Project 0: Counts

Due: Tuesday, Jan. 28th, 11:59pm

The General Idea

This is an utterly trivial assignment which should take between ten minutes and one hour to complete, hence it will not affect your final grade except in very borderline cases. Nevertheless, you should take this opportunity to (1) choose a language and make yourself comfortable with the sorts of tasks you'll be doing for the rest of the course, (2) write a nice tidy bit of code which you'll be re-using and adapting for most of the other assignments, and (3) acquaint yourself with the format of the data.

All your program has to do is **read the data from an arbitrary input file, tally up the occurrences of all the different words in the file, and spit out the list of all words with their respective counts into a specified output file.**

You can find a sample data file at `/course/cs146/asgn/counts/data.txt`. It's been pre-formatted, or *tokenized*, for you. Tokenizing is a thankless and mind-numbing task that we don't really want to talk about. What matters here is the *result* of tokenizing: a file of tokenized data contains *exactly one sentence per line*, and *words and punctuation are separated by single spaces.* Have a look at the first few lines for yourself. The data files that you will be using throughout this course will all follow this convention.

You'll notice some important conventions regarding what are interchangeably called "words" or "tokens": case matters - we consider "`The`" and "`the`" as two distinct words; punctuation (periods, commas, quotation marks etc.) are words; sometimes what we would normally consider whole words are split into fragments ("`don't`" becomes "`do n't`"). As we get into language modelling you'll understand why things are done this way.

The file generated by your program should have one word and its corresponding count on each line:

```
the 173
my 146
dog 31
cat 51
bit 22
...
```

or whatever the actual counts are. The words don't have to be in any particular order.

Depending on how you re-use your code, you may not need to generate an output file for subsequent assignments, but for the purpose of this warmup assignment, *make sure that you adhere to our specified output format or we won't be able to grade you properly!*

The template script is `/course/cs146/asgn/counts/counts`. *Copy this file and fill in the specified line with the command that runs your code and include this with your handin.* To hand in, run `/course/cs146/bin/cs146_handin counts` from the directory that contains your code.

Finally: Look through the word counts for the number of occurrences of "`the`" (lowercase) and be prepared to report this number in class the day after this assignment is due.