COMP10200: File I/O & MatPlotLib Exercises

Files

1. Write a program that reads `cities.csv` or `cities2.csv` and uses a dictionary to add up the total population of each country, then outputs the population of each country in alphabetical order by country code.

   Tips:
   - Use the `int()` function to convert the population strings to integers
   - Use `for tuple in dict.items()` to get each item in the dictionary as a tuple
   - If you sort a list of tuples, it will sort by the first item in the tuple first.

2. Write a program that reads `spamfile.txt` and uses a dictionary to keep track of how many times each word occurs. Then print the 10 most frequent words. Note that spamfile.txt uses a "latin-1" encoding, so you must open it like this: `open("spamfile.txt", encoding="latin-1")`. Failure to do so will cause unknown character errors when parsing the file.

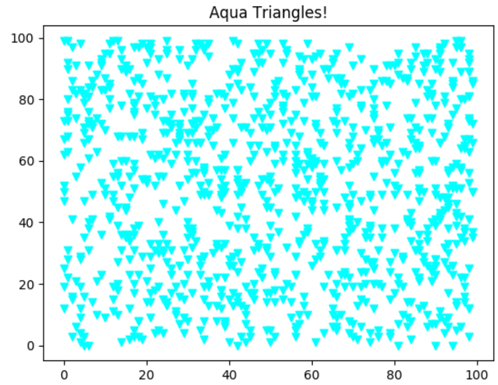
   Tips:
   - Use `line.split()` to split the line into a list of words.
   - Use `str.isalpha(s)` to decide whether `s` is a word.
   - Use `dict.items()` to turn a dictionary into a list of tuples
   - If you sort a list of tuples, it will sort by the first item in the tuple first.

   Extra Challenge: Once you're finished, you'll probably notice that the list of most frequent words is not very useful. One way to make the list better is to filter out "stop words". Stop words are the most common words in a language (e.g. "a", "the", "can", etc.). These words usually don't give much indication of the content of a document. Find a list of English stop words on line, and add code to your solution so that you remove the stop words from consideration.

3. Write a program using NumPy arrays that reads cities.csv, prints the largest, smallest, average, and total population, and then prints the names of the 10 biggest cities.

MatPlotLib

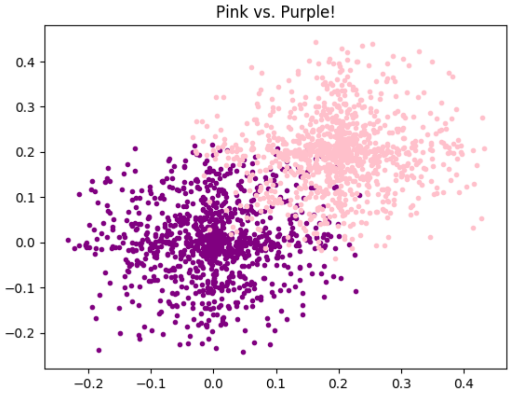
4. Use MatPlotLib to generate 1000 random integer coordinates with x and y values in the range 0 to 100 and plot them as shown below. Make sure you match the title, marker, and color shown on the right.



5. Use NumPy and MatPlotLib to recreate the scatter plot shown on the right. The purple blob is a set of 1000 random points centered at (0, 0) while the pink blob is a set of 1000 random points centered at (0.2, 0.2).

   To get a distribution of random values that looks like this, generate two random sets of coordinates with x and y values in the range –0.5 to +0.5, and multiply them together (i.e. (0.3, -0.1) * (0.5, 0.2) = (0.15, -0.02)).

   This can be done in six lines (not including imports).


Pink vs. Purple!