firm's workflow to any user anywhere in the world at anytime.

Figure 5: SaaS Diagram (image source: atlantic.net) (Matteo Duò, 2020)

6. Natural Language Processing (NLP)

NLP is a field of computing that enables computers to understand natural language similarly to humans, whether the language is spoken or written. NLP relies on AI to take real-world input, process it, make sense of it in a way a computer can understand, and generate a text output that a human can realise (Bhardwaj, 2021).

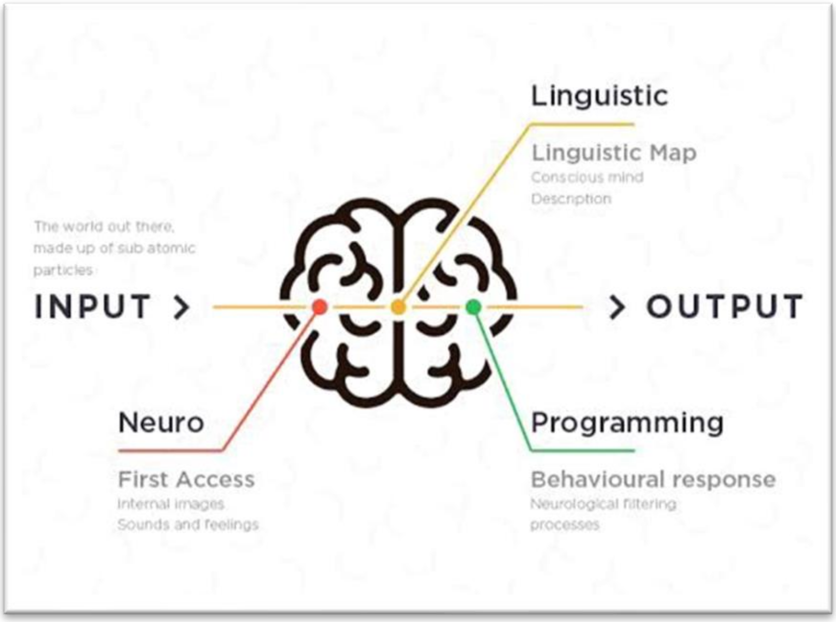


Figure 6 Human language interpretation process (Bhardwaj, 2021)

An essential tool for this project and through NLP subsets such as Natural Language Understanding (NLU) and Natural Language Generation (NLG), applications can learn to distinguish and accurately manage the meaning behind words, sentences, and paragraphs (TechTarget, 2021).

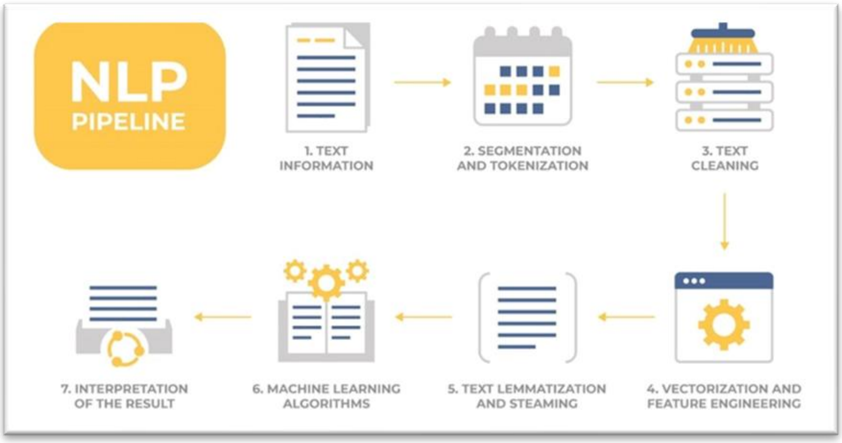

NLP PIPELINE
1. TEXT INFORMATION
2. SEGMENTATION AND TOKENIZATION
3. TEXT CLEANING
4. VECTORIZATION AND FEATURE ENGINEERING
5. TEXT LEMMATIZATION AND STEAMING
6. MACHINE LEARNING ALGORITHMS
7. INTERPRETATION OF THE RESULT

Figure 7 NLP Pipeline structure (IogrBobriakov, 2019)

7.2 How does NLP work?

Syntax Analysis and Semantic Analysis are the two main techniques used by Natural language Processing to "understand" a language via computer algorithms. Syntax Analysis refers to the arrangement of words in a sentence to make grammatical sense. In contrast, Semantic Analysis refers to the meaning that is conveyed by a text when computer algorithms are applied to understand the meaning and interpretation of words as well as how sentences are structured (Lutkevich & Burns, 2021).

The most used Syntax Analysis techniques by NLP systems are:

- Parsing: this is a grammatical analysis of a sentence and is helpful for more complex downstream processing tasks;
- Word segmentation: this is the act of taking a string of text and deriving word forms from it;
- Stemming: It involves cutting the inflected words to their root form and used as a heuristic procedure that crops off the ends of the words;
- Tokenisation: Assign numbers to terms present in the text, denominating those as tokens. Used to understand the meaning of the text via computational processing algorithms;

Semantic Analysis is a challenging aspect of NLP that is not fully resolved yet (Garbade, 2018); however, there are some techniques available such as:

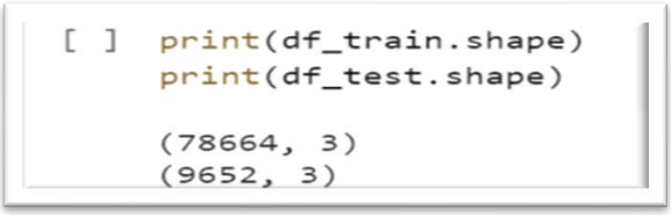
- Named Entity Recognition (NER): It involves determining the parts of a text that can identify and categorise pre-sets of groups;
- Word sense disambiguation: It consists in giving meaning to a word based on the context;

7. Data Understanding

Created by Rajpurkar in 2016, The Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset consisting of 107,785 observations from 536 different articles. The dataset contains context, question, and answer attributes, where the answer to every question is a segment of a span of text from the corresponding reading passage (Rajpurkar, 2022).

7.1 Summary

As seen in the image below, the SQuAD dataset contains 78664 observations in the train set and 9652 observations in the validation set, with both having three attributes described in ***Table 3.***



```
[ ]  print(df_train.shape)
     print(df_test.shape)

     (78664, 3)
     (9652, 3)
```

Figure 8: Dataset size

Attribute	Description	Type
context	The context on which questions are based.	Object
question	The question.	Object
answer	The answer to the question.	Object

Table 3: Dataset Column descriptions

To compile SQuAD, the creators will split the dataset by articles which 80% of articles went into the training set, 10% into a development set and 10% into a testing set.

7.2 Usability of text-based datasets

The SQuAD dataset focuses on developing models for question answering systems; however, it tests a model's ability to read a text passage and answer a relatively straightforward question (Wei, J., 2020).

NLP Pre-Trained Models (PTMs) are transformers models trained on a large dataset to perform specific NLP tasks and learn universal language representations that can benefit downstream NLP tasks and avoid introducing a new model from scratch.

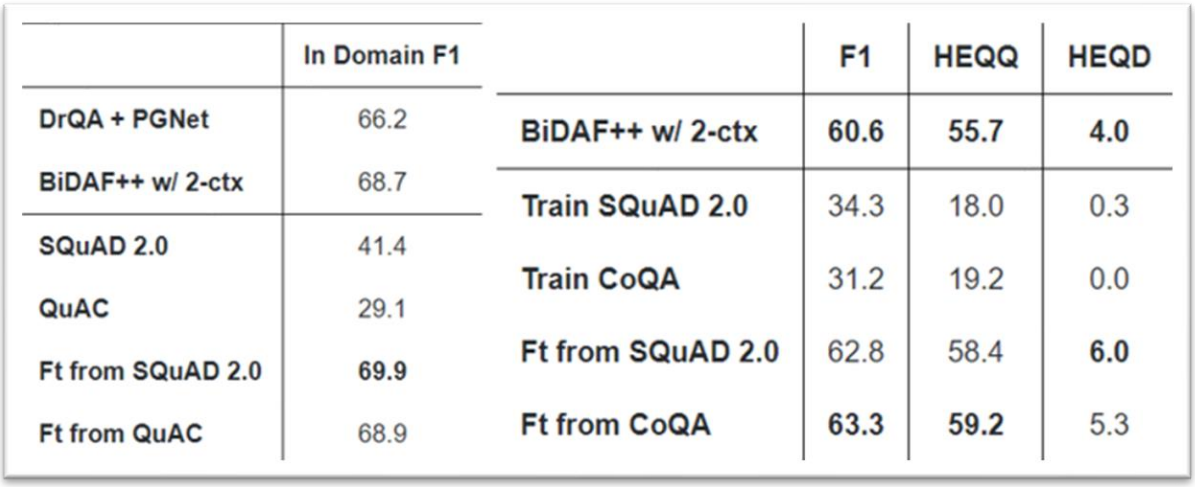
The reusable NLP models can be used to build an NLP application quickly. The transformers provide a suite of pre-trained NLP models across different NLP tasks such as text classification, question answering, machine translation etc. (Kumar, 2021).

The PTMs can be quickly loaded into Machine Learning libraries such as PyTorch, Tensorflow, etc. and are often easy to implement, providing high accuracy and less training time than custom-built models (Kumar, 2021).

7.3 Why the SQUAD dataset?

In summary, SQuAD is an excellent dataset for training models of language understanding, well-created and improved on many aspects compared to other datasets such as the Question Answering in Context (QuAC) or the Conversational Question Answering (CoQA) datasets.

A qualitative analysis showcasing the efficiency of each of the most popular question-answering datasets was performed by doing cross-dataset experiments with the pre-trained Bi-Directional Attention Flow for Machine Comprehension (BiDAF++) model. The study concludes that the SQuAD dataset was better than the other datasets when used to initialise models that use context as one of their training features (Yatskar, 2019).



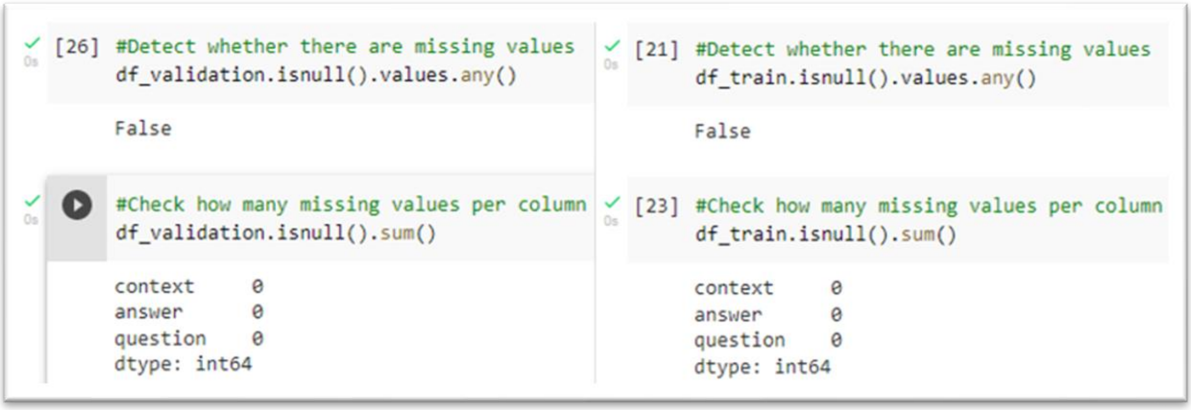
	In Domain F1
DrQA + PGNet	66.2
BiDAF++ w/ 2-ctx	68.7
SQuAD 2.0	41.4
QuAC	29.1
Ft from SQuAD 2.0	**69.9**
Ft from QuAC	68.9

	F1	HEQQ	HEQD
BiDAF++ w/ 2-ctx	60.6	55.7	4.0
Train SQuAD 2.0	34.3	18.0	0.3
Train CoQA	31.2	19.2	0.0
Ft from SQuAD 2.0	62.8	58.4	**6.0**
Ft from CoQA	**63.3**	**59.2**	5.3

Figure 9: Datasets qualitative analysis (Yatskar, 2019)

Another reason to choose this dataset was that cleaning the data is perhaps one of the most critical factors of any IT project. The SQUAD dataset provides a pretty well-cleaned dataset that can save time for other development tasks.

As proof of concept, Figure 10 shows that the dataset looks clean for a project starting point once there are no missing values.



```
[26] #Detect whether there are missing values
     df_validation.isnull().values.any()

     False
```

```
[21] #Detect whether there are missing values
     df_train.isnull().values.any()

     False
```

```
     #Check how many missing values per column
     df_validation.isnull().sum()

     context     0
     answer      0
     question    0
     dtype: int64
```

```
[23] #Check how many missing values per column
     df_train.isnull().sum()

     context     0
     answer      0
     question    0
     dtype: int64
```

Figure 10: Missing values check on SQUAD

8. Data Preparation

This section will be diving deeper into preparing a question answering dataset for the modelling section by using an NLP tool called Text-to-Text Transfer Transformer (T5) transformer with a technique called "Transfer Learning".

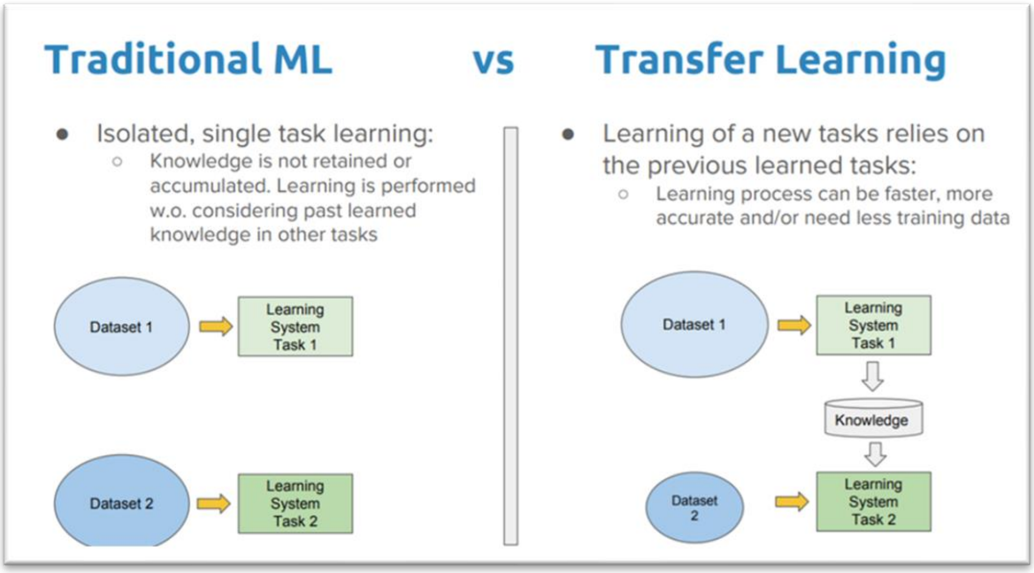
8.1 Transfer Learning

Unlike a numerical dataset, and as seen above, the English language has many parameters, variances, and nuances that are very difficult to calculate; therefore, the creation of a model from a text-based dataset from zero every time becomes quite a complex task.

For that reason, it is more interesting to have a starting point where we can have all those rules implicit. Pre-trained models possess such an advantage that they exclude all the grammatical guesswork that the production model would have, leaving a so-called "fine-tuned model" with the task of figuring out how to achieve the output (similarly to what happens on a numerical-based model) (V7labs.com, 2022).

Pre-trained models in a data-rich environment have become powerful tools in NLP. Primarily used in the industry, models such as GPT-2 and BERT (commonly used for translation and speech-to-text applications) are discussed in more detail in the next section.

Transfer learning is where a model is first pre-trained on a data-rich task before being fine-tuned on a downstream task, and it has emerged as one of the most potent techniques in natural language processing (NLP) (V7labs.com, 2022).



Traditional ML vs Transfer Learning

- Isolated, single task learning:
  - Knowledge is not retained or accumulated. Learning is performed w.o. considering past learned knowledge in other tasks

- Learning of a new tasks relies on the previous learned tasks:
  - Learning process can be faster, more accurate and/or need less training data

Figure 11: Traditional ML vs Transfer Learning (V7labs.com, 2022)

In other words, a pre-trained model called "Text-to-Text Transfer Transformer", or T5, will be used as the primary model of this project, which contains all the necessary tools to prepare the data and deploy the model.

8.2 Transformer models overview

There are many other transformer pre-trained models available for fine-tuning an NLP model. However, there are many aspects and applications for NLP models, which necessitates further analysing of a few options.

The Artificial Intelligence community states that generative Pre-trained Transformers such as GPT-2 and, more recently, the GTP-3 are considered the best AI pre-trained models. However, both are pretty robust models with sizes of 3.09 GB for the GTP-2 against 45TB for the GTP-3, which might increase depending on the application. It will also need a lot of extra GPU to process such enormous information (Markowitz, 2021).

This project context lies in having a light platform that can be uploaded to the cloud and quickly deliver an output. The most used pre-trained models for low to medium scale applications are the Bidirectional Encoder Representations from Transformers (BERT) and the Text-to text Transformer (T5) models (Markowitz, 2021).

The main difference between these models is that the T5 can encode and decode the information delivering the expected text output. In contrast, the BERT model is powered only by an encoder that requires an extra decoder to generate human-readable output (Mustafa, 2020).

This fact alone already gives us the conclusion of discarding the BERT model for this application; however, it is interesting to point out a comparison of both models at a pre-trained stage.

The image below consists of an experiment using the MS MARCO (Microsoft Machine Reading Comprehension Dataset) with results measured in Mean Reciprocal Rank. The investigation concludes that the T5 model performs better at the training stage, with results reporting means and 95% confidence intervals over five trials (Nogueira et al., 2020).

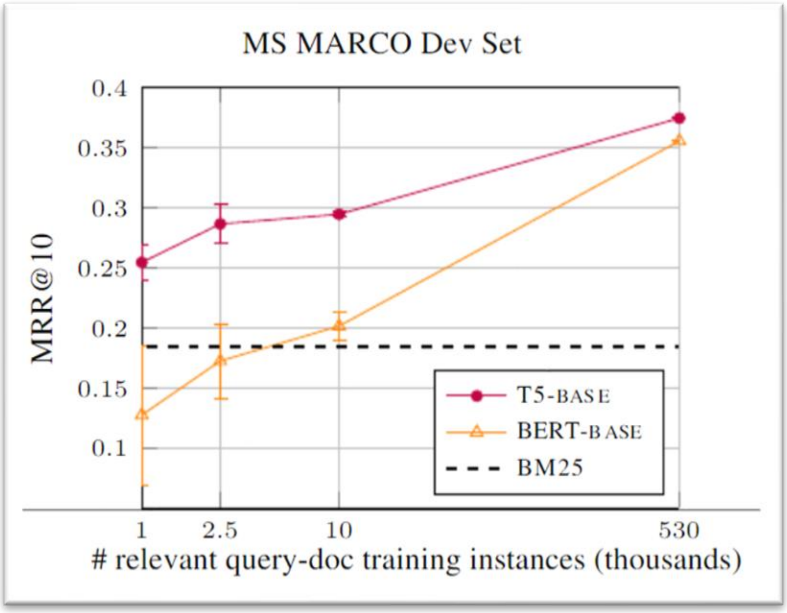


Figure 12: t5-base vs bert-base pre-trained models (Nogueira et al., 2020)

8.3 Text-to-Text Transfer Transformer (T5)

"The T5 transformer is an encoder-decoder and a type of neural network architecture model pre-trained on a multi-task mixture of unsupervised and supervised tasks where each task is converted into a text-to-text format."

It was first introduced to the IT industry in the article *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer* written by Colin Raffel, Noam Shazeer, and the remaining authors (Raffel et al., 2020) on 28th July 2020.

The T5 transformer encodes word by word the data given using some NLP concepts such as self-attention, enabling the model to numerically "understand" the context of a sentence. Whenever the model provides the output, it then reverses engineers the encoded words to form the output result, which could be a translation, phrase, or a question in the case of this project (Alammar, 2020).

In either way of the information flow, one essential aspect of the model is the usage of neural networks to perform the encoding-decoding of the inputs/outputs. The illustration below describes the whole encoding-decoding process in more detail:

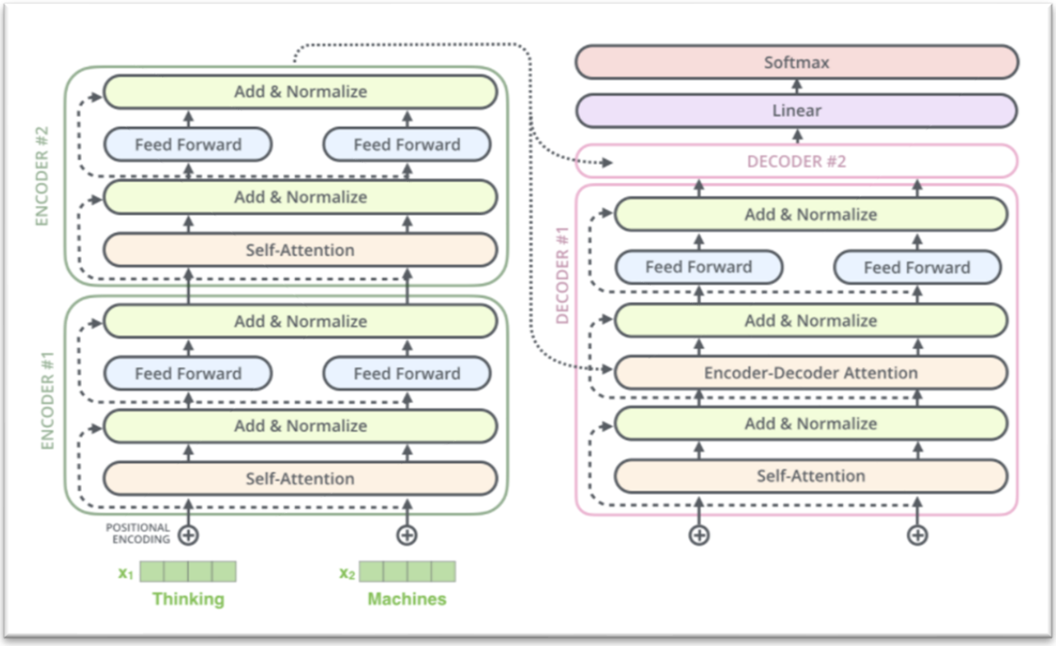


Figure 13:T5 model structure (Alammar, 2020)

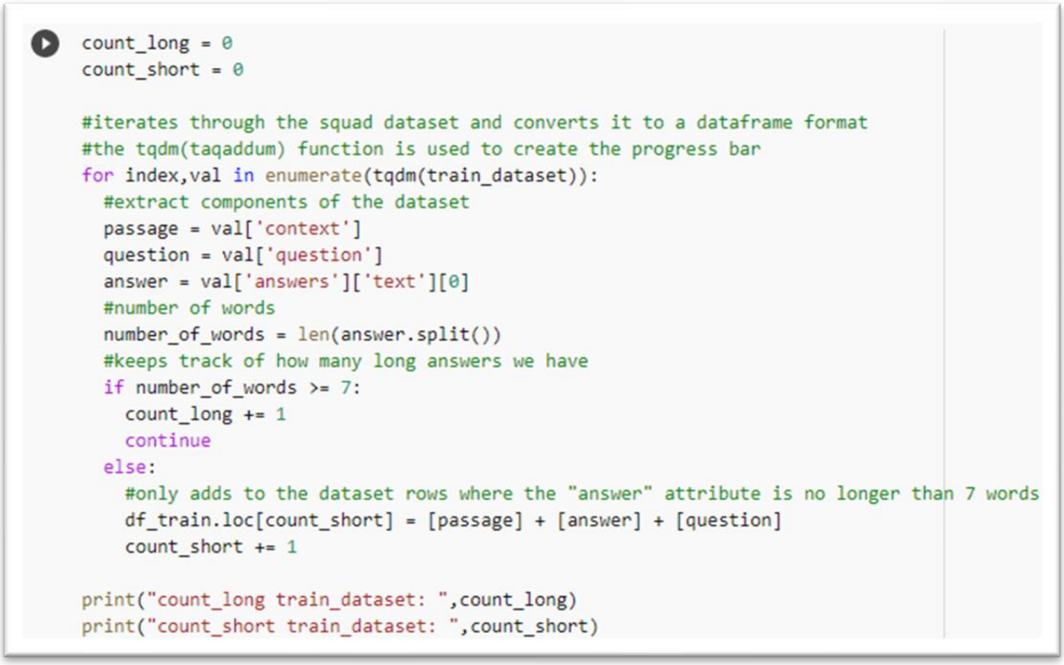
8.4 Filtering the dataset by answer length

According to (Ebstein, 2021), the average attention span of a young adult ranges from 8 to 9 seconds. Mainly caused due to the constant usage of smartphones and social media applications, those tools make usage of AI to "persuade" a user to spend more time on their applications.

One crucial aspect of getting the dataset prepared for the training process is filtering the "answers" column by length. The project goal is to offer a solution that will make students more likely to stick to the quiz rather than get distracted by other activities.

Be that as it may, the idea behind eliminating observations where the answer attribute is very long relies on the premise that quiz questions tend to be shorter. Therefore, the dataset must have a threshold regarding the length of the answer so that the model trains to generate primarily short answer questions.

The code below filters the dataset for answers no longer than seven words:



```
count_long = 0
count_short = 0

#iterates through the squad dataset and converts it to a dataframe format
#the tqdm(taqaddum) function is used to create the progress bar
for index,val in enumerate(tqdm(train_dataset)):
  #extract components of the dataset
  passage = val['context']
  question = val['question']
  answer = val['answers']['text'][0]
  #number of words
  number_of_words = len(answer.split())
  #keeps track of how many long answers we have
  if number_of_words >= 7:
    count_long += 1
    continue
  else:
    #only adds to the dataset rows where the "answer" attribute is no longer than 7 words
    df_train.loc[count_short] = [passage] + [answer] + [question]
    count_short += 1

print("count_long train_dataset: ",count_long)
print("count_short train_dataset: ",count_short)
```

Figure 14: filtering answers column by length

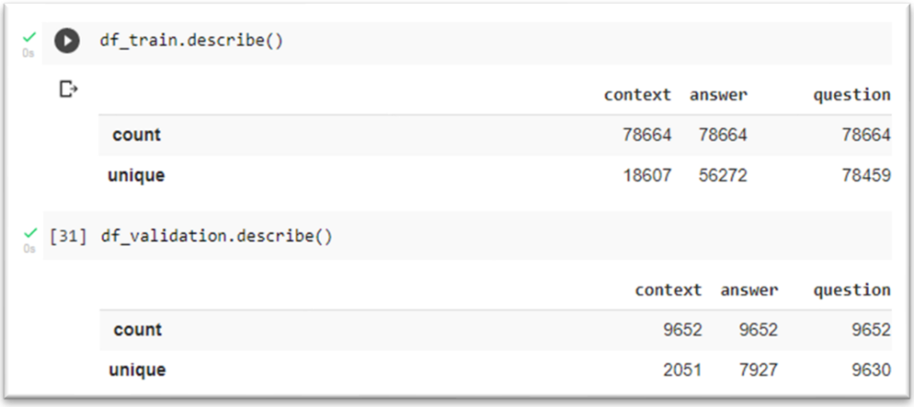
8.5 Removing duplicates

Another standard data preparation procedure for preparing text-based datasets is introducing bias by leaving similar questions on the dataset or even having duplicated observations with multiple distinct answers.

(Chorev, 2021) states that an entry that appears more than once during the training process can disproportionate the weight attribute, therefore, corrupting the model somehow. They can also ruin the train-test split stage by splitting similar entries throughout the sets or even occupying one of the sets, reducing the size of the partitioned set.

The removal of those observations can potentially lower the overall accuracy score of the model; however, it is more important to have a model that performs well with slightly lower accuracy than to open space for potential bias in the dataset.

First off, it is necessary to check whether there are any duplicated values on the dataset:



```
df_train.describe()
```

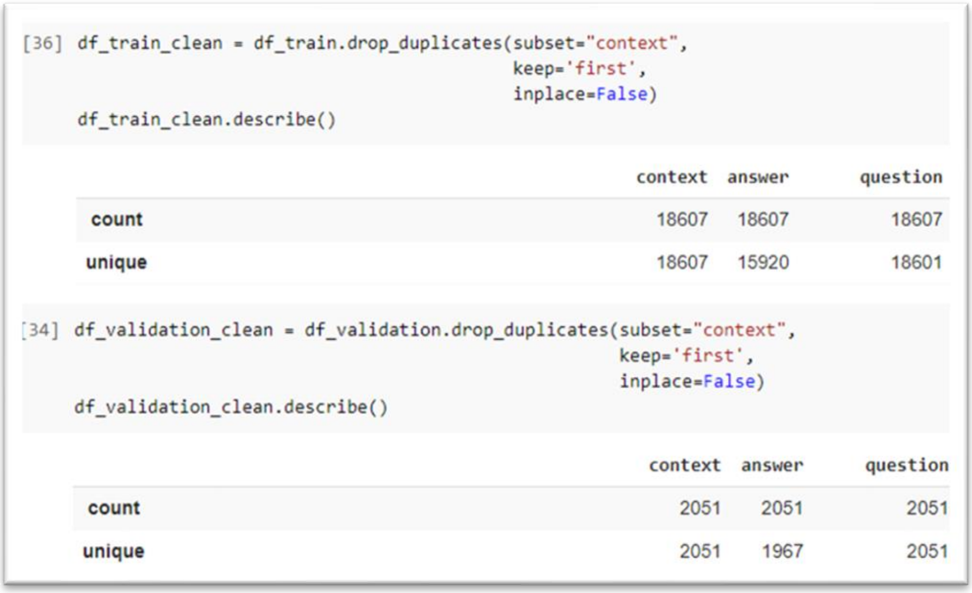
	context	answer	question
count	78664	78664	78664
unique	18607	56272	78459

```
[31] df_validation.describe()
```

	context	answer	question
count	9652	9652	9652
unique	2051	7927	9630

Figure 15: Duplicated data count

Once the model deals with question generation, the biggest concern is to have duplicated contexts with different answers, which can potentially "confuse" the model somehow; by removing the duplicates from the context attribute, we drastically reduce this possibility.



```
[36] df_train_clean = df_train.drop_duplicates(subset="context",
                                              keep='first',
                                              inplace=False)
     df_train_clean.describe()
```

	context	answer	question
count	18607	18607	18607
unique	18607	15920	18601

```
[34] df_validation_clean = df_validation.drop_duplicates(subset="context",
                                                        keep='first',
                                                        inplace=False)
     df_validation_clean.describe()
```

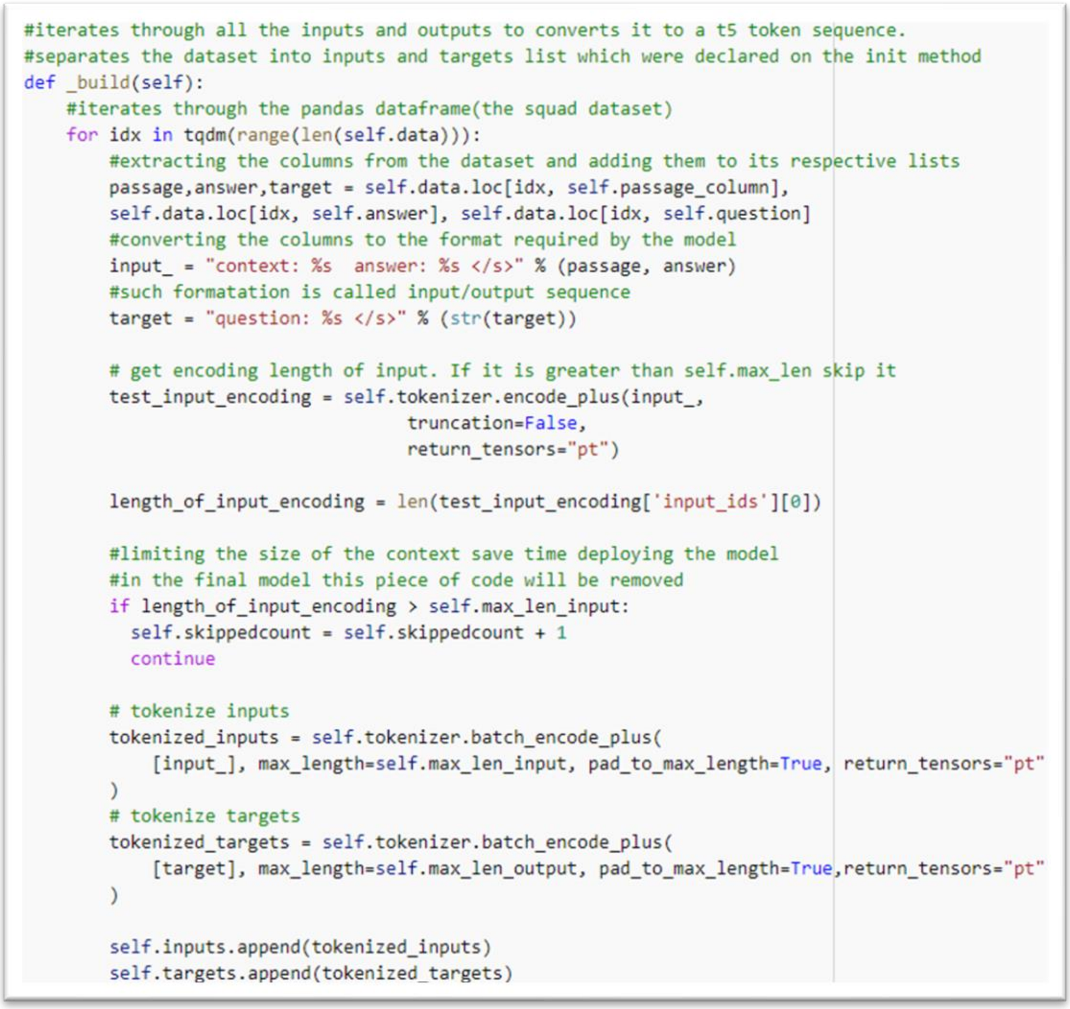
	context	answer	question
count	2051	2051	2051
unique	2051	1967	2051

Figure 16: after duplicated questions removal

8.6 Tokenizing the dataset

A transformer model does not comprehend plain text; it only understands data in a token format which is the final step in preparing the dataset for the modelling section.

The image below describes how the dataset is converted to the desired tokenised format by looping through each row and encoding each column from the dataset.



```
#iterates through all the inputs and outputs to converts it to a t5 token sequence.
#separates the dataset into inputs and targets list which were declared on the init method
def _build(self):
    #iterates through the pandas dataframe(the squad dataset)
    for idx in tqdm(range(len(self.data))):
        #extracting the columns from the dataset and adding them to its respective lists
        passage,answer,target = self.data.loc[idx, self.passage_column],
        self.data.loc[idx, self.answer], self.data.loc[idx, self.question]
        #converting the columns to the format required by the model
        input_ = "context: %s  answer: %s </s>" % (passage, answer)
        #such formatation is called input/output sequence
        target = "question: %s </s>" % (str(target))

        # get encoding length of input. If it is greater than self.max_len skip it
        test_input_encoding = self.tokenizer.encode_plus(input_,
                                    truncation=False,
                                    return_tensors="pt")

        length_of_input_encoding = len(test_input_encoding['input_ids'][0])

        #limiting the size of the context save time deploying the model
        #in the final model this piece of code will be removed
        if length_of_input_encoding > self.max_len_input:
          self.skippedcount = self.skippedcount + 1
          continue

        # tokenize inputs
        tokenized_inputs = self.tokenizer.batch_encode_plus(
            [input_], max_length=self.max_len_input, pad_to_max_length=True, return_tensors="pt"
        )
        # tokenize targets
        tokenized_targets = self.tokenizer.batch_encode_plus(
            [target], max_length=self.max_len_output, pad_to_max_length=True,return_tensors="pt"
        )

        self.inputs.append(tokenized_inputs)
        self.targets.append(tokenized_targets)
```

Figure 17: encoding the dataset

9. Modelling

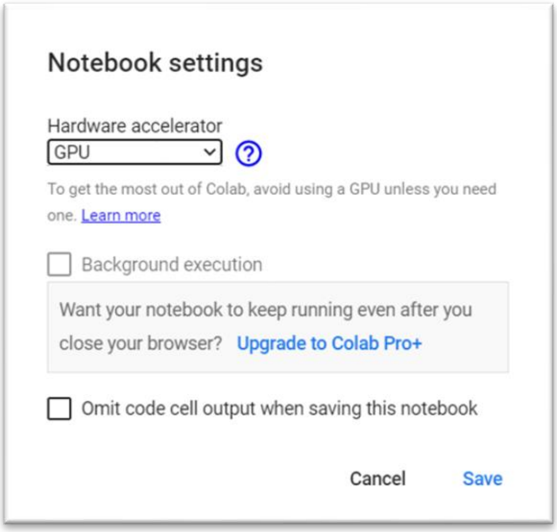
Now it is time to endeavour the process of applying the concepts previously discussed by training the model used on this project which is divided into:

- Setting up the environment
- Tokenising the dataset
- Fine-tuning the model
- Saving the model

9.1 Setting up the environment

Training an NLP model can take a lot of processing time due to its complexity. To mitigate this issue, Google Colab has an incredible feature where it is possible to assign a GPU (Graphics Processing Unit) to the runtime, making the training step a lot faster.

To enable GPU on Google Colab's runtime, go to Runtime → Change runtime type to GPU and hit save. To ensure an actual GPU is available at your runtime, use the command "!nvidia-smi".



Notebook settings

Hardware accelerator

GPU

To get the most out of Colab, avoid using a GPU unless you need one. Learn more

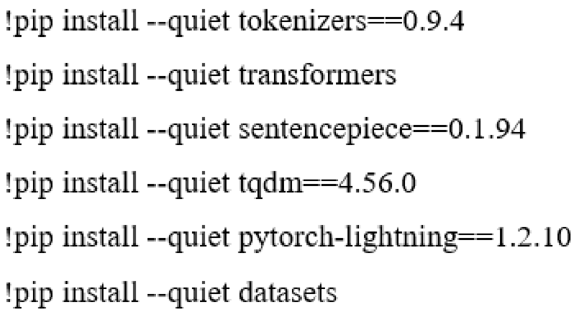
□ Background execution

Want your notebook to keep running even after you close your browser? Upgrade to Colab Pro+

□ Omit code cell output when saving this notebook

Figure 18: Enabling GPU

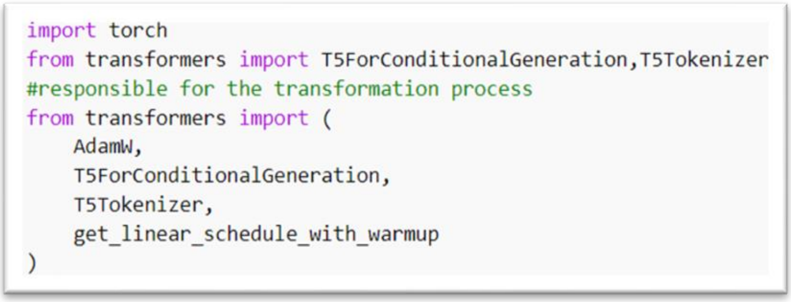
Another necessary setup is importing the libraries to load and work with the T5 transformer found in the image below. Those libraries need to be installed before their respective import. Here is the list of libraries necessary to be installed in Google Colab:



```
!pip install --quiet tokenizers==0.9.4
!pip install --quiet transformers
!pip install --quiet sentencepiece==0.1.94
!pip install --quiet tqdm==4.56.0
!pip install --quiet pytorch-lightning==1.2.10
!pip install --quiet datasets
```

Figure 19 Python Libraries needed for fine-tuning the model

The following imports described in the figure below also need to be installed to use the capabilities of the T5 transformer model.

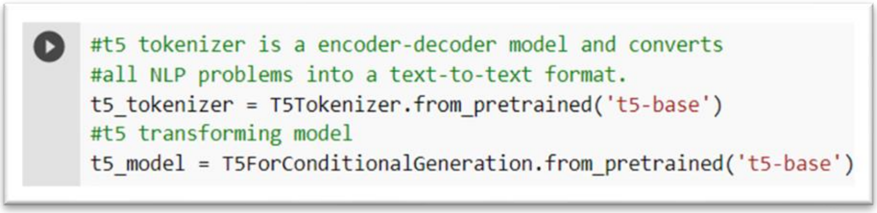


```
import torch
from transformers import T5ForConditionalGeneration,T5Tokenizer
#responsible for the transformation process
from transformers import (
    AdamW,
    T5ForConditionalGeneration,
    T5Tokenizer,
    get_linear_schedule_with_warmup
)
```

Figure 20: Imports

Deep learning uses a pre-trained model by default to apply another training step to it and generate a "better" model in regards to the desired output. For this project, the "t5-base" model is being used, created by Google and pre-trained for summarisation (patrickvonplaten, 2022).

Every NLP model also needs a tokeniser model, which will take care of applying tokens to the words coming from the dataset.



```
#t5 tokenizer is a encoder-decoder model and converts
#all NLP problems into a text-to-text format.
t5_tokenizer = T5Tokenizer.from_pretrained('t5-base')
#t5 transforming model
t5_model = T5ForConditionalGeneration.from_pretrained('t5-base')
```

Figure 21: Downloading the pre-trained models

9.2 Fine-Tuning the model

Because the model is already pre-trained, instead of "re-train" the model, the ideal approach is to fine-tune the model; therefore, adapting the model to our application. A neural network receives the input and feeds it through the processing node layers, passing the information forward to the next layer (Hargurjeet, 2021).

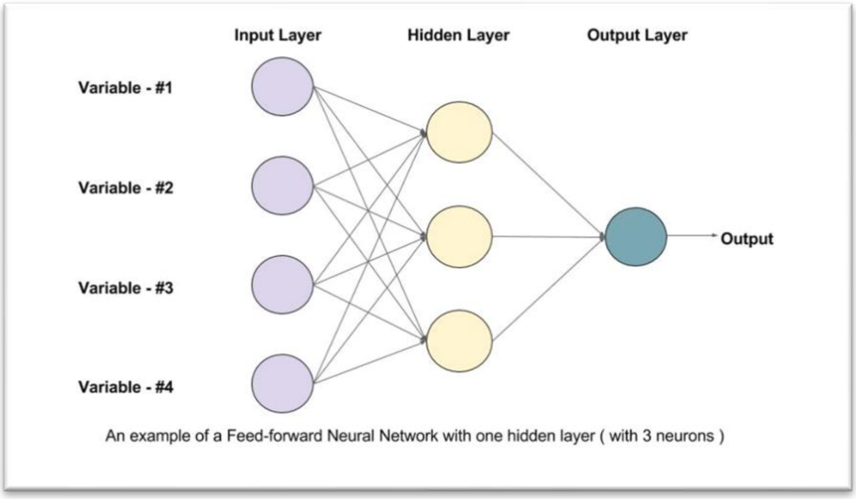
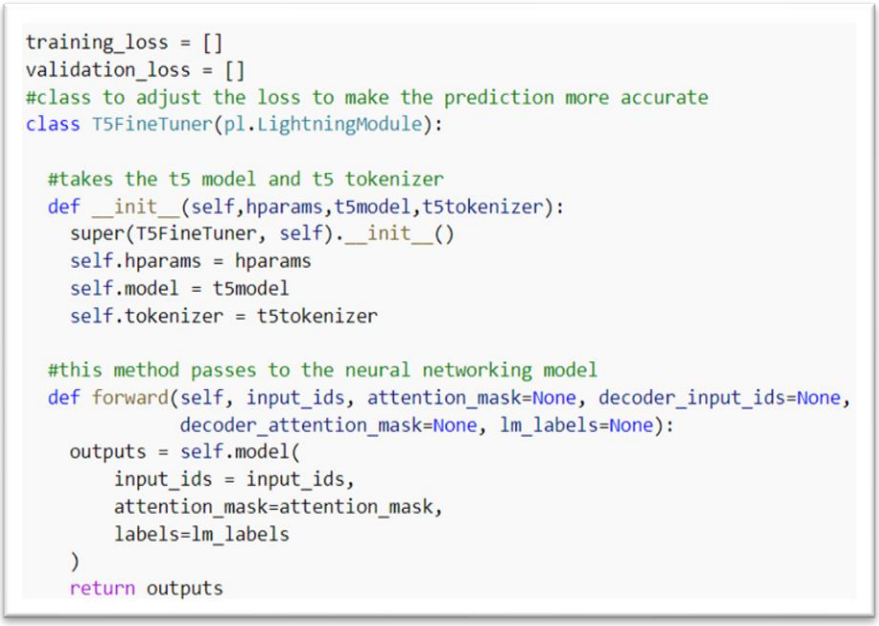


Figure 22: Feeding forward Neural Network example (Hargurjeet, 2021)

According to documentation, the T5 model has two feeding forward neural networks (one for the decoder and one for the actual predictions/training); both have six hidden layers by default (Huggingface.co, 2020).



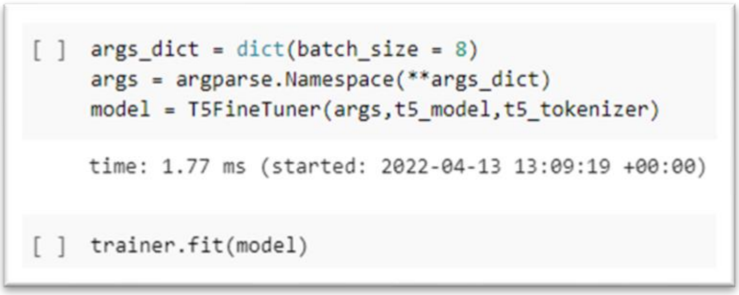
```
training_loss = []
validation_loss = []
#class to adjust the loss to make the prediction more accurate
class T5FineTuner(pl.LightningModule):

  #takes the t5 model and t5 tokenizer
  def __init__(self,hparams,t5model,t5tokenizer):
    super(T5FineTuner, self).__init__()
    self.hparams = hparams
    self.model = t5model
    self.tokenizer = t5tokenizer

  #this method passes to the neural networking model
  def forward(self, input_ids, attention_mask=None, decoder_input_ids=None,
              decoder_attention_mask=None, lm_labels=None):
    outputs = self.model(
        input_ids = input_ids,
        attention_mask=attention_mask,
        labels=lm_labels
    )
    return outputs
```

Figure 23: Feeding forward method to fine-tune the T5 model

The batch size tells the model how many inputs the training step will have at a time; the number 8 in the image below indicates that the training is working with eight rows at a time (after the conversion to tokens).



```
[ ]  args_dict = dict(batch_size = 8)
     args = argparse.Namespace(**args_dict)
     model = T5FineTuner(args,t5_model,t5_tokenizer)

     time: 1.77 ms (started: 2022-04-13 13:09:19 +00:00)

[ ]  trainer.fit(model)
```

Figure 24: Training the model

The image below implies that the feeding forward method worked once the neural network managed to optimise its loss function. A noticeable exponential decrease in the loss value during the training step confirms that the model is fine-tuned.

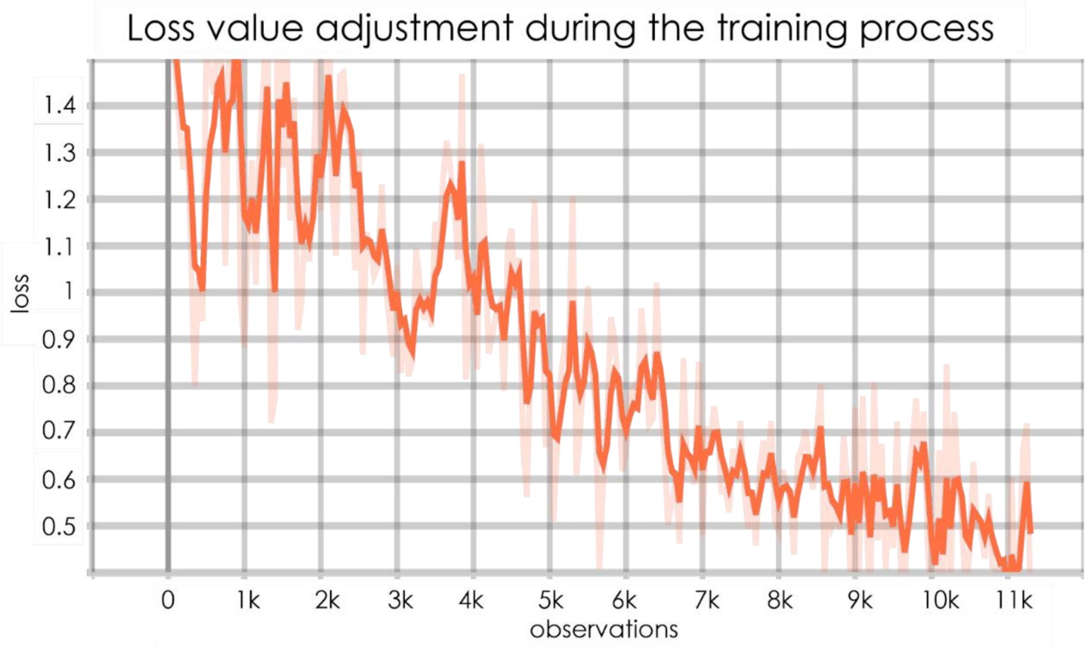
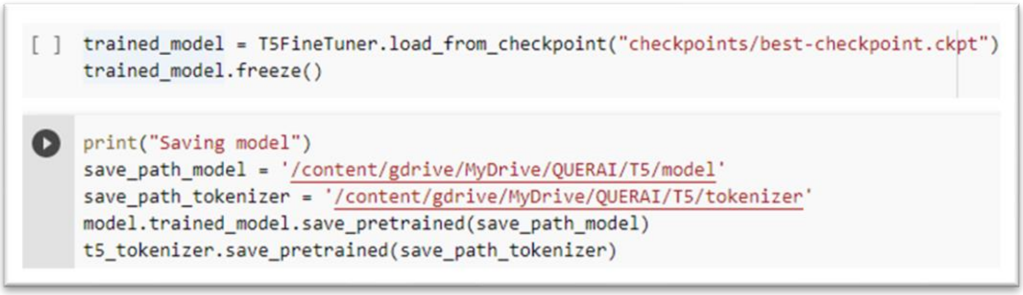


Figure 25: Loss value plot of the training step

9.3 Saving the model

Google Colab does not support saving files automatically to a local directory, so saving the model or any file requires a google drive account. To save a T5 model to your Google Drive account, please follow the instructions below:



```
trained_model = T5FineTuner.load_from_checkpoint("checkpoints/best-checkpoint.ckpt")
trained_model.freeze()
```

```
print("Saving model")
save_path_model = '/content/gdrive/MyDrive/QUERAI/T5/model'
save_path_tokenizer = '/content/gdrive/MyDrive/QUERAI/T5/tokenizer'
model.trained_model.save_pretrained(save_path_model)
t5_tokenizer.save_pretrained(save_path_tokenizer)
```

Figure 26: Saving the T5 model with the best checkpoint

10. Evaluation

The evaluation step for NLP processes requires more aspects than just comparing the model's output against the validation set output. The definitive test of an NLP model is to analyse the context of the model's output by checking if it is related to the context somehow.

(Pykes, 2021) states that there are two niches of evaluation metrics for NLP models described as follows:

- Intrinsic evaluation: focuses on the performance of an NLP component on a defined subtask, for example, text generation or translation.
- Extrinsic evaluation: focuses on the performance of the whole application, for example, the Natural Language Generation and Evaluation (nlg-eval) library discussed in *section 11.2*.

Typical metrics of a machine learning model such as accuracy, precision, f1 score, and recall are not ideal for testing an NLP model; however, they are still valid. For that matter, while going through the validation set, the loss value experienced changes during the testing step.

The image below displays an increase in the loss as the model walks through the validation set:

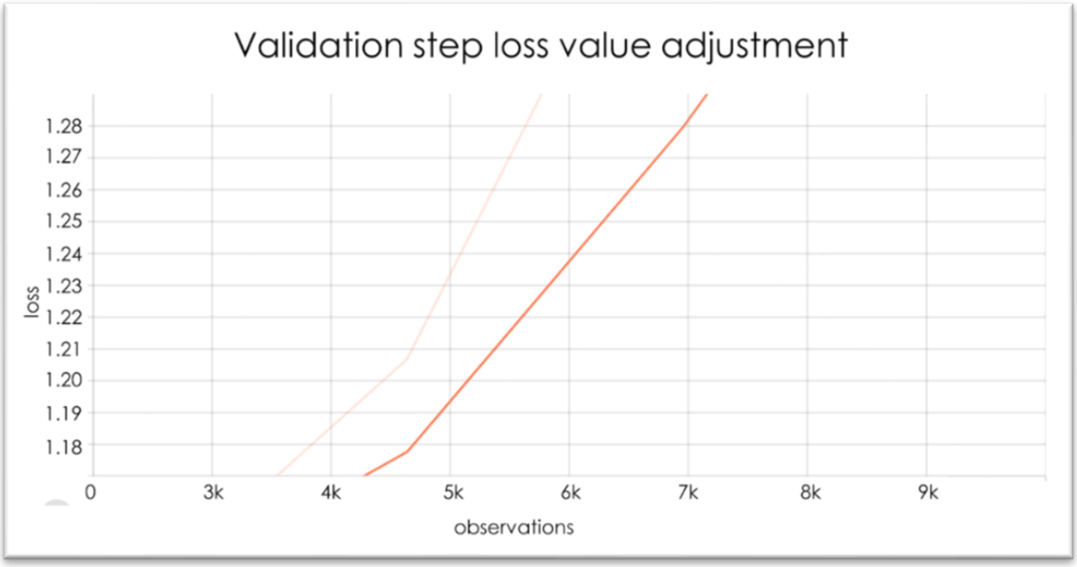
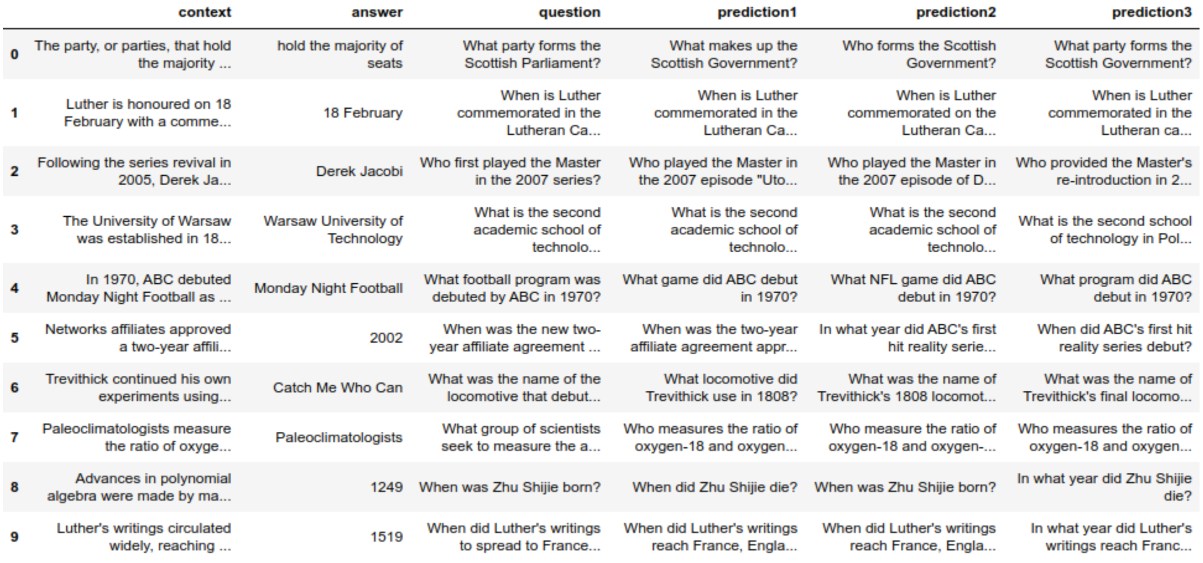


Figure 27: validation step loss adjustment

10.1 Predictions

The best way to analyse an NLP model's qualitative results is by having a human read and interpret the predictions. However, such an approach is quite expensive and time-consuming, causing delays to the development process's flow.

In such a case, the image below contains three specific questions generated by the model from a given context and answer. Those generated outputs showcase that the model is capable of generating human-readable quality text.



	context	answer	question	prediction1	prediction2	prediction3
0	The party, or parties, that hold the majority ...	hold the majority of seats	What party forms the Scottish Parliament?	What makes up the Scottish Government?	Who forms the Scottish Government?	What party forms the Scottish Government?
1	Luther is honoured on 18 February with a comme...	18 February	When is Luther commemorated in the Lutheran Ca...	When is Luther commemorated in the Lutheran Ca...	When is Luther commemorated on the Lutheran Ca...	When is Luther commemorated in the Lutheran ca...
2	Following the series revival in 2005, Derek Ja...	Derek Jacobi	Who first played the Master in the 2007 series?	Who played the Master in the 2007 episode "Uto...	Who played the Master in the 2007 episode of D...	Who provided the Master's re-introduction in 2...
3	The University of Warsaw was established in 18...	Warsaw University of Technology	What is the second academic school of technolo...	What is the second academic school of technolo...	What is the second academic school of technolo...	What is the second school of technology in Pol...
4	In 1970, ABC debuted Monday Night Football as ...	Monday Night Football	What football program was debuted by ABC in 1970?	What game did ABC debut in 1970?	What NFL game did ABC debut in 1970?	What program did ABC debut in 1970?
5	Networks affiliates approved a two-year affili...	2002	When was the new two-year affiliate agreement ...	When was the two-year affiliate agreement appr...	In what year did ABC's first hit reality serie...	When did ABC's first hit reality series debut?
6	Trevithick continued his own experiments using...	Catch Me Who Can	What was the name of the locomotive that debut...	What locomotive did Trevithick use in 1808?	What was the name of Trevithick's 1808 locomot...	What was the name of Trevithick's final locomo...
7	Paleoclimatologists measure the ratio of oxyge...	Paleoclimatologists	What group of scientists seek to measure the a...	Who measures the ratio of oxygen-18 and oxygen...	Who measure the ratio of oxygen-18 and oxygen-...	Who measures the ratio of oxygen-18 and oxygen...
8	Advances in polynomial algebra were made by ma...	1249	When was Zhu Shijie born?	When did Zhu Shijie die?	When was Zhu Shijie born?	In what year did Zhu Shijie die?
9	Luther's writings circulated widely, reaching ...	1519	When did Luther's writings to spread to France...	When did Luther's writings reach France, Engla...	When did Luther's writings reach France, Engla...	In what year did Luther's writings reach Franc...

Figure 28 Predictions overview

10.2 Natural Language Generation and Evaluation (nlg-eval)

There are many aspects to be analysed in a text, such as context, grammar, and many other elements that need to be considered. The nlg-eval Python library is convenient once it does most of those tasks under the hood. It consists of 8 scoring system types for NLP models (Maluuba, 2021).

For the matter of this project, only four of those metric systems are suitable for the type of output generated by the model, which are:

Skip Thought Vectors (STV): It consists of a powerful Neural Network model for learning fixed-length representations of sentences in any NLP model. Without any supervised learning data, the supervision signal used is only the ordering of sentences (Agarwal, 2017).

Replacing fixed representations from a sentence into an equivalent vector of numbers is the primary function of this model, enabling the computer to mathematically understand language to classify whether the Skip Thought generated sentence has any semantical similarity in meaning against the sentence rendered by the QUERAI model.

Embedding Average Cosine Similarity (EACS): words embedding relates to the idea that this model will transform words from the target and the output into vectors and analyse their similarity in morphological meaning, context, and lemmatisation which is the process of producing variants to the words being interpreted and comparing them against the target and the output (Intellica.AI, 2019).

Questions generated by the QUERAI model are in a similar context as the questions given by the dataset. The goal of this scoring system is to coherently compare words from the QUERAI model against observations from the question attribute of the validation set. To categorise its findings, the EACS gives a score from 0 to 1 regarding the quality of the text generated by the QUERAI model.

Vector Extrema Cosine Similarity (VECS): metric used to determine how similar the vectors converted from the sentences are regarding their size (similar size means a better score) (Prabhakaran, 2018).

Cosine similarity is essential when evaluating sentiment analysis and natural language generation models. This approach evaluates the model in its core, where the neurons from the neural networks connect their assumptions to develop the model.

Greedy Matching Score (GMS): The greedy matching score takes advantage of the greedy algorithm for its evaluation. By choosing the closest/nearest/most optimal node in a spanning-tree graph and keep picking the next best node, the algorithm tries to match its parameters with the algorithm's parameters (Glen, 2017).

By having a "top-down" approach to solve the problem, a greedy algorithm chooses one node after another, therefore, scaling down the problem and eventually optimising the solution.

In a nutshell, the greedy matching score goal is to produce matched samples with balanced characteristics between the validation set and the output generated by the QUERAI model, which can be matched as one-to-one or one-to-many pairs (Glen, 2017).

The pairs generated by the greedy matching are irreplaceable, meaning that once a node has been matched, it cannot be used again. Therefore, the more outputs matched with the control group (validation set), the higher score will be assigned to the generated text by the algorithm.

10.2.1 Applying nlg-eval

When implementing this library, it is recommended to use a Linux Operating System once its core relies on the Linux shell structure. Installing the dependency and then deploying the library is pretty much all that needs to be done to use this powerful tool. In the Linux terminal, type the following commands (Maluuba, 2021):

- pip install git+https://github.com/Maluuba/nlg-eval.git@master
- nlg-eval --setup

Considering that the QUERAI model can generate multiple questions of the same context and answer inputs and as seen above in *section 11.1,* the plot below compares the scores of 4 of the most adequate nlg evaluation algorithms against three distinct predictions of the same observation.

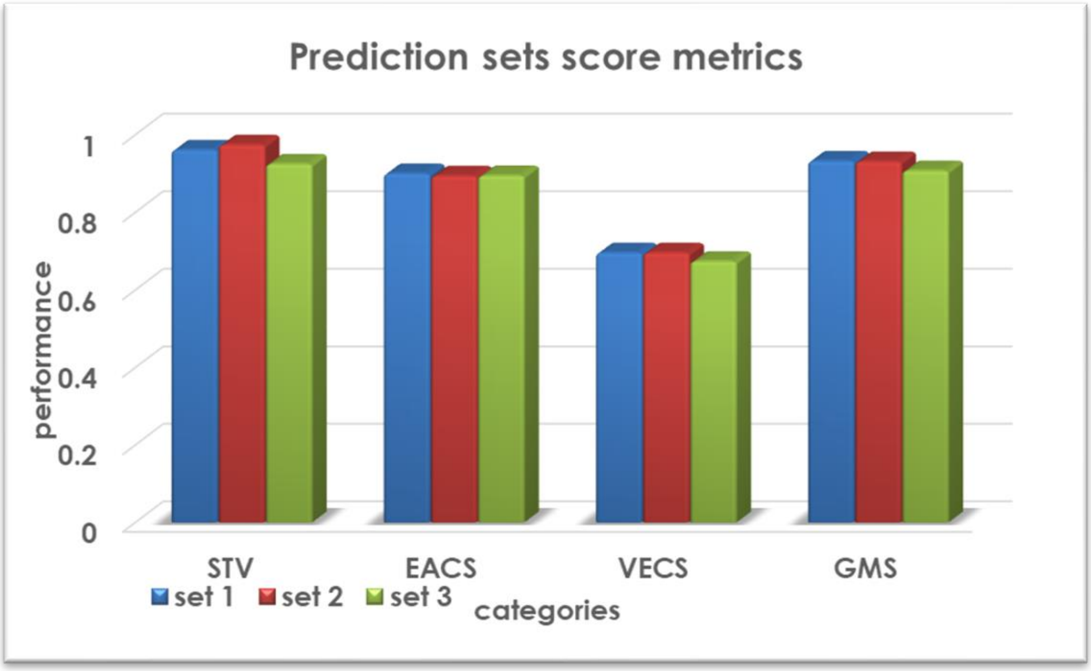


Figure 29: Prediction sets score metrics barplot

Notice that the model performed well in most aspects of the evaluation, taking the median score of those categories and concluding this evaluation analysis by naming the median of such types the "accuracy" of the model.

There is no proof that the average of these metrics can be interpreted as accuracy; however, as far as our findings go regarding evaluating an NLP model, it seems plausible to raise such an assumption that the final accuracy value is 91%.

11. Deployment

The deployment section will cover how to deploy all components necessary for the completion of the project. The image below gives an idea of how each section of the project was intended to work as independent components, which can be replaced at any future point, giving more flexibility to the development team.

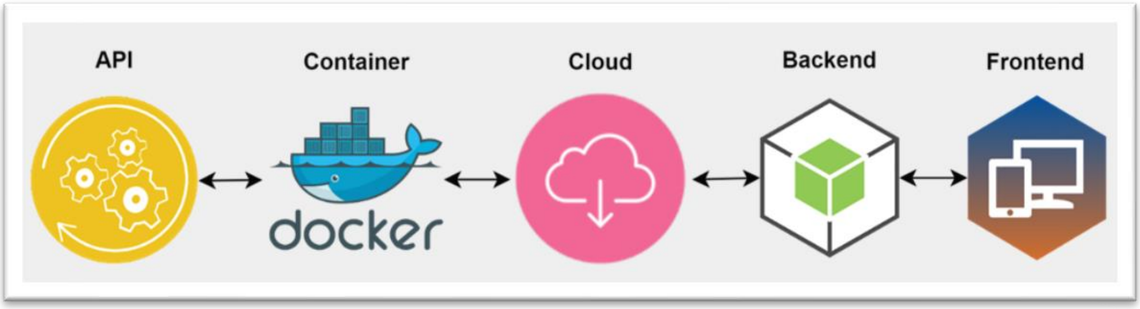


Figure 30: Web application workflow

11.1 Application Program Interface (API)

The idea behind the API is to have an easy access point to the model, which will generate the output and facilitate the work as a group. Other team members can efficiently work separately on different project sections by storing the API in the cloud. The API workflow happens as follows:

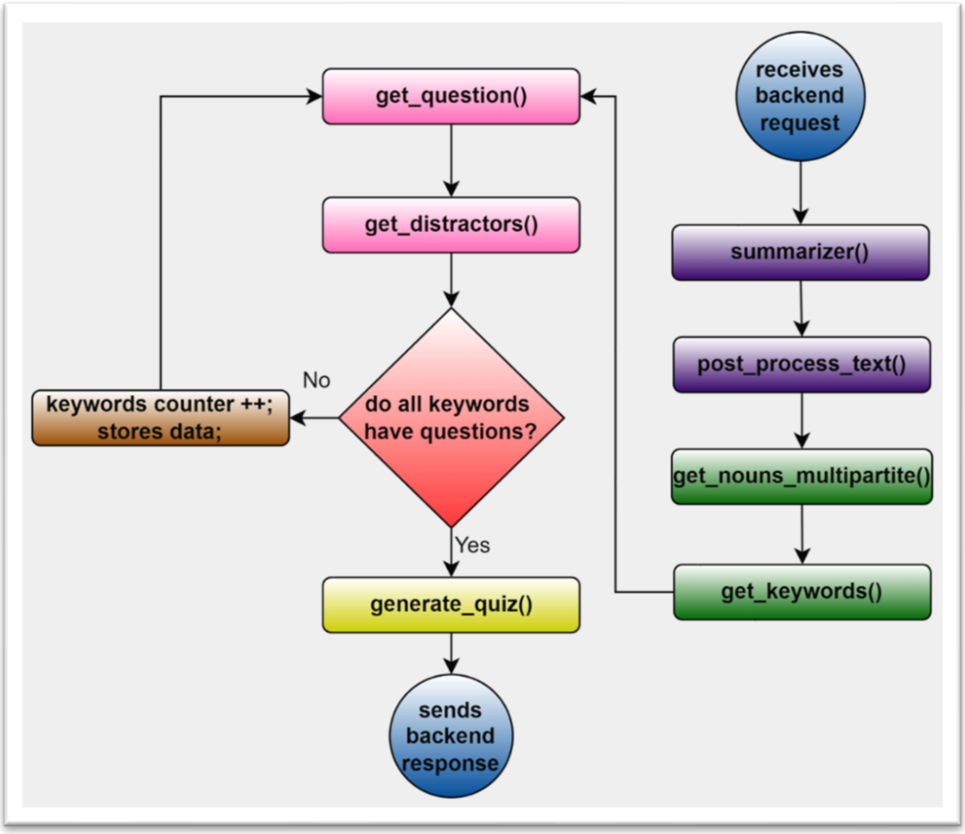


Figure 31: API workflow

In summary, the flowchart explanation step by step as follows:

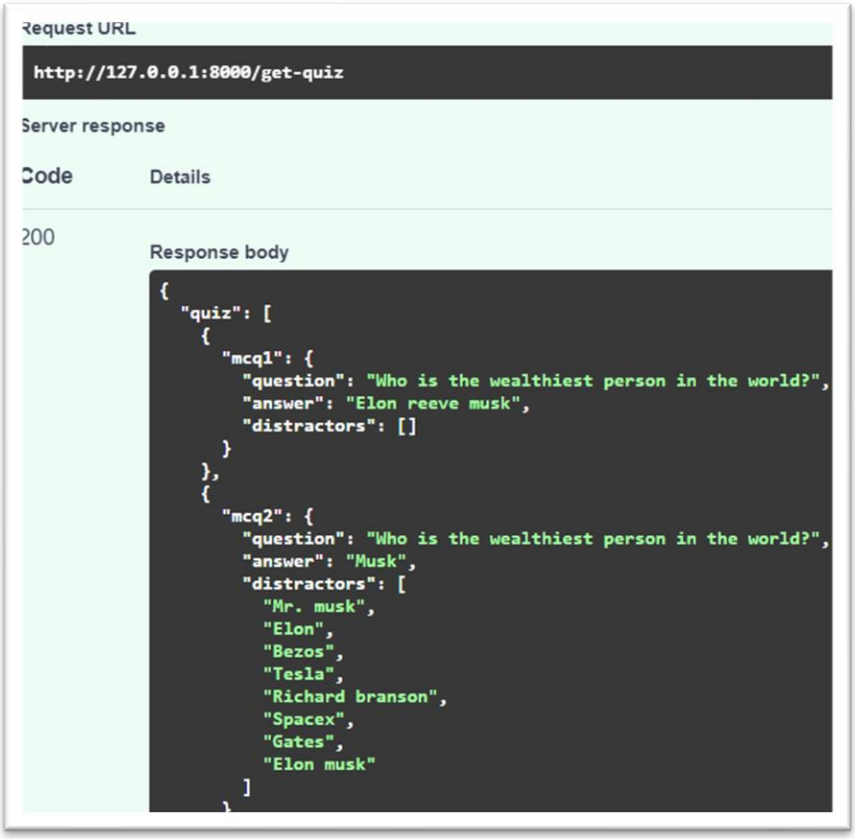
- 1: Summarises the text sent by the user.
- 2: Tokenizes the summarised text.
- 3: Uses Python keyphrase extraction library to retrieve keywords from the summarised and original text (boudinfl, 2022).
- 4: Post-process the keywords by comparing keywords found in both summarised and original text to enhance the quality of the keywords.
- 5: Calls the t5 model to generate a question.
- 6: Uses the sense2vec model to generate distractors (Mukesh-mehta, 2021). If sense2vec cannot find any distractors, the API automatically look for distractors with the wordnet lexical database (Princeton.edu, 2010).
- 7: Keeps generating questions until there are keywords available. In case there are more keywords, it stores the question, answer and distractors, while in case there are no more keywords, the API generates and returns the final quiz.

The fast API Python library (tiangolo, 2018) facilitates the work of setting up an API. It is also compatible with the Docker container, which will be very helpful when creating the image container (tiangolo, 2021). The first step is to install some dependencies:

- pip install --upgrade pip --user
- pip install fastapi
- pip install uvicorn

The final setting step is to download the sense2vec the part 1 of a neural network pre-trained vector called sense2vec (explosion, 2021) on your project folder, which will be used to generate distractors for the MCQ questions and uncompress the file with the following command:

- tar -xvf s2v_reddit_2015_md.tar.gz



Request URL

```
http://127.0.0.1:8000/get-quiz
```

Server response

Code Details

200

Response body

```
{
  "quiz": [
    {
      "mcq1": {
        "question": "Who is the wealthiest person in the world?",
        "answer": "Elon reeve musk",
        "distractors": []
      }
    },
    {
      "mcq2": {
        "question": "Who is the wealthiest person in the world?",
        "answer": "Musk",
        "distractors": [
          "Mr. musk",
          "Elon",
          "Bezos",
          "Tesla",
          "Richard branson",
          "Spacex",
          "Gates",
          "Elon musk"
        ]
```

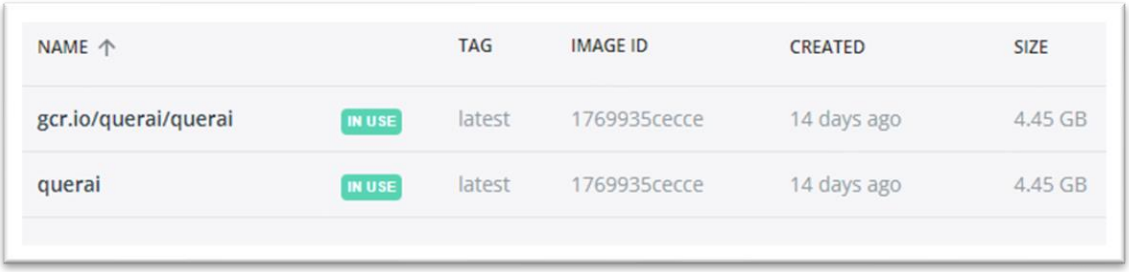
Figure 32: API running locally

11.2 Docker image

A Docker container is an excellent tool for uploading functional APIs to the cloud and facilitates scalability. These are the main reasons this tool was implemented in this project, consolidating the "separation of concerns" (Docker Documentation 2013).

After installing and setting up the Docker environment, inside of the terminal with admin privileges, enter the following commands:

- docker build -t querai .
- docker run --name querai -p 80:80 querai

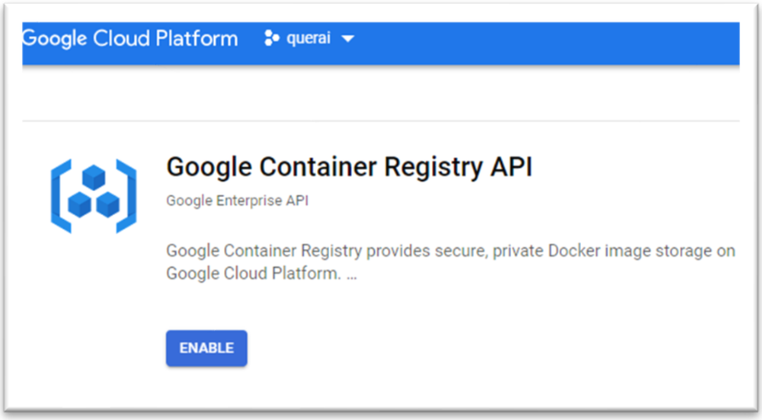


NAME ↑		TAG	IMAGE ID	CREATED	SIZE
gcr.io/querai/querai	IN USE	latest	1769935cecce	14 days ago	4.45 GB
querai	IN USE	latest	1769935cecce	14 days ago	4.45 GB

Figure 33: Docker image running

11.3 Google Cloud setup

The Google Cloud Platform is the most straightforward process for uploading and using docker containers (Google.com, 2022). The setup is straightforward: on the main dashboard, create a project, and on the search bar, enter the container registry to enable containers to be uploaded.



Google Container Registry provides secure, private Docker image storage on Google Cloud Platform. ...

Figure 34: Enabling Container Registry (Google.com, 2022)

By installing the Google Cloud SDK and opening it with administrative privileges, we can now start the Docker container upload process with the following commands:

- gcloud init
- docker build . --tag gcr.io/querai/querai:latest
- gcloud auth configure-docker
- docker push gcr.io/querai/querai:latest
- gcloud run deploy --image gcr.io/querai/querai:latest --cpu 2 --concurrency 1 --memory 8Gi --platform managed --min-instances 0 --timeout 1m --port 80

Note that enabling authentication with the gcloud command is very important; otherwise, the upload will not be successful. A link for the API is generated, which can be used by the back-end to manage all the requests to the API.

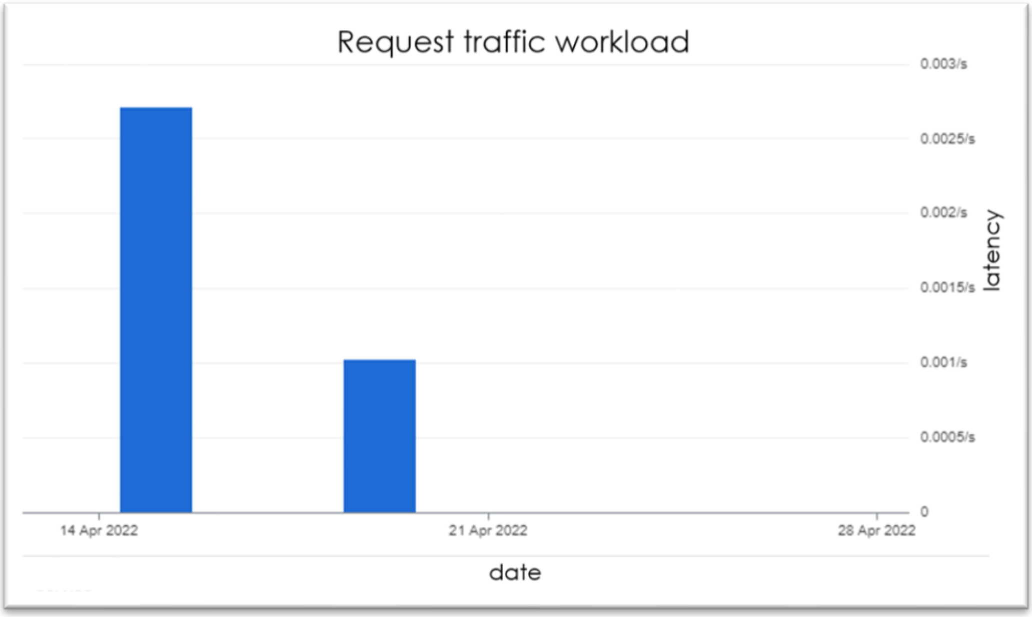
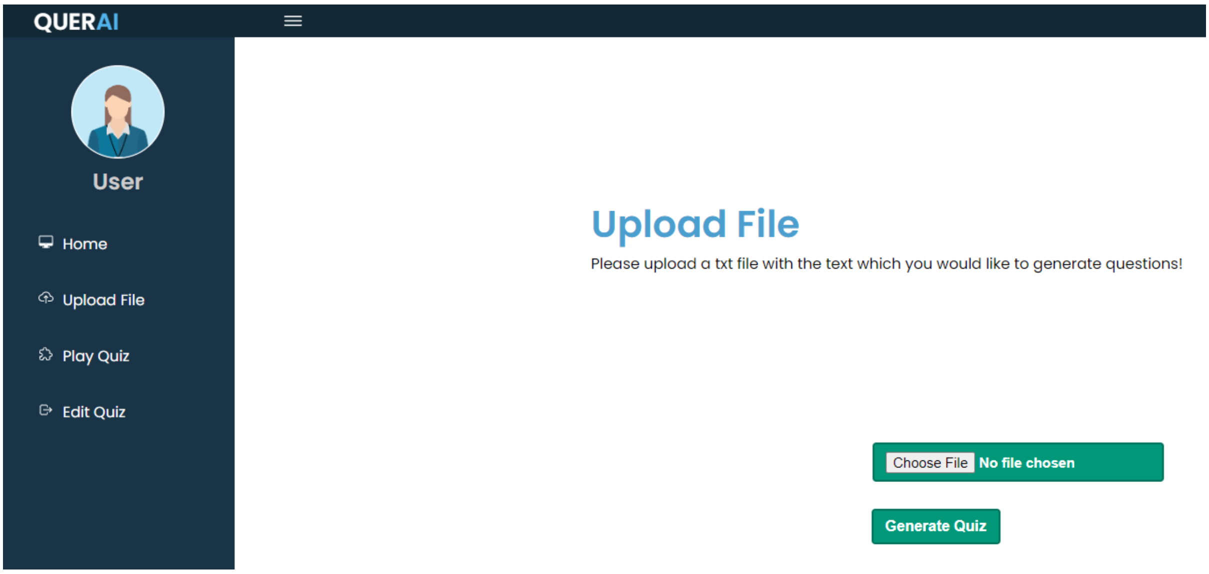


Figure 35: API request traffic workload

11.3 User experience

A desktop design-based, the QUERAI web app enables users to upload a text file with some content and generate a quiz, which helps ease the time to prepare a quiz from scratch.

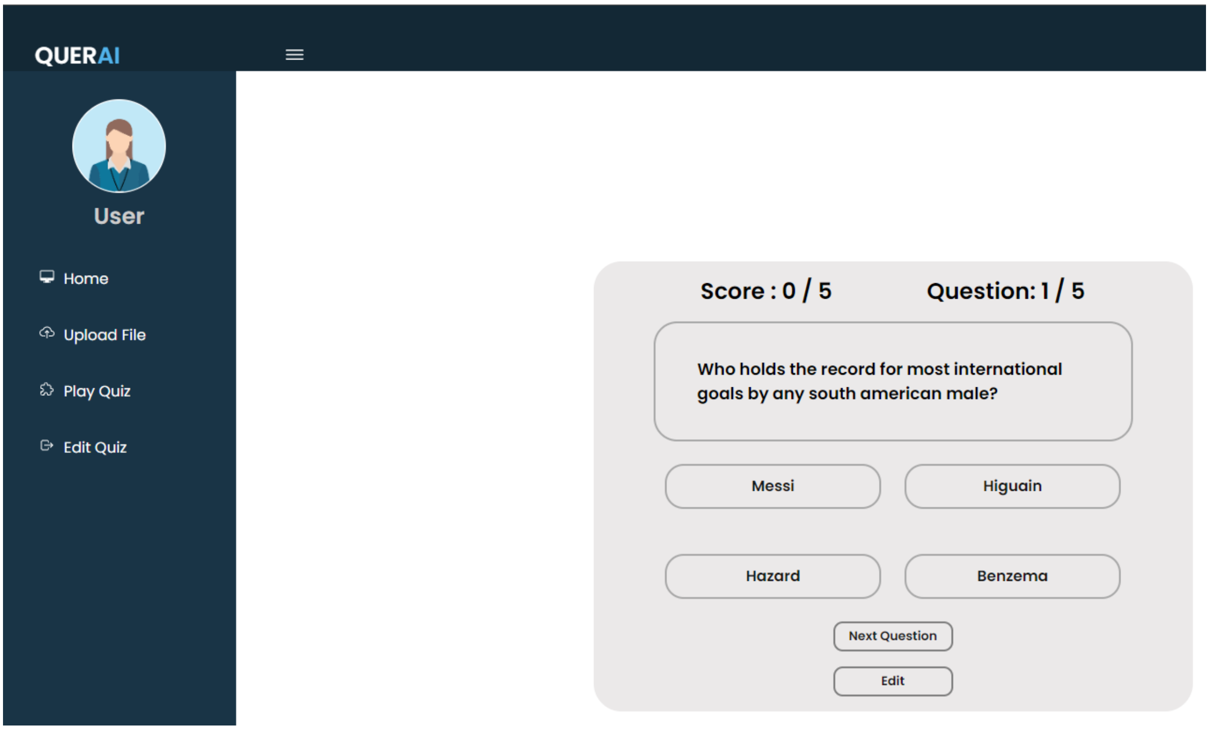


Upload File

Please upload a txt file with the text which you would like to generate questions!

Figure 36 Upload file to generate quiz

After uploading the document, the page will automatically redirect the user to the quiz tab, where the user can attempt the quiz.



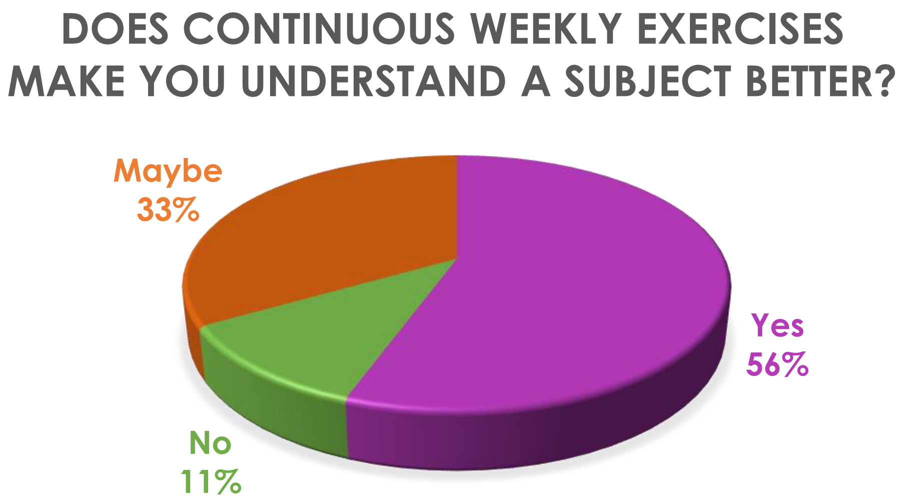
Who holds the record for most international goals by any south american male?

Figure 37 Attempting the quiz

12. Conclusions

It is crucial to have as many tools as students can use to help them go through the beginning of their academic careers, facilitating their inclusion and transition from a high school environment.

Adopting a personalised quiz system for college students can interfere positively with their academic results. Confirming such an assumption, a little research conducted among current graduate students affirms that a routine of weekly exercises can impact the learning experience.



Using the T5 transformer instead of other pre-trained models was motivated by models such as BERT and GTP-2 needing a separate decoder to transform the output into actual text. This aspect resulted in complications when making a detailed model comparison, which was impossible within the time frame imposed on this project.

Bing is very user-friendly and incredible to work with; the T5 transformer comes with many exciting features that create many possibilities regarding future updates for the project. Features such as implementing different types of questions such as true or false, fill in the blanks, etc., are in the plans for a QUERAI version 2.

NLP systems require extensive analysis of their outputs regarding the aspects of the language once computers are yet not fully capable of replicating all the nuances and manoeuvres that we as humans can perpetuate with language.

Finally, learning about the topic and the content absorbed by the group with this project are satisfactory and widely applicable in the industry. The NLP industry is still in constant research and development due to being a very recent topic in terms of technology, giving space for many job opportunities and career growth.

13. References

A. Richards, B. and W. Frankland, P. (2017). *The Persistence and Transience of Memory*. 94th ed. [ebook] Neuron, pp.2-5. Available at: <https://www.cell.com/neuron/fulltext/S0896-6273(17)303653?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0896627317303653%3Fshowall%3Dtrue> [Accessed 2 March 2022].

Agarwal, S. (2017). *My thoughts on Skip-Thoughts - Sanyam Agarwal - Medium*. [online] Medium. Available at: https://medium.com/@sanyamagarwal/my-thoughts-on-skip-thoughts-a3e773605efa [Accessed 4 May 2022].

Agues, N. (2020). *What is crisp DM methodology?* [online] Treehozz.com. Available at: https://treehozz.com/what-is-crisp-dm-methodology [Accessed 11 March 2022].

Alammar, J. (2020). The Illustrated Transformer. [online] Github.io. Available at: http://jalammar.github.io/illustrated-transformer/ [Accessed 5 Apr. 2022].

Alexa, G. (2008). *9 Advantages of using CSS | Webmaster Tips*. [online] Wmtips.com. Available at: https://www.wmtips.com/css/advantages-using-css/ [Accessed 9 March 2022].

Analytics Vidhya. (2021). *NLTK: A Beginners Hands-on Guide to Natural Language Processing*. [online] Available at: https://www.analyticsvidhya.com/blog/2021/07/nltk-a-beginners-hands-on-guide-to-natural-language-processing/ [Accessed 14 Apr. 2022].

Baheti, P. (2021). What Is Transfer Learning? [Examples & Newbie-Friendly Guide]. [online] V7labs.com. Available at: https://www.v7labs.com/blog/transfer-learning-guide [Accessed 30 Mar. 2022].

Bansal, L. (2020). *Google Cloud - Top Advantages And Why You Should Use It In 2020*. [online] Available at: https://www.c-sharpcorner.com/article/google-cloud-top-advantages-and-why-you-should-use-it-in-2020/ [Accessed 27 Apr. 2022].

boudinfl (2022). *boudinfl/pke: Python Keyphrase Extraction module*. [online] GitHub. Available at: https://github.com/boudinfl/pke [Accessed 2 Apr. 2022].

Burns, E., Laskowski, N. and Tucci, L. (2022). *What is artificial intelligence (AI)?* [online] SearchEnterpriseAI. Available at: https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence [Accessed 11 March 2022].

Cameron, B. (2019). *7 Essential Features of Visual Studio Code for Web Developers*. [online] Medium. Available at: https://bretcameron.medium.com/7-essential-features-of-visual-studio-code-for-web-developers-be77e235bf62 [Accessed 27 Apr. 2022].

Chorev, S. (2021). What Is Data Cleaning: A Practical Guide | Deepchecks. [online] Deepchecks. Available at: https://deepchecks.com/what-is-data-cleaning/ [Accessed 5 Apr. 2022].

Cloke, H., 2022. *What Is The Forgetting Curve (And How Do You Combat It)?*. [online] eLearning Industry. Available at: <https://elearningindustry.com/forgetting-curve-combat> [Accessed 2 March 2022].

Data Science Process Alliance. (2022). *CRISP-DM - Data Science Process Alliance*. [online] Available at: https://www.datascience-pm.com/crisp-dm-2/ [Accessed 17 Mar. 2022].

Docker Documentation. (2013). *Docker Documentation*. [online] Available at: https://docs.docker.com/ [Accessed 3 Apr. 2022].

Dziuba, A. (2021). *7 Advantages of Node.js for startups*. [online] Available at: https://relevant.software/blog/7-benefits-of-node-js-for-startups/ [Accessed 27 Apr. 2022].

Ebstein, J. (2021). Our ever-dwindling attention span. [online] Lowell Sun. Available at: https://www.lowellsun.com/2021/07/06/our-ever-dwindling-attention-span/ [Accessed 5 Apr. 2022].

Edu, L. (2021). *Branches of Linguistics, PDF & Branches of Phonetics | Leverage Edu*. [online] Leverage Edu. Available at: https://leverageedu.com/blog/branches-of-linguistics/ [Accessed 22 Apr. 2022].

explosion (2021). *GitHub - explosion/sense2vec: Contextually-keyed word vectors*. [online] GitHub. Available at: https://github.com/explosion/sense2vec [Accessed 4 May 2022].

Frankenfield, J. and Chavarria, A. (2022). Machine Learning. [online] Investopedia. Available at: https://www.investopedia.com/terms/m/machine-learning.asp [Accessed 11 March 2022].

Garbade, M.J. (2018). *A Simple Introduction to Natural Language Processing*. [online] Medium. Available at: https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32 [Accessed 13 Apr. 2022].

Github.io. (2021). *The Stanford Question Answering Dataset*. [online] Available at: https://rajpurkar.github.io/SQuAD-explorer/ [Accessed 11 Apr. 2022].

Glen, S. (2017). *Greedy Algorithm & Greedy Matching in Statistics*. [online] Statistics How To. Available at: https://www.statisticshowto.com/greedy-algorithm-matching/ [Accessed 5 May 2022].

GoIreland (2018). *Public & Private Universities in Ireland | Best Irish Universities | GoIreland*. [online] https://www.goireland.in. Available at: https://www.goireland.in/private-public-universities-ireland [Accessed 10 March. 2022].

Google's T5 (2022). *t5-base · Hugging Face*. [online] Huggingface.co. Available at: https://huggingface.co/t5-base [Accessed 6 Apr. 2022].

Gupta, J. (2016). *Spaced repetition: a hack to make your brain store information*. [online] the Guardian. Available at: https://www.theguardian.com/education/2016/jan/23/spaced-repetition-a-hack-to-make-your-brain-store-information [Accessed 9 March 2022].

Gurudrasil - An Online teach & learn platform. (2021). *Top 10 Challenges Faced By Teachers* - Gurudrasil. [online] Available at: https://gurudrasil.com/top-10-challenges-faced-by-teachers [Accessed 9 March 2022].

Hanson, M. (2021). *College Dropout Rate [2022]: by year + Demographics*. [online] Education Data Initiative. Available at: https://educationdata.org/college-dropout-rates [Accessed 20 Apr. 2022].

Hardesty, L. (2017). *Explained: Neural networks.* [online] MIT News | Massachusetts Institute of Technology. Available at: https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414 [Accessed 6 Apr. 2022].

Hargurjeet (2021). *Training Feed Forward Neural Network(FFNN) on GPU — Beginners Guide*. [online] Medium. Available at: https://medium.com/mlearning-ai/training-feed-forward-neural-network-ffnn-on-gpu-beginners-guide-2d04254deca9 [Accessed 26 Apr. 2022].

Hemmendinger, D. (2022). *HTML | Definition & Facts* | Britannica. In: Encyclopædia Britannica. [online] Available at: https://www.britannica.com/technology/HTML[Accessed. Accessed 2 March 2022].

Huggingface.co. (2020). *T5*. [online] Available at: https://huggingface.co/transformers/v4.10.1/model_doc/t5.html [Accessed 26 Apr. 2022].

Ibanez, D. (2021). *Encoder-Decoder Models for Natural Language Processing | Baeldung on Computer Science.* [online] Baeldung on Computer Science. Available at: https://www.baeldung.com/cs/nlp-encoder-decoder-models [Accessed 31 Mar. 2022].

Intellica.AI (2019). *Comparison of different Word Embeddings on Text Similarity — A use case in NLP*. [online] Medium. Available at: https://intellica-ai.medium.com/comparison-of-different-word-embeddings-on-text-similarity-a-use-case-in-nlp-e83e08469c1c [Accessed 4 May 2022].

IogrBobriakov (2019). *Top NLP Algorithms & Concepts - DataScienceCentral.com*. [online] Data Science Central. Available at: https://www.datasciencecentral.com/top-nlp-algorithms-amp-concepts/ [Accessed 13 Apr. 2022].

iSixSigma. (2010). *Determine The Root Cause: 5 Whys*. [online] Available at: https://www.isixsigma.com/tools-templates/cause-effect/determine-root-cause-5-whys [Accessed 21 March 2022].

Jansen, D. and Warren, K. (2020). *What Is Research Methodology? Definition + Examples - Grad Coach*. [online] Grad Coach. Available at: https://gradcoach.com/what-is-research-methodology/ [Accessed 20 Apr. 2022].

Khushboo (2021). *Types of Linguistics with Examples - EnglishBix*. [online] EnglishBix. Available at: https://www.englishbix.com/types-of-linguistics-with-examples/ [Accessed 13 Apr. 2022].

Korolev, S. (2019). *Prioritisation with MoSCoW: Rules and How to Use* | Railsware Blog. [online] Blog by Railsware | Blog on Engineering, Product Management, Transparency, Culture and many more... Available at: https://railsware.com/blog/moscow-prioritization/ [Accessed 3 March 2022].

Kumar, A. (2021). *NLP Pre-trained Models Explained with Examples - Data Analytics*. [online] Data Analytics. Available at: https://vitalflux.com/nlp-pre-trained-models-explained-with-examples/ [Accessed 14 Apr. 2022].

Levity.ai. (2022). *How Natural Language Processing works in 2022*. [online] Available at: https://levity.ai/blog/how-natural-language-processing-works [Accessed 11 Apr. 2022].

Linguisticsociety.org. (2022). *What is Linguistics? | Linguistic Society of America*. [online] Available at: https://www.linguisticsociety.org/what-linguistics [Accessed 13 Apr. 2022].

Lutkevich, B. and Burns, E. (2021). *natural language processing (NLP)*. [online] SearchEnterpriseAI. Available at: https://www.techtarget.com/searchenterpriseai/definition/natural-language-processing-NLP [Accessed 11 Apr. 2022].

Lynch, M. (2018). *7 Ways That Artificial Intelligence Helps Students Learn* - The Edvocate. [online] The Edvocate. Available at: https://www.theedadvocate.org/7-ways-that-artificial-intelligence-helps-students-learn/[Accessed 2 March 2022].

Maluuba (2021). *Maluuba/nlg-eval: Evaluation code for various unsupervised automated metrics for Natural Language Generation.* [online] GitHub. Available at: https://github.com/Maluuba/nlg-eval [Accessed 4 May 2022].

Markowitz, D. (2021). *Transformers, Explained: Understand the Model Behind GPT-3, BERT, and T5*. [online] Daleonai.com. Available at: https://daleonai.com/transformers-explained [Accessed 26 Apr. 2022].

MasterClass (2020). *How to Perform a Root Cause Analysis*. [online] MasterClass. Available at: https://www.masterclass.com/articles/how-to-perform-a-root-cause-analysis#what-is-a-root-cause-analysis [Accessed 21 March 2022].

Matteo Duò (2020). *40+ SaaS Products We Use to Grow Our Web Hosting Company*. [online] Kinsta®. Available at: https://kinsta.com/blog/saas-products/ [Accessed 27 Apr. 2022].

Meera (2020). *What is Linguistics: Meaning, Scope, Branches, Types and Career*. [online] Sociology Group: Sociology and Other Social Sciences Blog. Available at: https://www.sociologygroup.com/linguistics-meaning-branches-types-scope-career/ [Accessed 13 Apr. 2022].

Megida, D. (2021). *What is JavaScript? A Definition of the JS Programming Language*. [online] freeCodeCamp.org. Available at: https://www.freecodecamp.org/news/what-is-javascript-definition-of-js [Accessed 6 March 2022].

Millidge, S. (2018). *Benefits of Open Source vs. Proprietary Software.* [online] dzone.com. Available at: https://dzone.com/articles/benefits-of-open-source-vs-proprietary-software [Accessed 27 Apr. 2022].

Mindfire Solutions (2017). *Python: 7 Important Reasons Why You Should Use Python*. [online] Medium. Available at: https://medium.com/@mindfiresolutions.usa/python-7-important-reasons-why-you-should-use-python-5801a98a0d0b [Accessed 9 Mar. 2022].

Mindtools.com. (2021). SWOT *Analysis: Strengths, Weaknesses, Opportunities, Threats*. [online] Available at: https://www.mindtools.com/pages/article/newTMC_05.htm [Accessed 3 March 2022].

mukesh-mehta (2021). *explosion/sense2vec: Contextually-keyed word vectors*. [online] GitHub. Available at: https://github.com/explosion/sense2vec [Accessed 2 Apr. 2022].

Mustafa, N. (2020). *Getting the Most Out of Pre-trained Models*. [online] Toptal Engineering Blog. Available at: https://www.toptal.com/deep-learning/exploring-pre-trained-models#:~:text=Google's%20T5%20is%20one%20of,decoder%20blocks%2C%20T5%20uses%20both. [Accessed 26 Apr. 2022].

Nicholas, J. (2021). *The 5 Steps in Problem Analysis*. [online] BusinessAnalystMentor.com. Available at: https://businessanalystmentor.com/problem-analysis/#:~:text=Problem%20analysis%20is%20the%20process,solved%20before%20developing%20a%20solution [Accessed 19 March 2022].

Nogueira, R., Jiang, Z., Ronak Pradeep and Lin, J. (2020). *Document Ranking with a Pretrained Sequence-to-Sequence Model*. [online] ResearchGate. Available at: https://www.researchgate.net/publication/347235946_Document_Ranking_with_a_Pretrained_Sequence-to-Sequence_Model [Accessed 26 Apr. 2022].

Paradiso eLearning Blog. (2017). *Moodle vs Blackboard, which is better for your organisation?* [online] Available at: https://www.paradisosolutions.com/blog/moodle-vs-blackboard-lms-comparison/ [Accessed 9 March 2022].

Parvez, F. (2021). *Introduction to CSS | CSS Tutorial for Beginners*. [online] GreatLearning Blog: Free Resources what Matters to shape your career! Available at: https://www.mygreatlearning.com/blog/css-tutorial [Accessed 8 March 2022].

patrickvonplaten (2022). *t5-base at main*. [online] Huggingface.co. Available at: https://huggingface.co/t5-base/tree/main [Accessed 20 Apr. 2022].

Prabhakaran, S. (2018). *Cosine Similarity – Understanding the math and how it works (with python codes)*. [online] Machinelearningplus.com. Available at: https://www.machinelearningplus.com/nlp/cosine-similarity/#:~:text=Cosine%20similarity%20is%20a%20metric%20used%20to%20determine%20how%20similar,in%20a%20multi%2Ddimensional%20space. [Accessed 4 May 2022].

Princeton.edu. (2010). *WordNet | A Lexical Database for English*. [online] Available at: https://wordnet.princeton.edu/ [Accessed 2 Apr. 2022].

Pykes, K. (2021). *The Most Common Evaluation Metrics In NLP - Towards Data Science*. [online] Medium. Available at: https://towardsdatascience.com/the-most-common-evaluation-metrics-in-nlp-ced6a763ac8b [Accessed 4 May 2022].

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu Google, P. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. [online] Available at: https://arxiv.org/pdf/1910.10683.pdf [Accessed 30 Mar. 2022].

Rajpurkar, P., Zhang, J., Lopyrev, K. and Liang, P. (n.d.). *SQuAD: 100,000+ Questions for Machine Comprehension of Text*. [online] Available at: https://nlp.stanford.edu/pubs/rajpurkar2016squad.pdf. [Accessed 13 Apr. 2022].

Reactor, H. (2021). *What is JavaScript used for?* [online] Hackreactor.com. Available at: https://www.hackreactor.com/blog/what-is-javascript-used-for [Accessed 9 Mar. 2022].

Redhat.com. (2022). *What is Docker?* [online] Available at: https://www.redhat.com/en/topics/containers/what-is-docker?sc_cid=7013a000002wLvxAAE&gclid=Cj0KCQjw06OTBhC_ARIsAAU1yOUQK5XTOpHE-CN3nw8M2lycGHillc80ZrZhFUuuKi6Xj326w62Cq-8aAt-HEALw_wcB&gclsrc=aw.ds [Accessed 27 Apr. 2022].

Reyes, K. (2022). *What is Deep Learning and How Does It Works [Explained]*. [online] Simplilearn.com. Available at: https://www.simplilearn.com/tutorials/deep-learning-tutorial/what-is-deep-learning [Accessed 11 March 2022].

Sas.com. (2022). *What is Natural Language Processing (NLP)?* [online] Available at: https://www.sas.com/en_ie/insights/analytics/what-is-natural-language-processing-nlp.html [Accessed 11 Apr. 2022].

Schawbel, D. (2013). *Josh Kaufman: It Takes 20 Hours Not 10,000 Hours To Learn A Skill*. Forbes. [online] 17 Dec. Available at: https://www.forbes.com/sites/danschawbel/2013/05/30/josh-kaufman-it-takes-20-hours-not-10000-hours-to-learn-a-skill/?sh=785ebbf1363d [Accessed 29 March 2022].

Singh, V. (2020). *What is PyCharm IDE? What is PyCharm used for?* [online] TechGeekBuzz. Available at: https://www.techgeekbuzz.com/what-is-pycharm/ [Accessed 9 March 2022].

Square. (2020). *How to Implement a Subscription Service into Your Business*. [online] Available at: https://squareup.com/ie/en/townsquare/how-to-create-a-subscription-service?country_redirection=true [Accessed 27 Apr. 2022].

Tamm, S. (2021). *Spaced Repetition: A Guide to the Technique - E-Student*. [online] E-Student. Available at: https://e-student.org/spaced-repetition [Accessed 30 April 2021].

TechTarget (2021). *natural language understanding (NLU)*. [online] SearchEnterpriseAI. Available at: https://www.techtarget.com/searchenterpriseai/definition/natural-language-understanding-NLU [Accessed 5 Apr. 2022].

tiangolo (2018). *tiangolo/fastapi: FastAPI framework, high performance, easy to learn, fast to code, ready for production*. [online] GitHub. Available at: https://github.com/tiangolo/fastapi [Accessed 3 Apr. 2022].

tiangolo (2021). *tiangolo/uvicorn-gunicorn-fastapi-docker: Docker image with Uvicorn managed by Gunicorn for high-performance FastAPI web applications in Python 3.6 and above with performance auto-tuning. Optionally with Alpine Linux*. [online] GitHub. Available at: https://github.com/tiangolo/uvicorn-gunicorn-fastapi-docker [Accessed 4 Apr. 2022].

Trinity College Dublin (2022). *Linguistics (M.Phil. / P.Grad.Dip.) | Trinity College Dublin*. [online] Www.tcd.ie. Available at: https://www.tcd.ie/courses/postgraduate/courses/linguistics-mphil--pgraddip/ [Accessed 13 Apr. 2022].

Ucdavis.edu. (2022). *What is Linguistics? — Linguistics*. [online] Available at: https://linguistics.ucdavis.edu/undergraduate/what-linguistics [Accessed 13 Apr. 2022].

Ucsc.edu. (2022). *What is Linguistics?* [online] Available at: https://linguistics.ucsc.edu/about/what-is-linguistics.html [Accessed 13 Apr. 2022].

Wei, J. (2020). *The Quick Guide to SQuAD - Towards Data Science*. [online] Medium. Available at: https://towardsdatascience.com/the-quick-guide-to-squad-cae08047ebee [Accessed 13 Apr. 2022].

Witman, E. (2021). *What is Python? The programming language, explained*. [online] Business Insider. Available at: https://www.businessinsider.com/what-is-python?r=US&IR=T [Accessed 6 March 2022].

Wolframcloud.com. (2019). *SQuAD v2.0 | Wolfram Data Repository*. [online] Available at: https://datarepository.wolframcloud.com/resources/SQuAD-v2.0 [Accessed 11 Apr. 2022].

Yalçın, O.G. (2020). 4 Reasons Why You Should Use Google Colab for Your Next Project. [online] Medium. Available at: https://towardsdatascience.com/4-reasons-why-you-should-use-google-colab-for-your-next-project-b0c4aaad39ed [Accessed 27 Apr. 2022].

Yatskar, M. (2019). *A Qualitative Comparison of CoQA, SQuAD 2.0 and QuAC*. [online] Association for Computational Linguistics, pp.2318–2323. Available at: https://aclanthology.org/N19-1241.pdf [Accessed 26 Apr. 2022].

Zublenko, E. (2016). *Why Django is the Best Web Framework for Your Project*. [online] Steelkiwi.com. Available at: https://steelkiwi.com/blog/why-django-best-web-framework-your-project/ [Accessed 11 March 2022].

14. Appendix A: Definitions

Before we jump into processing all these stages, we will go through a few essential concepts that will guide us in implementing an NLP model.

Machine Learning is the concept that a computer program can learn and adapt to new data without human intervention. Machine learning is a field of artificial intelligence (AI) that keeps a computer's built-in algorithms current regardless of changes in the worldwide economy (Frankenfield and Chavarria, 2022).

Deep Learning can be considered a subset of the machine learning field, where the learning and improvement of the model process happen on its own by examining computer algorithms. While machine learning uses more straightforward concepts, deep learning works with artificial neural networks, which are designed to imitate how humans think and learn (Reyes, 2022).

Neural Networks, in simple words, consist of the idea of reproducing the human brain's neuron workflow via processing nodes that are densely interconnected to each other. Those nodes communicate with each other within a "feed-forward" manner, meaning that the data moves through the nodes in only one direction where a node will be interconnected to many other nodes beneath it to receive data and above it to send data(Hardesty, 2017).

Artificial Intelligence is the simulation of human intelligence processes by machines, especially computer systems. Specific applications of AI include expert systems, natural language processing, speech recognition and machine vision. (Burns, E. et al. 2022).

15. Appendix B: Technologies and Tools

Some technologies and tools are of crucial importance for the achievement of the project goals; they can together be used to implement a teamwork project management strategy where an explanation and the reasons why they were chosen are provided as follows:

15.1 Open-Source vs Proprietary Software

Open-source software (Millidge, 2018) offers a few advantages over proprietary software. Bugs and defects are found and corrected proactively by developers worldwide.

The reason for using Open-Source software over proprietary software is that the second ones are built with hidden source code and offered via a licensing system that users must pay for to use the software.

15.2 Hypertext Markup Language (HTML)

HTML is used to display content from the internet on a personal computer's web browser by being written with tags that specify elements such as headings, paragraphs, and tables (Hemmendinger, 2020).

Considered industry-standard for content delivery to users, HTML can be easily implemented with other languages to create fully responsive websites which are used in front-end web development, and the majority of browsers support it.

15.3 CSS (Cascading Style Sheets)

CSS handles the attributes of a webpage presentation, such as colours, layouts, and fonts by enabling automatic adjustment of component sizes for several types of screen sizes by adapting the website accordingly (Parvez, 2021).

CSS can easily be used with all types of XML-based markup languages, such as HTML, and used to create friendly and customisable webpages by having functions that allow element positioning and animations (Parvez, 2021).

15.4 JavaScript

JavaScript is a dynamic programming language that can be used for all types of development that exist to this date; for example, it allows the implementation of dynamic features on web pages that cannot be done with only HTML and CSS alone (Megida, 2022).

Considered one of the most famous programming languages, troubleshooting content is widely available on the internet. Also, being the only programming language native to the web browser, it is one of the most desired skills in the industry nowadays (Reactor, 2021).

This language was chosen for the front-end and back-end website development as it is an open-source programming language with a massive amount of troubleshooting content available. It is a skill very desired in the market.

15.5 Python

Python is considered the pioneer and an industry-standard programming language of Artificial Intelligence (AI) and Machine Learning (ML) for the implementation and deployment of models (Mindfire Solutions, 2017).

A multipurpose programming language that has applicability anywhere that uses data or mathematical computation. Unlike Javascript, Python is not confined to being used only for web development (Witman, 2021).

15.6 PyCharm

PyCharm was developed by JetBrains (Singh, 2022) and is one of the most popular integrated development environment (IDE) frameworks in the market for Python development by offering better support for Data Science and Machine Learning libraries than its competitors such as VSCode, Anaconda, and others (Singh, 2022).

This software was used as the primary IDE for the deployment section as it contains a brilliant Code Editor that facilitates writing high-quality Python code with integrated tools.

15.7 Google Colab

According to (Yalçın, 2020), Google Colab is an excellent tool for deep learning and neural network tasks because it allows developers to write and execute Python code through their browser.

It also gives free access to Google computing resources such as GPUs and TPUs, contributing to a faster training process of the models, which is the main advantage over its competitor Jupyter Notebook.

15.8 Docker vs Linux Containers

Docker is an open-source project that offers a software development solution known as containers. For (Redhat.com, 2022), the disadvantage of Traditional Linux Containers is that it uses an init system that can manage multiple processes where the entire applications can run as one. The Docker technology allows applications to be broken down into separate processes.

The main advantage of using Docker over Linux Containers is that it eases creating and building containers, shipping images, and versioning images, among other things.

15.9 Google Cloud vs AWS

Google Cloud is a highly reliable and scalable cloud platform offering the best cloud computing services on the web that allows users to store and compute data while also helping developers test, build, and deploy applications (Bansal, 2020).

Google Cloud was chosen over others cloud platforms because it has the most accessible service to deploy docker images and better pricing plans compared to its competitors.

15.10 Node.js vs AngularJS

Node is an open-source, cross-platform runtime environment for executing JavaScript code at the server-side and outside a web browser. Node.js (Dziuba, 2021) offers fast building, scalable network applications, benefits in performance, more rapid development, and other perks. These processing and consuming real-time information requirements are paramount and are exceptionally fast for multi-user real-time data situations.

Node.js was chosen because it is suitable for building server-side applications, while AngularJS is suited for building single-page client-side web applications.

15.11 VSCode vs Other Editors

(Cameron, 2019) mention that Visual Studio Code comes with great features built-in, added in a large (and growing) pool of extensions and end up with thousands of ways to customise the programming experience.

VSCoe was chosen to develop the website even though there are many other editor options because its extensions make the development process a lot easier.

16. Appendix C: Reflective Journal: Individual Contributions

16.1 Elton da Silva

Since the first meeting, while deciding on the idea for the project, the group was always active and willing to discuss ideas that would eventually complement what the project became at the end of the semester.

As per my contribution to the group, I was intensively involved with managing the tasks in both practical and theoretical sections by giving the direction in which the project should be take.

Teamwork and communication among the group were critical to successfully implementing the final product. By making weekly meetings outside of class, we strived for something challenging, rewarding, and, nonetheless, a product that can be showcased to employers as a professional accomplishment.

I genuinely believe that I have improved my soft and hard skills by completing this project with such complexity involved in a short timed-manner by overcoming most obstacles during the implementation phase.

I would like to especially thank Dr Muhammad Iqbal and Dr David McQuaid for giving intensive assistance by sourcing enough study materials and by being always willing to help in moments where we were struggling.

Finally, I would like to thank my peers for their commitment and the focus in achieving a professional level of work and all the professors involved in my four years studying at CCT College.

16.2 Fernando Aires da Silva

My contribution for the project started when we were deciding about what to do, giving my suggestions of how I imagined the project might be following the assignment requests, researching machine learning and talking with the group proving ideas on how we could use artificial intelligence to start execution.

The project at the beginning seemed quite confusing, as we did not have much knowledge about Natural Language Processing. In order to help us to achieve it was as divided into several tasks such as studies through online courses, reading books, online research, theoretical and practical application, I researched some videos and courses on platforms such as Udemy, YouTube amont others. I also contributed to obtaining important materials such as books that helped us to better understand and put the idea into practice.

I started doing implementation of the code using the Jupyter notebook tool, but it was not successful due system limitations. After several weeks trying to make the project work, we agreed I could focus on the development of the front-end design and website, although we had a module about that, I felt a little lost at first, especially using CSS and JavaScript. So, after some research on the subject and tutorials I was able to develop a responsive website to make it easy and intuitive thinking about the usability for the end customer, in addition to being able to be used on different devices in order to popularise as we provide a plataform for questions and answers using Artificial Intelligence.

The technological part of the project demanded that we seek to understand different technologies. I prepared the written part with the technologies used and their usefulness in the project.

First, I would like to thank you the lecturers Dr Muhammad Iqbal and Dr. David McQuaid and Ken Healy for all the support and effort during this years, as well as, my groupmates Elton, Tai Teei and King Jang, who worked brilliantly for us to achieve the final result.

Finally, I would like to thank you for CCT College as it was an incredible journey during these 4 years of learning and improving my soft and hard skills that I feel it was the best decision I could have made and I will always remember and be grateful.

16.3 Kim Jang Wong

It is my privilege to work with a group of teams like Elton, Tai Teei and Fernando in this project. Each of us have dedicated ourselves to different tasks to complete this project effectively.

My contribution to this project, consist of the following:

1) Business Strategy – analyse the needs of business strategy

2) CRISP-DM – using the framework to identify the important aspect in CRISP-DM methodology

3) Usability of text-based datasets – the usability of the SQuAD dataset, Natural Language Processing and Pre-Trained Models

4) Summary – concluding the summary of the data understanding.

5) User experience – provides use of user experiences toward the goal of this project and manages user feedback.

6) Create a README file on GitHub – documenting the idea of this project and how to use it and what to expect on this project.

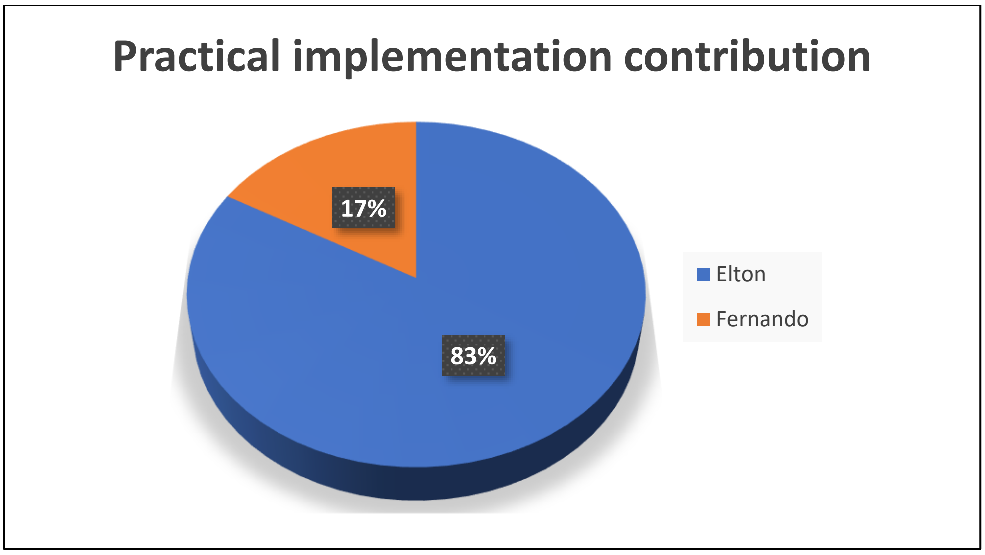
7) Appendix C: Excel development plan spreadsheet – composes the index of each chapter on this project and provides acknowledgement on this project.

16.4 Tai Teei Ho

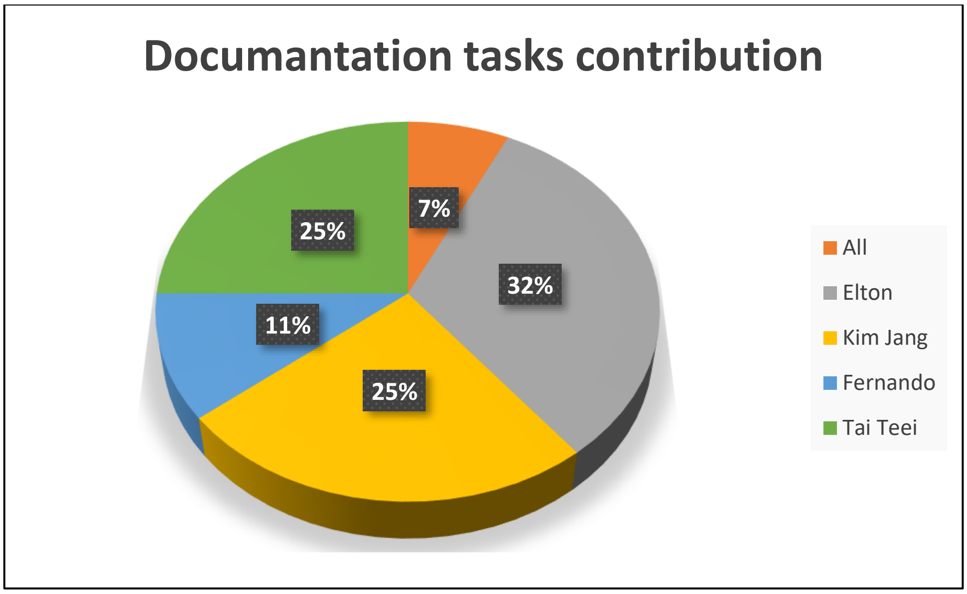
My task is more on research and documentation part. For example, Problem Analysis, What is Natural Language Processing and How does it work on our project and SaaS Pricing strategy. I was assigned to do the Figures/Table numbering and the general formatting for the report. I would like to thank my teammates, Elton, Fernando and Kim Jang. I am grateful and happy to learn and work together with them. They always encourage me when I was in difficulty on the research. Without them, I would not have made it through my Bachelor (Hons) course. I appreciate that I have an opportunity to communicate with people as I used to be an introverted person. Doing this project, I realised working with a group was a challenge as we had to arrange online meetings weekly and everyone had their own ideas and opinions on the project. However this is a good way to learn team working skills and gain experience from others. Last but not least, the completion of this project could not have been possible without the assistance of our lecturers, Dr. Muhammad Iqbal, Ken Healy and David McQuaid.

17. Appendix D: Reflective Journal: Group Contributions

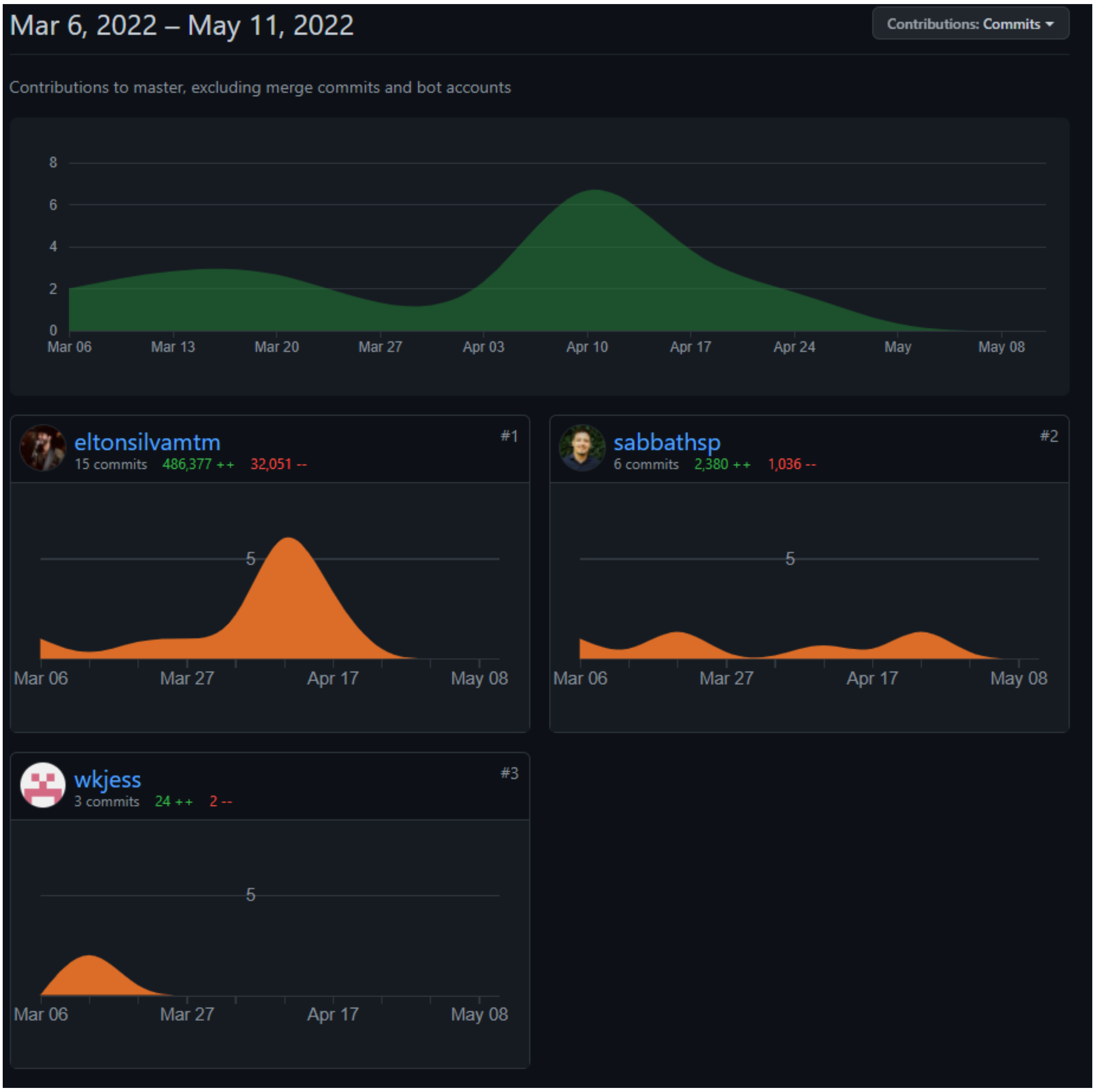
The graph below represents the distribution of the practical work throughout the semester. The visualisation was achievable by retrieving the information added to basecamp on the to-dos section.



The graph below displays how the group distributed the documentation work, the data acquired for the chart was generated by the to-dos tasks on basecamp.



The image below reflects the amount of practical work pushed to the GitHub website throughout the semester. It also contains information such as the number of commits and the number of lines of code each member uploaded to the main repository.



Mar 6, 2022 – May 11, 2022

Contributions: Commits

Contributions to master, excluding merge commits and bot accounts

eltonsilvamtm #1
15 commits 486,377 ++ 32,051 --

sabbathsp #2
6 commits 2,380 ++ 1,036 --

wkjess #3
3 commits 24 ++ 2 --

The table below represents all the workflow from both practical and theoretical work into a development plan worksheet that tracks the process of developing the project on a weekly basis.