Week 12

Extra Question: [***]

In this question, you need to process a file that contains many documents. Each line of the file contains two columns separated by tabs. The first column is a string, which is the ID of a document. The second column is the content of the corresponding document, shown as a sequence of words, separated by spaces.

You can open "documents.txt" to see an example file of this format.

In q4.py, define a function called `get_document_pair()`. The function takes in the name of a file as its input. The function tries to find out which pair of two documents in the given file shares the largest number of common words. For example, if "D1" and "D2" share 10 words in common, and "D1" and "D3" share 15 words in common, then the pair ("D1", "D3") has more common words than the pair ("D1", "D2").

When counting common words, if a word appears more than once in a document, it should be counted only once. For example, if the word "apple" appears 3 times in "D1" and 4 times in "D2", it should still be counted as 1 common word shared by "D1" and "D2".

The function should return a tuple of three elements: the IDs of the two documents and the number of common words they share.

When you call the function `get_document_pair()` with the argument 'documents.txt' you should get the following output: `('D2', 'D4', 3)`.

Run q4.py to test your code.