

## HW6: Parsing Information from Files

In this homework you will read and parse a file with text and numbers. You will write three functions. First write a function to extract the numbers in the file and return the sum of the numbers. Then write a function to return a count of the number of times a given word appears in the file. Count the word even if it starts a sentence (begins with a capital letter). Don't count the word if it has additional characters after it (like an 's'). Finally, write a function to return a list of all of the URLs in the file of the form something.something.something like "www.cnn.com". The something should be any number of alphanumeric characters (but at least one alphanumeric character).

### Data Files

We provide two files for this assignment.

- Sample data: [http://py4e-data.dr-chuck.net/regex\\_sum\\_42.txt](http://py4e-data.dr-chuck.net/regex_sum_42.txt) (There are 90 values with a sum=445833)
- Actual data: [http://py4e-data.dr-chuck.net/regex\\_sum\\_132198.txt](http://py4e-data.dr-chuck.net/regex_sum_132198.txt) (There are 82 values and the sum ends with 566)

Open these links open in a new window. Make sure to save the file into the same folder as you will be writing your Python program.

### Data Format

The file contains text from the introduction of a textbook except that random numbers are inserted throughout the text. Here is a sample of the output you might see:

```
Why should you learn to write programs? 7746
12 1929 8827
Writing programs (or programming) is a very creative
7 and rewarding activity. You can write programs for
many reasons, ranging from making your living to solving
8837 a difficult data analysis problem to having fun to helping 128
someone else solve a problem. This book assumes that
everyone needs to know how to program ...
```

The sum for the sample text above is **27486**. The numbers can appear anywhere in the line. There can be any number of numbers in each line (including none).

### Handling the Data

- 1) Write a function `sumNums(filename)` to read from a file when given the filename and look for integers (that are not phone numbers) using the `re.findall()`, and then convert the extracted strings to integers and return the sum of the integers.
- 2) Write a function `countWord(filename, word)` to return a count of the number of times a specified word appears in a file. It should match the word when it starts a sentence also (starts with a capital letter). It should not match any additional letters after the word. For example, if called on "computer" it should match "Computer" and "computer" but not

“computers”. For file `regex_sum_42.txt` it will return 21 when called with the word “computer”.

- 3) Write a function `listURLs(fileName)` to return a list of the URLs in the file when given the file name. It should match URLs like [www.cnn.com](http://www.cnn.com). It doesn't have to return the `http://` part or the `https://` part of the URL, but it can. For file `regex_sum_42.txt` it will return a list of three URLs.

Use `hw6.py` to start. It has unit tests to test your code. Turn in a link to your github repo.

### Grading

- 15 points for passing the `test_sumNums1`
- 15 points for passing the `test_sumNums2`
- 10 points for passing the `test_countWord` with “computer”
- 5 points for passing one of our tests that you haven't been given for `countWord`.
- 10 points for passing the `test_listURLs`.
- 5 points for passing one of our tests with `listURLs` that you haven't been given for `listURLs`.

### Total 60

You can earn 1 point of extra credit for a possible total of 3 points for each non-trivial commit that you make before Friday Oct 19<sup>th</sup> at 10pm. Each commit must be at least 3 hours apart.