Programming Language Learning Series
Mastery of Python Language
(Document Search Engine)

Document retrieval is the task of finding documents that meet the search criteria input by a user. The most well-known example is web search, where a user types in a set of key words and the search engine finds web pages that are relevant to their search query. True document retrieval can be quite difficult, as it needs to take into account many different factors. In this project you will implement a very simple document retrieval engine.

DataSet: The file “documents.txt” contains several old newswire articles and you can download the file from project directory. We will use this as our document collection. Each article in the collection is separated by a line that contains only the “<NEW DOCUMENT>”token.

Implement the following subtasks:

Build Index: In this function, you must read in the documents from the given directory and builds an index of key-value pairs with each word as key. The word’s value is the set of documents that this word appears in. This arrangement allows you to look up a keyword in the dictionary and immediately get all the documents that it appears in, making it easy to figure out documents that might meet a search query.

- *Implementation Hints:* You will likely need to store two types of data in two different data structures. In one, you will have a dictionary that stores single words as the key with the value as the set of documents(numbers) that the word appears in. In the other data structure, you will store the actual text of the documents, so that you can display it for the user when they ask. You can use a list or a dictionary for the second data structure.

Search Documents: In this function, you will get a group of search words as input and return the documents(i.e,document numbers) that contain all of those keywords. If no documents in the collection contain every keyword input by the user, you return an empty group. Search queries should not be case sensitive, i.e. searching for “Stocks” should give all documents that contain ‘stocks’, ‘STOCKS’, etc. You should remove punctuation from the start and end of a word as well. If the string “stock,” appears in the document, this should be counted as an instance of the word “stock”(without the comma).

- *Implementation Hints:* You might find the stringmodule and sets to be useful.

Print/Display Document: In this function, you must print out the entire document that corresponds to the given document number.

Unit Testing: Unit test each function you have written sothat you can be more confident on the project outcome

Main Method: Write the code to integrate subtasks and test it on provided dataset.