ASSIGNMENT 3

50 points

This assignment focuses on the implementation of Python programs to read files and process data using dictionaries and sets.

Assignment Background

What is co-occurrence problem?

You will write a Python program to solve the co-occurrence problem. The co-occurrence problem is stated as follows. We have a file containing English sentences, one sentence per line. Given a list of query words, your program should output the line number of lines that have *all* those words. While there are many ways to do this, the most efficient way is to use sets and dictionaries.
Here is one example. Assume that the following is the content of the file. Line numbers are included for clarity; the actual file doesn't have the line numbers.

1. Try not to become a man of success, but rather try to become a man of value.
2. Look deep into nature, and then you will understand everything better.
3. The true sign of intelligence is not knowledge but imagination.
4. The difference between stupidity and genius is that genius has its limits.

(These are quotes from Albert.txt)

If we are asked to find all the lines that contain this set of words: {"true", "knowledge", "imagination"} the answer will be line 3 because all three words appeared in line 3. If they appear in more than one line, your program should report all of them. For example, co-occurrence of {"the", "is"} will be lines 3 and 4.

Assumptions

1. All the lines in the file are in lowercase
2. File does not contain words with punctuation, such as ",", ".", "!", etc.
3. File does not contain apostrophes and hyphens, e.g."can't" ,"first-born"
4. File contains only alphabetic characters

Implementation

You need to implement the following functions:

1) `open_file()`

The `open_file` function will prompt the user for a file-name, and try to open that file. If the file exists, it will return the file object; otherwise it will re-prompt until it can successfully open the file. This feature should be implemented using a `while` loop, and a `try-except` clause.

2) `read_data(fp)`

This function has one parameter: a file object.This function will read the contents of that file line by line, process them and store them in a dictionary. The dictionary is returned.

You add the words into a `dictionary` with the key being the word and the value is a `set` of line numbers where this word has appeared. For example, after processing the first line, your dictionary should look like:

```
Mydict = {"try":[1], "not":[1], "to":[1], "become":[1], "man":[1],
"of":[1], "success":[1], "but":[1], "rather":[1], "value":[1]}
```

This should be repeated for all the lines; the new keys are added to the dictionary, and if a key already exists, its value is updated. At the end of processing all these 7 lines, the value in the dictionary associated with key `"the"` will be the list [3, 4, 7]. (Note: the line numbers start from 1.)

Use the following code to store list of line numbers as the value of a dictionary.

```
mydict={}
mydict.setdefault('the', [])
mydict['the'].append(3)
mydict['the'].append(4)
```

3) `find_cooccurance(D, inp_str)`

The first parameter is the dictionary returned by `read_file`; the second one is a string called `inp_str`. This `inp_str` contains zero or more words separated by white space. You need to split them into a list of words, and find the line numbers for each word. Convert the list of line numbers into set.Then, use the *intersection* or *union* operation on the sets from `D` (you need to figure out which operation is appropriate).

Convert the list of line numbers into set using the sample code given below

```
s1=set(mydict['the'])
s2=set(mydict['is'])
```

4) `main()`

The `main` function of your program should call the three functions above. Loop, prompting the user to enter space-separated words. Use that input to find the co-occurrence and print the results. Continue prompting for input until “q” or “Q” is input.

Very important considerations

Every time you want to look up a key in a dictionary, first you need to make sure that the key exists. Otherwise it will result in an error.

Sample Output

```
Enter a file name: albert.txt

Enter space-separated words: the
The co-occurance for: the
Lines: 3, 4, 7

Enter space-separated words: the is
The co-occurance for: the, is
Lines: 3, 7

Enter space-separated words: true knowledge imagination
The co-occurance for: true, knowledge, imagination
Lines: 3

Enter space-separated words: q
```