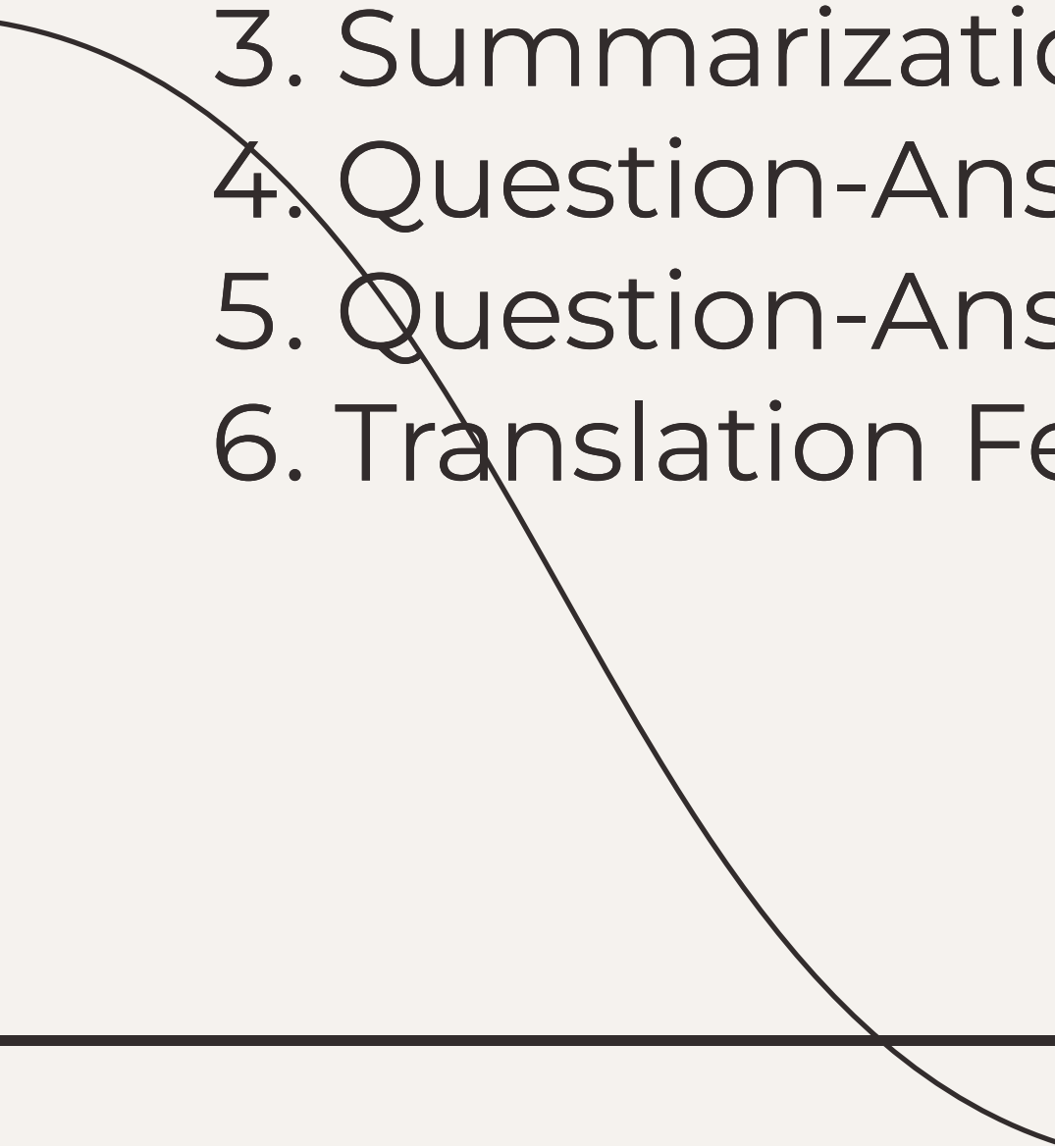# Revolutionizing Text Processing: Advanced OCR with Summarization, QA, and Translation Capabilities

table of contents

1. previous work
2. what is OCR
3. Summarization Techniques
4. Question-Answering In English
5. Question-Answering In Arabic
6. Translation Features

previous work

```python
pipe_ar = pipeline("question-answering",model='ZeyadAhmed/AraElectra-Arabic-SQuADv2-QA')

pipe_en = pipeline("question-answering",model='deepset/roberta-base-squad2')
```

previous work

previous work

What is OCR?

Optical Character Recognition (OCR) is the technology that enables the conversion of different types of documents, such as scanned paper documents and PDFs, into editable and searchable data. It plays a crucial role in digitizing information.

Summarization Techniques

```python
# summarization model in english
summarizer = pipeline("summarization", model="facebook/bart-large-cnn")


def summarize(text):
    summary = summarizer(text, max_length=200, min_length=30, do_sample=False)
    return summary[0]['summary_text']
```

Question-Answering In English


```python
model_name = "deepset/roberta-base-squad2"

pipe_qa_en = pipeline('question-answering', model=model_name, tokenizer=model_name)

model = AutoModelForQuestionAnswering.from_pretrained(model_name)
tokenizer = AutoTokenizer.from_pretrained(model_name)


def question_answering(question,context):
    QA_input = {
        'question': question,
        'context': context,
    }
    res = pipe_qa_en(QA_input)
    return res['answer']
```

Question-Answering In Arabic

```python
from transformers import pipeline
# question-answering model in arabic
pipe_qa_ar = pipeline("question-answering", model="gp-tar4/QA_FineTuned")

def question_answering_ar(question,context):
    QA_input = {
        'question': question,
        'context': context,
    }
    res = pipe_qa_ar(QA_input)
    return res['answer']
```

Translation Features

```python
# translation english to arabic
pipe_to_arabic = pipeline("translation", model="Helsinki-NLP/opus-mt-en-ar")

def to_arabic(text, max_length=512):
    # Split the text into chunks if it is more than 512
    chunks = [text[i:i+max_length] for i in range(0, len(text), max_length)]

    translated_chunks = []
    for chunk in chunks:
        translated_chunk = pipe_to_arabic(chunk, max_length=max_length)[0]['translation_text']
        translated_chunks.append(translated_chunk)

    # Join the translated chunks
    return ' '.join(translated_chunks)
```

example of summrizaton

example of question_answering

Thanks!
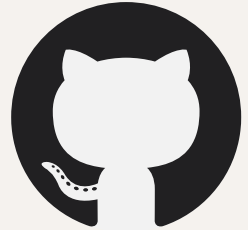
https://github.com/khaled0625/ocr_summarization_and_question_answering-AI/tree/main

https://huggingface.co/spaces/khaled06/ocr_summarization_and_question_answering