Computer Science I Bill Griffith PhD
HW10 10 pts **Due Sunday, December 8, 11:59 PM**

To start, this function reads a large text file, ENGLISH_LIT, that has already been pre-processed (see Canvas) to remove punctuation and convert to lower case, and returns a list of the words in the file.

```
def read_word_list():
    davidC =open('ENGLISH_LIT.txt',"r")  # This file will be posted in Canvas
    print ("OPEN")
    book=davidC.read()
    print ("DONE")
    return book.split()
```

The list is a long sample of English text--the novel 'David Copperfield',which contains more than 380,000 words. NOTE: While you are debugging, use the file ELShort which is posted in Canvas.

How many distinct words does it contain? How many times does each word occur? It's easy enough with built-in methods for lists to find out, say, that 'the' occurs 13599 times; but how would we tabulate the number of occurrences for every word in the file, without knowing in advance what they are?

Doing this with Python lists is quite inefficient. Instead we build a Python dictionary, **whose keys are the words themselves, and whose values are the number of occurrences of each word**.

We can use the dictionary to answer some basic queries: Number of distinct words and most frequently-occurring words.

```
def build_dictionary(word_list):
    davidDict={}
```

You will write this section. Build a dictionary with word as key and occurrences as value.
Recall how you built lists with unique values (no repeats)

```
    return davidDict
```

Create a list of all the words that appear at least n times.
#
Try this out with values of n around 1000.

We will make a list of tuples:

(freq,word)

so we can sort it by frequency. You can use the built in SORT function.

Sample Run:

```
This program prints word occurrences in a long list

OPEN
DONE
FILE READ
DICTIONARY CREATED
Enter freq: 1000
TUPLE LIST
[(13599, 'the'), (12931, 'i'), (11993, 'and'), (10359, 'to'), (8619, 'of'), (7779, 'a'), (6140, 'in'), (5279, 'was'),
(5176, 'my'), (5147, 'that'), (4663, 'it'), (3832, 'her'), (3560, 'me'), (3505, 'he'), (3456, 'you'), (3341, 'with'),
(3152, 'as'), (3050, 'had'), (2939, 'said'), (2924, 'his'), (2665, 'she'), (2638, 'at'), (2558, 'for'), (2472, 'mr'),
(2382, 'have'), (2319, 'on'), (2142, 'but'), (1999, 'be'), (1966, 'not'), (1703, 'is'), (1681, 'so'), (1680, 'him'),
(1643, 'when'), (1521, 'if'), (1497, 'all'), (1461, 'we'), (1396, 'by'), (1379, 'this'), (1292, 'what'), (1265,
'which'), (1257, 'were'), (1206, 'no'), (1104, 'been'), (1092, 'there'), (1091, 'out'), (1088, 'little'), (1050,
'from')]
```

```
#HW10 Dictionaries -- Word Counts

def read_word_list():
    davidC =open('ENGLISH_LIT.txt',"r")
    print ("OPEN")
    book=davidC.read()
    print ("DONE")
    return book.split()

def build_dictionary(word_list):
    davidDict={}
```

You will write this section.

```
    return davidDict

def most_frequent(          ,          ):
```

You will write this section

```
# main section begins here

print ("\n This program prints word occurrences in a long list \n")

book=read_word_list()
print ("FILE READ")
#print (book)

dict_book=build_dictionary(book)
print ("DICTIONARY CREATED")
#print (dict_book)
frequency_limit = int(input("Enter freq: "))
tuple_list = most_frequent(dict_book, frequency_limit)
print ("TUPLE LIST")
print (tuple_list)
```