

## BACK-END ASSIGNMENT

### Submission Deadline

Please submit your completed answer **within 5 days** of receiving this assignment and send to the HR representative (email: [hr@thinkprompt.com](mailto:hr@thinkprompt.com))

### File Manipulation Task

1. Write a script to extract all text and images from the provided PDF/DOCX. Ensure that all images are saved to the disk.  
\* Pay attention to how the files/folders are named during this process.
2. For each paragraph in the PDF/DOCX, extract the following details:
  - Text content
  - Font type
  - Font size
  - Styling elements (such as italics, bold, etc.)
  - Text color
3. Convert the text of each extracted paragraph to **UPPERCASE**. Subsequently, compile all the **UPPERCASE** paragraphs into a new PDF/DOCX, maintaining the original formatting (font type and styling) as closely as possible.
4. Write a script to extract all text and images from the provided PPTX and then **translate** all the text in file to English and then **append** the translated text under the original text back in slides, please try to keep the font size as reasonable as possible.

#### Note:

- If possible, use container for deployment
- Submit code via Git (Gitlab or Github is highly recommended)

### Reference

[1] A high performance Python library for data extraction, analysis, conversion & manipulation of PDF (and other) documents. <https://pypi.org/project/PyMuPDF/>

[2] A flexible free and unlimited python tool to translate between different languages in a simple way using multiple translators. <https://pypi.org/project/deep-translator/>