CMPE230 Systems Programming Homework 2

This project is about developing a program that will traverse the directories and search the files and directories that are duplicates of each other based on their contents (not names).

It will select the duplicates for which an action is taken. They can be filtered with their types (file or directory) and whether their names match the given pattern.

The action will be given as argument or be print as default. Also, directories to traverse can be given as arguments (default directory is the one which the program is running).

For solution, I decided to take on a simple algorithm. Steps are the followings:

- Initialize the required global variables (detailed explanations of them is below).
- Parse the program arguments using argparse module.
- Check the results of argument parsing and modify the global variables according to the results.
- Traverse all directories with DFS.
  - Calculate hashes of files from their contents.
  - Calculate hashes of directories from sorted and concatenated hashes of their files and directories.
  - Add all these hashes to a dictionary.
- Process duplicates.
  - Find duplicates from dictionary.
  - Filter them with their types and names.
  - Apply the command in program arguments or print them.

Global Variables and Their Functionalities

- all_hashes (dict) : stores all hash to list (files and directories) mappings for checking duplicates
- check_pattern (boolean) : stores whether file names or directory names will be matched to a pattern
- current_dir (str) : stores the directory where program is running
- dirs (list) : list of directories to traverse while searching for duplicates

Methods and Their Functionalities

- def dfs(pathname)
  - Recursive method for traversing all contents of the directory.
  - param pathname (str) : directory to traverse
  - returns (str) : the hash of sorted and concatenated hashes of contents