Programming, Probability, and the Modern Mathematics Classroom Exercises — Part 7

Manan Shah
Mathematician-At-Large

June 11, 2013

Please make sure to have read the blog post with the same topic name on the Math Misery website, otherwise this will be out of context.

Some reminders

- Have students to comment the code, line by line.

- Let students tweak, modify, and otherwise alter the code.

- Don't worry about programming formalities.

- Have fun, relax, and learn.

The distribution of the letters in Anna Karenina

In this part we will build a histogram for the letters used in Anna Karenina by Leo Tolstoy. A copy of the book seems to be available for download for free here: archive.org.

The lesson is inline below. Take note of the method `makeStars`.

```
>>> import os
>>> os.chdir("C:/myBlog")
>>> lettersdictionary = {}
>>> with open("annakarenina.txt","r") as f:
        letters = set(list("abcdefghijklmnopqrstuvwxyz")) ## what does this do?
        for line in f:
                if "CHAPTER I" in line:
                        break
        for line in f:
                if "ANNA KARENINA" in line:
                        continue ## what does continue do?
                if "CHAPTER" in line:
                        continue
                if "PART" in line:
                        continue
                if "THE END" in line:
                        break
```

```
                    ## why did we have all the if statements above?
                    for character in line: ## what are we iterating through?
                            c = character.lower() ##what does .lower do? and why do we do this?
                            if c in letters: ##why is it better that letters is a set rather than a list?
                                    lettersdictionary[c] = lettersdictionary.get(c,0) + 1


>>> len(lettersdictionary)
26
>>> for letter in sorted(lettersdictionary):
        print(letter, lettersdictionary[letter])


a 123980
b 21587
c 35869
d 68704
e 189152
f 31105
g 30209
h 108312
i 112230
j 1581
k 17217
l 61225
m 35330
n 107906
o 112246
p 23667
q 1147
r 80409
s 101518
t 138476
u 39327
v 19150
w 37451
x 2318
y 31465
z 1370
>>> def makeStars(number, divisor):
        numstars = int(number/divisor) + 1
        stars = ""
        while len(stars) != numstars:
                stars += "*"
        return stars

>>> with open("annak_lettersdist.txt","w") as g:
        for letter in sorted(lettersdictionary):
                z = g.write(letter + ": " + makeStars(lettersdictionary[letter],500) + "\n")
```

Summary

In this part, we went through an example of generating counting all the letters in Anna Karenina and building a histogram which we wrote to a text file. So here are some questions that we can ask:

- Since this is a story by Leo Tolstoy, does that affect the letter distribution?

- Would the letter distribution be different if we did the same analysis on "War & Peace"?

- What about something written by Mark Twain? Would the "y" be as prominent?

Here is a fun little exercise to do. Spell out all the integers from zero to one hundred and do the same frequency analysis. What do you notice?

If you need help, have questions, or would like to set up a workshop at your school get in touch with me at help@mathmisery.com.