

Homework 8

Introductory Programming
Fall 2004

Allen B. Downey

The reading for this assignment is Chapters 10–11 of *How to think....*

8.1 Be unique!

1. Start with either your `analyze.py` from the previous homework or my `ip_hw07_soln.py`, which is available on the class web page. Either way, you should have a program that reads a file and counts the number of words.
2. If you are working with a big file, you will have a hard time debugging. It's a good idea to start with a smaller dataset, which you can create with `head`. In Linux, type

```
head -100 gatsby.txt > short.txt
```

Assuming that you have a file named `gatsby.txt`, this command should copy the first 100 lines and create the file `short.txt`. Now modify your program to read the shorter file and run it again.

3. Modify the program to print each list of words after cleaning, and run it again. Now you have a concrete idea of what data the program is working with.
4. Modify `process_file` so that it creates a new dictionary before the loop and prints the dictionary after the loop. Run the program and confirm that it prints an empty dictionary.
5. Inside the loop, add another loop that traverses the list of words and adds each word to the dictionary. Use the words as keys and give every key the same value, say 1. Now, when you run the program, it should print a big dictionary with lots of keys and values.
6. Modify the program so that instead of printing the dictionary, it prints the length of the dictionary (using `len`). Run the program again. What does the output mean?
7. Now you have a function that prints one value (the length of the dictionary) and returns another value (the total number of words). This is not usually a good idea. It would be more useful to make it a fruitful function that returns a tuple containing both values. Modify the program to do that.
8. Challenge: Read Section 10.7 and then modify your program so that it counts the number of times each word appears.
9. Double Challenge: Modify the program again so that it finds the word that appears most often.