The Problem –

We are presented with an array of pages of text. Each page is exactly 24 lines of x 40 characters.

Any page may be missing one or more whole lines of text.

But each page may be repeated multiple times.

But probably different lines will be missing in each example.

There may be several different texts represented within the file, so it is not possible to simply replace missing lines from just any later example, as it that may represent a different text.

Example:

|  |  |  |  |
| --- | --- | --- | --- |
| Blah Blah Foo Bar | Foo Bar  Something else | The quick Brown fox  Something else | George  Peppa  Sally |

We need to end up with an array that says

|  |  |
| --- | --- |
| Blah Blah  Foo Bar  The quick Brown fox  Something else | George  Peppa  Sally |

To complicate matters, there may be (single or multiple bit) data errors in the text, so lines may not compare exactly. Where multiple lines are, however, very similar, the most consistent match (either by line or by character) should be chosen. It is accepted that this may not result in perfection.

|  |  |  |  |
| --- | --- | --- | --- |
| The quicK Brown fox | The quhck Brown fox | The quick Brown fox | The quick brown fow |

Should give

|  |
| --- |
| The quick Brown fox |

To clarify: If a line is either identical or very similar to the same line in another page, it is likely that the pages contain the same text. If a line is very different (excluding by reason of being blank) it is almost certain they are different texts.

Some lines may be common across different texts. These will mostly be within the first 5 and last 2 lines. If comparing pages to ascertain similarity, these lines may be scored lower.

Now, write a PHP function that accepts the first array and returns the cleaned up version.