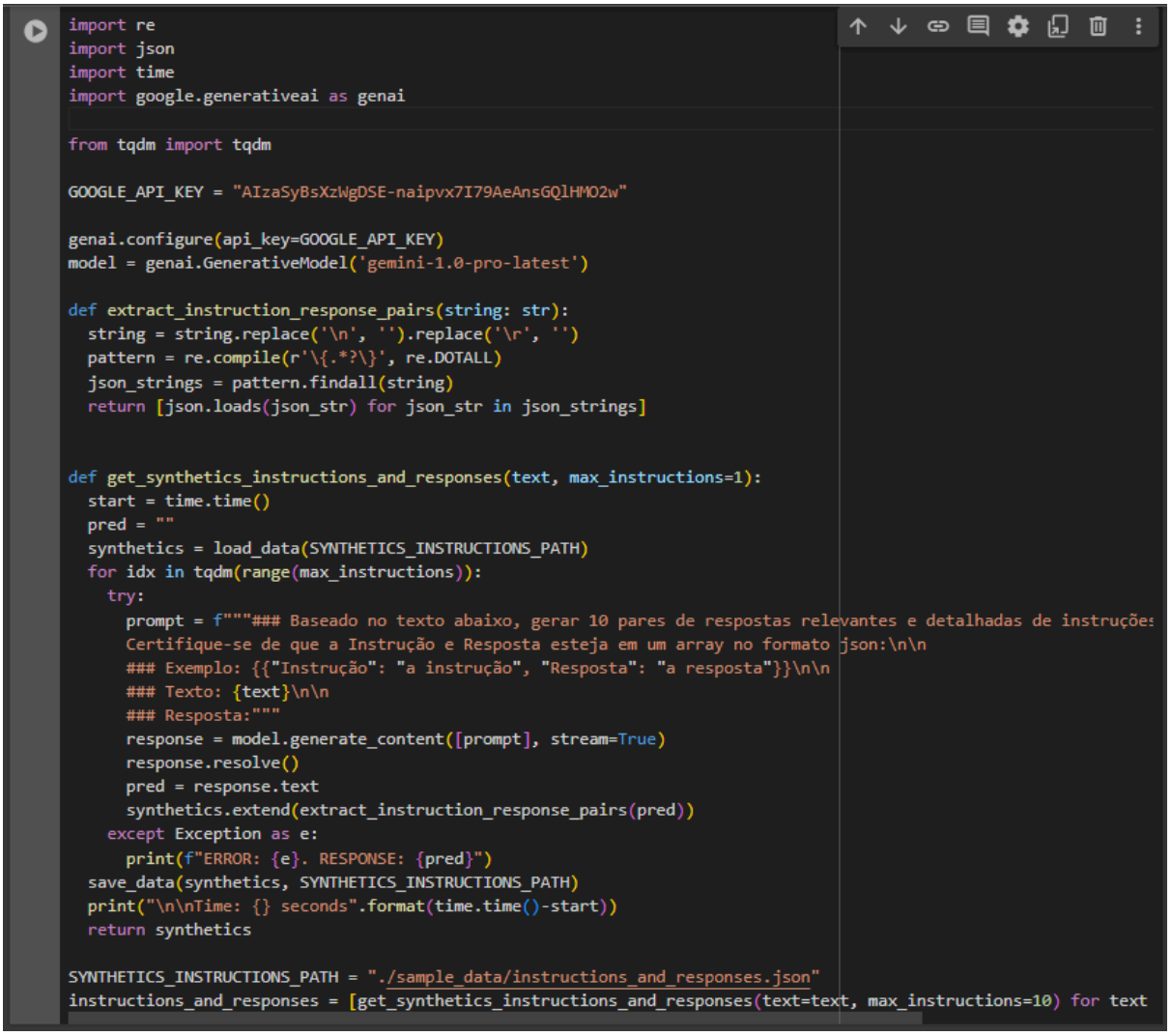
2. Base de Dados Sintética

2.1. Arquivo contendo os 1000 exemplos de perguntas e respostas gerados.

- A base de dados sintética com os 1000 exemplos está no arquivo “sample_data/instructions_and_responses.json”

2.2. Metodologia utilizada para a geração dos exemplos.

- Para gerar a base de dados foi utilizado o modelo Gemini do Google, utilizando um prompt de comando passando o texto extraído do PDF e pedido para o modelo criar os exemplos de perguntas e respostas, conforme o código abaixo:



```python
import re
import json
import time
import google.generativeai as genai

from tqdm import tqdm

GOOGLE_API_KEY = "AIzaSyBsXzWgDSE-naipvx7I79AeAnsGQlHMO2w"

genai.configure(api_key=GOOGLE_API_KEY)
model = genai.GenerativeModel('gemini-1.0-pro-latest')

def extract_instruction_response_pairs(string: str):
  string = string.replace('\n', '').replace('\r', '')
  pattern = re.compile(r'\{.*?\}', re.DOTALL)
  json_strings = pattern.findall(string)
  return [json.loads(json_str) for json_str in json_strings]


def get_synthetics_instructions_and_responses(text, max_instructions=1):
  start = time.time()
  pred = ""
  synthetics = load_data(SYNTHETICS_INSTRUCTIONS_PATH)
  for idx in tqdm(range(max_instructions)):
    try:
      prompt = f"""### Baseado no texto abaixo, gerar 10 pares de respostas relevantes e detalhadas de instruções
      Certifique-se de que a Instrução e Resposta esteja em um array no formato json:\n\n
      ### Exemplo: {{"Instrução": "a instrução", "Resposta": "a resposta"}}\n\n
      ### Texto: {text}\n\n
      ### Resposta:"""
      response = model.generate_content([prompt], stream=True)
      response.resolve()
      pred = response.text
      synthetics.extend(extract_instruction_response_pairs(pred))
    except Exception as e:
      print(f"ERROR: {e}. RESPONSE: {pred}")
  save_data(synthetics, SYNTHETICS_INSTRUCTIONS_PATH)
  print("\n\nTime: {} seconds".format(time.time()-start))
  return synthetics

SYNTHETICS_INSTRUCTIONS_PATH = "./sample_data/instructions_and_responses.json"
instructions_and_responses = [get_synthetics_instructions_and_responses(text=text, max_instructions=10) for text
```