Ain Shams University
Faculty of Computer and Information Sciences
Data Strucutres Project

Title	**Duplicate file finder**
Description	People usually make many copies of the same file for backup purposes. Some times they will give different names for these copies. It is useful in many occasions to find and report such duplicates. A frequent example is when copying a folder to some external memory (e.g., usb stick), then changing some files. One would like to copy back the folder to PC, without duplicating files. There is a need for a utility to check what is duplicate and what is different. The files can be of any type: office documents, images, binary files, etc. There are many duplicate file checkers on the Internet, ask Google. Still, it is a good programming exercise to write your own duplicate file finder. A duplicate file finder can search for files with the same name, the same size, the same check-sum, and/or the exact same contents. A naive program would compare every pair of files, which is $O(n^2)$. One should utilize an index to avoid such a high complexity. The index type depends on the way of comparison. A B-Tree is suitable if you are comparing the check-sum, for example. **The goal of this project is to develop a duplicate file finder. The program should scan a given directory, search the directory, and all sub directories for duplicate files, list the full path of the duplicate files, and give the user the option to delete duplicates. The program should do the comparison based on the check-sum. You should use B-tree for indexing. You are free to choose the programming language. You are expected to write the check-sum algorithm and the B-tree code yourself.**
Group size	3 – 5 members.
Deliverables	1- A program doing the specified task. 2- A short (1 page) user manual.
Bonus extensions	- Extend your program to find duplicate photos, even if they have different scales (i.e., stretched).
Mentor	TA. Mourad Aly
Notes	This project requires good programming and research skills. Unless you are an A-team, it might be risky to apply for this idea.