Name:_____

Today I'd like to think about how we might use some of the built in Python collections (list, tuples, dictionaries, etc) to solve the following concordance-production application. A Webster's dictionary definition of concordance is: "an alphabetical list of the main words in a work." In addition to the main words, I want you to keep track of all the line numbers where these words occur.

WORD & LINE CONCORDANCE

Since the concordance should only keep track of the "main" words, there will actually be two files of textual information. The first will be a list of stop words--these words will not be included in the concordance even if they do appear in the data file. The stop words are to be stored in a second collection which will be accessed to ensure its contents do not appear in the concordance. The second file will be the actual data file. Words in this file are to be extracted, compared with the stop words, and, when appropriate, added to the concordance list along with the line number of the current occurrence of the word. Often, a word will be encountered several times--each line of encounter is to be recorded in the concordance list, but each word is to appear only once. Finally, the words are to be printed out in alphabetical order along with the numbers of the lines in which they appear.

NOTES:

a) Words are defined to be sequences of letters that are delimited by any white space, punctuation, brackets, parentheses, dashes (two hyphens in a row), double quotes, etc. but not an apostrophe or single hyphens. For example, "it's" and "end-of-line-characters" should be considered words.

b) There is to be no distinction made between upper and lower case characters, i.e., "ADT" is the same word as "adt".

c) The line numbers are to relate to non-empty lines. Blank lines are not to be counted.

Sample Output:

academy: 28, 444
acres: 1119
adam: 33
adieus: 1427
ale: 903, 932
allen: 2032
almonds: 893, 1096
american: 899
amy: 517
apology: 1701
apple: 175, 926, 927
.
.
.