



Project Title :

MCQ Generator System



Technologies: NLP, Deep Learning



B22HS02 : Team members

Bharghav Sai. P
(Leader)

Akhul Chiranjeevi

Guru Nithin

Instructor : Dr. Himangshu Sir

Project Idea :



“Automatically Generate Multiple Choice Questions (MCQs) from any content like textbooks/Wikipedia page/pdf format/raw text using Natural Language Processing and Deep Learning”.

Problem Statement and Motivation :



- All institutes, colleges, and schools have been switched to online learning.
- Assessment is an essential tool to test the knowledge of the students.
- And the pattern of the assessment has changed from subjective-based to objective-based i.e. Multiple Choice Questions (MCQs).

Problem Statement and Motivation :



- So the problem is, it is very difficult for the teachers to set the questions.
- As well as for the students who are preparing for competitive exams.
- The current method involves the setting of questions manually..

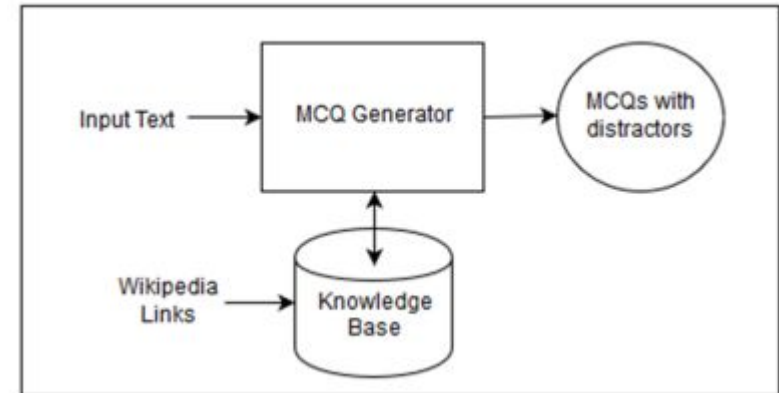
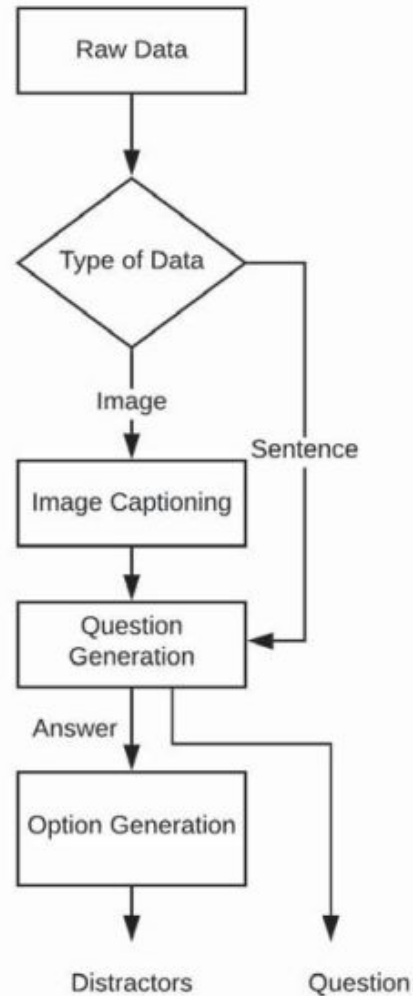
Problem Statement and Motivation :



- lot of human intervention and time.
- So there is a growing need for a system that can create questions with ease
- And less amount of time and requires less human effort.

State of the art :

Source : [A. Srivastava, S. Shinde, N. Patel, S. Deshpande, A. Dalvi and S. Tripathi, "Questionator-Automated Question Generation using Deep Learning," 2020 International Conference on Emerging Trends in Information Technology and Engineering \(ic-ETITE\), 2020, pp. 1-5, doi: 10.1109/ic-ETITE47903.2020.212.](#)



SIMPLE EXAMPLE WITH APPROACH !

Chapter 6 Social Studies The Amazon River

The Amazon in South America is the largest river by discharge volume of water in the world, and the second-longest river in the world.

The total length of the river as measured from the headwaters of the Ucayali-Apurímac river system in southern Peru is at least 4,000 miles (6,400 km), which makes it slightly shorter than the Nile River but still the equivalent of the distance from New York City to Rome. Its westernmost source is high in the Andes Mountains, within 100 miles (160 km) of the Pacific Ocean, and its mouth is in the Atlantic Ocean, on the northeastern coast of Brazil. However, both the length of the Amazon and its ultimate source have been subjects of debate since the mid-20th century, and some claim that the Amazon is longer than the Nile.


Find important
Sentences
(Abstractive or
Extractive
Summarization)

The **Amazon** in South America is the largest river by discharge volume of water in the world, and the second-longest river in the world.



Sentence2MCQ

Which is the world's second longest river ?

a Mississippi

b Amazon

c Nile

d Yangtze

ABSTRACTIVE / EXTRACTIVE SUMMARIZATION

- Summarization refers to the task of creating a short summary of the whole text. Summarization can be done in two ways, abstractive summarization, and extractive summarization.
- While extractive summarization extracts words and word phrases from the original text to create a summary, abstractive summarization learns an internal language representation to generate more human-like summaries, paraphrasing the intent of the original text.

Output that we got after summarization

original Text >>

Natural Language Processing (NLP) is a subfield of artificial intelligence (AI). It helps machines process and understand the human language so that they can automatically perform repetitive tasks. Examples include machine translation, summarization, ticket classification, and spell check. Take sentiment analysis, for example, which uses natural language processing to detect emotions in text. This classification task is one of the most popular tasks of NLP, often used by businesses to automatically detect brand sentiment on social media. Analyzing these interactions can help brands detect urgent customer issues that they need to respond to right away, or monitor overall customer satisfaction. One of the main reasons natural language processing is so critical to businesses is that it can be used to analyze large volumes of text data, like social media comments, customer support tickets, online reviews, news reports, and more. All this business data contains a wealth of valuable insights, and NLP can quickly help businesses discover what those insights are. It does this by helping machines make sense of human language in a faster, more accurate, and more consistent way than human agents. NLP tools process data in real time, 24/7, and apply the same criteria to all your data, so you can ensure the results you receive are accurate – and not riddled with inconsistencies. Once NLP tools can understand what a piece of text is about, and even measure things like sentiment, businesses can start to prioritize and organize their data in a way that suits their needs. While there are many challenges in natural language processing, the benefits of NLP for businesses are huge making NLP a worthwhile investment. However, it's important to know what those challenges are before getting started with NLP. Human language is complex, ambiguous, disorganized, and diverse. There are more than 6,500 languages in the world, all of them with their own syntactic and semantic rules. Even humans struggle to make sense of language. So for machines to understand natural language, it first needs to be transformed into something that they can interpret. In NLP, syntax and semantic analysis are key to understanding the grammatical structure of a text and identifying how words relate to each other in a given context. But, transforming text into something machines can process is complicated. Data scientists need to teach NLP tools to look beyond definitions and word order, to understand context, word ambiguities, and other complex concepts connected to human language.

Summarized Text >>

Natural language processing (nlp) is a subfield of artificial intelligence (ai) it helps machines process and understand the human language so that they can automatically perform repetitive tasks. Examples include machine translation, summarization, ticket classification, and spell check. Sentiment analysis is one of the most popular tasks of nlp, often used by businesses to automatically detect brand sentiment on social media.

time: 14.9 s (started: 2022-04-25 06:51:33 +00:00)

Extract keywords

News Article

Elon Musk and Dogecoin

Elon Musk has shown again he can influence the digital currency market with just his tweets.

After saying that his electric vehicle-making company Tesla will not accept payments in Bitcoin because of environmental concerns, he tweeted that he was working with developers of Dogecoin to improve system transaction efficiency.

Following the two distinct statements from him, the world's largest cryptocurrency hit a two-month low, while Dogecoin rallied by about 20 percent.



Extract Keyphrases using
Python Keyword Extraction Library

Bitcoin, Dogecoin, Tesla, Cryptocurrency

MultipartiteRank

Keyword/Keyphrase extraction

Given the paraphrased sentence, next, we would extract the keywords. These keywords would serve as the basis for questions that will be generated in the next step.

Multipartite rank is one useful algorithm, used in keyword extraction. We can leverage the open-source library python-keyphrase-extraction (pk3) for this task.

Output that we got after keyword extraction

```
↳ keywords unsummarized: ['natural language processing', 'nlp', 'businesses', 'machines', 'text', 'language', 'tasks', 'brand sentiment', 'data', 'customer issues', 'subfield', 'examples']  
keywords_found in summarized: ['media', 'businesses', 'language', 'brand sentiment', 'tasks', 'nlp', 'examples', 'natural language processing', 'machines', 'subfield']  
['natural language processing', 'nlp', 'businesses', 'machines']  
time: 1.01 s (started: 2022-04-25 06:52:07 +00:00)
```

Generate a question using T5 Transformer model

Context:

The **Amazon** in South America is the largest river by discharge volume of water in the world, and the second-longest river in the world.

Answer:

Amazon



**Text to text transfer
transformer**

Question

Which is the world's second longest river ?

https://drive.google.com/file/d/1Dc6W3F_okw1q6GxhKs46lvgeeBsP0iG/view?usp=sharing

QUESTION GENERATION

→ We have the keywords and the text, which will now be used to generate questions related to each keyword.

→ We can use the **T5 Transformer model** that is explicitly trained to take some context and a relevant keyword and generate an appropriate question as shown in prev slide.

Output we got after Question generation with correct ans

```
/usr/local/lib/python3.7/dist-packages/transformers/generation_utils.py:1839:  
    next_indices = next_tokens // vocab_size  
What does nlp stand for?  
Natural language processing
```

```
What is a subfield of artificial intelligence called?  
Nlp
```

```
Who uses sentiment analysis to detect brand sentiment on social media?  
Businesses
```

```
What does natural language processing help?  
Machines
```

```
time: 1.79 s (started: 2022-04-25 06:54:17 +00:00)
```



Visualization with gradio

Running on public URL: <https://23597.gradio.app>

This share link expires in 72 hours. For free permanent hosting, check out Spaces: <https://huggingface.co/spaces>

context

Natural Language Processing (NLP) is a subfield of artificial intelligence (AI). It helps machines process and understand the human language so that they can automatically perform repetitive tasks. Examples include machine translation, summarization, ticket classification, and spell check. Take sentiment analysis, for example, which uses natural language processing to detect emotions in text. This classification task is one of the most popular tasks of NLP, often used by businesses to automatically detect brand sentiment on social media. Analyzing these interactions can help brands detect urgent customer issues that they need to respond to right away, or monitor overall customer satisfaction. One of the main reasons natural language processing is so critical to businesses is that it can be used to analyze large volumes of text data, like social media comments, customer support tickets, online reviews, news reports, and more. All this business data contains a wealth of valuable insights, and NLP can quickly help businesses

What does nlp stand for?Ans: Natural language processing

What is a subfield of artificial intelligence called?

Ans: Nlp

Who uses sentiment analysis to detect brand sentiment on social media?Ans: Businesses

What does natural language processing help machines do?Ans: Machines process

Summary: Natural language processing (nlp) is a subfield of artificial intelligence (ai) it helps machines process and understand the human language so that they can automatically perform repetitive tasks. Examples include machine translation, summarization, ticket classification, and spell check. Sentiment analysis is one of the most popular tasks of nlp, often used by businesses to automatically detect brand sentiment on social media.

Flag

(<gradio.routes.App at 0x7fdbd5e4a110>,

Generate wrong choices for MCQ options

Chapter 6 Social Studies

The Amazon River

The Amazon in South America is the largest river by discharge volume of water in the world, and the second-longest river in the world.

The total length of the river as measured from the headwaters of the Ucayali-Apurimac river system in southern Peru is at least 4,000 miles (6,400 km), which makes it slightly shorter than the Nile River but still the equivalent of the distance from New York City to Rome. Its westernmost source is high in the Andes Mountains, within 100 miles (160 km) of the Pacific Ocean, and its mouth is in the Atlantic Ocean, on the northeastern coast of Brazil. However, both the length of the Amazon and its ultimate source have been subjects of debate since the mid-20th century, and some claim that the Amazon is longer than the Nile.



Multiple Choice Question

Which is the world's longest river ?

a

Mississippi

b

Amazon

c

Nile

d

Yangtze

Given the **Amazon** as input, we will see how we can generate **distractors** (wrong choices) like **Mississippi, Nile, Yangtze**, using **Wordnet**.

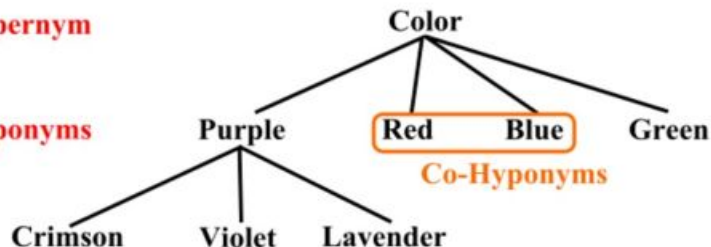
<https://blog.questgen.ai/complete-guide-to-generating-mcq>


Generate distractors using Wordnet

<https://wordnet.princeton.edu/>

Hypernym

Hyponyms

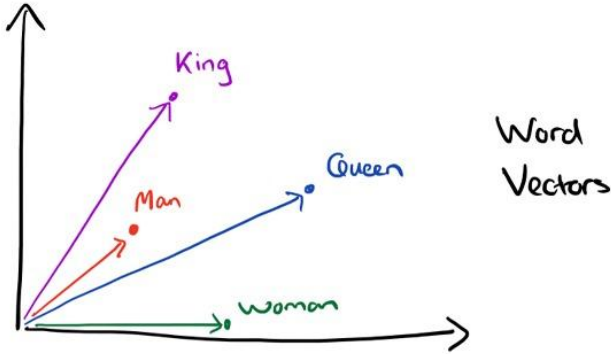


An example of the relationship between hyponyms and hypernym 

https://en.wikipedia.org/wiki/Hyponymy_and_hypernymy

- WordNet® captures relations.
- A hyponym is in a type-of relationship with its hypernym.
- A hypernym is an umbrella term and blanket term. Eg: Color is the hypernym for Purple.
- Red, Blue, Green are the hyponyms of "Color". Since Red is a type of color.
- Our goal is to extract Co-Hyponyms as distractors.

Introduction to Word vectors








$$\text{King} - \text{Man} + \text{Woman} = \text{Queen}$$



- Word vectors are words converted into a vector or array representation.
- These vectors capture associations among different kinds of words.
- Unlike Wordnet and Conceptnet these are not human curated and automatically generated from a text corpus.
- A neural network algorithm is trained with predicting the focus word given other words or predicting surrounding words given a focus word.
- The limitations are that all senses of a given word have only one vector. Word vectors are context-independent.

Generate distractors using Sense2vec

Term	Sense 	Year 
natural language processing 	auto 	2015 
machine learning >		90%
computer vision >		86%
deep learning >		86%
data analysis >		84%
neural nets >		83%
relational databases >		82%

Source: <https://explosion.ai/demos/sense2vec>

<https://explosion.ai/blog/sense2vec-reloaded>

- Contextual information is captured. Trained on Reddit comments.
- The words with the same senses are differentiated with parts of speech eg: duck | VERB and duck | NOUN
- Noun phrases and named entities are annotated during training so multiword phrases like "natural language processing" also have an entry as opposed to some word vector algorithms which are trained with only single words.
- We will use 2015 trained vectors as opposed to 2019 as the results were slightly better in my experimentation

Output we got for distractors after using wordnet

```
original word: Lion
```

```
['Cheetah', 'Jaguar', 'Leopard', 'Liger', 'Saber-toothed Tiger', 'Snow Leopard', 'Tiger', 'Tiglon']
```

```
original word: Green
```

```
['Blond', 'Blue', 'Brown', 'Complementary Color', 'Olive', 'Orange', 'Pastel', 'Pink', 'Purple', 'Red', 'Salmon', 'Yellow']
```

```
time: 31.7 ms (started: 2022-04-25 06:54:29 +00:00)
```

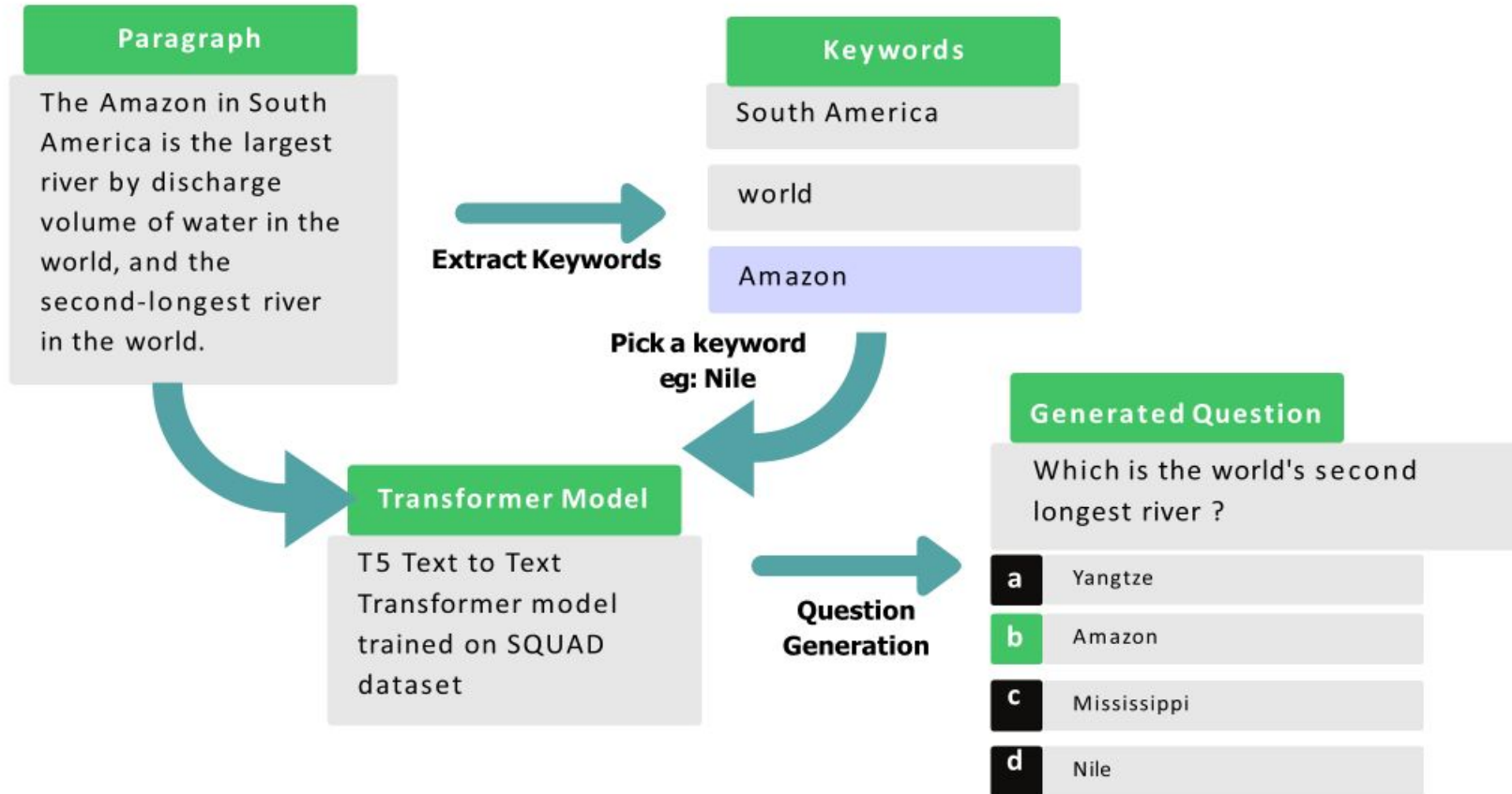
Output we got for distractors after using Sense2vec

```
word donald_trump
```

```
Best sense Donald_Trump|PERSON
```

```
[('Sarah_Palin|PERSON', 0.8547), ('Mitt_Romney|PERSON', 0.8246), ('Barrack_Obama|PERSON', 0.8082), ('Bill_Clinton|PERSON', 0.8046), ('Oprah|GPE', 0.8042), ('Paris_Hilton|ORG', 0.7963),  
time: 1.07 s (started: 2022-04-25 06:55:08 +00:00)]
```

Overview of MCQ generation from any content



<https://datasciencemilan.medium.com/overview>

What we have done so far :

We have Implemented the following:

Step 1 : Identify key Sentences/Concepts

Step 2 : Identify keywords from sentences

Step 3 : Generate Questions using T5 transformer

Step 4 : Generate options for MCQs

Posing it as an NLP problem

Abstractive/Extraction summarization (T5 Transformer)



Paraphrasing of sentences (T5)



Keyword/Keyphrase Extraction (YAKE, TopicRank, KeyBERT etc)



Question Generation (T5)



Wrong choices/distractors generation(wordnet,, Word2vec, Sense2vec etc)

Future work :

- **As of now we have an overview of how all the submodules are working.**
- **In the final review we will be working on Integrating them and getting a proper Question with the 4 best possible options for it.**
- **Expanding the domain like T/F or fill in the blanks or match the following.**
- **Get better accuracy on Question formation and generate more relevant options**

Literature Review :

AUTHOR AND YEAR	TITLE	ABSTRACT	METHODOLOGY	CONCLUSION
Bidyut Das, ·Mukta Majumder,Santanu Phadikar, Arif Ahmed Sekh, Springer 2021	Multiple-choice question generation with auto-generated distractors for computer-assisted educational assessment	<ul style="list-style-type: none">● MCQs are used as instrumental tool for assessment, not only in various competitive examinations ,active learning, etc.● Therefore, automatic generation of multiple-choice test items from text-based learning material is a truly demanding task in computer aided-assessment	<ul style="list-style-type: none">● Keywords identification RAKE (Rapid Automatic Keyword Extraction)● Stem generation and answer key identification Top-ranked sentences are selected as informative sentences to generate MCQ stems.● Distractors K-means clustering algorithm K-means clustering algorithm	First,they extracted the topic-words from the corpus. Then, identified the simple sentences and ranked them based on the topic-words.D istractors are generated using k means clustering algorithm

AUTHOR AND YEAR	TITLE	ABSTRACT	METHODOLOGY	CONCLUSION
<p>Dhawaleswar Rao Ch and Sujan Kumar Saha, IEEE 2018</p>	<p>Automatic Multiple Choice Question Generation from Text : A Survey</p>	<p>MCQs are widely accepted for large-scale assessment in various domains and applications. However, manual generation of MCQs is expensive and time-consuming. So,, researchers were attracted towards automatic MCQ generation since the late 90's. Since then, many systems have been developed for MCQ generation</p>	<ul style="list-style-type: none"> ● KEY SELECTION Frequency Count, Parts-of-speech & Parse info, Semantic Info, Pattern matching ●QUESTION FORMATION Syntactic transformation, Discourse connectives ,Semantic information based ●DISTRACTOR SELECTION Frequency , WordNet , Pattern matching, Semantic analysis 	<p>Established a generic workflow consisting of six broadly classified dependent phases, namely, pre-processing, sentence selection, key selection, question formation, distractor generation, and post-processing. Various techniques have been employed for implementation of the individual phase.</p>

AUTHOR AND YEAR	TITLE	ABSTRACT	METHODOLOGY	CONCLUSION
<p>Indrashis Das, Bharat Sharma, Siddharth S. Rautaray, Manjusha Pandey, IEEE 2019</p>	<p>An Examination System Automation Using Natural Language Processing.</p>	<p>There are lot many examination portals that are deployed over several servers which are used to conduct online examination for various purposes. Aim is not to work on the technology that is already existing, rather some technology that is very rare. Here we talk of the descriptive online examination system.</p>	<ul style="list-style-type: none"> ● KEYWORDS Sparse matrix TF-IDF. ● QUESTION FORMATION Syntactic transformation, Discourse connectives ,Semantic information based By appropriate Wh-word selection , Subject-Verb-Object and their relationship , Knowledge in sentence , Dependency based patterns . 	<p>The algorithm is also efficient as the faculty conducting the test can himself set the question along with the desired keywords he is expecting in one's answer.It is easy to correct the answer scripts and the student will know were he did the mistake</p>

Project Modules :

- Finding Data Set :

- We will be creating the dataset by scrapping from different platforms.
- Like wikipedia or any other info providing website

- Data Cleaning / Transformation:

- After getting the raw data we will later transform this data into useful format
- By using preprocessing techniques and NLP techniques.

Project Modules :

- Model Development/Evaluation using BERT/T5 and other DL techniques:

- In this module, We extract keywords and
- Also we use Wordnet(to generate distractors and Word Sense Disambiguation)
- T5 transformer for question formation and some machine learning algorithms to train the model.

- Testing and Deployment:

- In this module we test our model's results and
- Try to improvise the model to get better accuracy on real world data.

Tools / Techniques / Libraries :

Data collection :

- Stanford Q/A data set or any other raw data.

Link to SQuAD : <https://arxiv.org/pdf/1606.05250.pdf>

Data cleaning and processing using NLP :

- Natural Language Toolkit (NLTK) and PorterStemmer for Stemming.
- Stopwords, Text Categorization, Text Summarization.
- Word Co-occurrence Matrix, Sentence Mapping, Wordnet, Generation distractors.
- Numpy, Pandas, Re ,Spacy, AllenNLP, HuggingFace/BERT transformers, etc.

Tools / Techniques / Libraries :

Model Development:

- BERT/T5 transformers
- Tensorflow/Keras/Scikit-learn

Application Development:

- Dockerize T5 model with FastAPI
- and create a local API

Model Deployment:

- **Google cloud run**



Supporting articles/papers/open source project for the MCQ Generator System using NLP and Deep Learning.

- ❑ [Teaching machines to read](#)
- ❑ [Natural Language Understanding](#)
- ❑ [Natural Language Processing \(almost\) from Scratch](#)
- ❑ [TensorFlow API Documentation](#)
- ❑ [Keras API reference](#)

Supporting articles/papers/open source project for the MCQ Generator System using NLP and Deep Learning.

- [Das, B., Majumder, M., Phadikar, S., & Sekh, A. A. \(2021\). Multiple-choice question generation with auto-generated distractors for computer-assisted educational assessment. Multimedia Tools and Applications, 80\(21\), 31907-31925.](#)
- [Ch, D. R., & Saha, S. K. \(2018\). Automatic multiple choice question generation from text: A survey. IEEE Transactions on Learning Technologies, 13\(1\), 14-25.](#)
- [Das, I., Sharma, B., Rautaray, S. S., & Pandey, M. \(2019, July\). An examination system automation using natural language processing. In 2019 International Conference on Communication and Electronics Systems \(ICCES\) \(pp. 1064-1069\). IEEE.](#)
- [Aldabe, I., & Maritxalar, M. \(2014\). Semantic similarity measures for the generation of science tests in basque. IEEE transactions on Learning Technologies, 7\(4\), 375-387.](#)

THANK YOU