DirScan Project

This project extends the final assignment (Section 6 Assignment 1) of the excellent Udemy course "The Complete Python Developer Certification Course", by Imtiaz Ahmad.

The assignment was to write python code that, given a root directory and a list of words:

- walks through the directory tree, starting at the root directory
- searches each text file encountered for the words in the word list
- returns the count of occurrences for each word

The Default Solution

The solution provided in the course follows an algorithm in which, for each word, a recursive search for file names is performed and, as files are located, they are searched for the word. This approach works, but it requires a recursive descent of the directory tree for each word, and opening/closing each file for each word.

My Solution

My approach, I believe, reduces both processor load and file I/O:

- The code recurses down the directory tree only **once**, building a list of files to be searched as it goes.
- Once the file list is built, the code then opens each file in the list and searches that file for occurrences of each word in the list.

Project Structure

The program "**search_text_files.py**" is the driver program. It is invoked from the command line with at least 2 arguments. The first is the root directory. The second (and, optionally, other) arguments are words to be searched for (search terms). **search_text_files.py** prints, for each search term, the total number of times it appeared in the files.

Details

The project also:

- Defines 2 utility functions, each of which has a specific job, and which are packaged in a python module.
    - **dir_walk** performs a top-down, recursive scan of a directory tree. It returns a list of all the files it encounters that match a regular expression. It ignores directories that match another regular expression. The directory tree to be scanned and the two regular expressions are arguments to this function.
    - **count_all_occurrences** utilizes a "defaultdict" object to track the running totals of occurrences of each of the search terms. This object has a key for each of the search words. Each key references a list.

- As each file is searched for each search term, the count of occurrences in the file is appended to that word's list.
- When all files have been searched for all search terms, we have a dictionary with, for each search term, a list of the number of occurrences of the term in each file. The sum() function is used on these lists to find the total number of occurrences for each search term.

Test Data

The Lexham English Bible(LEB) is freely available from Logos Bible Software for download and non-commercial use. I downloaded several books to use as test data. The license information for the LEB is below.

Acknowledgement

The Udemy course "The Complete Python Developer Certification Course", was produced and recorded by Imtiaz Ahmad. The presentation and content of the course are quite good. While my approach to this project differs from that in the course, please note that this course provided me with the tools and test data to create my own solution.

License Information

The software and sample data in this repository is licensed under GNU General Public License v3.0. See license.md in this repository. No closed-source or commercial use is allowed.

The sample data are taken from the Lexham English Bible. This text is copyrighted by Logos Bible Software, 1313 Commercial St. Bellingham, WA 98225. Their license statement reads:

License

You can give away the Lexham English Bible, but you can't sell it on its own. If the LEB comprises less than 25% of the content of a larger work, you can sell it as part of that work.

If you give away the LEB for use with a commercial product, or sell a work containing more than 1,000 verses from the LEB, you must annually report the number of units sold, distributed, and/or downloaded.

You must always attribute quotations of the LEB.

If you quote less than 100 verses of the LEB in a single work you can attribute it by simply adding (LEB) after the quotation. Longer quotations, or use of 100 or more verses in a single work, must be accompanied by the following statement: Scripture quotations marked (LEB) are from the Lexham English Bible. Copyright 2012 Logos Bible Software. Lexham is a registered trademark of Logos Bible Software.

In electronic use, link "LEB" and "Lexham English Bible" to http://lexhamenglishbible.com, and "Logos Bible Software" to http://logos.com. If all quotations are unmarked and from the LEB, you may remove "marked (LEB) are" from the statement.

In support of non-English Bible translation, non-profit organizations may use 50% as the maximum portion the LEB may comprise of a work offered for sale. (This specifically allows the creation and commercial sale of diglot Bibles.)