24th January, 2023
Task | # 01

PYTHON

Text Document Keyword Analyzer

Duration: 1 Day | Marks: 10

Scenario

You are working on a project where you need to develop a Python program to analyze a collection of text documents and extract relevant keywords. The program should read multiple text files, process the text data, and generate a list of the most frequently occurring words. Additionally, it should calculate the frequency of each word and display the top N keywords based on their occurrence count.

Question

You have been assigned the task of developing a Python program to analyze a collection of text documents and extract keywords. The program should take a folder path as input and process all the text files within that folder. Your program should extract the most frequently occurring words and display the top 10 keywords along with their frequency count.

Task Description

- Develop a Python program that accepts a folder path as input.
- Read all the text files within the specified folder.
- Process the text data to remove punctuation and convert all text to lowercase.
- Tokenize the text into individual words.
- Count the frequency of each word.
- Display the top 10 keywords along with their frequency count.

Requirements:

- The program should be able to handle multiple text files within the specified folder.
- Text processing should include removing punctuation and converting text to lowercase.
- The program should tokenize the text into words and count the frequency of each word.
- The output should display the top 10 keywords along with their frequency count.

Condition:

- The program should handle different file formats, such as .txt or .csv, within the specified folder.
- The keyword extraction should be case-insensitive, meaning "hello" and "Hello" should be considered the same word.
- Stop words (common words like "the," "and," "is," etc.) should be excluded from the keyword extraction process.

Explanation

You are tasked with developing a Python program that analyzes a collection of text documents and extracts keywords. The program should take a folder path as input, process the text files within that folder, and display the top 10 keywords along with their frequency count. It should handle different file formats, remove punctuation, convert text to lowercase, and exclude common stop words. The program's functionality can be completed within one day, making it suitable for an intermediate-level Python project.

NOTE: - The application can be implemented as either a GUI or console-based program, depending on your preference and specific requirements.