Directory Structure Overview:

1. Input

   This folder contains input data. Any document or image you want to be analyzed should go into this folder.

   Let's say for example you want to analyze a contract titled 'Lease Agreement.pdf' then you can place this document inside the input folder.

   Limitations: For text-based documents only pdf files are supported. For images png, jpeg, and jpg are supported.

2. Code

   This folder contains the relevant code. There are three files here described below:

   a. Analyze_expense.py:

      This code file is supposed to be run when dealing with image-based documents such as invoices and receipts. It is responsible for extracting data from the invoice or receipt and storing it in a JSON file.

   b. Analyze_text.py:

      This code file is supposed to be run when dealing with text-based documents such as contracts and NDAs. It is responsible for extracting data from the contract or NDA and storing it in a txt file.

   c. Model.py

      This code file is supposed to be run when you have extracted the relevant data and want to ask questions in Natural Language about that particular data.

3. Output

   This folder contains output data. Any document or image that we extracted data from using the scripts mentioned in 2a (analyze_expense.py) and 2b (analyze_text.py) should be present here, either in JSON format (for images such as png and jpeg) or txt format (for text documents such as PDF).

Changes for New Documents:

As mentioned earlier our script can handle 2 types of documents. Let's discuss how to handle both of these below:

1. Image-based documents (Invoices and Receipts):

In order to process image-based documents, the following sequence of steps needs to be performed:

a. Place the file in the input folder.

   Make sure the file has a png or jpeg extension.

b. Go inside the code folder and open the file analyze_expense.py
   i. On line 91 change the value of the invoice_file_path to the relevant file.
      For example, in the above step if you placed the file abc.png inside the input folder, then in this step you need to set the invoice_file_path to '..\input_data\abc.png'

   ii. Then run the file using this command:
      Python analyze_expense.py

c. Now you would be able to see a new file added in the output folder which contains the extracted data from the file we provided in Step A.

   For example, if we provided the file abc.png in Step A, then in the output folder we would be able to see abc.json after performing Step B above.

d. Now go inside the code folder and open file models.py
   i. On line 28 change the value of the context to the relevant file.
      For example, continuing the example of abc.png, in this step you need to set context to context = get_context("abc.json")

   ii. On line 32 change the value of the question to whatever question you want to ask about the data.
      For example, set question to "What is the most expensive item in this receipt?"

   iii. Then run the file using this command:
      Python model.py

2. Text-based documents (Contracts and NDAs):

In order to process text-based documents, the following sequence of steps needs to be performed:

e. Place the file in the input folder.

   Make sure the file has a pdf extension.

f. Go inside the code folder and open the file extract_text.py
   i. On line 3 change the value of the file to the relevant file.
      For example, in the above step if you placed the file abc.pdf inside the input folder, then in this step you need to set the file to '..\input_data\abc.pdf

   ii. Then run the file using this command:
      Python extract_text.py

g. Now you would be able to see a new file added in the output folder which contains the extracted data from the file we provided in Step A.

   For example, if we provided the file abc.pdf in Step A, then in the output folder we would be able to see abc.txt after performing Step B above.

h. Now go inside the code folder and open file models.py
   i. On line 28 change the value of the context to the relevant file.
      For example, continuing the example of abc.pdf, in this step you need to set context to context = get_context("abc.txt")

   ii. On line 32 change the value of the question to whatever question you want to ask about the data.
      For example, set question to “What type of contract is this?”

   iii. Then run the file using this command:
      Python model.py