Module 12 — Interacting with Files: Activities

(1) Write a function `get_number( )` that asks a user for a number. If the user does not input a number, the program prints an error message and keeps asking for a number.

(2) Download the file "alice.txt" from the following link
http://tschwarz.mscs.mu.edu/Classes/Python_Big_Data/Module12/alice.txt .
Write programs that
(a) calculate the number of lines, words, and letters ("abcd…zABC…Z") in "Alice in Wonderland".
(b) print out all words that have at least 10 letters can contain both y and z (which might be in capital letters).
(c) print out all lines that do not contain the letter "a".

(3) Download the following file
http://tschwarz.mscs.mu.edu/Classes/Python_Big_Data/Module12/abq10001.dat .
This file contains solar radiation data from the SOLRAD station in Albuquerque, New Mexico. The first line gives the station name and the second its latitude, longitude, elevation, and the hours displaced from the local standard time, as well as the version of the file format. The next data is the important one. Each line contains the following data:

year: integer year, i.e., 2002		
jday: integer	Julian day (1 through 365 [or 366])	
month	integer	number of the month (1-12)
day	integer	day of the month(1-31)
hour	integer	hour of the day (0-23)
min	integer	minute of the hour (0-59)
dt	real	decimal time (hour.decimalminutes),e. g., 23.5 = 2330
zen	real	solar zenith angle (degrees)
dw_psp	real	downwelling global solar (Watts m^-2)
direct	real	direct solar (Watts m^-2)
diffuse	real	downwelling diffuse solar (Watts m^-2)
uvb	real	global UVB (milliWatts m^-2)
uvb_temp	real	UVB temperature (C) -- 25 deg. C is normal.

Write a program that calculates the mean of the ninth column, the one that starts out with -5.6.
(Hint: This is actually quite easy to do. You open the file in read only format, and then read line by line. You observe that the columns are separated by white spaces, therefore you can use split in order to split each line into an array. Then take the ninth element in this array, convert it into a float and add it to a sum. You also need to count the number of lines, since the SOLRAD data does not contain obviously flaky data and so the number of lines varies from file to file.

Once you are done, congratulations! You just did your first data scrubbing with real data.

4. Download the file "iris.cvs" and place it in your current directory. This file contains the famous iris data set originally used by the British statistician Ronald Fisher for a 1936 paper and used heavily in data mining text books. The data is about flowers from three different types of Iris, measuring sepals and petals. The second column contains the Sepal length and the sixth column the name of the species.
    a. Write a program that only prints out the lines belonging to Iris Variegata.
    b. Write a program that calculates the average sepal length for each of the three flower types. Your program can assume the three values.

Hint: After opening your file, you read in line by line (with the exception of the first line. You might observe that the values are separated by commas, so you use a split with parameter string ",". This results in an array of strings. You maintain three different sums for the three types of Iris. You compare the last element of the array in order to see whether this is data for a setosa, a variegata, or a virginica. Depending on this, you add the second column after conversion to a float to the corresponding sum. You also need to count the number of occurrences of each flower type, so that you can calculate the average.