# Lab Exercise #10

**Assignment Overview**

This lab exercise provides practice with sets in Python.

You will work with a partner on this exercise during your lab session. Two people should work at one computer. Occasionally switch the person who is typing. Talk to each other about what you are doing and why so that both of you understand each step.

**Part A: Programming with Sets**

Consider the file named "lab10a.py". That file contains the skeleton of a Python program to do a simple analysis of two files: it will display the number of unique words which appear in the two files (the union of those two sets of words), as well as the number of unique words which are common to both files (the intersection of those two sets of words).

Case does not matter: the words "pumpkin", "Pumpkin" and "PUMPKIN" should be treated as the same word. Only unique words should be counted: if a word appears more than once in a file, it should only be counted once. Note: remember to remove punctuation from words, e.g. "it," should be "it"

a. Replace the comments labeled "YOUR COMMENT" in function "build_word_set" with meaningful comments to describe the work being done in the next statement. Use more than one comment line, if necessary.

b. Revise function "compare_files" to accomplish the work described in the comments.

c. Test the revised program. There are two sample documents available: "document1.txt" (The Declaration of Independence) and "document2.txt" (The Gettysburg Address).

★ **Demonstrate your completed program to your TA. On-line students should submit the completed program (named "lab10a.py") for grading via the CSE handin system.**

**Part B: Programming with Dictionaries and Sets**

Consider the file named "lab10b.py". That file contains the skeleton of a Python program to display information about the words in a document.

Function "main" is complete. It handles the interaction with the user and calls other functions to perform the appropriate tasks.

Function "print_word_index" is complete. It receives a dictionary, where each element is a word and a set of line numbers where that word appears in a document. It displays all of the words (in alphabetic order), along with the lines numbers for each word (in ascending order).

Function “build_word_index” is incomplete. It receives an input file and builds a dictionary containing the unique words which appear in the input file, along with the line numbers where each word appears. The first line of the input file should be considered to be line 1.

a. Revise function “build_word_index” to accomplish the specified work. You may wish to review function “build_word_set” (above) for ideas about how to handle upper and lower case letters, as well as punctuation.

b. Test the revised program using the sample documents.

★ **Demonstrate your completed program to your TA. On-line students should submit the completed program (named “lab10b.py”) for grading via the CSE handin system.**