

# CSCI 544 Applied NLP: Homework 1

**Due Date: September 7, 2015 (11:59 PM PST)**

Your objective is to write a program which will read in a file, count the number of tokens (as defined below) on each line, and write an output file with the number of tokens on each line of the input. The goal is to install and check your setup with the required tools for the class (Bitbucket, Python 3, Ubuntu (VirtualBox)) and make sure you're comfortable with basic programming.

You will write a Python3 program (assignment1.py) which will take a path to an input file (absolute path name) as the first parameter and a path to an output file (absolute path) as a second parameter. It will read the lines from the input and write the number of tokens in each input line as a separate line in the output. We use the term, "token" to refer to sequences of non-space characters separated by spaces. These could be words but could also be punctuation. Good tokenization is a nontrivial problem, but for this assignment, simply use spaces to break input lines into tokens.

Your program should work as follows:

```
>python3 assignment1.py /path/to/inputfile /path/to/outputfile
```

If the line "People love to read about Nelson ." occurred in the input file, you would write the line: "7" to the output file. You can test your code using this sample input ([dev.data](#)) and the corresponding results file ([dev.results](#)). The actual test file will be similar.

All submissions will be completed through your Bitbucket account.

- Create an assignment repository ("csci544-hw1" (case sensitive)).
- In your repository, create a file "assignment1.py", this will be your submission.