You are given two files invoice.txt and suppliernames.txt. The first file (invoice.txt) contains the words of an invoice (see invoice.pdf) generated by the OCR engine. Each line has the text of the word and information about the location of the word in the document. For example, the line:

{'pos_id': 1, 'cspan_id': 1, 'rspan_id': 0, 'right': 87.06, 'word': 'INVOICE', 'line_id': 0, 'top': 4.0, 'height': 1.52, 'width': 10.51, 'left': 76.55, 'page_id': 1, 'word_id': 1}

contains the word 'INVOICE', the page it came from (page_id = 1), the line number (line_id = 6), the position of the word in that line (pos_id = 0).

You might not need the other information. The second file, suppliernames.txt, contains a list of supplier names.
The requirement is to find the supplier name of that invoice by matching the given list of supplier names to the invoice.

Although only a limited number of supplier names are given, make your solution scalable to hundreds of thousands of supplier names.

You can use the programming language that you prefer.

It is prefer that you submit your solution through GitHub, with instruction how to run it.

Please do not hesitate to contact us at ammar.mohemmed@xtracta.com if you have any inquiry.

Good Luck