KEYWORD EXTRACTOR OR COUNTER OF A PDF FILE

SOURCE CODE:

```
import fitz  # PyMuPDF for reading PDFs
import re
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
# Download necessary NLTK resources
nltk.download('punkt')
nltk.download('stopwords')
# Function to extract text from a PDF file
def extract_text_from_pdf(pdf_path):
    doc = fitz.open(pdf_path)
    text = " ".join(page.get_text() for page in doc)  # Extract text from all pages
    return text
# Function to check if a given keyword exists in the text
def check_keyword_in_pdf(text, keyword):
    words = word_tokenize(text.lower())  # Convert text to lowercase and tokenize words
    words = [re.sub(r'\W+', '', word) for word in words if word.isalpha()]  # Remove special characters
    keyword = keyword.lower()  # Convert keyword to lowercase for case-insensitive matching
    count = words.count(keyword)  # Count occurrences of the keyword
    return count
# Path to your PDF file
pdf_path = "D:/Saketh/Python Programming/sampleprograms/NLP/NLPLabPrograms/sample.pdf"
# Extract text from PDF
pdf_text = extract_text_from_pdf(pdf_path)
# Get user input for the keyword
keyword = input("Enter the keyword to search in PDF: ")
# Check if the keyword exists in the PDF
occurrences = check_keyword_in_pdf(pdf_text, keyword)
# Display results
if occurrences > 0:
    print(f"\nThe keyword '{keyword}' was found {occurrences} times in the PDF.")
else:
    print(f"\nThe keyword '{keyword}' was NOT found in the PDF.")
```

SAMPLE OUTPUT:

```
Enter the keyword to search in PDF: machine
The keyword 'machine' was found 5 times in the PDF.
Enter the keyword to search in PDF: deep
The keyword 'deep' was NOT found in the PDF.
```

DEPENDENCIES TO BE INSTALLED:

```
pip install fitz
pip install pymupdf
pip install nltk
```