Interview Cake

← course home (/table-of-contents#section_system-design_question_find-duplicate-files)

You left your computer unlocked and your friend decided to troll you by copying a lot of your files to random spots all over your file system.

Even worse, she saved the duplicate files with random, embarrassing names ("this_is_like_a_digital_wedgie.txt" was clever, I'll give her that).

Write a function that returns a list of all the duplicate files. We'll check them by hand before actually deleting them, since programmatically deleting files is really scary. To help us confirm that two files are actually duplicates, return a list of tuples where:

- the **first** item is the **duplicate** file
- the **second** item is the **original** file

For example:

```
[('/tmp/parker_is_dumb.mpg', '/home/parker/secret_puppy_dance.mpg'),
 ('/home/trololol.mov', '/etc/apache2/httpd.conf')]
```

You can assume each file was only duplicated once.

Gotchas

Are you correctly handling child folders as well as sibling folders? Be careful that you're traversing your file tree correctly...

When you find two files that are the same, don't just choose a random one to mark as the "duplicate." Try to figure out which one your friend made!

Does your solution work correctly if it's an empty file system (meaning the root directory is empty)?

Our solution takes $O(n)$ time and space, where n is the *number of files*. Is your solution order of the *total size on disk of all the files*? If so, you can do better!

To get our time and space costs down, we took a small hit on accuracy—we might get a small number of false positives. We're okay with that since we'll double-check before actually deleting files.

Breakdown

> **No idea where to start?** Try writing something that just walks through a file system and prints all the file names. If you're not sure how to do that, look it up! Or just *make it up*. Remember, even if you can't implement *working code*, your interviewer will still want to see you *think through* the problem.

One brute force solution is to loop over all files in the file system, and for each file look at every *other* file to see if it's a duplicate. This means n^2 file comparisons, where n is the number of files. That seems like a lot.

Let's try to save some time. Can we do this in *one* walk through our file system?

Instead of holding onto one file and looking for files that are the same, we can just keep track of *all* the files we've seen so far. What data structure could help us with that?

We'll use a dictionary. When we see a new file, we first check to see if it's in our dictionary. If it's not, we add it. If it is, we have a duplicate!

Once we have two duplicate files, how do we know which one is the original? It's hard to be sure, but try to come up with a reasonable heuristic that will probably work most of the time.

Most file systems store the time a file was last edited as metadata on each file. The more recently edited file will *probably* be the duplicate!

One exception here: lots of processes like to regularly save their state to a file on disk, so that if your computer suddenly crashes the processes can pick up more or less where they left off (this is how Word is able to say "looks like you had unsaved changes last time, want to restore them?").

If your friend duplicated some of *those* files, the most-recently-edited one may *not* be the duplicate. But at the risk of breaking our system (we'll make a backup first, obviously.) we'll run with this "most-recently-edited copy of a file is probably the copy our friend made" heuristic.

So our function will walk through the file system, store files in a dictionary, and identify the more recently edited file as the copied one when it finds a duplicate. Can you implement this in code?

Here's a start. We'll initialize:

1. a **dictionary** to hold the files we've already seen
2. a **stack** (we'll implement ours with a list) to hold directories and files as we go through them
3. a **list** to hold our output tuples

Python 2.7

```
def find_duplicate_files(starting_directory):
    files_seen_already = {}
    stack = [starting_directory]

    # We'll track tuples of (duplicate_file, original_file)
    duplicates = []

    while len(stack) > 0:
        current_path = stack.pop()

    return duplicates
```

(We're going to make our function iterative instead of recursive to avoid stack overflow.)

Here's one solution:

Python 2.7

```
import os

def find_duplicate_files(starting_directory):
    files_seen_already = {}
    stack = [starting_directory]

    # We'll track tuples of (duplicate_file, original_file)
    duplicates = []

    while len(stack) > 0:
        current_path = stack.pop()

        # If it's a directory,
        # put the contents in our stack
        if os.path.isdir(current_path):
            for path in os.listdir(current_path):
                full_path = os.path.join(current_path, path)
                stack.append(full_path)

        # If it's a file
        else:
            # Get its contents
            with open(current_path) as file:
                file_contents = file.read()

            # Get its last edited time
            current_last_edited_time = os.path.getmtime(current_path)

            # If we've seen it before
            if file_contents in files_seen_already:
                existing_last_edited_time, existing_path = files_seen_already[file_contents]
                if current_last_edited_time > existing_last_edited_time:
                    # Current file is the dupe!
                    duplicates.append((current_path, existing_path))
                else:
                    # Old file is the dupe!
                    # So delete it
                    duplicates.append((existing_path, current_path))
                    # But also update files_seen_already to have
                    # the new file's info
```

```
                    files_seen_already[file_contents] = (current_last_edited_time, current_path)


            # If it's a new file, throw it in files_seen_already
            # and record the path and the last edited time,
            # so we can delete it later if it's a dupe
            else:
                files_seen_already[file_contents] = (current_last_edited_time, current_path)

    return duplicates
```

Okay, this'll work! What are our time and space costs?

We're putting the full contents of every file in our dictionary! This costs $O(b)$ time and space, where b is the *total amount of space taken up by all the files on the file system*.

That space cost is pretty unwieldy—we need to store a duplicate copy of our entire filesystem (like, several gigabytes of cat videos alone) in working memory!

Can we trim that space cost down? What if we're okay with losing a bit of accuracy (as in, we do a more "fuzzy" match to see if two files are the same)?

What if instead of making our dictionary keys *the entire file contents*, we hashed those contents first? So we'd store a constant-size "fingerprint" of the file in our dictionary, instead of the whole file itself. This would give us $O(1)$ space per file ($O(n)$ space overall, where n is the number of files)!

That's a huge improvement. But we can take this a step further! While we're making the file matching "fuzzy," can we use a similar idea to save some *time*? Notice that our time cost is still order of the total *size* of our files on disk, while our space cost is order of the *number* of files.

For each file, we have to look at every bit that the file occupies in order to hash it and take a "fingerprint." That's why our time cost is high. Can we fingerprint a file in *constant* time instead?

What if instead of hashing the *whole* contents of each file, we hashed three fixed-size "samples" from each file made of the first x bytes, the middle x bytes, and the last x bytes? This would let us fingerprint a file in constant time!

How big should we make our samples?

When your disk does a read, it grabs contents in constant-size chunks, called "blocks."

How big are the blocks? It depends on the file system. My super-hip Macintosh uses a file system called HFS+, which has a default block size of 4Kb (4,000 bytes) per block.

So we could use just 100 bytes each from the beginning middle and end of our files, but each time we grabbed those bytes, our disk would actually be grabbing 4000 bytes, not just 100 bytes. We'd just be throwing the rest away. We might as well use all of them, since having a bigger picture of the file helps us ensure that the fingerprints are unique. So our samples should be the the size of our file system's block size.

Solution

We walk through our whole file system iteratively. As we go, we take a "fingerprint" of each file in constant time by hashing the first few, middle few, and last few bytes. We store each file's fingerprint in a *dictionary* as we go.

If a given file's fingerprint is already in our dictionary, we assume we have a duplicate. In that case, we assume the file edited most recently is the one created by our friend.

Python 2.7

```
import os
import hashlib

def find_duplicate_files(starting_directory):
    files_seen_already = {}
    stack = [starting_directory]

    # We'll track tuples of (duplicate_file, original_file)
    duplicates = []

    while len(stack) > 0:
        current_path = stack.pop()

        # If it's a directory,
        # put the contents in our stack
        if os.path.isdir(current_path):
            for path in os.listdir(current_path):
                full_path = os.path.join(current_path, path)
                stack.append(full_path)

        # If it's a file
        else:
            # Get its hash
            file_hash = sample_hash_file(current_path)

            # Get its last edited time
            current_last_edited_time = os.path.getmtime(current_path)

            # If we've seen it before
            if file_hash in files_seen_already:
                existing_last_edited_time, existing_path = files_seen_already[file_hash]
                if current_last_edited_time > existing_last_edited_time:
                    # Current file is the dupe!
                    duplicates.append((current_path, existing_path))
                else:
                    # Old file is the dupe!
                    duplicates.append((existing_path, current_path))
                    # But also update files_seen_already to have
                    # the new file's info
                    files_seen_already[file_hash] = (current_last_edited_time, current_path)
```

```
            # If it's a new file, throw it in files_seen_already
            # and record its path and last edited time,
            # so we can tell later if it's a dupe
            else:
                files_seen_already[file_hash] = (current_last_edited_time, current_path)

    return duplicates


def sample_hash_file(path):
    num_bytes_to_read_per_sample = 4000
    total_bytes = os.path.getsize(path)
    hasher = hashlib.sha512()

    with open(path, 'rb') as file:
        # If the file is too short to take 3 samples, hash the entire file
        if total_bytes < num_bytes_to_read_per_sample * 3:
            hasher.update(file.read())
        else:
            num_bytes_between_samples = (
                (total_bytes - num_bytes_to_read_per_sample * 3) / 2
            )

            # Read first, middle, and last bytes
            for offset_multiplier in xrange(3):
                start_of_sample = (
                    offset_multiplier
                    * (num_bytes_to_read_per_sample + num_bytes_between_samples)
                )
                file.seek(start_of_sample)
                sample = file.read(num_bytes_to_read_per_sample)
                hasher.update(sample)

    return hasher.hexdigest()
```

We've made a few assumptions here:

Two *different* files won't have the same fingerprints. It's not impossible that two files with different contents will have the same beginning, middle, and end bytes so they'll have the same fingerprints. Or they may even have different sample bytes but still hash to the same value (this is called a "hash collision"). To mitigate this, we could do a last-minute check whenever we find two "matching" files where we actually scan the full file contents to see if they're the same.

The *most recently edited* file is the duplicate. This seems reasonable, but it *might* be wrong—for example, there might be files which have been edited by daemons (programs that run in the background) *after* our friend finished duplicating them.

Two files with the same contents are the same file. This seems trivially true, but it could cause some problems. For example, we might have empty files in multiple places in our file system that aren't duplicates of each-other.

Given these potential issues, we definitely want a human to confirm before we delete any files. Still, it's much better than combing through our whole file system by hand!

Some ideas for further improvements:

1. If a file wasn't last edited around the time your friend got a hold of your computer, you know it probably wasn't created by your friend. Similarly, if a file wasn't *accessed* (sometimes your filesystem stores the last accessed time for a file as well) around that time, you know it wasn't copied by your friend. You can use these facts to skip some files.
2. Make the file *size* the fingerprint—it should be available cheaply as metadata on the file (so you don't need to walk through the whole file to see how long it is). You'll get lots of false positives, but that's fine if you treat this as a "preprocessing" step. Maybe you *then* take hash-based fingerprints only on the files which which have matching sizes. *Then* you fully compare file contents if they have the same hash.
3. Some file systems also keep track of when a file was *created*. If your filesystem supports this, you could use this as a potentially-stronger heuristic for telling which of two copies of a file is the dupe.
4. When you *do* compare full file contents to ensure two files are the same, no need to read the entire files into memory. Open both files and read them one block at a time. You can short-circuit as soon as you find two blocks that don't match, and you only ever need to store a couple blocks in memory.

Complexity

Each "fingerprint" takes $O(1)$ time and space, so our total time and space costs are $O(n)$ where n is the *number of files* on the file system.

If we add the last-minute check to see if two files with the same fingerprints are *actually* the same files (which we probably should), then in the worst case *all the files are the same* and we have to read their full contents to confirm this, giving us a runtime that's order of the total size of our files on disk.

Bonus

If we wanted to get this code ready for a production system, we might want to make it a bit more modular. Try separating the file traversal code from the duplicate detection code. Try implementing the file traversal with a generator!

What about concurrency? Can we go faster by splitting this procedure into multiple threads? Also, what if a background process edits a file *while our script is running*? Will this cause problems?

What about link files (files that point to other files or folders)? One gotcha here is that a link file can point *back up the file tree*. How do we keep our file traversal from going in circles?

What We Learned

The main insight was to save time and space by "fingerprinting" each file.

This question is a good example of a "messy" interview problem. Instead of one optimal solution, there's a big knot of optimizations and trade-offs. For example, our hashing-based approach wins us a faster runtime, but it can give us false positives.

For messy problems like this, focus on clearly explaining to your interviewer what the trade-offs are for each decision you make. The actual choices you make probably don't matter that much, as long as you show a strong ability to understand and compare your options.

← course home (/table-of-contents)

Next up: Short Circuit Evaluation → (/concept/short-circuit-evaluation?course=fc1&section=general-programming)