Since there are a large number of files, and each of them can be very lengthy, comparing line by line is NOT a very good move. One of the ways to do this is precompute a hash for each file and store them in a hash table, we can use SHA256 for instance, then files that are the same, would have the same hash. Also, the hash table can be optimized to have the worst-case average seek time in O(lgn) time instead of O(n) (precompute each hash then sort them and store them in an array and use binary search).