Orientações:

- O código será rodado com um número grande de arquivos pdf, portanto é importante que esteja otimizado.
- Ao rodar uma vez o código, ele deve gerar os 3 JSONs definidos em baixo para otimizar.

1. Extrair o Executive Summary e salvar em um JSON em que a chave é o nome do arquivo pdf.
    - O texto a ser extraído vem imediatamente depois de "EXECUTIVE SUMMARY" e deve parar antes do proximo título em negrito e capitalizado (exemplo. APPENDIX).
    - No caso de tabelas dentro do executive summary, elas não devem ser incluídas.
    - Não é necessário manter as tags para formatação exata do pdf, a string do texto é suficiente.
    - Em alguns casos o executive summary se extende por mais de uma pagina e não devem ser incluidos elementos de formatação da página, como por exemplo o cabeçalho e o rodapé.

2. Criar um JSON em que a chave é o nome do arquivo pdf e o conteúdo indica as palavras chaves presentes (da lista) e frequência.

    Exemplo:
    key_words = { pdf_name1 = {key_word1 = 3, key_word2 = 5},
    pdf_name2 =....
    }

    - Objetivo: Relacionar pdf a temas/riscos
    - O arquivo com as palavras chaves deve permanecer como um arquivo editável (excel) para que possa ser modificado pelo usuário.
    - Para cada pdf, só incluir as palavras chaves com frequencia maior do que 0.
    - As palavras chaves devem bater com o texto independente da capitalização, espaço ou outro caractér especial (ex. covid-19 deve bater com COVID e com Covid 19).

3. Criar um JSON em que a chave é o nome do arqruivo pdf e o conteúdo indica palavras chaves novas e frequencia. Top 3 de maior frequencia por pdf.
    - Objetivo: indentificar novos temas/riscos
    - Palavras chaves novas são os substantivos significativos que mais aparecem no pdf que não estão contidos na lista de palavras chave e não estão na lista de não significativos.

- Criar também um arquivo editável (excel) para o usuário poder colocar palavras que ele está observando nos resultados, mas que não são significativos. O objetivo disso é que quando uma palavra é adicionada na lista, na próxima vez que o código for rodado, essa palavra seja excluida, aumentando a precisão.