Project Title: "DataMining for Sustainability Analysis in Brazilian Port Reports"

Description:
This data science project involves the analysis of sustainability reports from Brazilian ports using data mining techniques. The primary objective is to count the occurrences of certain critical phrases and words related to sustainability in these reports.

The reports, stored as PDF files, contain vital information about the ports' commitment to various global sustainability initiatives. The key phrases and words being tracked include but are not limited to 'sustainable development goals', 'SDGs', 'global pact', 'green port initiative', 'World Ports Sustainability Program', 'Global Reporting Initiative', 'GRI', 'Sustainability Accounting Standards Board', 'ECOPORTS', and various ISO standards.

The Python script leverages the PyMuPDF and regular expression (re) libraries for text extraction and analysis. PyMuPDF is used to open the PDF files and extract the text content. The extracted text is split into paragraphs to facilitate granular analysis. For each tracked word or phrase, the script creates a case-insensitive regular expression. This regular expression is used to search through each paragraph of text. When a match is found, the count for the paragraph and word is increased by one. Additionally, the script also prints the content of the paragraph where the match was found, providing useful context.

At the end of the analysis, the script prints the total number of times each word or phrase appears in the entire report. This output serves as a quantitative measure of the frequency of these important sustainability terms, providing valuable insights into the content and focus of the port's sustainability reports.

This project serves as a practical example of using data mining techniques to analyze text data in a real-world context, and demonstrates how Python can be used to extract, process, and analyze data from PDF files.