

Que Extractor Using Python

```
import PyPDF2
import re

def extract_questions_from_pdf(pdf_files):
    # Initialize an empty set to store the questions
    questions = set()

    for pdf_file in pdf_files:
        # Open the PDF file
        with open(pdf_file, "rb") as file:
            # Create a PDF file reader object
            reader = PyPDF2.PdfReader(file)

            # Loop through each page in the PDF
            for page_num in range(len(reader.pages)):
                # Extract the text from the page
                text = reader.pages[page_num].extract_text()

                # Use a regular expression to find questions
                # This assumes that all questions start with
                # indicating the marks for the question
                found_questions = re.findall(
                    r"Q\.\d\s\(\w\).*?(?=\nQ\.\d|\nOR|\Z)", text, re.DOTALL
                )

                # Add the found questions to the set of questions
                questions.update(found_questions)

            # Write the unique questions to a text file
            with open("questions2.txt", "w") as file:
                for question in questions:
                    file.write(question + "\n")

# List of PDF files to extract questions from
pdf_files = [
    "D:\python\paperExtractor\CN-BE-SUMMER-2018.pdf",
    "D:\python\paperExtractor\CN-BE-SUMMER-2019.pdf",
    "D:\python\paperExtractor\CN-BE-WINTER-2018.pdf",
    "D:\python\paperExtractor\CN-BE-WINTER-2019.pdf",
]

# Extract questions from the PDF files
extract_questions_from_pdf(pdf_files)
```

- paperExtractor
 - CN-BE-SUMMER-2018.pdf
 - CN-BE-SUMMER-2019.pdf
 - CN-BE-WINTER-2018.pdf
 - CN-BE-WINTER-2019.pdf
 - paperExtractor.py
 - pds-BE-SUMMER-2021.pdf
 - pds-BE-SUMMER-2022.pdf
 - pds-BE-WINTER-2021.pdf
 - pds-BE-WINTER-2022.pdf
- timepass
- translator_apk
 - questions CN.txt
 - questions PDS.txt

file is created as name and path specified

Enter your own file here, that accept as many pdfs file as you want
2,3,4,5....,