Lesson 09: Benchmarking and Evaluating LLM Capabilities

Overview:

In this activity, you will create a PDF summarizer and benchmark the summary using ROUGE. This guided practice aims to reinforce your skills and familiarity with these technologies, ensuring a more comprehensive understanding of future practical applications in the respective field.

Instructions:

1. Read the tasks carefully.
2. Generate the summary for any research paper (only PDF).
3. Benchmark the generated summary using ROUGE.
4. Check your answers with the key answers provided at the end of the activity.

Tasks:

Task 1: Generate the summary for any research paper (only PDF):

1. Choose any research paper of your choice.
2. Create a text summarizer application to read the PDF in one shot and generate a summary.

Task 2: Benchmark the generated summary using ROUGE

1. Read the PDF and summary of that file.
2. Compute ROUGE scores for the summary.

Discussion Questions (Optional)

If time permits, discuss the below question:

1. How do you think the summarization process might differ when applied to different types of documents or content, and what considerations should be considered?
2. What insights did you gain from benchmarking the summary using ROUGE, and how do these metrics influence the evaluation of summarization models?

Key Answers

Task 1:

1. Use **Lesson 11 Demo 01: Text Summarizer.ipynb** file.
2. Change the **PAPER_PATH** value to your PDF file name in **Step 2: Download and Read the PDF.**

```
# Define the path of the paper
PAPER_PATH = "Your_PDF_File_Name.pdf"

# Read the PDF
reader = PdfReader(PAPER_PATH)

# Print the number of pages in the PDF
print(f"Number of pages: {len(reader.pages)}")
```

3. Keep all other steps unchanged.

Task 2:

1. Use **Lesson 12 Demo 01: ROUGE Benchmark.ipynb** file.
2. Change the **pdf_file** and **pdf_reader** values to your PDF file name in **Step2: Read the File.**

```
# Read the PDF file from a URL
pdf_file = open(Your_PDF_File_Name.pdf', 'rb')

# Extract the text from the PDF file
pdf_reader = PyPDF2.PdfReader(Your_PDF_File_Name.pdf')
num_pages = len(pdf_reader.pages)
document_text = ""
for page in range(num_pages):
    document_text += pdf_reader.pages[page].extract_text()
```

3. Keep all other steps unchanged.